ECML/PKDD 2006 Workshop on

Practical Data Mining:

Applications, Experiences and Challenges

Organized by:
Markus Ackermann,
Carlos Soares,
Bettina Guidemann

In Partnership with
SAS Deutschland, Heidelberg

Berlin, September 22nd, 2006

Preface

General Description

The Workshop on Practical Data Mining is jointly organized by the Institute for Computer Science of University of Leipzig, the Artificial Intelligence and Computer Science Laboratory of the University of Porto and SAS Deutschland with the goal of gathering researchers and practitioners to discuss relevant experiences and issues in the application of data mining technology in practice and to identify important challenges to be addressed.

We are sure that this workshop will provide a forum for the fruitful interaction between participants from universities and companies, but we aim to go beyond that! We hope that this workshop will become the starting point for practical projects that involve people from the two communities. The future will tell if we succeeded.

Motivation

Business, government and science organizations are increasingly moving toward decision-making processes that are based on information. In parallel, the amount of data representing the activities of organizations that is stored in databases is also growing. Therefore, the pressure to extract as much useful information as possible from this data is very strong.

Many methods for this purpose have been developed in research areas such as data mining, machine learning and statistics. These methods are available not only in data mining and business intelligence tools but are increasingly being integrated into other information systems and tools (e.g., customer relationship management, database management systems and network security). Despite the maturity of the field, new problems and applications are continuously challenging both researchers and practitioners. However, the dialog between these two communities is not always easy.

Topics

The program includes contributions from researchers and practitioners both in industry as in academia. The papers describe case studies with application areas including: finance, telecom, retail, government, bio-informatics, e-business, transportation, electricity and health care. The problems and techniques addressed also include data warehousing, customer profiling, decision support, churn prediction, credit risk management, fraud and fault detection, and quality control.

Invited Talks

We are very pleased to have on the workshop two invited talks. The first one is by Stefan Wrobel of Fraunhofer Institute for Autonomous Intelligent Systems on *Geo Intelligence – New Business Opportunities and Research Challenges in Spatial Mining and Business Intelligence*. And the second talk is by Ulrich Reincke of SAS Institute, who speaks about *Directions of Analytics, Data and Text Mining – a Software Vendor's View*

Paper Acceptance

There were 24 papers submitted to this workshop. Each paper was reviewed by at least two reviewers. Based on the reviews, 10 papers were selected for oral presentation at the workshop and 9 for poster presentation.

Acknowledgments

We wish to thank the Program Chairs and the Workshop Chairs of the ECML-PKDD 2006 conference, in particular Tobias Scheffer, for their support.

We thank the members of the Program Committee for the timely and thorough reviews and for the comments which we believe will be very useful to the authors.

We are also grateful to DMReview and KMining for their help in reaching audiences which we otherwise would not be able to.

Thanks go also to the SAS Deutschland, Heidelberg, for their support and partnership, and to our sponsors, Project Triana and Project Site-O-Matic, both of University of Porto.

September 2006

Markus Ackermann
Carlos Soares
Bettina Guidemann
Workshop Chairs
Practical Data Mining'06

Organization

Program Chairs and Organizing Committee

- Markus Ackermann, University of Leipzig, Germany
- Carlos Soares, University of Porto, Portugal
- Bettina Guidemann, SAS Deutschland, Heidelberg, Germany

Program Committee

- Alípio Jorge, University of Porto, Portugal
- André Carvalho, University of São Paulo, Brazil
- Arno Knobbe, Kiminkii/Utrecht University, The Netherlands
- Carlos Soares, University of Porto, Portugal
- Dietrich Wettschereck, Recommind, Germany
- Dirk Arndt, DaimlerChrysler AG, Germany
- Donato Malerba, University of Bari, Italy
- Fátima Rodrigues, Polytechnical Institute of Porto, Portugal
- Fernanda Gomes, BANIF, Portugal
- Floriana Esposito, Universitá degli Studi di Bari, Italy
- Gerhard Heyer, University of Leipzig, Germany
- Gerhard Paaß, Fraunhofer Institute, Germany
- Lubos Popelinsky, Masaryk University, Czech Republic,
- Luc Dehaspe, PharmaDM
- Manuel Filipe Santos, University of Minho, Portugal
- Mário Fernandes, Portgás, Portugal
- Marko Grobelnik, Josef Stefan Institute, Slovenia
- Markus Ackermann, University of Leipzig, Germany
- Mehmet Göker, PricewaterhouseCoopers, USA
- Michael Berthold, University of Konstanz, Germany
- Miguel Calejo, Declarativa/Universidade do Minho, Portugal
- Mykola Pechenizkiy, University of Jyväskylä, Finland
- Paula Brito, University of Porto, Portugal
- Paulo Cortez, University of Minho, Portugal
- Pavel Brazdil, University of Porto, Portgual
- Peter van der Putten, Chordiant Software/Leiden University, The Netherlands
- Petr Berka, University of Economics of Prague, Czech Republic
- Pieter Adriaans, Robosail
- Raul Domingos, SPSS, Portugal
- Reza Nakhaeizadeh, DaimlerChrysler AG, Germany
- Robert Engels, CognIT, Germany
- Rüdiger Wirth, DaimlerChrysler AG, Germany

– Rui Camacho, University of Porto, Portugal
– Ruy Ramos, University of Porto, Portugal
– Sascha Schulz, Humboldt University of Berlin, Germany
– Stefano Ferilli, University of Bari, Italy
– Steve Moyle, Secerno, United Kingdom
– Teresa Godinho, Allianz Portugal, Portugal
– Timm Euler, University of Dortmund, Germany

In Partnership with

SAS Deutschland, Heidelberg

Sponsors

Project Triana and Project Site-o-Matic of University of Porto.

Table of Contents

Health Care and Medical Applications

Industry

Public Authorities and Law Enforcement

Tools

Geo Intelligence – New Business Opportunities and Research Challenges in Spatial Mining and Business Intelligence (Invited Talk)

Stefan Wrobel

Fraunhofer IAIS & University of Bonn
stefan.wrobel@iais.fraunhofer.de

Abstract

Every customer has an address, every store has a location, and traffic networks are a decisive factor in accessibility and logistics. Even in classical business data analysis, a large majority of data have a spatial component, and optimal business decisions must take geographical context into account. In the talk, we will present several examples of real world customer projects ranging from location selection and geo-marketing to outdoor media. We will then move on to the new challenges and opportunities brought about by the widespread availability of localisation technology that allows tracking of people and objects in time and space.

Professor Dr. Stefan Wrobel is a professor of computer science at university of Bonn and one of the three directors of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS (created in July 2006 as a merger of Fraunhofer Institutes AIS and IMK).

Directions of Analytics, Data and Text Mining – A software vendor's view (Invited Talk)

Ulrich Reincke

Executive Expert Analytics, SAS Institute Germany
Ulrich.Reincke@ger.sas.com

Abstract

After the years of hype at the turn of the millennium, immediately followed by the crush of the dot.com-bubble, data mining has become a mature market. New business applications are continuously developed even for remote industries and total new data sources are becoming increasingly available to be explored with new Data Mining methods. The common type of data sources moved initially from numerical over time-stamped to categorical and text, while the latest challenges are geographic, biological and chemical information, that are both of text and numerical type coupled with very complex geometric structures.

If you take a closer look at the concrete modelling options of both freeware and commercial data mining tools, there is pretty little difference between them. They all claim to provide their users with the latest analysis models that are consensus within the discussions of the research community. However, what makes a big difference, is the ability to map the data mining process into a continuous IT-flow, that controls the full information from the raw data, cleaning aggregation and transformation, analytic modelling, operative scoring, and last but not least final deployment. This IT process needs to be set up as to secure that the original business question is solved and the resulting policy actions are applied appropriately in the real world. This ability constitutes a critical success factor in any data mining project of larger scale. Among other environmental parameters of a mining project it depends mainly on clean and efficient metadata administration and the ability to cover and administer the whole project information flow with one software platform: data access, data integration, data mining, scoring and business intelligence. SAS is putting considerable effort to pursue the development of its data mining solutions in this direction. Examples of real life projects will be given.

Sequence Mining for Customer Behaviour Predictions in Telecommunications

Frank Eichinger[1], Detlef D. Nauck[2], and Frank Klawonn[3]

[1] Universität Karlsruhe (TH), Institute for Program Structures and Data Organisation (IPD), Karlsruhe, Germany, eichinger@ipd.uka.de
[2] BT Group plc, Intelligent Systems Research Centre, Ipswich, UK, detlef.nauck@bt.com
[3] University of Applied Sciences Braunschweig/Wolfenbüttel, Department of Computer Science, Wolfenbüttel, Germany, klawonn@fh-wolfenbuettel.de

Abstract Predicting the behaviour of customers is challenging, but important for service oriented businesses. Data mining techniques are used to make such predictions, typically using only recent static data. In this paper, a sequence mining approach is proposed, which allows taking historic data and temporal developments into account as well. In order to form a combined classifier, sequence mining is combined with decision tree analysis. In the area of sequence mining, a tree data structure is extended with hashing techniques and a variation of a classic algorithm is presented. The combined classifier is applied to real customer data and produces promising results.

1 Introduction

Predicting churn, i.e. if a customer is about to leave for a competitor, is an important application of analysing customer behaviour. It is typically much more expensive to acquire new customers then to retain existing ones. In the telecommunication industry, for example, this factor is in the range of about five to eight [1]. Correctly predicting that a customer is going to churn and then successfully convincing him to stay can substantially increase the revenue of a company, even if a churn prediction model produces a certain number of false positives.

Beside the prediction of churn, other customer-related events like faults, purchases or complaints can be predicted in order to be able to resolve some problems before the actual event occurs. The prediction of sales events can be used for cross-selling, where a certain product is offered just to customers who have an increased likelihood to buy it.

In the telecommunications industry the available customer data is typically timestamped transactional data and some static data (e.g. address, demographics and contract details). Transactional data are sequences of timestamped events which can easily be stored in relational database tables. Events can be any kind of service usage or interaction, particularly calls to the company's call centre, for example, complaints or orders.

In the area of data mining, many approaches have been investigated and implemented for predictions about customer behaviour, including neural networks, decision trees and naïve Bayes classifiers (e.g., [1, 2, 3]). All these classifiers work with static data. Temporal information, like the number of complaints in the last year, can only be integrated by using aggregation. Temporal developments, like a decreasing monthly billing amount, are lost after aggregation. In this paper, sequence mining as a data mining approach that is sensitive to temporal developments is investigated for the prediction of customer events.

In Chapter 2 we present sequence mining and its adoptions for customer data in telecommunications along with an extended tree data structure. In Chapter 3, a combined classification framework is proposed. Chapter 4 describes some results with real customer data and Chapter 5 concludes this paper and points out the lessons learned.

2 Sequence Mining

Sequence mining was originally introduced for market basket analysis [4] where temporal relations between retail transactions are mined. Therefore, most sequence mining algorithms like AprioriAll [4], GSP [5] and SPADE [6] were designed for mining frequent sequences of itemsets. In market basket analysis, an itemset is the set of different products bought within one transaction. In telecommunications, customer events do not occur together with other events. Therefore, one has to deal with mining frequent event sequences, which is a specialisation of itemset sequences. Following the Apriori principle [7], frequent sequences are generated iteratively. A sequence of two events is generated from frequent sequences consisting of one event and so on. After generating a new candidate sequence, its support is checked in a database of customer histories. The support is defined as the ratio of customers in a database who contain the candidate sequence in their history.

2.1 Sequence Mining for Customer Behaviour Predictions

A crucial question in sequence mining is the definition of the relationship "S is contained in T" (denoted as $S \prec T$), which is decisive for determining the support of a sequence. Originally, a sequence S is contained in a sequence T, if all elements of S occur in T in the same order [4]. It does not matter if S and T are equal or if one or more additional events are contained in T as well. A strict definition would not allow any extra events in between the events of sequence T, but at its beginning and end. For example, $\langle C \leftarrow B \leftarrow A\rangle \prec \langle X \leftarrow X \leftarrow X \leftarrow C \leftarrow Y \leftarrow B \leftarrow Y \leftarrow A \leftarrow Z\rangle$ is true in the original definition, but not in the strict one as there are two events Y which are not allowed.

In this work, we want to use sequence mining for classification. If a certain sequence of events was identified leading to a certain event with a high confidence, we want to use this sequence for classifying customers displaying the same sequence. If we chose the strict definition of "is contained in", we would not classify

customers correctly who contain a very significant sequence but with an extra event in between. This extra event could be a simple call centre enquiry which is not related to the other events in the sequence. The original definition would allow many extra events occurring after a matched sequence. In the application to customer behaviour prediction, a high number of more recent events after a significant sequence might lower its impact. Therefore, we introduce two new sequence mining parameters: *maxGap*, the maximum number of allowed extra events in between a sequence and *maxSkip*, the maximum number of events at the end of a sequence before the occurrence of the event to be predicted. With these two parameters, it is possible to determine the support of a candidate sequence very flexibly and appropriately for customer behaviour predictions. For instance, the presented example is true if $maxGap = 2$ and $maxSkip = 3$. It is not true any more, if one of the parameters is decreased.

2.2 The Sequence Tree Data Structure

Multiple database scans, which are necessary after every generation of candidate sequences, are considered to be one of the main bottlenecks of Apriori-based algorithms [8, 9]. Such expensive scans can be avoided by storing the database of customer histories efficiently in main memory. In association rule mining, tree structures are used frequently to store mining databases (e.g., [8]). In the area of sequence mining, trees are not as attractive as lattice and bitmap data structures (e.g., [6, 9]). This is due to smaller compressing effects in the presence of itemsets. In our case, as well as in the application of sequence mining to web log analysis (e.g., [10]) where frequent sequences of single events are mined, tree structures seem to be an efficient data structure. In this paper, such a tree structure or more precisely trie memory[4] [11] as known from string matching [12], is employed to store sequences compressed in main memory. We call our data structure *SequenceTree*.

In the *SequenceTree*, every element of a sequence is represented in an inner- or leaf node. The root node and all inner nodes contain maps of all direct successor nodes. Each child represents one possible extension of the prefix sequence defined by the parent node. The root node is not representing such an element, it just contains a map of all successors, which are the first elements from all sequences. Every node, except the root node, has an integer counter attached which indicates how many sequences are ending there.

An example for a *SequenceTree* containing five sequences is given in Figure 1. To retrieve the sequences from the tree, one can start at every node with a counter greater than zero and follow the branch in the tree towards the root node. Note that if the sequence $\langle A \leftarrow B \leftarrow C \rangle$ is stored already, just a counter needs to be increased if the same sequence is added again. If one wants to add $\langle A \leftarrow B \leftarrow C \leftarrow D \rangle$, the last node with the C becomes an inner node and a new leaf node containing the event D with a count of one is added.

[4] Tries are also called prefix trees or keyword trees.

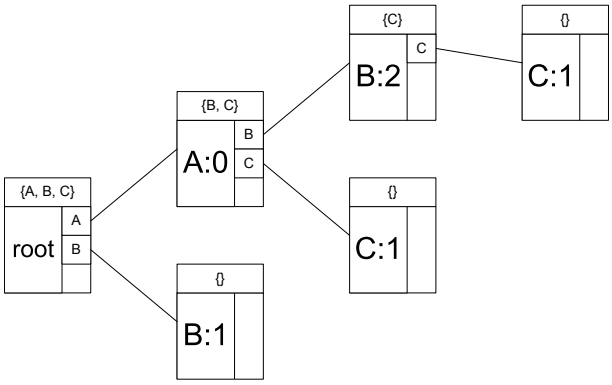


Figure 1. A *SequenceTree* containing the sequences $\langle A \leftarrow B \leftarrow C \rangle$, $\langle A \leftarrow B \rangle$ (twice), $\langle A \leftarrow C \rangle$ and $\langle B \rangle$. The number after ":" indicates the count how many sequences are ending in the node. The hash tables with all subsequent events are denoted by {}.

The compact storage of sequences achieved using the *SequenceTree* is due to two compressing effects:

1. The usage of counters as in [8, 9, 10] avoids the multiple storage of the same sequences. Obviously, the compression ratio depends very much on the kind and amount of data. Experiments with real customer data showed that the usage of counters reduces the memory necessary to store the sequences by a factor of four to ten.
2. Sequences with the same prefix sequence are stored in the same branch of a tree as done in the finite state machines known from string pattern matching [12]. Especially if sequences are long, this technique can reduce the memory needed significantly.

In sequence mining algorithms like in [4, 5] or in the one described in the following subsection, it happens very frequently that a candidate sequence is being searched in a database in order to determine its support. These searches can be very time consuming, even if the database is stored in an efficient data structure. In order to speed up searches in the *SequenceTree*, hash tables are used in every node which contain all events occurring in all succeeding nodes. If a candidate sequence is searched in the tree, the search can be pruned at an early stage if not all events in the searched sequence are included in the hash table of the current node. For example, we want to count the support of the sequence $\langle A \leftarrow B \leftarrow D \rangle$ in the *SequenceTree* from Figure 1. The search algorithm would check the hash table of the root node first. As D is not contained in this table, the search could be stopped immediately. As hash tables provide constant time performance for inserting and locating [13], the maintenance of hashtables as well as lookups do not require much extra time. Also the memory overhead is marginal as it is sufficient to store small pointers to events in the hash tables. In our experiments we measured a speed up of three by utilising hash tables in a *SequenceTree* during a real churn prediction scenario.

2.3 Sequence Mining Using the Sequence Tree

In Figure 2 a sequence mining algorithm taking advantage of the *SequenceTree* is described. This algorithm is based on AprioriAll [4], adopts its candidate generation, but avoids multiple database scans as the database is being loaded into a *SequenceTree* first. In every iteration, candidate sequences $candidates_k$ are generated. Afterwards, the support of every candidate *cand* is calculated in the *SequenceTree* $\mathcal{C}$. Only Sequences exceeding a user defined minimum support *minSup* are kept and returned at the end.

Require: $\mathcal{C}$ (database of customers), *minSup*(sequence mining parameter)
  $L_1 = \{\langle E \rangle \mid support(\langle E \rangle) \geq minSup\}$
  for $(k = 2;\ L_{k-1} \neq \emptyset;\ k{+}{+})$ **do**
    $candidates_k = generate\_candidates(L_{k-1})$
    for all $(cand \in candidates_k)$ **do**
      $cand.count = \mathcal{C}.determineSupport(cand)$
    end for
    $L_k = \{cand \mid support(cand) \geq minSup \wedge cand \in candidates_k\}$
  end for
  return $\bigcup_k \{S \mid S \in L_k\}$

Figure 2. Sequence mining algorithm with $\mathcal{C}$ stored in a *SequenceTree*.

The method *determineSupport*() is responsible for calculating the support of a sequence. This is done by performing a greedy depth first search of the candidate sequence in the $SequenceTree$[5]. Due to the introduced parameters *maxGap* and *maxSkip* which allow a flexible definition of support, the search is not as easy as searches in string matching [12]. The parameter *maxGap* is implemented by skipping up to *maxGap* nodes during the search and backtracking afterwards. Backtracking and searching in all possible branches is necessary, as a candidate sequence can occur in several branches if gaps are allowed. The parameter *maxSkip* requires to perform up to *maxSkip* searches in parallel. Up to *maxSkip* events can be skipped at the beginning of a search. Therefore, a new parallel search is started at every node which is reached by the search algorithm by traversing deeper into the tree.

2.4 Experimental Sequence Mining Results

Sequence mining as described in the previous subsection was applied to real customer data in a churn prediction scenario. The dataset used was artificially sampled in order to obtain an overall churn rate of exactly 4%. A number of sequences were found and for every sequence a confidence value was calculated. The confidence value is a likelihood for the occurrence - in this case - of a churn

[5] All algorithms traversing the tree were implemented iteratively as our experiments showed a significant performance gain compared to recursive implementations.

event. The result was a set of sequential association rules like the following one: "$\langle ENQUIRY \leftarrow ENQUIRY \leftarrow REPAIR \rangle$, confidence = 4.4%, support = 1.2%", meaning that 1,2% of all customers display the pattern with a specific repair first, than an enquiry followed by another enquiry in their event history. 4.4% of all customers with this pattern are having a churn event afterwards. Therefore, we know that customers displaying this pattern have a slightly higher churn probability than the average of all customers. On the other hand, a support of 1,2% means that just a small fraction of all customers is affected. This rule is just one rule in a set of around hundred rules (depending on the predefined minimum support). Only some rules exist with a higher confidence of around 10%, but they affect even smaller fractions of customers. Even if such a rule with a high confidence of e.g. 10% is used to identify churners, this rule would still classify 90% of the customers incorrectly. Therefore, even a large set of sequential association rules was not suitable for churn prediction.

3 A Framework for Customer Behaviour Prediction

Given that more information than just the sequential order of events was available in our application scenario, we built a classifier which is based on sequence mining, but analyses additional attributes with decision trees. These additional attributes are such associated with the customer (e.g., the contract duration), the sequence (e.g., the number of days between two events) and the events itself (e.g., the time to resolve a repair). A similar combination of sequence mining and other classifiers has been successfully implemented in bio-informatics [14]. In the following, we describe a prediction framework (Figure 3) consisting of a model building process and a classification process.

In the model building process, sequence mining as described in the previous section is applied first. Afterwards, a decision tree is induced and pruned for each detected sequence incorporating a number of further attributes. The sequences are saved together with the corresponding decision trees building a combined classification model.

In the classification process, single customers are classified using the classification model. At first, sequences that are supported by the customer's event history are selected from the model. Subsequently, the customer is classified by the decision trees associated with these sequences. The final classification of the customer is computed by averaging all results and applying a threshold value.

4 Experimental Results

The combined classifier was applied to real customer data from a major European telecommunication provider. In this paper, just some results from a churn prediction scenario are presented, even if the model was tested in a number of different scenarios for different events. For reasons of data protection, non-representative random test samples with a predefined churn rate had to be generated. In a three months churn prediction scenario, the combined classifier was first trained

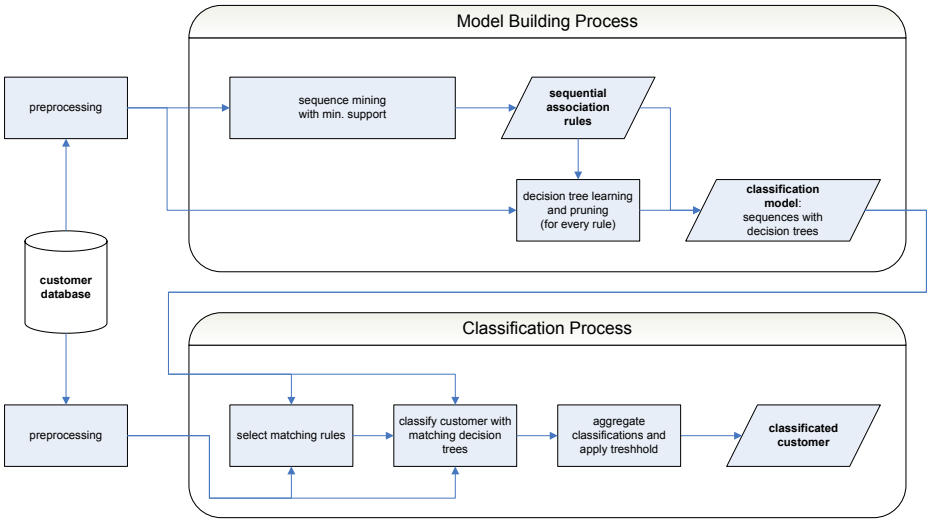


Figure 3. The framework for customer behaviour prediction.

with historic data including all events within one year and then applied to a test set from a more recent time window. The classifier found 19.4% of all churners with a false positive rate of only 2.6%. The gain[6] of this test result – the ratio how much the classifier is better than a random classifier – is 5.5.

It is hard to compare our test results. On the one hand, all results related to customer data are usually confident and therefore they are not published. On the other hand, published results are hardly comparable due to differences in data and test scenarios. Furthermore, most published results were achieved by applying the predictive model to a test set from the same time window (e.g., [3]) instead of making future predictions.

5 Conclusion and Lessons Learned

In this paper we extended a tree data structure and approach for sequence mining. This approach was combined with decision trees in order to form a combined classifier which is able to predict any desired customer event.

In the area of sequence mining, we showed that traditional definitions of support and especially of the "is contained in" relationship are not feasible for customer behaviour predictions. We introduced two new parameters to flexibly specify this relation.

As multiple events at the same time are unusual in telecommunication customer data, we introduced an extended tree data structure and algorithm for

[6] The gain measure is defined as the predictor's churn rate (the ratio of all correctly predicted churners to all customers predicted as churners) divided by the a priori churn rate (the rate of churners in the test set).

mining sequences of single events. We showed that our tree structure in combination with hashing techniques is very efficient.

Our investigations showed that sequence mining alone is not suitable for making valuable predictions about the behaviour of customers based on typically rare events like churn. However, it is capable of discovering potentially interesting relationships concerning the occurrence of events.

Furthermore, our study showed that it is more promising to analyse temporal developments by employing sequence mining in combination with other classifiers than to use only static classification approaches.

References

[1] Yan, L., Miller, D.J., Mozer, M.C., Wolniewicz, R.: Improving Prediction of Customer Behavior in Nonstationary Environments. In: Proc. International Joint Conference on Neural Networks (IJCNN). (2001)

[2] Buckinx, W., Baesens, B., den Poel, D., van Kenhove, P., Vanthienen, J.: Using Machine Learning Techniques to Predict Defection of Top Clients. In: Proc. 3rd International Conference on Data Mining Methods and Databases. (2002) 509–517

[3] Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., Mason, C.H.: Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research **43**(2) (2006) 204–211

[4] Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: Proc. 11th International Conference on Data Engineering (ICDE). (1995) 3–14

[5] Srikant, R., Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements. In: Proc. 5th International Conference Extending Database Technology (EDBT). (1996) 3–17

[6] Zaki, M.J.: SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning **42**(1–2) (2001) 31–60

[7] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th International Conference Very Large Data Bases (VLDB). (1994) 487–499

[8] Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation. In: Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD). (2000) 1–12

[9] Savary, L., Zeitouni, K.: Indexed Bit Map (IBM) for Mining Frequent Sequences. In: Proc. 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). (2005) 659–666

[10] El-Sayed, M., Ruiz, C., Rundensteiner, E.A.: FS-Miner: Efficient and Incremental Mining of Frequent Sequence Patterns in Web Logs. In: Proc. 6th ACM Workshop on Web Information and Data Management (WIDM). (2004) 128–135

[11] de la Briandais, R.: File Searching Using Variable Length Keys. In: Proc. Western Joint Computer Conference. (1959) 295–298

[12] Aho, A.V., Corasick, M.J.: Efficient String Matching: An Aid to Bibliographic Search. Communications of the ACM **18**(6) (1975) 333–340

[13] Aho, A.V., Hopcroft, J.E., Ullman, J.D.: Data Structures and Algorithms. Series in Computer Science and Information Processing. Addison-Wesley (1982)

[14] Ferreira, P.G., Azevedo, P.J.: Protein Sequence Classification Through Relevant Sequence Mining and Bayes Classifiers. In: Proc. 12th Portuguese Conference on Artificial Intelligence (EPIA). (2005) 236–247

Machine Learning for Network-based Marketing

Shawndra Hill, Foster Provost and Chris Volinsky

New York University and AT&T Labs-Research

Abstract

Traditionally data mining and statistical modeling have been conducted assuming that data points are independent of one another. In fact, businesses increasingly are realizing that data points are fundamentally and inextricably interconnected. Consumers interact with other consumers. Documents link to other documents. Criminals committing fraud interact with each other.

Modeling for decision-making can capitalize on such interconnections. In our paper appearing in the May issue of Statistical Science (Hill et al. 2006), we provide strong evidence that "network-based marketing" can be immensely more effective than traditional targeted marketing. With network-based marketing, consumer targeting models take into account links among consumers. We concentrate on the consumer networks formed using direct interactions (e.g., communications) between consumers. Because of inadequate data, prior studies have not been able to provide direct, statistical support for the hypothesis that network linkage can directly affect product/service adoption. Using a new data set representing the adoption of a new telecommunications service, we show very strong support for the hypothesis. Specifically, we show three main results:

1) "Network neighbors"—those consumers linked to a prior customer—adopt the service at a rate 3-5 times greater than baseline groups selected by the best practices of the firm's marketing team. In addition, analyzing the network allows the firm to acquire new customers that otherwise would have fallen through the cracks, because they would not have been identified by models learned using only traditional attributes.

2) Statistical models, learned using a very large number of geographic, demographic, and prior purchase attributes, are significantly and substantially improved by including network-based attributes. In the simplest case, including an indicator of whether or not each consumer has communicated with an existing customer improves the learned targeting models significantly.

3) More detailed network information allows the ranking of the network-neighbors so as to permit the selection of small sets of individuals with very high probabilities of adoption. We include graph-based and social-network features, as well as features that quantify (weight) the network relationships (e.g., the amount of communication between two consumers).

In general, network-based marketing can be undertaken by initiating "viral marketing," where customers are given incentives to propagate information themselves. Alternatively, network-based marketing can be performed directly

by any business that has data on interactions between consumers, such as phone calls or email. Telecommunications firms are obvious candidates.

These results also provide an interesting perspective on recent initiations and acquisitions of network communication services by non-telecom companies, for example, gmail for Google, Skype for Ebay. Would targeting the social neighbors of consumers who respond favorably to ads help Google? Clearly that depends on the type of product. In addition, the results emphasize the intrinsic business value of electronic community systems that provide explicit linkages between acquaintances, such as MySpace, Friendster, Facebook, etc.

Based on: S. Hill, F. Provost, and C. Volinsky. "Network-based marketing: Identifying likely adopters via consumer networks." Statistical Science 21(2), 2006.

Customer churn prediction - a case study in retail banking

Teemu Mutanen, Jussi Ahola, and Sami Nousiainen

VTT Technical Research Centre of Finland
teemu.mutanen@vtt.fi

Abstract. This work focuses on one of the central topics in customer relationship management (CRM): transfer of valuable customers to a competitor. Customer retention rate has a strong impact on customer lifetime value, and understanding the true value of a possible customer churn will help the company in its customer relationship management. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers. We predict customer churn with logistic regression techniques and analyze the churning and nonchurning customers by using data from a consumer retail banking company. The result of the case study show that using conventional statistical methods to identify possible churners can be successful.

1 Introduction

This paper will present a customer churn analysis in consumer retail banking sector. The focus on customer churn is to determinate the customers who are at risk of leaving and if possible on the analysis whether those customers are worth retaining. A company will therefore have a sense of how much is really being lost because of the customer churn and the scale of the efforts that would be appropriate for retention campaign.

The customer churn is closely related to the customer retention rate and loyalty. Hwang et al. [8] defines the customer defection the hottest issue in highly competitive wireless telecom industry. Their LTV model suggests that churn rate of a customer has strong impact to the LTV because it affects to the length of service and to the future revenue. Hwang et al. also defines the customer loyalty as the index that customers would like to stay with the company. Churn describes the number or percentage of regular customers who abandon relationship with service provider [8].

$$\text{Customer loyalty} = 1 - \text{Churn rate} \tag{1}$$

Modeling customer churn in pure parametric perspective is not appropriate for LTV context because the retention function tends to be *spiky* and non-smooth, with spikes at the contract ending dates [14]. And usually on the marketing perspective the sufficient information about the churn is the probability of

Table 1. Examples of churn prediction in literature.

article	market sector	case data	methods used %
Au et al.[1]	wireless telecom	100 000 subscribers	DMEL-method (data mining by evolutionary learning)
Buckinx et al.[2]	retail business	158 884 customers	Logistic regression, ARD (automatic relevance determination), decision tree
Ferreira et al.[6]	wireless telecom	100 000 subscribers	Neural network, decision tree HNFS, rule evolver
Garland [7]	retail banking	1 100 customers	multiple regression
Hwang et al.[8]	wireless telecom	16 384 customers	logistic regression, neural network, decision tree
Mozer et al.[12]	wireless telecom	46 744 subscribers	logistic regression, neural network, decision tree
Keaveney et al.[9]	online service	28 217 records	descriptive statistics based on the questionnaires sent to the customers

possible churn. This enables the marketing department so that, given the limited resources, the high probability churners can be contacted first [1].

Lester explains the segmentation approach in customer churn analysis [11]. She also points out the importance of the right characteristics studied in the customer churn analysis. For example in the banking context those signals studied might include decreasing account balance or decreasing number of credit card purchases. Similar type of descriptive analysis has been conducted by Keveney et al. [9]. They studied customer switching behavior in online services based on questionnaires sent out to the customers. Garland has done research on customer profitability in personal retail banking [7]. Although their main focus is on the customers' value to the study bank, they also investigate the duration and age of customer relationship based on profitability. His study is based on customer survey by mail which helped him to determine the customer's share of wallet, satisfaction and loyalty from the qualitative factors.

Table 1 presents examples of the churn prediction studies found in literature: the analysis of the churning customers have been conducted on various fields. However, based on our best understanding, no practical studies have been published related to retail banking sector focused on the difference between continuers and churners.

2 Case study

Consumer retail banking sector is characterized by customers who stays with a company very long time. Customers usually give their financial business to one company and they won't switch the provider of their financial help very often. In the company's perspective this produces a stabile environment for the customer relationship management. Although the continuous relationships with the customers the potential loss of revenue because of customer churn in this case can be

huge. The mass marketing approach cannot succeed in the diversity of consumer business today. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers.

In this study a customer database from a Finnish bank was used and analyzed. The data consisted only of personal customers. The data at hand was collected from time period December 2002 till September 2005. The sampling interval was three months, so for this study we had relevant data of 12 points of time [t(0)-t(11)]. In logistic regression analysis we used a sample of 151 000 customers.

In total, 75 variables were collected from the customer database. These variables are related to the topics as follows: (1) account transactions IN, (2) account transactions OUT, (3) service indicators, (4) personal profile information, and (5) customer level combined information.

The data had 30 service indicators in total, (e.g. 0/1 indicator for housing loan). One of these indicators, C1 tells whether the customer has a current account in the time period at hand or not, and the definition of churn in the case study is based on it. This simple definition is adequate for the study and makes it easy to detect the exact moment of churn. The customers without a C1 indicator before the time period were not included in the analysis. Their volume in the dataset is small. In banking sector a customer who does leave, may leave an active customer id behind because bank record formats are dictated by legislative requirements.

The definition of churn, presented above, produced relatively small amount of customers to be considered churners. On average there were less than 0.5% customers in each time step to be considered churners.

This problem has been identified in the literature under term class imbalance problem [10] and it occurs when one class is represented by a large number of examples while the other is represented by only a few. The problem is particularly crucial in an application, such as the present one, where the goal is to maximize recognition of the minority class [4]. In this study a down-sizing method was used to avoid all predictions turn out as nonchurners. The down-sizing (undersampling) method consists of the randomly removed samples from the majority class population until the minority class becomes some specific percentage of the majority class [3]. We used this procedure to produce two different datasets for each time step: one with a churner/nonchurner ratio 1/1 and the other with a ratio 2/3.

In this study we use binary predictions, *churn* and *no churn*. A logistic regression method [5] was used to formulate the predictions. The logistic regression model generates a value between bounds 0 and 1 based on the estimated model. The predictive performances of the models were evaluated by using lift curve and by counting the number of correct predictions.

3 Results

A collection of six different regression models was estimated and validated. Models were estimated by using six different training sets: three time periods (4, 6, and 8) with two datasets each. Three time periods (t = 4, 6, 8) were selected for the logistic regression analysis. This produced six regression models which were validated by using data sample 3 (115 000 customers with the current account indicator). In the models we used several independent variables, these variables for each model are presented in the table 2. The number of correct predictions is presented in each model in the table 3. In the validation we used the same sample with the churners before the time period t=9 removed and the data for validation was collected from time periods t(9) - t(11).

Table 2. Predictive variables that were used in each of the logistic regression models. Notion X_1 marks for training dataset with a churner/nonchurner ratio 1/1 and X_2 for a dataset with a ratio 2/3. The coefficients of variable in each of the models are presented in the table.

Model	4_1	4_2	6_1	6_2	8_1	8_2
Constant	-	-	0.663	-	0.417	-
Customer age	0.023	0.012	0.008	0.015	0.015	0.013
Customer bank age	-0.018	-0.013	-0.017	-0.014	-0.013	-0.014
Vol. of (phone) payments in t=i-1	-	-	-	-	0.000	0.000
Num. of trasactions (ATM) in t=i-1	0.037	0.054	-	-	0.053	0.062
Num. of trasactions (ATM) in t=i	-0.059	-0.071	-	-	-0.069	-0.085
Num. of transactions (card payments) t=i-1	0.011	0.013	-	0.016	0.020	0.021
Num. of transactions (card payments) t=i	-0.014	-0.017	-	-0.017	-0.027	-0.026
Num. of transactions (direct debit) t=i-1	0.296	0.243	0.439	0.395	-	-
Num. of transactions (direct debit) t=i	-0.408	-0.335	-0.352	-0.409	-	-
Num. services, (not current account)	-1.178	-1.197	-1.323	-1.297	-0.393	-0.391
Salary on logarithmic scale in t=i	0.075	0.054	-	-	-	-

Although all the variables in each of the models presented in the table 2 were significant there could still be correlation between the variables. For example in this study the variables *Num. of transactions (ATM)* are correlated in some degree because they represent the same variable only from different time period. This problem that arises when two or more variables are correlated with each other is known as multicollinearity. Multicollinearity does not change the estimates of the coefficients, only their reliability so the interpretation of the coefficients will be quite difficult [13]. One of the indicators of multicollinearity is high standard error values with low significance statistics. A number of formal tests for multicollinearity have been proposed over the years, but none has found widespread acceptance [13].

The lift curve will help to analyze the amount of true churners that are discriminated in each subset of customers. In the figure 1 the % identified churners

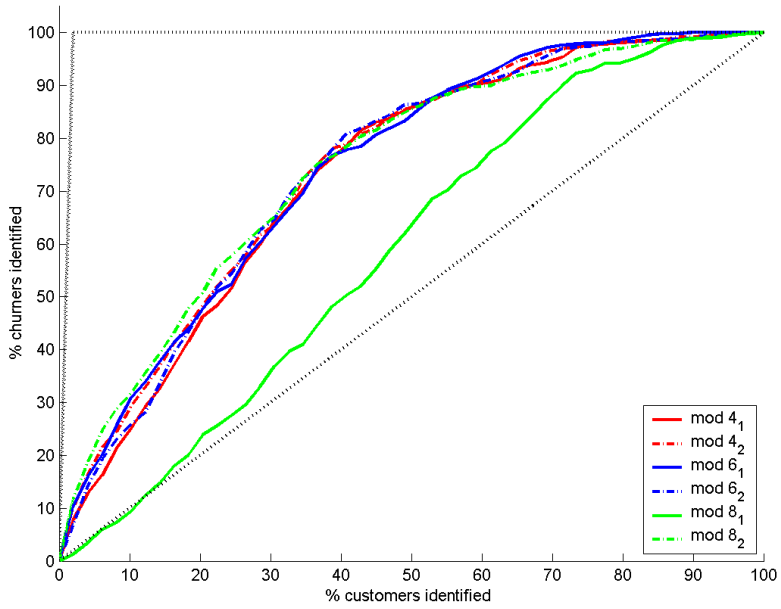


Fig. 1. Lift curves from the validation-set (t=9) performance of six logistic regression models. Model number (4, 6, and 8) represents the time period of the training set and (1 and 2) represent the down-sizing ratio.

Table 3. Number and % share of the correct predictions (mean from the time periods t=9, 10, 11). In the validation sample there were a 111 861 cases. The results were produced by the models when the threshold value 0.5 was used.

Model	Number of correct predictions	% correct predictions	% churners in the predicted set	% true churners identified as churners
model 4_1	69670	62	0.8	75.6
model 4_2	81361	72	0.9	60.5
model 6_1	66346	59	0.8	79.5
model 6_2	72654	65	0.8	73.4
model 8_1	15384	14	0.5	97.5
model 8_2	81701	73	0.9	61.3

are presented based on each logistic regression models. The lift curves were calculated from the validation set performance. In the table 3 the models 4_1, 6_1, and 6_2 have correct predictions close to 60% where models 4_2 and 8_2 have above 70% of correct predictions. This difference between the five models has vanished when amount of correct predictions is analyzed in the subsets as is presented in the figure 1.

4 Conclusions

In this paper a customer churn analysis was presented in consumer retail banking sector. The different churn prediction models predicted the actual churners relatively well. The findings of this study indicate that, in case of logistic regression model, the user should update the model to be able to produce predictions with high accuracy since the independent variables of the models varies. The customer profiles of the predicted churners weren't included in the study.

It is interesting for a company's perspective whether the churning customers are worth retaining or not. And also in marketing perspective what can be done to retain them. Is a three month time span of predictions enough to make positive impact so that the customer is retained? Or should the prediction be made for example six months ahead?

The customer churn analysis in this study might not be interesting if the customers are valued based on the customer lifetime value. The churn definition in this study was based on the current account. But if the churn definition was based on for example loyalty program account or active use of the internet service. Then the customers at focus could possibly have greater lifetime value and thus it would be more important to retain these customers.

References

1. Au W., Chan C.C., Yao X.: A Novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. on evolutionary comp. **7** (2003) 532–545
2. Buckinx W., Van den Poel D.: Customer base analysis: partial detection of behaviorally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research **164** (2005) 252–268
3. Chawla N., Boyer K., Hall L., Kegelmeyer P.: SMOTE: Synhetic minority oversampling technique. Journal of Artificial Research **16** (2002) 321–357
4. Cohen G., Hilario M., Sax H., Hugonnet S., Geissbuhler A.: Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine **37** (2006) 7–18
5. Cramer J.S.: The Logit Model: An Introduction. Edward Arnold (1991). ISBN 0-304-54111-3
6. Ferreira J., Vellasco M., Pachecco M., Barbosa C.: Data mining techniques on the evaluation of wireless churn. ESANN2004 proceedings - European Symposium on Artificial Neural Networks Bruges (2004) 483–488
7. Garland R.: Investigating indicators of customer profitability in personal retail banking. Proc. of the Third Annual Hawaii Int. Conf. on Business (2003) 18–21

8. Hwang H., Jung T., Suh E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications **26** (2004) 181–188
9. Keaveney S., Parthasarathy M.: Customer Switching Behaviour in Online Services: An Exploratory Study of the Role of Selected Attitudinal, Behavioral, and Demographic Factors. Journal of the Academy of Marketing Science **29** (2001) 374–390
10. Japkowicz N., Stephen S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6** (2002) 429–449
11. Lester L.: Read the Signals. Target Marketing **28** (2005) 45–47
12. Mozer M. C., Wolniewicz R., Grimes D.B., Johnson E., Kaushansky H.: Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunication Industry. IEEE Transactions on Neural Networks, (2000)
13. Pindyck R., Rubinfeld D.: Econometric models and econometric forecasts. Irwin/McGraw-Hill (1998). ISBN 0-07-118831-2.
14. Rosset S., Neumann E., Eick U., Vatnik N., Idan Y.: Customer lifetime value modeling and its use for customer retention planning. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Canada (2002) 332-340

Predictive Marketing

Kampagnenübergreifende Marketingoptimierung und Echtzeitempfehlungen im Inbound

Jochen Werner

SPSS GmbH Software
Rosenheimer Strasse 30
81669 München
www.spss.com/de

Das traditionelle Marketing fokussiert in der Regel auf das Produkt, bzw. auf die einzelnen Kampagnen. Dort wird bereits häufig mit Data Mining eine recht hohe Optimierung von einzelnen Kampagnen erreicht, allerdings wird in der Regel keine übergreifende Optimierung mehrerer Kampagnen über die Zeit vorgenommen, noch werden Inboundkanäle wie Call Center und Web für zielgerichtete Kampagnen genutzt. Gerade dort liegt aber noch ein sehr hohes Optimierungspotential, welches SPSS mit Predictive Marketing adressiert!

1. Der Wechsel vom traditionellen Marketing zum Multi-Channel Marketing

Das traditionelle Marketing ist in der Regel stark produktorientiert und versucht im Rahmen der Kampagnenoptimierung den besten Kunden für die nächste durchzuführende Marketingaktion zu finden. Diese Aktionen werden sequentiell durchgeführt und optimiert (first comes, first serves). Dabei werden die klassischen Outbound Kanäle, wie Direkt Mail, Telemarketing und eMail genutzt.

Diese Vorgehensweise lässt viel Potential ungenutzt, da durch den Fokus auf Produkt und Aktion sequentiell die jeweils besten Kunden pro Aktion selektiert und adressiert werden (Bester Kunde pro Aktion), aber nicht die kommenden Aktionen optimiert für die jeweiligen Kunden zum optimalen Zeitpunkt durchgeführt werden (Beste Aktion pro Kunde)!

In Abhängigkeit der jeweiligen Umsetzung können sich daraus die folgenden Probleme ergeben:

- Der Kunde erhält zu viele und überlappende Aktionen
  - Kunden werden mit zu vielen Inhalten überflutet und reagieren überhaupt nicht mehr
- Der Zeitpunkt für die Aktion ist nicht optimal
  - Obwohl für die Aktion affin, reagiert der Kunde nicht, bzw. hat bereits auf eine andere Aktion reagiert!
- Nicht die Aktion mit dem höchsten Gewinn und der höchsten Affinität wird zuerst gefahren
  - Umsatzpotential geht verloren

Traditionelles Marketing ➡ **Multi-Channel Marketing**

Traditionelles Marketing	Multi-Channel Marketing
▪ Produkt fokussiert	▪ Fokus auf den Kunden
▪ Suche nach dem besten Kunden für ein Produkt / Aktion	▪ Suche nach dem besten Produkt / Aktion für jeden einzelnen Kunden
▪ Ausschließlich Outbound Kanäle ▪ Direkt Mail ▪ Telemarketing ▪ eMail	▪ Inbound und Outbound Kanäle: ▪ Direkt Mail ▪ Telemarketing ▪ eMail ▪ Service Call Centers ▪ Web ▪ Chat

2. Übergreifende Kampagnenoptimierung

2.1. Traditionelle Kampagnenoptimierung

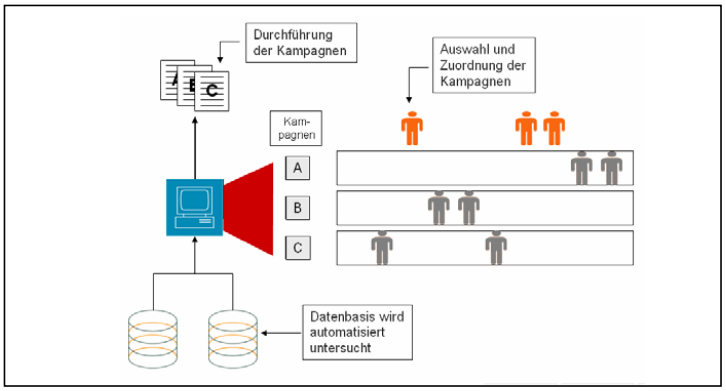
Die aufgeführte Problemstellung ist im heutigen Marketing durchaus bekannt und man versucht dies durch die Definition von Regeln (z.B. kein Kunde wird öfter als 1x im Quartal angeschrieben) und/oder händischer Optimierung zu adressieren. Die Kampagnen werden priorisiert und sequentiell durchgeführt. Dies führt zu einem sehr hohen Potentialverlust, da nun zwar keine Überlappungen mehr existieren, aber durch die Selektion nach bester Kunde pro Kampagne, nicht die besten (gewinnbringendsten) Kampagnen den Kunden zugeordnet werden.

2.2. Übergreifende Kampagnenoptimierung

Bei der übergreifenden Kampagnenoptimierung geht es um die Nutzung von Prognosemodellen für eine weitere kampagnenübergreifende Optimierung.

Im Einzelnen werden:

- Die richtige Kampagne für jeden einzelnen Kunden bestimmt
- Kampagnen nach dem gewünschten Ergebnis ausgewählt – z. B. maximaler Return on Investment (RoI) oder eine beliebige Kombination benutzerdefinierter Kriterien
- Die richtigen Kanäle für die einzelnen Kampagnen und Kunden festgelegt
- Jeden Kunden zum optimalen Zeitpunkt kontaktiert
- Alle Kampagnen kanalübergreifend koordiniert, um Überschneidungen zu vermeiden und die Responsequoten zu steigern



2.3. Kombination aus Geschäftswissen und leistungsstarker Analyse

Die kampagnen- und kanalübergreifende Steuerung bedingt eine Berücksichtigung von betriebswirtschaftlichen Rahmenbedingungen. Es werden Kundeninformationen und Vorhersagemodelle aus Data Mining oder Statistik Anwendungen genutzt und darüber hinaus nach Bedarf weitere betriebswirtschaftliche Regeln, wie etwa Einschränkungen für die Kanalauslastung, das Budget und die Kundenkontakte, sowie zusätzliche Informationen, wie Zielvorgaben oder bevorzugte Gruppen festgelegt. Als Ergebnis können die Auswirkungen von Kampagnen auf Kosten und Erträge simuliert und prognostiziert werden. Die Ergebnisse müssen dann in Form von leicht verständlichen betriebswirtschaftlichen Konzepten präsentiert werden.

2.4. Zusammenfassung

Die übergreifende Kampagnenoptimierung setzt auf bestehende Prognosemodelle auf und ist als Ergänzung zu einem vorhandenen Kampagnenmanagement-System zu sehen. Es können deutlich gesteigerte Responsequoten bei unverändertem Kampagnenvolumen erzielt werden.

2.5. Automatisierte Auswahl des richtigen Kanals und des richtigen Zeitpunktes

Der Kanal, über den ein Kunde ein Angebot erhält, kann ebenso wichtig sein, wie das Angebot selbst. Der eine reagiert vielleicht eher auf eine E-Mail, der andere klickt sofort auf die Löschtaste, würde dasselbe Angebot im Geschäft aber sofort annehmen. Bei der übergreifenden Optimierung wird der beste Kanal für jeden Kunden und jede Kampagne ermittelt, um die Wahrscheinlichkeit einer positiven Reaktion zu erhöhen. Die kanalübergreifende Optimierung verhindert, dass ein Kunde mehrere Angebote erhält. Stößt ein Kanal an seine Grenzen, wird die Kampagne auf einem Alternativkanal zu Ende geführt.

Neben der kanalübergreifenden Optimierung müssen interne Regeln für den Kundenkontakt sowie interne und externe Auflagen (beispielsweise behördliche Ausschlusslisten) eingehalten werden. Damit wird ausgeschlossen, dass Kunden mit Angeboten überhäuft oder über unerwünschte Kanäle kontaktiert werden.

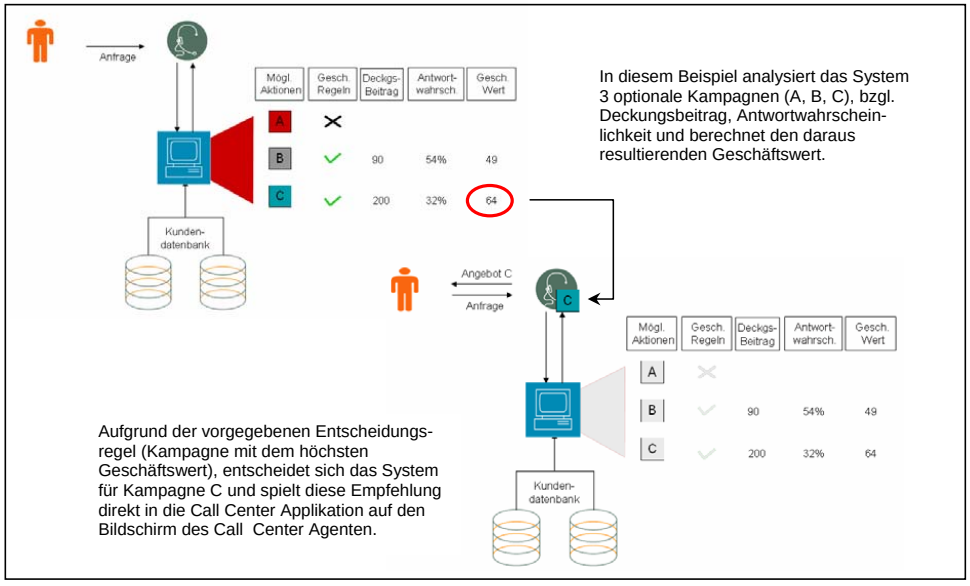
Ebenso hat der richtige Zeitpunkt einen entscheidenden Einfluss auf den Erfolg einer Aktion. Eine intelligent getimte Kampagne kann eine Abwanderung zur Konkurrenz verhindern oder den Verkauf eines höherwertigen Produkts (Up-Sell-Geschäft) ermöglichen. Änderungen im Kundenverhalten, wie etwa eine sinkende Anzahl von Einkäufen, oder ein wiederholtes Überschreiten des monatlichen Minutenkontingents sind wichtige Signale. Es wird automatisch nach Änderungen im Kundenverhalten („Ereignisse“ genannt) gesucht, die auf eine unerfüllte Nachfrage oder einen möglichen Wertverlust hindeuten, und situationsbezogen die beste Kampagne ausgewählt. Die Kunden erhalten solcherart zeitnah Angebote, die auf ihren speziellen Bedarf zugeschnitten sind.

3. Nutzung von Inbound Kanälen für gezielte Marketingkampagnen

Die Nutzung von prädiktiven Modellen erfolgt heute primär für Outbound Maßnahmen wie z.B. Optimierung von Kampagnen und Kundenbindungsmaßnahmen. Hauptgrund hierfür ist die mangelnde zeitnahe Verfügbarkeit von Analysen und Prognosen in den klassischen Inboundbereichen wie Call Center oder Webseite. Mit Echtzeitlösungen (z.B. SPSS) können Handlungsanweisungen und Empfehlungen an Call Center Mitarbeiter oder einen Besucher der Unternehmenswebseite gegeben werden. Damit können diese Kanäle für gezieltes Cross- und Upselling, Kundenbindung, Betrugserkennung und ähnliches genutzt und deutliche Umsatzsteigerungen oder Kostenreduktionen erzielt werden.

3.1. Call Center – von einem Cost Center zum Profit Center

Bei der Echtzeitanalyse werden eingehende Anrufer unmittelbar identifiziert und durch weitere in diesem Moment gestellte Qualifizierungsfragen (z.B. geht es um eine Produktanfrage, Beschwerde, Auskunft, etc.) ein Echtzeitscoring vorgenommen. Dies bedeutet, dass der Anrufer als Ergebnis der in Echtzeit ablaufenden Analyse, auf Basis seiner historischen und der aktuellen Informationen einem Kundensegment zugeordnet wird. Das System spielt dann dem Call Center Agenten eine Handlungsempfehlung auf den Bildschirm, die dieser Person und der aktuellen Situation entspricht. Dies kann z.B. für Produktempfehlungen, Cross Selling, Kundenbindungsmaßnahmen, etc. genutzt werden. Wesentlich ist, dass der Anrufer nicht irgendein Angebot erhält, sondern ein auf ihn und die jeweilige Situation zugeschnittenes Angebot, auf welches er mit sehr hoher Wahrscheinlichkeit positiv reagiert.



Darüber hinaus ist es gegebenenfalls sinnvoll neben der höchsten Antwortwahrscheinlichkeit, auch noch Geschäftsregeln als Entscheidungskriterium mit einfließen zu lassen. Das in der obigen Abbildung gezeigte Beispiel hat z.B. als Entscheidungskriterium den Geschäftswert, welcher sich aus dem Deckungsbeitrag und der Antwortwahrscheinlichkeit berechnet. Aufgrund der hinterlegten Präferenz für den höchsten Geschäftswert, entscheidet sich das System nun für Aktion C. Hier ist die Antwortwahrscheinlichkeit zwar niedriger als bei Aktion B, allerdings ist der Deckungsbeitrag mehr als doppelt so hoch, so dass diese Aktion sinnvoller ist als Kampagne B.

3.2. Nutzung der Webseite für eine gezielte Kundenansprache

Dieselbe Systematik kann auch für Empfehlungen über die Webseite eines Unternehmens genutzt werden. Vorraussetzung ist ein personalisierter Bereich, auf welchem sich die Besucher vorab identifizieren (z.B. Kundennummer) müssen. Analog zum Vorgehen im Call Center blendet das System eine auf diese Person optimierte Aktion ein. Ebenfalls besteht die Möglichkeit durch vorab gestellte Qualifizierungsfragen, das Angebot weiter zu verfeinern bzw. situationsgerecht zu machen.

4. Fazit

Um weitere Effizienzsteigerungen bei der Kampagnenoptimierung zu erzielen und das mittlerweile immer stärker überhand nehmende ‚Overmailing' zu vermeiden ist das Marketing gefordert über neue Ansätze nachzudenken. Die aufgeführten Möglichkeiten bieten Unternehmen die Möglichkeit deutlich höhere Responseraten zu erzielen, Inboundkanäle effektiv für Kampagnen zu nutzen und damit erhöhte Umsätze und Wettbewerbsvorteile zu erzielen.

Customer churn prediction using knowledge extraction from emergent structure maps

Alfred Ultsch and Lutz Herrmann

Philipps-University of Marburg
Department of Mathematics and Computer Science
{ultsch,lherrmann}@informatik.uni-marburg.de

Abstract. Typical approaches to customer relationship management (CRM) construct a black box model for the prediction of churning. The approach presented here aims at understandable insights into the phenomena of the particular customer-business relationship. Emergent Structure Maps (ESM) are a visualization technique for distance and density structures in high dimensional data spaces. We report the application of a combination of ESM visualization methods and knowledge extraction techniques to mobile phone customer data. Aim of this approach is to discover understandable knowledge which can be used for the prevention of churn. Output of our approach are human intellegibe descriptions of who the churning and non churning customers are, what profitable customers are and why churning is happening.

1 Introduction

Churn means the discontinuation of a contract with a business. In many European countries market structures of wireless telecommunications have changed in the last years from monopolistic to very competitive markets. To be able to control and reduce the customer churn rate may be a vital factor for the survival of mobile phone companies. If it is possible to predict whether a customer is likely to churn in two months, given today's customers record, appropriate action on the side of the business may prevent the loss of the customer. Central issues for churn management are:

- What types of customers are likely to quit?
- Are there subgroups of potential churners that are so promising that further efforts to keep the customers are worthwhile?

To know *why* a profitable customer decides to discontinue a contract is essential to better tailor products to profitable marked segments. Business objectives for churn predictions are to concentrate marketing activities on such customers that are likely to produce profit by remaining with a company for longer periods of time and to retain profitable customers.
In this short paper we report the application of our ESM knowledge discovery method to data from a Swiss mobile phone company. Main focus was to find knowledge which can be used to predict and ultimately prevent churning.

2 Emergent Stucture Maps

Emergent Stucture Maps (ESM, [4]) show distance and density structures of high dimensional data. ESM are constructed on top of a projection of high dimensional data onto a two dimensional retina. A retina is a set of so called neurons arranged on a grid. Each neuron is associated with a fixed position on the finite grid and a vector in the data space (prototype vector). Emergent self organizing maps (ESOM) can be used to construct the projection, but other projection methods can also be used too [4]. The usage of ESOM has the advantage that potential clusters in the data are preserved [2]. On top of the retina, distance and density structures are visualized using U- and P-Matrix technologies. Clusters in the data can be found using U*C, a density and distance driven segmentation algorithm [5].

Distance visualization: The canonical distance visualization technique is the U-Matrix. The local distances between neighbouriong neurons are displayed on top of each neuron as an elevation creating a 3D landscape. U-Matrices have the following properties:

- neighbouring data in the high-dimensional input data space lie in a common valley
- gaps in the distribution of input samples produce hills on the U-Matrix

Density visualization: Sometimes distances are not enough to see structures in the data. In mixture data sets clusters are rather defined by local maxima in densites than by distances. A P-Matrix [4] displays a local density measure using Pareto Density Estimation (PDE), an information optimal kernel density estimation. In analogy to the U-Matrix, P-Matrices are also displayed as 3D landscapes.

Combined distance and density visualization: The U*-Matrix combines distance and demnsity information by modifying the heights of the U-Matrix according to the P-Matrix [4]. The values of the U-Matrix are lowerd in highly dense regions, unchanged in regions of average density, and emphasized in sparse regions.

Clustering: The U*C cluster algorithm [5] is used for detection of clusters. U*C uses the information of U-Matrix and P-Matrix to construct cluster borders and cluster cores. This results in a distance and density based clusteralgorithm that is able to cluster even linear not separable clusters. The number of clusters is automatically detected and outliers are identified.

For more applications of U*-Matrices and the U*C algorithm see [4] [5]. In the application described below 47 clusters of customers could be found.

3 Knowledge Extraction from ESM

Knowledge Discovery (KD) can be defined as the process of discovering new, understandable and useful knowledge in data sets [3]. Applying KD to data sources created by the customer/business interaction might

therefore be effectively used to acquire valuable knowledge about a customer. Of particular interest is to know in beforehand which customer will quit and what the motives for quitting are.
Structures discovered by ESM might be new, but in general they are not immediately useful for CRM. The structures by themselves do not represent knowledge yet. As we understand it, knowledge means statements about the data set that are understandable by humans, that are also interpretable by a knowledge based system and, in particular, have a meaning to the business. I.e. knowledge must lead to a non-trivial understanding of important features of the data generating process. Knowledge, in this sense, can be extracted from ESM structures using sig* techniques [3].
The algorithm sig* operates on the groups identified by U*C and produces a description of these groups in the form of understandable decision rules [3]. In contrast to other decision rule algorithms, like decision tree inference, the understandability of the generated is the main aim of this algorithm. With the knowledge extraction algorithm sig* rules for each of these groups could be extracted from the ESM. Sig* produces a description for each group in the form of characterising and differentiating rules.
The Data used for this study consisted of a sample of 300.000 customer data records. The data was provided by Swisscom AG, Bern Switzerland. We used 21 variables concerning usage of the Swisscom Mobile telecommunication network. The variables described aspects of customer behaviour like spent money, usage of services (like SMS), usage of different networks, usage times and destinations of long distance calls. With ESM, U*C and sig* technology, 47 meaningful groups could be found in the data.

4 Churn Prediction & Discovered Knowledge

Churners were defined as customers that discontinued at least one contract with Swisscom Mobile in month m. For these customers the data of the second preceding month $m-2$ was selected and called Churn Data. This Churn Data was also classified using ESM.
The sig* rules for the different groups were analysed and it was found that the same variables were the most important for several groups. So these groups could be aggregated to classes with common features. The 47 groups could be aggregated to 9 classes having common usage profiles.
The rules generated by sig* for the Churn Data turned out to be effective predictors for churning. An overall accuracy for the prediction rules was measured to be 99.8%. The overall churn rate is not to be published in here in order to keep confidentiality but it should be mentioned that our method turned out to create effective predictors for churning.
A knowledge based system can interpret the rules generated by sig*. This results in a classifier for mobile phone customers. Sensitivity and specifity of the classifier is close to 100%for all rules.
The description of the 47 groups respectively 9 classes in the form of rules allows to understand what type of customers are using the Swisscom mobile phone network. The conclusion was drawn that customers

of Swisscom Mobile are rather characterised by the usage of the network than by their social or economic background. Finally, a comparison of the distribution of the groups among Churners and Non Churners reveals, however, interesting properties. Since all groups have a meaningful description in terms of network usage the reasons of a churning decision can be understood.

5 Discussion & Summary

The presented approach aims at the understanding of the phenomenon churning in a particular business. Product of the method presented here using ESM, U*C and knowledge extraction (sig*) is a set of rules that lead to a better characterisation of the customer groups that are likely to churn or not to churn. Furthermore among these groups segments that are more profitable than others can be identified. In the case presented here concrete characterisations of a large numbers of churners could be derived by the interpretation of the rules. Furthermore product groups could be identified that were likely to attract new customers but are not attractive to build up a longer lasting business customer relationship. In particular customers that used (or did hardly use) certain services and/or made phone calls from/to certain countries could be identified to be churners. Special offers could be identified that were bought by persons who remained only short time customers became therefore unprofitable customers.
Therefore, the presented method is knowledge discovery in the sense of producing new (formerly unknown) knowledge about a business' customers out of the data generated by customer-business interaction.

References

1. S.A. Brown (Ed) : Customer Relationship Management: A Strategic Imperative in the World of E-Business, John Wiley & Sons, 2000
2. T. Kohonen: Self-Organized Formation of Topologically Correct FeatureMaps, Biological Cybernetics Vol. 43, pp 59 - 69, 1982
3. A. Ultsch: The Integration of Neural Networks with Symbolic Knowledge Processing, in Diday et al. "New Approaches in Classification and Data Analysis", pp 445 - 454, Springer Verlag 1994
4. A. Ultsch, F. Mörchen: ESOM-Maps - tools for clustering, visualization, and classification with Emergent SOM, Technical Report No. 46, Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2005
5. A. Ultsch: Clustering with SOM: U*C, In Proceedings Workshop on Self-Organizing Maps (WSOM 2005), Paris, France, 2005

Online Ensembles for Financial Trading

Jorge Barbosa[1] and Luis Torgo[2]

[1] MADSAD/FEP, University of Porto,
R. Dr. Roberto Frias, 4200-464 Porto, Portugal
jorgebarbosa@iol.pt
[2] LIACC-FEP, University of Porto, R. de Ceuta, 118, 6., 4050-190 Porto, Portugal
ltorgo@liacc.up.pt,
WWW home page: http://www.liacc.up.pt/~ltorgo

Abstract. This paper describes an application of online ensembles to financial trading. The work is motivated by the research hypothesis that it is possible to improve the performance of a large set of individual models by using a weighted average of their predictions. Moreover, we explore the use of statistics that characterize the recent performance of models as a means to obtain the weights in the ensemble prediction. The motivation for this weighting schema lies on the observation that, on our application, the performance ranking of the models varies along the time, i.e. a model that is considered the "best" at a time t will frequently loose this position in the future. This work tries to explore this diverse and dynamic behavior of models in order to achieve an improved performance by means of an online ensemble. The results of our experiments on a large set of experimental configurations provide good indications regarding the validity of the dynamic weighting schema we propose. In effect, the resulting ensembles are able to capture a much larger number of trading opportunities with a signal accuracy similar to the best individual models that take part on the ensemble.

1 Introduction

Financial markets are highly dynamic and complex systems that have long been the object of different modelling approaches with the goal of trying to anticipate their future behavior so that trading actions can be carried out in a profitable way. The complexity of the task as well as some theoretical research, have lead many to consider it an impossible task. Without entering this never ending argumentation, in this paper we try to experimentally verify a hypothesis based on empirical evidence that confirms the difficulty of the modelling task. In effect, we have tried many different modelling approaches on several financial time series and we have observed strong fluctuations on the predictive performance of these models. Figure 1 illustrates this observation by showing the performance of a set of models on the task of predicting the 1-day future returns of the Microsoft stock. As we can observe the ranking of models, according to the used performance measure (NMSE measured on the previous 15 days), varies a lot,

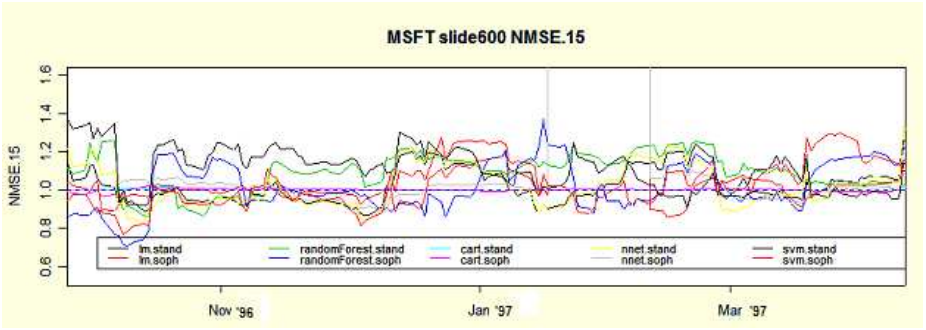


Fig. 1. The performance of different models on a particular financial time series.

which shows that at any given time t, the model that is predicting better can be different.

Similar patterns of performance where observed on other time series and using several other experimental setups. In effect, we have carried out a very large set of experiments varying: the modelling techniques (regression trees, support vector machines, random forests, etc.); the input variables used in the models (only lagged values of the returns, technical indicators, etc.); the way the past data was used (different sliding windows, growing windows, etc.). Still, the goal of this paper is not the selection of the best modelling approach. The starting point of this work are the predictions of this large set of approaches that are regarded (from the perspective of this paper) as black boxes. Provided these models behave differently (i.e. we can observe effects like those shown in Figure 1), we have a good setup for experimentally testing our hypothesis.

Our working hypothesis is that through the combination of the different predictions we can overcome the limitations that some models show at some stages. In order to achieve this, we claim that the combination should have dynamic weights so that it is adaptable to the current ranking of the models. This means that we will use weights that are a function of the recent past performance of the respective models. This way we obtain some sort of dynamic online ensembles, that keep adjusting the ensemble to the current pattern of performance exhibited by the individual models.

There are a few assumptions behind this hypothesis. First of all, we are assuming that we can devise a good characterization of the recent past performance of the models and more over that this statistic is a good indicator (i.e. can serve as a proxy) of the near future performance of the models. Secondly, we must assume that there is always diversity among the models in terms of the performance at any time t. Thirdly, we must also assume that there will always be some model performing well at any time t, otherwise the resulting combination could not perform well either.

Obviously, the above assumptions are what we could consider an ideal setup for the hypothesis to be verified in practice. Several of these assumptions are

difficult to meet in a real world complex problem like financial trading, left alone proving that they always hold. The work we present here can be regarded as a first attempt to experimentally test our hypothesis. Namely, we propose a statistic for describing the past performance of the models and then evaluate an online ensemble based on a weighing schema that is a function of this statistic on a set of real world financial time series.

2 The Dynamic Weighting Schema

Traders do not look for models that achieve a good average predictive accuracy. The reason lies on the fact that the most common return on a stock price is around zero. As such, predicting well on average usually means being very good at predicting these zero-like future returns. However, these predictions are useless for a trader as no one can earn money with such short variations on prices due to the transactions costs. As such, traders are more interested in models that predict well the larger but rare variations [3]. Based on these arguments we have decided not to use a standard statistic of prediction error, like for instance the Normalized Mean Squared Error (NMSE), as an indicator of the past performance of a model. Instead, we have used a measure that is more useful for trading. We have used two thresholds on the predicted return that determine when buy (sell) signals would be issued by an hypothetical trader. These thresholds can be seen as creating 3 bins on the range of returns. In effect, if the predicted future return $\hat{R}$ is above (below) a threshold $\alpha(\mu)$ we generate a **buy** (**sell**) signal, otherwise we have a **hold** signal. This process creates a discretized target variable, the predicted signal.

Using this discretization process we can then calculate statistics regarding the signals predicted by the models. Namely, we have calculate the *Precision* of the predictions as the proportion of buy (sell) signals that are correct (i.e. correspond to returns that really overcame the thresholds). We have also calculated the *Recall* as the proportion of true signals (i.e. real returns that were above (below) the threshold) that are signaled as such by the models. Finally, we have combined these two measures into a single statistic of performance using the F-measure [2].

The F-measure calculated on the previous 15 days was the measure of past performance that we have used to obtain the weights of each model in the ensemble. Namely, the combined prediction at any time t was obtained by,

$$\hat{R}_{t+1,ens} = \frac{\sum_{k=1}^{S} \hat{R}_{t+1,k} \times F.15_k}{\sum_{k=1}^{S} F.15_k} \tag{1}$$

where $F.15_k$ is the value of the F-measure calculated using the predictions of model k for the time window $[t-15..t]$.

3 Experiments and Results

We have carried out a very large set of experiments designed to test several instantiations of our working hypothesis [1]. Due to space restrictions we will

Table 1. The results for the DELL stock.

	±1%		±1.5%		±2%	
	Prec	Rec	Prec	Rec	Prec	Rec
lm.stand	0.405	0.077	**0.472**	0.017	0.364	0.006
lm.soph	**0.453**	0.173	0.460	0.045	0.519	0.014
randomForest.stand	0.434	0.227	0.402	0.080	0.328	0.026
randomForest.soph	0.420	0.321	0.399	0.154	0.345	0.075
cart.stand	0.444	0.004	0.444	0.004	**0.600**	0.003
cart.soph	0.231	0.004	0.154	0.004	0.154	0.004
nnet.stand	0.388	0.246	0.353	0.128	0.289	0.075
nnet.soph	**0.453**	0.063	0.360	0.037	0.323	0.030
svm.stand	0.391	0.380	0.357	0.190	0.326	0.075
svm.soph	0.415	0.360	0.397	0.189	0.373	0.097
Ensemble	0.438	**0.451**	0.354	**0.287**	0.296	**0.210**

limit our description to the ensembles formed by a weighted combination of predictions using Equation (1).

In our experiments we have tried three different setups in terms of thresholds on returns for generating trading signals, namely, $\pm 1.0\%, \pm 1.5\%$ and $\pm\, 2\%$.

We have compared our dynamic online ensembles with several individual models that participated on the ensemble. We present the results in terms of *Precision* and *Recall*, as different traders may require different tradeoffs between these two conflicting statistics. We have carried out this and other experiments on 9 different financial time series containing daily data for more than 10 years. The experiments were carried out over this large time period of time using a sliding window approach. Due to space limitations we can only show the results for one stock, DELL, which involves around 10 years of testing (approximately 2500 daily predictions). The reader is referred to [1] for full details.

Table 1 shows the results of the individual models and of the ensemble for the DELL stock. The first observation we can make is that the results are generally poor. The individual methods achieve quite low values of *Recall* and usually less than 50% of *Precision*. A possible explanation for these poor results is the fact that all methods are obtained by optimizing a criterion (usually some form of mean squared error) that is an average error estimator and thus will not be optimal for predicting the extreme low and high returns [3].

Although we can observe a few high *Precision* scores these are usually obtained with too few trades (very low *Recall*) to make these strategies worth investing[3].

Regarding the results of our dynamic ensembles we can generally state that they are interesting. In effect, we have a much higher value of *Recall* with a *Precision* that is most of the times near the best individual scores. The results

[3] Although *Recall* and *Precision* do not directly translate into trading results, they still provide good indications as they are more related to trading decisions than pure prediction error metrics like mean squared error, for instance.

in terms of *Recall* get more impressive as we increase the thresholds, i.e. make the problem even harder because increasing the thresholds means that trading signals are even more rare and thus harder to capture by models that are not designed to be accurate at predicting rare values. Still, the *Precision* of the ensemble signals also suffers on these cases, which means that the resulting signals could hardly be considered for real trading. Nevertheless, the results with the thresholds set to $\pm 1\%$ can be considered worth exploring in terms of trading as we get both *Precision* and *Recall* around 45%. The best individual model has a similar *Precision* but with only 17.3% *Recall*.

The same general pattern of results were observed on the experiments with other stocks. In summary, this first set of experiments provides good indications towards the hypothesis of using indicators of the recent performance of models as dynamic weights in an online ensemble in the context of financial markets.

4 Conclusions

In this paper we have explored the possibility of using online ensembles to improve the performance of models in a financial trading application. Our proposal was motivated on the observation that the performance of a large set of individual models varied a lot along the testing period we have used. Based on this empirical observation we have proposed to use dynamic (calculated on moving windows) statistics of the performance of individual models as weights in an online ensemble.

The application we are addressing has some particularities, namely the increased interest on accurately predicting rare extreme values of the stock returns. This fact, lead us to transform the initial numerical prediction problem into a discretized version where we could focus our attention on the classes of interest (the high and low returns that lead to trading actions). As a follow up we have decided to use *Precision* and *Recall* to measure the performance of the models at accurately predicting these trading opportunities. Moreover, as statistical indicator of recent past performance we have used the F-measure that combines these two statistics.

The results of our initial approach to dynamic online ensembles in the context of financial trading are promising.We have observed an increased capacity of the ensembles in terms of signaling the trading opportunities. Moreover, this result was achieved without compromising the accuracy of these signals. This means that the resulting ensembles are able to capture much more trading opportunities than the individual models.

As main lessons learned from this application we can refer:

- Diverse behavior of models on predicting complex dynamic systems: in problems with so many unrecorded factors influencing the dynamics of a system is hard to find a predictive model that is good for all regimes. We have observed similar beahviors on other problems like for instance predicting algae blooms in freshwater systems.

– Predicting extreme values is a completly different problem: we have confirmed our previous observations [3] that the problem of predicting rare extreme values is very different from the standard regression setup, and this demands for specific measures of predictive performance.
– Dynamic online ensembles are a good means to fight instability of individual models: whenever the characteristics of a problem lead to a certain instability on models' performance, the use of dynamic online ensembles is a good approach to explore the advantages of the best models at each time step.

As future work we plan to extend our study of recent past performance statistics. We also plan to explore this hypothesis using base models that are better tuned towards the prediction of rare extreme values.

References

1. J. Barbosa. Metodos para lidar com mudancas de regime em series temporais financeiras - utilizacao de modelos multiplos na combinacao de previsoes (in portuguese). Master's thesis, Faculty of Economics, University of Porto, 2006.
2. C. Van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.
3. L. Torgo and R. Ribeiro. Predicting rare extreme values. In W. Ng, editor, *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2006)*, number 3918 in Lecture Notes in Artificial Intelligence. Springer, 2006.

Bayesian Network Classifiers for Country Risk Forecasting

Ernesto Coutinho Colla[1], Jaime Shinsuke Ide[1], and Fabio Gagliardi Cozman[2]

[1] Escola de Economia de São Paulo, Fundação Getulio Vargas
Rua Itapeva, 474, 12nd floor - CEP 01332-000, São Paulo, SP - Brazil
[2] Escola Politécnica, Universidade de São Paulo
Av. Prof. Mello Moraes, 2231 - CEP 05508900, São Paulo, SP - Brazil

Abstract. We present a probabilistic model, learned with machine learning tools, to predict and describe Brazil's risk trends. Our main goal is to verify whether machine learning offers useful techniques to build economic models. We use Bayesian Networks to perform pattern recognition in observed macroeconomics and financial data, with promising results We start from expected theoretical relationships between country risk and economic variables, international economic context and market expectations; using those, we are able to classify Brazil's risk trend with at least 80%.

1 Introduction

The forecast of inflation, employment, economic activity and interest rates (that is, of macroeconomic variables) is extremely important to decide on corporate investment decisions, government policy and family consumption. Decision support tools that can handle such variables and construct economic forecasts are clearly useful.

Country risk ratings are important macroeconomic indicators as they summarize the *perceptions* of economic agent about economic and political stability [6]. This perception affects the country's direct investment flows and international loans [16], thus impacting on its domestic economics activities.

Our goal in this work is to forecast daily country risk ratings behavior using pattern recognition from economic and financial variables. We believe that this forecasting is very useful to build corporate, government and financial decision support tool. As the risk of a country cannot be directly measured, we have chosen to forecast one of the most adopted indicators of country risk, called the *Emergent Market Bound Index Plus* (EMBI+). This indicator is provided by J.P. Morgan and tracks total returns for traded external debt instruments in the emergent market. We use data about Brazil.

We employ Bayesian networks and their machine learning algorithms [10, 14] for economic forecasting based on pattern recognition. The present paper is a preliminary report on the promising (and challenging) use of Bayesian networks to rate country risks. We should note that other machine learning tools have been

investigated in the literature on economics. For example, neural networks have been extensively studied and successfully applied in finance [17] and economic forecasting [13, 3, 15]. We are interested in Bayesian networks as they have some advantages over neural networks, such as the ability to combine expert opinion and experimental data [10] — possibly the reason why Bayesian networks have been successfully applied in many areas such as medical diagnosis, equipment failure identification, information retrieval.

2 Bayesian network classifiers for risk rating

Our prediction task consists of producing a nominal value for the *class* variable Y (prediction of EMBI+ trend) given a set of *attribute* variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ that are economic indicators. We encode the joint distribution of Y and $\mathbf{X}$ into a Bayesian network and then compute $\arg\max_Y P(Y|\mathbf{X})$ for classification. The Bayesian network is learned from a data set D consisting of samples over $(\mathbf{X}, Y)$. There are in fact several learning approaches to learn the structure and the probabilities of a Bayesian network, including fixed and non-fixed structures, and different *score metrics* of distributions [2].

A simple Bayesian network classifier that performs rather well in practice is the Naive Bayes (NB) classifier [7]. The structure is fixed so that attributes are conditionally independent, given the class variable: $P(\mathbf{X}|Y) = \prod_{i=1}^{n} P(X_i|Y)$. An extension of the NB is the *Tree Augmented Naive Bayes* (TAN) classifier [14], which allows dependencies between attribute variables. The structure of a TAN is learned through a variant of the Chow and Liu algorithm [4].

3 Methodology and Data

We can summarize our methodology as follows.

1. Data acquisition. Our database consisted of values of 117 attributes between January 1999 and March 2006. It included daily, monthly data focused on Brazil (but it also included relevant international economic variables) and their variants like differences, variances, lags. The set of explanatory variables was derived from earlier empirical studies on currency risk, country risk, debt servicing capacity of countries and theoretical models of international borrowing in the presence of default risk [6, 8, 1]. For this initial work we used only quantitative (numerical) variables that can be obtained directly from central banks or other public means. No subjective variables were included despite the fact that research on this subject often includes indicators that reflect nonobservable variables such as indices of political stability. Our hypothesis is that quantitative variables reflect the agents perception about this nonobservable variables, so these variables were included implicitly.

In broad terms, our variables describe the country's fiscal responsibility, its exports and imports flows, its debt stocks, exchange levels and volatility, domestic and international markets expectations (each one reflects different aspect of macroeconomic environment). In Figure 1, we describe some of these variables.

The database included raw variables and built variables, for example, ratios and first differences obtained from original variables — hence obtaining at the end 117 different attribute variables.

Name	Description
Economical activity	
PIMBR	Industrial production index based on the monthly industrial physical production
PIMBR_DZ	Industrial production index based on the monthly industrial physical production seasonally adjusted
Fiscal balance	
DLSPTOT_PIB	Public sector net debt (% of GDP)
NFSP_TOT_NOM	Total nominal fiscal deficit of the public sector (% of GDP)
NFSP_TOT_PRIM	Primary result of the public sector (% of GDP)
Monetary variables	
SELIC_12M	Annual Interest Rate (SELIC) (% year)
IPCA_12M	Annual inflation rate (IPCA) 12 month accumulated
External variables	
USTR_3M_D	U.S. government securities/Treasury - 3 months
OILD	Oil Price: Brent Europe Cash Price
Financial market expectation	
IBOVESPA	São Paulo Stock Exchange
DJONES	Monthly variation Dow Jones
interest curve	Defined using interest rates for different periods: DI30, DI60, DI90, DI180 and DI360
Trade balance and Exchange rates	
TC_PIB	Current account result (% of GDP)
PTAX	Official exchange rate
EXPORT_D1M_12M	Export average variation (last 12 months)
DIVEXT_EXPORT	Total Government external debt (% of export)
DIVTOT_EXPORT	Total Government debt (% of export)

Fig. 1. Description of collected attributes by group according to it's economical characteristics.

As our purpose was to predict the trend of EMBI+, we took special care in organizing the dataset so as to guarantee that the data used to prediction was exactly the information that the decision maker would have in a real situation. Figure 2 shows the EMBI+ time-series. The period between June to December of 2002 is specially volatile because of political factors (election period).

2. Data cleaning. We cleaned the data (using a Perl script) by removing samples with missing values, obtaining in the end 1483 instances (note that we could have used all data by processing the missing values say with the EM algorithm [2]).

3. Filtering and discretization. The exploratory data analysis showed $EMBI+$ high daily volatility, so the target classes for the classification process were created from a nominal discretization over a filtered $EMBI+$ series. The main reason for using the filtered dataset instead of the raw values is based on our intuition that for corporate investment and policymakers decisions it would

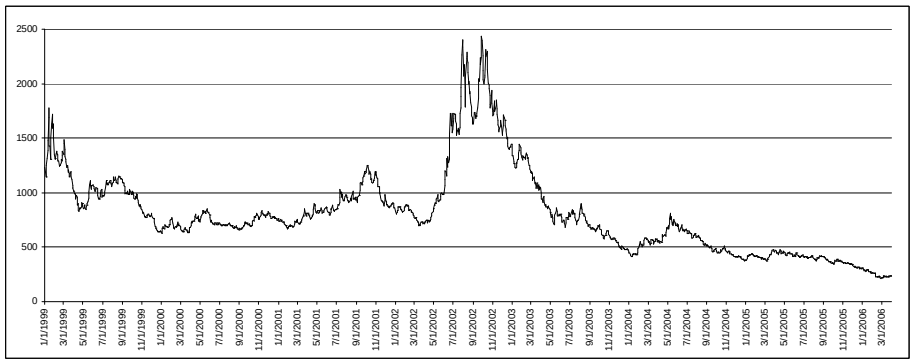


Fig. 2. EMBI+ series of period January/1999 to December/2005.

be useful to predict $EMBI+$ trend. As smoothing method we chose Hodrick-Prescott filter [11] once it is widely used among macroeconomists to estimate trend component of a series typically applied to estimate long-trend component and business cycle. It is worth mentioning that the Hodrick-Prescott filter allows adjustments to the level of smoothing controlled by parameter λ. Greater values of λ means stronger smoothing, and greater distance between the original and the Hodrick-Prescott filtered series. In Figure 3, it is possible to visualize how the filter Hodrick-Prescott smoothens the original EMBI+ series. As part of experiments, the classification of $EMBI+$ raw values demonstrated bad classifier performance with $\Delta EMBI_t$ prediction task[3] because the high volatility of EMBI+.

Once that our goal was to predict the trend we tried to forecast the *i-th percentual difference* of the filtered $EMBI+$ value (from now on $EMBIHP$) calculated as showed above:

$$\Delta EMBIHP_{t+i} = \frac{EMBIHP_{t+i} - EMBIHP_t}{EMBIHP_t} \tag{1}$$

where $EMBIHP_t$ is the value of current period and $EMBIHP_{t+i}$ is the value for i working days from current period.

As categorization criteria we adopted predefined *intervals of percentual variation* to categorize the direction and the magnitude of the $\Delta EMBIHP_{t+i}$. Considering the direction, $\Delta EMBIHP_{t+i}$ was categorized as *stable* if its absolute value is smaller than the adopted interval. If the absolute value exceed the predefined interval it was categorized as *down*, if its sign was negative, and *up*, otherwise. Considering the magnitude, $\Delta EMBIHP_{t+i}$ was categorized as *down* or *2down* if it had negative sign and its absolute value was respectively at least greater or one time greater than the predefined interval. Analogous categorization was done to positive values of $\Delta EMBIHP_{t+i}$. For empirical purposes based on this criteria the $\Delta EMBIHP_{t+i}$ was divided into 3 and 5 categories as Tables 1 and 2.

[3] An alternative task is to predict the $EMBI$'s level, as in [5] and good results are obtained.

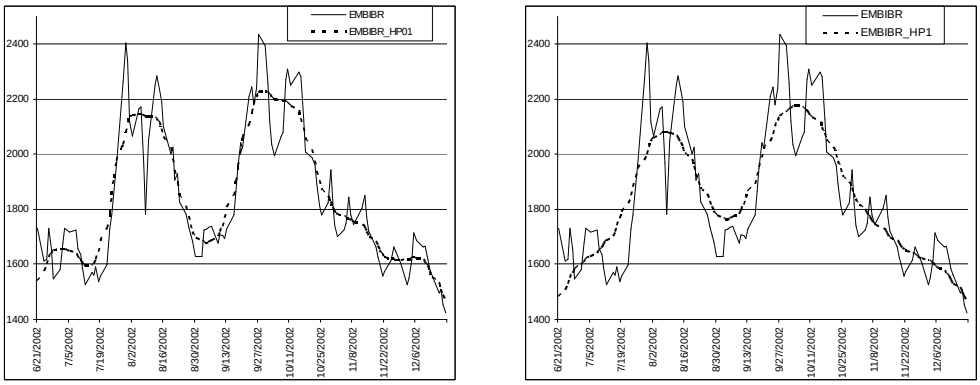


Fig. 3. EMBI+ series of period (June-December of 2002) and its filtered series with smoothing parameter $\lambda = 100$ (left side) and $\lambda = 1000$ (right side).

As can be seen in Tables 1 and 2, both the standard deviation of $\Delta EMBIHP_{t+i}$ and a fraction of it were used to define the intervals of nominal discretization. The main reason to use standard deviation was the well know association between variance and the uncertainty of predictions. Note that for each leading time i, we have different values for the standard-deviation sd. The variance of $\Delta EMBIHP_{t+i}$ increases in time, as does the uncertainty of a prediction, as can be checked in Table 3. We kept the same criterion of one standard deviation (and half of its value) to specify the level of uncertainty, no matter how far the prediction in time. The result of this decision was the loose of prediction quality for further periods.

Table 1. Class variable discretized into 3 categories. The value $\Delta EMBIHP_{t+i}$ is indicated by 'x'. The standard deviation of x is indicated by 'sd'.

Category	**down**	**stable**	**up**
Interval 1:	$x < -0.5sd$	$-0.5sd \leq x \leq 0.5sd$	$x > 0.5sd$
Interval 2:	$x < -1sd$	$-1sd \leq x \leq 1sd$	$x > 1sd$
Numerical example	$x < -0.3\%$	$-0.3\% \leq x \leq 0.3\%$	$x > 0.3\%$

Table 2. Class variable discretized into 5 categories. The value $\Delta EMBIHP_{t+i}$ is represented as 'x' in short. The standard-deviation of x is represented as 'sd'.

Category	**2down**	**down**	**stable**	**up**	**2up**
Interval 3:	$x < -1sd$	$-1sd \leq x < -0.5sd$	$-0.5sd \leq x \leq 0.5sd$	$0.5sd < x \leq 1sd$	$x > 1sd$
Interval 4:	$x < -2sd$	$-2sd \leq x < -1sd$	$-1sd \leq x \leq 1sd$	$1sd < x \leq 2sd$	$x > 2sd$

In Table 1 a numerical example is presented for $i = 0$, considering Hodrick-Prescott filtered series using $\lambda = 1000$; the standard-deviation of $\Delta EMBIHP_t$ is $sd = 0.63\%$.

Table 3. Standard deviation of $\Delta EMBIHP_{t+i}$ for different smoothing λ and leading times i.

Standard deviation(%)	$i=0$	$i=5$	$i=10$	$i=15$	$i=20$
$\lambda=100$	0.83%	4.07%	7.58%	10.35%	13.32%
$\lambda=500$	0.68%	3.40%	6.58%	9.39%	12.69%
$\lambda=1000$	0.63%	3.13%	6.13%	8.89%	12.30%

That is, if we consider $EMBIHP = 1000$ and Interval 1, the category *stable* corresponds to values between 997 and 1003. For $i = 20$ (prediction for the next month), the standard deviation of $\Delta EMBIHP_{t+20}$ is $sd = 12.3\%$; and then, the category *stable* corresponds to values between 877 and 1123.

4. Training and testing. After the previous preprocessing stages, we used the data set D to train a Bayesian network classifier. We used the free software *WEKA* [2] to train and test D, including the continuous data discretization. The dataset D was divided into training and testing data. Fayyad and Irani's supervised discretization method [9] was applied to the training data and the same discretization rule was used for testing data. We used *10-fold cross-validation* [12] as evaluation criterion.

4 Experiments and Results

Our experiments were parameterized by: number of class discretization (3 or 5 categories), prediction accuracy (0.5 or 1 standard-deviation), leading times ($t = 0, 5, 10, 15, 20$), smoothing level (Hodrick-Prescott with parameter $\lambda = 100, 500, 1000$) and classification method: Naive Bayes(NB), TAN, Radial Basis Function (RBF) and C4.5 of Quinlan (description of these methods can be found at [18]). Hence, we had 240 different experiments. For each experiment we obtained a classification rate and confusion matrix; here we describe some of most relevant results.

In Tables 4 and 5, we compare classification rates for different methods and two class discretizations. RBF classifier is a standard neural network classifier that uses a normalized Gaussian radial basis function. C4.5 is a decision tree based classifier. The best results are obtained for C4.5 classifiers followed by TAN classifier. Class discretization into five nominal categories leads us to more accurate prediction, but with slight decrease in classification rates. Keeping the same criterion of 1 standard-deviation to specify uncertainty level, classification rates increase with leading time

Table 4. Classification rate results for different methods. Class $\Delta EMBIHP_{t+i}$ is discretized in 3 nominal categories (as described at Table 1). $CR(t+i)$ is the classification rate for $\Delta EMBIHP_{t+i}$ with different leading times i. Accuracy of 1 standard-deviation and smoothing parameter λ=1000.

Method	$CR(t)$	$CR(t+5)$	$CR(t+10)$	$CR(t+15)$	$CR(t+20)$
TAN	90.36%	90.83%	89.62%	92.58%	93.12%
NB	86.45%	86.18%	86.92%	88.06%	88.94%
RBF	86.04%	86.31%	86.24%	88.06%	88.60%
C4.5	92.52%	93.73%	93.26%	94.00%	95.41%

Table 5. Classification rate results for different methods. Class $\Delta EMBIHP_{t+i}$ is discretized in 5 nominal categories (as described at Table 2). $CR(t+i)$ is the classification rate for $\Delta EMBIHP_{t+i}$ with different leading times i. Accuracy of 1 standard-deviation and smoothing parameter λ=1000.

Method	$CR(t)$	$CR(t+5)$	$CR(t+10)$	$CR(t+15)$	$CR(t+20)$
TAN	87.19%	88.94%	88.20%	91.30%	92.04%
NB	80.98%	81.86%	82.94%	84.29%	86.04%
RBF	80.98%	83.21%	84.22%	85.23%	87.05%
C4.5	89.55%	91.37%	91.50%	91.64%	94.07%

— probably because of increases in standard deviations. Classification rates (Tables 4 and 5) increase with leading time, but uncertainty of this prediction (measured by the standard deviation associated for each leading time i, Table 3) increase much faster. This result emphasizes the tradeoff between classification rate and uncertainty.

Table 6. Classification rate results for different smoothing parameters λ. Class $\Delta EMBIHP_{t+i}$ is discretized in 5 nominal categories (as described at Table 2). Results are for leading time $i = 0$, accuracy of 0.5 standard-deviation and for the two best methods.

Smoothing parameter	$\lambda = 100$	$\lambda = 500$	$\lambda = 1000$
TAN	62.98%	75.46%	79.37%
C4.5	68.85%	78.49%	82.67%

At Table 6, we have classification rate results for different smoothing parameters — rates increase as we increase λ. From this result we conclude that prediction task is more difficult as we approximate to the real $EMBI+$ observations, that are very volatile ($\lambda = 100$ is the closest series as we can see at Figure 3). From Tables 5 and 6, it is possible to observe the effect of reducing the interval of percentual variation: 1 to 0.5 standard-deviation. For the experiment with $i = 0$ and $\lambda = 1000$, we have for TAN classifier: 87.19% and 79.37% and , for C4.5 classifier: 89.55% and 82.67%.

5 Conclusion

We have presented preliminary results on the use of Bayesian network classifiers for rating country risk (measured by EMBI+). We have compared different methods (Tables 4 and 5), and examined the effect of various smoothing parameters (Table 6). We get rather good forecasts on the EMBI+ trend (for Brazil) — up to 95% classification rate using TAN and C4.5 classifiers. While C4.5 produces the best classification rates, it has a somewhat less explanatory capacity. Probabilistic models such as TAN and NB instead produce a probability distribution. TAN classifier outperforms NB, probably because it relaxes independence assumptions. Using the Bayesian network provided by TAN classifier, one can also perform a sensitivity analysis so as to investigate causal relationships between macroeconomics variables.

Acknowledgements

The first author was supported by CNPq. The second author was supported by FAPESP (through grant 05/57451-8). This work was (partially) developed in collaboration with CEMAP/EESP.

References

1. J. Andrade and V.K. Teles, *An empirical model of the brazilian country risk - an extension of the beta country risk model*, Applied Economics, vol. 38, 2006, pp. 1271–1278.
2. R. Bouckaert, *Bayesian Network Classifiers in WEKA*, Tech. Report 14/2004, Computer Science Department, University of Waikato, New Zeland, 2004.
3. X. Chen, J.S. Racine, and N.R. Swanson, *Semiparametric ARX Neural Network Models with an Application to Forecasting Inflation*, IEEE Transactions on Neural Networks: Special Issue on Neural Networks in Financial Engineering **12** (2001), no. 4, 674–684.
4. C. Chow and C. Liu, *Approximating Discrete Probability Distributions with Dependence Trees*, IEEE Transactions on Information Theory **14** (1968), no. 3, 462–467.
5. E.C. Colla and J.S. Ide, *Modelo empirico probabilistico de reconhecimento de padroes para risco pais*, Proceedings of XXXIV Encontro Nacional de Economia - ANPEC, 2006 (submited).
6. J.C. Cosset and J. Roy, *The Determinant of Country Risk Ratings*, Journal of International Business Studies **22** (1991), no. 1, 135–142.
7. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Willey and Sons, New York, 1973.
8. C. Erb, C. Harvey, and T. Viskanta, *Political Risk, Economic Risk and Financial Risk*, Finance Analysts Journal - FAJ **52** (1996), no. 6, 29–46.
9. U.M. Fayyad and K.B. Irani, *Multi-interval Discretization of Continuous-valued Attributes for Classification Learning.*, IJCAI, 1993, pp. 1022–1029.
10. D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, Tech. Report MSR-TR-95-06, Microsoft Research, Redmond, Washington, 1995.
11. R. Hodrick and E.C. Prescott, *Postwar U.S. Business Cycles: An Empirical Investigation*, Journal of Money, Credit and Banking **29** (1997), no. 1, 1–16.
12. R. Kohavi, *A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection*, IJCAI, 1995, pp. 1137–1145.
13. J.E. Moody, *Economic Forecasting: Challenges and Neural Network Solutions*, Proceedings of the International Symposium on Artificial Neural Networks, Hsinchu, Taiwan, 1995.
14. D. Geiger N. Friedman and M. Goldszmidt *Bayesian Network Classifiers*, Machine Learning **29** (1997), 131–163.
15. E. Nakamura, *Inflation Forecasting using a Neural Network*, Economic Letters **86** (2005), no. 3, 373–378.
16. A.C. Shapiro, *Currency Risk and Country Risk in International Banking*, Journal of Finance **40** (1985), no. 3, 881–891.
17. E. Turban, *Neural Networks in Finance and Investment: Using artificial intelligence to improve real-world performance*, McGraw-Hill, 1995.
18. I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2 ed., Morgan Kaufmann, 2005.

Experiences from Automatic Summarization of IMF Staff Reports

Shuhua Liu[1] and Johnny Lindroos[2]

[1]Academy of Finland and IAMSR, Abo Akademi University
[2]IAMSR, Abo Akademi University
Lemminkaisenkatu 14 B, 20520 Turku, Finland
sliu@abo.fi, johnny.lindroos@abo.fi

Abstract: So far, the majority of the studies in text summarization have focused on developing generic methods and techniques for summarization that are often designed and heavily evaluated on news texts. However, little has been reported on how these generic systems will perform on text genres that will be rather different from news. In this study, we report our experience with applying the MEAD system, a text summarization toolkit developed at the University of Michigan, to the summarization of IMF staff reports, which represent an inherently very different type of documents comparing to new texts.

1 Introduction

Automated text summarization tools and systems are highly anticipated information processing instruments in the business world, in government agencies or in the everyday life of the general public. While human beings have proven to be extremely capable summarizers, computer based automated abstracting and summarizing has proven to be extremely challenging tasks.

Since 1990s there have been very active research efforts on exploring a variety of text summarization methods and techniques such as statistical sentence scoring methods, discourse analysis methods based on rhetorical structure theory, and the use of lexical and ontology resources such as WordNet as an aid for improvements of other methods (Mani and Maybury, 1999). Great progress has been made (Mani and Maybury, 1999; Moens and Szpakowicz, 2004) and a number of rather impressive text summarization systems have appeared such as the MEAD system from University of Michigan (Radev et al, 2003; 2004), the SUMMARIST system from University of Southern California (Hovy and Lin, 1998), and the Newsblaster from University of Columbia, all employ a collection of summarization techniques.

With the rapid progress in the development of automated text summarization methods and related fields, it also opens up many new grounds for further investigation. So far, the majority of the studies in text summarization have focused on developing generic methods and techniques for summarization that are often designed and heavily evaluated on news texts (DUC 2001-2005, http://duc.nist.gov/). However, little has been reported on how these generic systems will perform on text genres that will be rather different from news. In this study, we report our experience with applying the MEAD system, a text summarization toolkit developed at the University of Michigan, to the summarization of IMF staff reports, which are inherently very different from new texts in terms of the substance, length and writing style. Four sets of summarization experiments are carried out and the system sum-

maries are evaluated according to their similarity to the corresponding staff-written "Executive Summary" included in the original reports.

The IMF Staff Reports are an important source of information concerning macroeconomic development and policy issues for the member countries of the IMF. They are written by IMF mission teams (Fund economists) as the product of their missions and are carefully re-viewed through an elaborate process of review by relevant departments in the Fund (Harper, 1998). Although different missions have their unique kinds of concerns, and the corresponding staff reports will be addressing different policy issues, staff reports all present economic and policy issues in a historical perspective from past to present and future. The structure of staff reports also reflects a standard that includes such components: (i) General Economic Setting (ii) Policy Discussions and (iii) Staff Appraisal. The first part often contains the conclusion of the last mission and the economic developments since last mission while also points to problems and questions that would be the focus of the current mission. It will deliver an overall picture of the current economic situation, highlights the important aspects and critical issues. It will also point out what are the mid-term and long-term trends, and what are only the one-off events. The second part is a report on discussions held together with the authorities about monetary and exchange rate policy, fiscal policy and others. The focus will be on what the authorities perceived as the major obstacles in achieving their mid-term objectives, and elaborations on policy alternatives. The third part is perhaps the most important of all. It presents mission team's policy recommendations for the member country, to advice on what will be the policy strategy that will ensure the joint goals of long-term growth and balance of payments stability (Harper, 1998).

MEAD is a public domain multi-document summarization system developed at the CLAIR group led by Prof. Dragomir Radev at University of Michigan. MEAD offers a number of summarization (or in fact sentence extraction) methods such as position-based, query-based, centroid, and mostly recently the LexPageRank, plus two baselines: random and lead-based methods. In addition to the summarization methods, the MEAD package also includes with it tools for evaluating summaries (the MEAD Eval). MEAD Eval supports two classes of intrinsic evaluation metrics: co-selection based metrics (precision, recall, relative utility, kappa) and content-based metrics (cosine that uses TF*IDF, simple consine that does not use TF*IDF, and unigram-, bigram-overlap) (Radev et al, 2003, MEAD Documentation v3.08). The MEAD system is freely available and can be downloaded from the research group's website (http://www.summarization.com/mead/).

The overall architecture of MEAD system consists of five types of processing functions: Preprocessing, Features Scripts, Classifiers, Re-rankers and Evaluators (Radev et al, 2003). *Prerocessing* takes as input the documents to be summarized in text or HTML format, identifies sentence boundaries and transforms them into an XML representation of the original documents. Then, a set of *features* will be extracted for each sentence to support the applying of different summarization methods such as position-based, centroid based or query-based sentence extraction methods (Radev et al, 2003). Following feature calculation, a *classifier* is used to compute a composite score for each sentence. The composite score is based on weighted combination of the sentence features in a way specified by the classifier, which can potentially refer to any features that a sentence has. After the classifier, each sentence has been assigned a significance score. A *re-ranker* will then modify the sentence scores by considering possible cross-sentence dependencies, source preferences, and so on. Finally, MEAD Eval offers the instruments for evaluating summaries in pairs in terms of their lexical similarity.

2 Summarizing IMF Staff Reports Using MEAD

Our study started with summarizing five IMF staff reports: the Article IV Consultation reports for China, Finland, Sweden and Norway, during the year of 2004-2005. The experiments are further expanded to include 30 staff reports. All the documents are downloaded directly from the IMF publication database accessible via the IMF website (http://www.imf.org). In this paper we present our results from a series of further experiments in order to evaluate the effects of different summarization methods and to find out a good summarization scheme for IMF staff reports.

2.1 Source Documents

All of the reports contain text in several different forms, such as table of contents, appendices, tables, text "boxes", figures, footnotes, Staff Statement that presents information that becomes available since the report was circulated to the Executive Board, Statement by the Executive Director for the member country and Public Information Notice on the Executive Board Discussion. All staff reports have a one-page Executive Summary that we are particularly interested to use as a benchmark for evaluating the system-generated summaries. Before a report is fed into the system to produce summaries, tables, figures, formulas, footnotes and table of contents in the staff re-ports are removed. The executive summaries are removed also but saved in a separate file to be used for evaluation purpose. Figure 1 gives an overview of the word count in each report and the corresponding executive summary following preprocessing (more details in Lindroos, 2006).

As can be seen, the amount of words in the executive summary is 395–466, which is equivalent to a compression ratio of 5-10 % of words. Using MEAD, a compression ratio of 10 % of words is usually substantially less than 10 % of sentences, as MEAD tends to favor longer sentences. In the following set of experiments we tested two compression rates: 5% and 10% of sentences.

Figure 1 Length of Staff Reports and Executive Summaries (word count after pre-processing)

	Whole Report	Executive Summary	Compression Ratio
China	8470 words	414 words	4.9%
Finland	5211 words	449 words	8.6%
Norway	4687 words	451 words	9.6%
Sweden 04	5311 words	395 words	7.4%
Sweden 05	4713 words	466 words	9.9%

2.2 Summarization Scheme

A summarization schema specifies a method-parameter configuration for the system: the preferred *compression rate*, the *classifier* and *re-ranker* applied in the summarization process. Among the summarization techniques in MEAD system, the Centroid method is a novel representative. Based on tf*idf vector space model of documents, the Centroid method calculates a pseudo sentence that is the average of all the sentences in all documents, given a set of documents to be summarized. This pseudo sentence is regarded as the centroid sen-

tence of the document, and the significance of each sentence is determined by calculating how similar each sentence is to this "centroid" sentence.

MEAD also includes two baseline summarizers: *random* and *lead-based.* The random method simply put together sentences randomly selected from a document as a summary. It assigns a random value between 0-1 to each sentence. Lead-based method choose to select the first sentence of each document, then the second sentence of each document until the desired size of the summary is reached. It assigns a score of 1/n to each sentence where n is number of sentences in the document (Radev et al, 2003).

MEAD re-ranker orders the sentences by score from highest to lowest then iteratively decides whether to add each sentence to the summary or not. At each step, if the quota of words or sentences has not been filled, and the sentence is not too similar to any higher scoring sentence already selected for the summary, the sentence in question is added to the summary. The remaining sentences are discarded. MEAD *Cosine re-ranker* simply discards sentences above a certain similarity threshold. *MMRreranker,* on the other hand, adjusts sentence scores so that similar sentences end up receiving a lower score.

MEAD default classifier weights three methods: the centroid, position and length, as equally important features/methods for scoring sentences. In our experiments presented here, we basically adopted two summarization strategies: one *summarizes a staff report as a whole;* the other *summarizes a staff report broken into multiple parts* according to its natural sections. Since the sections that appear at a later part of the report (e.g. policy discussion, staff appraisal) are as important as (if not more than) the sections that appear earlier, the second strategy hopefully helps to achieve a balance of content in the output summary. In all cases the MMRreranker (with Lamda as 0.2) is used instead of MEADCosine as we found that it has a better control over redundancy (Liu, 2006). In summarizing a report as a whole, only one classifer – the Centroid method is applied. In summarizing section by section, two classifiers are tested: (i) Centroid, (ii) Centroid + Position, where position-based method assigns extra weighs to sentences that are at the beginning of a document. In addition, baselines are created using Lead-based and Random method.

```
Classifier : bin/default-classifier.pl
             Centroid 1.0
Reranker   : bin/default-reranker.pl
             MMR 0.2 enidf
Compression: 5% of sentences
```

```
Classifier : bin/default-classifier.pl
             Centroid 1.0 Position 1.0
Reranker   : bin/default-reranker.pl
             MMR 0.2 enidf
Compression: 5% of sentences
```

2.3 Summary Evaluation Metrics

Summary evaluation is as challenging a task as automated summarization itself, especially because different people can judge the quality of a summary differently, one summary can be of different value to different tasks, and there is no golden metrics for measuring the quality of a summary. Summary evaluations can be distinguished between intrinsic evaluation and extrinsic evaluation. Extrinsic evaluation is also referred to as task-based evaluation

and the extrinsic evaluation methods measure how well a summary help in performing a specific task. Intrinsic evaluation, on the other hand, judges the quality of a summary by comparing it to some model summaries (e.g. a manually-written summary).

In our study we are interested first in intrinsic evaluation of the system generated summaries by comparing them to the staff-written executive summary of a report. Given the assumption that the executive summary is an ideal summary, the most optimal re-sult of MEAD would be a summary identical to the executive summary. However, as the MEAD summary is an extract and the executive summary is an abstract, there will be some inevitable differences between the two summaries. The differences between two texts can be measured using text similarity measurements. Two different similarity measurements are applied: semantic simi-larity and lexical similarity. Lexical similarity measures the simi-larity of the actual words used, and does not concern itself with the meaning of the words. On the other hand, semantic similarity at-tempts to measure similarity in terms of meaning.

While the first and foremost goal is to produce summaries that are semantically similar to the original document, lexical similarities may potentially provide a valuable similarity measurement. In our case, the original report and its executive summary have been writ-ten by the same authors, similar words and expressions tend to be used to denote the same concepts, so it is possible that the lexical differences between the executive summary and the MEAD pro-duced summary are not as outstanding as they would have been if the executive summary and original document had been written by unrelated authors.

MEAD provides two types of evaluation tools for evaluating sum-maries. One is based on co-selection metrics and is used to com-pare two summaries that share identical sentences, i.e. extractive summaries. This kind of evaluation tools are of no interest in this experiment, as they are only usable in a system to system evalua-tion environment. The second and in this case more applicable tools are based on content-based metrics, which can be used to measure lexical similarities of two arbitrary summaries, i.e. they are not limited to evaluating summaries generated by MEAD (Radev et al. 2004b). These tools apply various word overlap algo-rithms (Cosine, Simple cosine, as well as unigram and bi-gram co-occurrence statistics) to measure the similarity of two texts (Radev et al. 2004b; Papineni et al. 2002):

- Simple cosine: cosine similarity measurement of word overlap; calculates the cosine similarity with a simple binary count.
- Cosine: Cosine similarity measurement of weighted word overlap; weights are defined by TF*IDF values of words.
- Token overlap: single word overlap measurement.
- Bigram overlap: bigram overlap measurement.
- Normalized LCS: measurement of the longest-common subsequence.
- BLEU: a measurement method based on an n-gram model.

It can be noted that most of the overlap algorithms use relatively similar approaches, and that the results from each algorithm seem to correlate with the results from the other overlap algorithms (Lin-droos, 2006).

The other type of similarity that will be evaluated in our experiments is semantic similarity. One approach that can be used for such evaluation is Latent Semantic Analysis (LSA), which is de-signed to capture similarities in terms of meaning between two texts. LSA makes it possible to approximate human judgments of meaning similarity between words. However, for LSA to function properly, the words in the texts being evaluated must be repre-sented in the semantic space. If they are not represented, the result will be an inaccurate estimation of the similarity of the two texts. LSA has been shown to be remarkably

effective given the right semantic space. LSA has been found to be more accurate than lexical methods when trained on documents related to the text being evaluated. However, lexical methods have outperformed LSA when LSA has been poorly trained, i.e. trained on unrelated or too general documents in comparison to the texts being evaluated (Landauer et al. 1998; Peck et al. 2004).

Making use of the semantic space and similarity analysis tools at the University of Colorado at Boulder (Latent Semantic Analysis @ CU Boulder, http://lsa.colorado.edu/), we carried out one-to-many, document-to-document comparisons between the different summary outputs for each staff report and the corresponding executive summary. This gave us indications of the semantic similarity between an “executive summary” and each of the system summaries for the same report.

As mentioned earlier, using the right semantic space is an impor-tant part of applying the LSA tool. The LSA tool available at the University of Colorado at Boulder website does not have a seman-tic space for texts related to economics, or other types of text that would be similar to IMF staff reports. As a result, the "General Reading up to 1st year college" semantic space is used in the ex-periments in this thesis. Consequently, it should be noted that LSA may not perform as well as it might have performed provided a more specialized semantic space had been available.

3 Results

In this section we present results from four sets of summarization experiments.

3.1 Pre-experiment: multi document summarization

Before the start of our more focused experiments, we tested the MEAD default summarizer on the five staff reports with a compression rate of 5% of sentences:

```
Classifier : bin/default-classifier.pl
             Centroid 1.0
Reranker   : bin/default-reranker.pl
             MMR 0.2 enidf
Compression: 5% of sentences
```

Summaries are generated separately for each of the staff reports. However the input text contains not only the staff report, but also text boxes and the supplements to it. The system output thus represents a summary of multi documents. The summary outputs were examined by compared to the content of the original reports.

The China report: At a compression rate of 5%, the system summary contains 29 full sentences: 12 sentences are from the first section (Recent Economic Development and Outlook); 5 sentences from the second section (Policy Discussion), with two of them from text boxes which are there to provide more historical and background information. Only one sentence is selected from section III, the Staff Appraisal. And, 11 sentences are picked up from the Public Information Notice part. As such, the system summary seems to be missing out much of the most important information of a staff report, that is, the policy discussions as well as analysis and recommendations from the mission team. The reason may be that, since the Section I and the Public Information Notice part both have a large portion of text describing recent economic development and outlook, the system captures this to be the

Centroid of the report, with the help of the Position method, which adds more weights to content from Section I.

Another observation is that the selected sentences tend to fall short in terms of the topic coverage. The sentences selected from Section I and IV seems to be concentrated on only one or two issues or aspects of the economy, often with redundancy, while neglect others aspects.

The Finland report: The Finland report is relatively shorter and the 5% system summary is an extract of all together 16 sentences, in which ten sentences fall into Section-I (Economic Background), two sentences from the "Public Information Notice" and four sentences from "Policy Discussions", with no sentences from "Policy Setting and Outlook", "Staff Appraisal" and "Statement by the Executive Director for Finland". Similar to with the China report, the system summary is still unbalanced with regard to the origins of the extracted sentences. The system summary seems have captured the core subjects, i.e. the population aging and fiscal policy. It also captures the positive flavor in the original report. The difference from the case with China report is that, it seems to be a bit better in terms of the coverage of different aspects of the Finnish Economy. The substance covered in the extracted parts touches upon more diverse aspects and policy issues.

The Sweden 2004 report: The system summary includes 19 sentences: eight from Economic Background; two from Policy Setting and Short-Term Outlook; four from Policy Discussions, four from Public Information Notice (Board Assessment), none from Staff Appraisal, perhaps because these two sections concern similar topics, and the sentences in Policy discussions are longer than those in the Appraisal part. There is a very long sentence in "Policy Discussion" that is selected. The sentence is in fact a composition of listed items that were not properly segmented because of none dot-ending, although the result is not bad in terms of summary quality because it captures the list of concerned issues for policy discussion. The Sweden report summary seems to be a bit more balanced in terms of sentence origins than the China and Finland reports. This long sentence that touches upon all the important issues may have played a role.

The Sweden 2005 report: The Sweden 2005 report is identical to the Sweden 2004 report in terms of length, format and content structure. The system output contains 18 sentences: five from Economic Background, five from Policy Setting and Short-Term Outlook, three from Policy Discussions, four from Public Information Notice, and none from Staff Appraisal, basically also identical to the system summary for the 2004 report in terms of the origins of the selected sentences. However, the result also reflects major differences in their substance, with its focus being more on fiscal issues, overall economic development, etc.

The Norway report: The system summary includes all together 16 sentences. Among these sentences, ten are from the first section Economic Back-ground, two from Policy Discussions, one from Staff Appraisal, and three from the Public Information Notice, all centered around the fiscal policy related with non-oil deficit, oil and gas revenues and GPF (Norway Government Petroleum Fund), strong economic performance. The selected sentences are unbalanced in an unfavorable way due to the effect of the use of the Position feature, i.e. sentences appearing at the beginning and early part of the report get significantly more weight than those in the Policy Discussion and Staff Appraisal parts. The human written executive summary has a much wider coverage.

Overall, the system has performed consistently with reports for different countries. The results reflect the higher weight of the parts "Economic Background" and "Public Information Notice", but tend to overlook the parts "Policy Discussion" and "Staff Appraisal". The

default classifier weights the centroid, position and length as equally important. The result of such a summarizer on the staff reports tend to pick up sentences that appear at the very beginning of the report as well as the earlier parts of the report. The default re-ranker favors redundancy over topic cover-age, which is why the results from this set of experiment see much redundancy in the extracted sentences.

3.2 Experiment 1

Learning from the pre-experiment, in this and the following experiments we devoted more pre-processing work on the input texts before summarizing them. All additional items in the staff reports are removed before input to the summarization system. In addition, customized summarizers are applied. Specifically, in this set of experiment, each of the five staff reports is summarized as a whole applying Centroid method and MMRreranker, with two compression rates, 5% and 10% of sentences. In addition, two baselines are generated for each report, one is lead-based, another Random based, all at a compression rate of 5% sentences. In total 18 summary outputs are produced.

Judging by the content in comparing to the original report, the result is significantly different from the pre-experiments. The 5% summary contains significantly more sentences from the policy discussion and staff appraisal than in the last experiment, while also has a much wider topic coverage. The MMR re-ranker seems working very well. There are much less redundancy than in the last experiment. However, the complete removing of Position feature results in the missing-out of some important sentences that often appear at the beginning of different sections.

3.3 Experiment 2

Next, the China report is broken into three parts by its natural sections and then summarized as one cluster. Two summarizers described earlier (Centroid vs Centroid+Position) are applied. Again the experiments are repeated with two compression rates, 5% and 10% of sentences. Four summaries are created.

At 5% compression rate, the Centroid method extracted 6 sen-tences from Part I (Economic Background), 11 sentences from Policy Discussions and 5 sentences form Staff Appraisal. The spe-cific summarization process guarantees a good balance of the ori-gin of the selected sentences. With a compression rate of 10%, although the length of the summary doubles (while always included all the sentences from the 5% compression rate experiment), there is not much redundancy in the result due to the effect of the MMR re-ranker.

All the system generated summaries from Experiment 1 and 2 are evaluated by examining their semantic similarity to the model summary – the staff written executive summary, using LSA analysis. The similarity measure can be e.g. cosine similarity of document vectors in the semantic space. The best available semantic space seems to be the General_Reading_up_to_1st_year_college. The results are shown in Table 1 below. Although this does not give us the precise similarity values, they give us an indication of the relative performance of different summarization scheme. (Note: The baselines for the China report were only created in the experiments of 10% sentence compression rate, so they are not shown here. Nonetheless, judging from the 10% result, there is not much surprise as compared with other reports).

Table 1 Comparing System Summaries with Executive Summaries using Latent Semantic Analysis (all with MMR re-ranker)					
	China04	Fin04	Swed.04	Swed.05	Nor.04
5%C	0.79	0.85	0.91	0.84	0.83
10%C	0.80	0.86	0.91	0.88	0.86
5%Lead		0.84	0.79	0.75	0.75
5%Rand.		0.79	0.89	0.77	0.78
	0.78	5%Multipart Centroid			
	0.79	10%Multi-part Centroid			
	0.83	5%Multi-part Centroid+Position			
	0.82	10%Multi-part Centroid+Position			

The results seem to indicate that, the performances of the different summarization approaches for the China report do not differ very much among the Centroid or "Multi-parts Centroid" methods. But the "Multi-parts Centroid + Position" approach looks slightly better. Overall, the compression rate of 5% and 10% does not make much difference to the evaluation result, although judging from the content of the summaries, the latter deliver much more rich information than the former.

The same Centroid method performs rather consistently on the Finland report, Sweden 2005 report and Norway report, but differs a bit with the China report and Sweden 2004 report, may be due to the slightly different ways the substance in the executive summary was drawn up. The Centroid method is shown performing better than Lead-based and Random methods in all cases, but only slightly. What is surprising is that the Random method seems to over-perform Lead-based method, except in the case of the Finland report.

The selection of the semantic space turns out to be not the best choice, as the results are accompanied by remarks of "terms in the reports that do not exist in the corpus selected". However, for the moment this seems to be the closest proximity to a semantic space in the Economy domain that is available on the site.

3.4 Experiment 3

In addition to the above evaluation, extensive summarization experiments are carried out on 30 IMF staff reports from the year 2003-2005, applying six different summarization configurations to each report: (1) Centroid method with MEAD default reranker (C-D), (2) Centroid+Position method with MEAD default reranker (CP-D), (3) Random method (Rand.); (4) Lead method (Lead); (5) Centroid method with MMR reranker (C-MMR); and (6) Centroid+Position method with MMR reranker (CP-MMR); all at a compression rate of 5% sentences. In total, 180 summaries were produced, 6 summaries for each report. All the system outputs are evaluated against the corresponding staff written Executive Summaries using five of the content based evaluation metrics: Simple cosine, Cosine (tf*idf), Token Overlap, Bigram Overlap and BLEU.

The performance of the six different summarization configurations along the five different lexical similarity measurements is given in the tables below (paired t-test: probability associated with a Student's paired t-test, with two tailed distribution).

The results again confirm the correlation among the different overlapping measurements. If a summarization scheme performs well at one similarity metric, it generally also performs well in terms of the other four similarity metrics. Overall, it can be noted that, among the four summarization schemes C-D, CP-D, C-MMR and CP-MMR, CP-D and CP-

MMR show somewhat better performance than the C-D and C-MMR respectively. When the summarization method is held unchanged (C or CP), the summarization schemes adopting the MEAD default re-ranker usually exhibit inferior performance comparing to the summarization schemes using MMRreranker. The two baselines are sometimes the top performer and sometimes the worst performer. Their performance is very unstable, which is also indicated by the highest standard deviations they usually have.

Simple Cosine Similarity						
	C-D	CP-D	Rand.	Lead	C-MMR	CP-MMR
Mean	0.3078	0.3162	0.0284	0.3243	0.3111	0.3235
Stdev	0.0408	0.0615	0.0427	0.1559	0.0407	0.0762
Paired T-test		0.3987	0.0178	0.1847	0.6466	0.2923

Cosine Similarity						
	C-D	CP-D	Rand.	Lead	C-MMR	CP-MMR
Mean	0.4010	0.3889	0.2860	0.3509	0.3926	0.3884
Stdev	0.1005	0.1039	0.0765	0.1698	0.0869	0.0979
Paired T-test		0.1699	2E-5	0.0564	0.1776	0.7062

Token Overlap						
	C-D	CP-D	Rand.	Lead	C-MMR	CP-MMR
Mean	0.1779	0.1861	0.1653	0.2072	0.1788	0.1909
Stdev	0.0296	0.0450	0.0296	0.1568	0.0291	0.0570
Paired T-test		0.2698	0.0338	0.1670	0.3353	0.1935

Bi-gram Overlap						
	C-D	CP-D	Rand.	Lead	C-MMR	CP-MMR
Mean	0.0577	0.0653	0.0456	0.0964	0.0566	0.0687
Stdev	0.0241	0.0473	0.0185	0.1733	0.0236	0.0590
Paired T-test		0.3083	0.2669	0.1159	0.2135	0.2138

BLEU						
	C-D	CP-D	Rand.	Lead	C-MMR	CP-MMR
Mean	0.0720	0.0870	0.0700	0.1239	0.0692	0.0893
Stdev	0.0490	0.0648	0.0458	0.1843	0.0471	0.0698
Paired T-test		0.0489	0.2057	0.0983	0.1188	0.0352

The paired t-tests (null hypothesis) show much varied results with different similarity measurements. Judging by *simple cosine similarity* and *token overlap*, only the performance difference between the Random method and CP-D is statistically significant ($p<0.05$). All the other differences are statistically insignificant. In terms of cosine similarity however, the

difference between Random method and CP-D as well as between Lead method and Random method are statistically significant. Judging by bi-gram overlap, the performance differences between the paired methods are all in-significant. Finally, using the BLEU metric, the paired t-test indicates significant performance difference between C-D and CP-D, between Lead and Random method, between CP-MMR and C-MMR. On the contrary, the performance difference between method Random and CP-D, between Lead and Random method, and between C-MMR and Lead method are not statistically significant, which is counter-intuitive. This seems to suggest that BLEU as a document similarity measurement is more different from the other four measurements. Human judgment needs to be incorporated to find out whether this may suggest that BLEU is a less appropriate evaluation metric than others, and which ones are the more appropriate summary evaluation metrics.

4 Conclusions

In this paper we report our experience with applying the MEAD system to summarize the IMF staff reports. Comparing to news articles that are usually simpler in format and much shorter in length, in summarizing the IMF staff reports, it is necessary to incorporate suitable preprocessing, and take advantage of the flexibility in selecting and combining multiple methods. Comparing with the rich variety in writing style for news text, the IMF Staff reports are much more "predictable" due to their very standardized format for structuring and presenting the content. This makes it easier to see the effects of different summarization schemes.

In order to achieve good summary results, it would require a good understanding of the summarization methods in the system. However, to find out the best combinations of methods requires non-trivial effort. Also, one insight gained from the experiments is that, the IMF staff reports are documents that contain what may be called multi-centroid or multi-subject that are equally important, while the MEAD Centroid based method is supposed to capture only one centroid.

Compression rate naturally should have a big impact on the output summaries. The LSA based similarity analysis of system summary and the corresponding executive summary, however, show basically little difference between 5% and 10% outputs. This may indicate that the 5% compression rate is a rather good choice for summarizing staff reports if outputs length is a critical measurement. However, what should also be noted is that, at 10% compression rate the system outputs deliver much more complete content than at a 5% compression rate. It is very hard for machine-generated summaries to be wide in coverage and at the same time short in length. The system generated summary will usually be considerably lengthy if it is expected to be as informative as the Executive Summary. The results also revealed that depending solely on current LSA tool to evaluate the summarization results based on similarity analysis is not enough.

Further evaluations have been carried out using a number of content-based evaluation metrics. The results confirm the correlation among the different overlapping measurements. In addition, it shows that the CP-D and CP-MMR schemes show somewhat better performance than C-D and C-MMR respectively. When the summarization method is held unchanged (C or CP), the summarization schemes adopting MMRreranker exhibits better performance comparing to the summarization schemes using the MEAD default re-ranker. The two baselines are sometimes the top performer and sometimes the worst performer. Their performances tend to be unstable. To find out a best summarization scheme and best evalua-

tion metrics for summarizing IMF staff reports, expert evaluation of summary quality will be incorporated in our future research.

Finally, the IMF staff reports have evident content structuring characteristics that could be utilized to benefit the summarization output. Each section of a staff report is divided into a set of numbered paragraphs. Each paragraph appears to start with a sentence describing the most essential information that is discussed or can be derived from that paragraph, with the rest of the paragraph aimed towards supporting the information presented in the initial sentence. This would not only indicate that the original document contains sentences suitable to be extracted for a summary, but also make it a rational approach to simply extract all such introductory sentences to produce a summary of reasonable quality. This approach is further explored in other experiments.

Acknowledgments Financial support from Academy of Finland is gratefully acknowledged.

References

[1] Mani I. and Maybury M. T. (Eds.) (1999). *Advances in Automatic Text Summarization*, MIT Press. Cambridge, MA.

[2] Moens M. and Szpakowicz S. (Eds.) (2004). *Text Summarization Braches Out*, Proceedings of the ACL-05 Workshop, July 25-26 2004, Barcelona, Spain

[3] Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Celebi A., Drabek E., Lam W., Liu D., Qi H., Saggion H., Teufel S., Topper M. and Winkel A. (2003). *The MEAD Multidocument Summarizer*. MEAD Documentation v3.08. available at http://www.summarization.com/mead/.

[4] Radev D., Jing H., Stys M. and Tam D. (2004). *Centroid-based summarization of multiple documents*. Information Processing and Management, 40, 919-938.

[5] Hovy E. and Lin C. (1999). *Automated Text Summarization in SUMMARIST*. Mani and Maybury (eds), Advances in Automatic Text Summarization, MIT Press, 1999.

[6] Harper, R. (1998). *Inside the IMF*, Academic Press. 1st Edition. Academic Press.

[7] Carbonell J. G. and J. Goldstein. *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*. In Alistair Moffat and Justin Zobel, editors, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 335-336, Melbourne, Australia, 1998.

[8] Liu S. and J. Lindroos, *"Automated Text Summarization using MEAD: Experience with the IMF Staff Reports"*, to be presented at IIIA2006 -International Workshop on Intelligent Information Access, July 6-8, 2006, Helsinki.

[9] Lindroos J. *Automated Text Summarization using MEAD: Experience with the IMF Staff Reports*, Master Thesis, Åbo Akademi University, 2006

Taping into the European Energy Exchange (www.EEX.de) to feed a Stochastic Portfolio Optimization Tool "SpOt" for Electric Utilities

Ulrich Reincke[1], Heiner Lake[1],
Michael Wigbels[2], André Röthig[2],
Michael Lucht[3]

[1]*SAS Institute GmbH*, In der Neckarhelle 162,
69118 Heidelberg, Germany
[2] *Capgemini Deutschland GmbH*, Hamborner Strasse 55
40472 Düsseldorf, Germany
[3]*Fraunhofer Institut für Umwelt-, Sicherheits- und Energietechnik UMSICHT*,
Osterfelder Straße 3, 46047 Oberhausen Germany

Abstract. Within the electric utility industry, portfolio optimization is the process of implementing strategy and information technology solutions to maximize the value and manage the risk of integrated energy portfolios over the near-, medium-, and long-term. While conventional portfolio management solutions rely on or heuristics and/or cumbersome trial and error iterations only a stochastic portfolio optimization model as implemented in "SpOt" can provide full transparency and find automatically the answers to the questions listed above. Robust mathematical approaches including jump diffusion and GARCH models as well as stochastic portfolio optimization techniques are applied to develop a dynamic portfolio management solution.

Keywords: Simulation, GARCH Models, Stochastic Programming, Portfolio Optimization, Portfolio Management, Risk Management

1 Introduction

The wealth of consumers around the globe depends on affordable electric energy. While in general energy prices have been on a roller coaster ride within the last 4 years, electric utilities need to find new ways to provide energy at low prices and low risk.

What is the best mix of energy assets to hold for the risk we are willing to absorb given our customers demand profiles? What are the key asset decisions we should make and when should we make them? What options should we exercise? What long term contracts should we negotiate? All these questions relate to changes in an electric utility's asset portfolio. Within the electric utility industry, portfolio optimization is the process of implementing strategy and information technology solutions to maximize the value and manage the risk of integrated energy portfolios over the near-, medium-, and long-term. While conventional portfolio management solutions rely on or heuristics and/or cumbersome trial and error iterations only a

stochastic portfolio optimization model can provide full transparency and find automatically the answers to the questions listed above.

For this purpose the Fraunhofer Institut für Umwelt-, Sicherheits- und Energietechnik UMSICHT developed in cooperation with the companies Capgemini Deutschland GmbH and SAS Institute GmbH the "Stochastic Portfolio Optimization Tool" SpOt. Robust mathematical approaches including jump diffusion and GARCH models as well as stochastic portfolio optimization techniques are applied to develop a dynamic portfolio management solution. Our approaches focus on delivering flexible portfolio management strategies that adapt to changing markets and regulatory parameters (for example carbon dioxide certificates). The solution SpOt can be applied in the following four areas: 1) Asset optimization providing support for a wide set of problems, ranging from long term investment decisions (e.g., the upgrade of a power plant or decision to install pollution controls) to medium-term strategy support (e.g., implementing the right mix of contracts and options). 2) Asset deployment characterizing the short-term (e.g., daily) operational decisions on assets. Examples include traders' daily nomination decisions on purchases to meet generation demand. 3) Asset valuation determines the value of a tangible or intangible asset (e.g., a tangible power plant or the intangible emissions allowances). 4) Company wide risk controlling and overall portfolio management

2 GARCH Models for Price Forward Curve Simulations

Since history never repeats itself, reliance on constant historical price volatility estimates and historical price curves will lead to sub optimal performance. Our approach estimates therefore a GARCH model to determine the underlying characteristic parameters of an endogenous price volatility process. User specified price forward curves that reflect fundamental expectations of future price levels are used as a basis to simulate thousands of price trajectories around the expected future price by means of the mean-reverting GARCH model. Mean reversion is a tendency for a stochastic process to remain near, or tend to return over time to a long-run average level. In this way we are able to generate a large number of forward-looking price trajectories to incorporate many alternate views on future price levels and price volatilities into an integrated valuation and decision-support framework for a subsequent holistic stochastic optimization.

3 Stochastic Optimization

Stochastic optimization is a framework for modelling decision problems that involve uncertainty. Whereas deterministic optimization problems are formulated with known parameters, real world problems almost invariably include to a certain extend

unknown parameters. Stochastic optimization models take advantage of the fact that probability distributions governing the data are known or can be estimated or simulated. The goal here is to find some policy that is feasible for all simulated data instances and maximizes the expectation of some utility function of the decisions and the random variables. More generally, such models are formulated, solved, and analyzed in order to provide useful information to a decision-maker in the real world with uncertain information.

The key modelling feature in our objective function is a convex combination of Cvar (risk) and average procurement cost over all price trajectories:

$$\text{Min! } \boldsymbol{a}\text{*mean(cost)+(1-}\boldsymbol{a}\text{)Cvar(cost)} \qquad (\mathbf{1})$$

where Cvar is the conditional Value at risk, $\boldsymbol{a}$ is the users risk preference. For $\boldsymbol{a}$=0 the user wants to reduce risk at any costs, and for $\boldsymbol{a}$=1 the user prefers to reduce cost no matter what the risk is.

4 Results: Efficiency Frontier as a Landmark for better Portfolio Restructuring Decisions

SpOt solves problem (1) for different values of $\boldsymbol{a}$ in [0,1] and prompts the user with a portfolio efficiency frontier. For any energy portfolio under consideration the corresponding point following along an imaginary horizontal line until the efficiency frontier will yield a better hedge with lower lover average cost at the same risk. Following instead an imaginary vertical line until we hit the efficiency frontier we can determine a better hedge with lower lover risk at same cost. Thus the resulting portfolio adjustment transactions to move to the efficiency frontier are easily determined.

Users of SpOt get full transparency over their portfolio restructuring process and are able to determine optimal portfolios that will better satisfy their risk preference. In this way SpOt helps to obtain lower operating cost and risk. These savings provide utilities with stronger competitiveness and will ultimately contribute to dampen electric energy prices for consumers.

References

1. Spangardt, G.; Lucht, M.; Handschin, E.: Applications for Stochastic Optimization in the Power Industry. Electrical Engineering (Archiv für Elektrotechnik), 2004, in Druck, online DOI 10.1007/s00202-004-0273-z

2. Spangardt, G.; Lucht, M.; Althaus, W.: Optimisation of Physical and Financial Power Purchase Portfolios. Central European Journal of Operations Research 11/2003, S. 335-350.
3. Spangardt, G; Lucht, M.; Althaus, W.: Automatisierte Portfolio-Optimierung für Stadtwerke und Industrie- kunden. VDI-Berichte 1792: Optimierung in der Energieversorgung, VDI Verlag, Düsseldorf, 2003, S. 13-22.
4. Spangardt, G; Lucht, M.; Althaus, W.; Handschin, E.: Vom Portfoliomanagement zur Portfolio-Optimierung. Marktplatz Energie, April 2003, S. 6-7.
5. Spangardt, G.: Mittelfristige risikoorientierte Optimierung von Strombezugsportfolios. Dissertation, Fakultät für Elektrotechnik und Informationstechnik, Universität Dortmund, UMSICHT-Schriftenreihe, Band 43, Fraunhofer IRB- Verlag, Stuttgart.

Mining Medical Administrative Data - The PKB System

Aaron Ceglar[1], Richard Morrall[2] and John F. Roddick[1]

[1] School of Informatics and Engineering
Flinders University,
PO Box 2100, Adelaide, South Australia 5001,
Email: {ceglar, roddick}@infoeng.flinders.edu.au

[2] PowerHealth Solutions
27 Angas Street, Adelaide, South Australia 5000,
Email: richard.morrall@powerhealthsolutions.com

Abstract. Hospitals are adept at capturing large volumes of highly multidimensional data about their activities including clinical, demographic, administrative, financial and, increasingly, outcome data (such as adverse events). Managing and understanding this data is difficult as hospitals typically do not have the staff and/or the expertise to assemble, query, analyse and report on the potential knowledge contained within such data. The Power Knowledge Builder project is investigating the adaptation of data mining algorithms to the domain of patient costing, with the aim of helping practitioners better understand their data and therefore facilitate best practice.

1 Introduction

Hospitals are driven by the twin constraints of maximising patient care while minimising the costs of doing so. For public hospitals in particular, the overall budget is generally fixed and thus the quantity (and quality) of the health care provided is dependent on the patient mix and the costs of provision.

Some of the issues that hospitals have to handle are frequently related to resource allocation. This requires decisions about how best to allocate those resources, and an understanding of the impacts of those decisions. Often the impact can be seen clearly (for example, increasing elective surgery puts more surgical patients in beds, as a result constraining admissions from the emergency department leading to an increase in waiting times in the emergency department) but the direct cause may not be apparent (is the cause simply more elective patients? is it discharge practices? has the average length of stay changed? is there a change in the casemix in emergency or elective patients? and so on). Analysing the data with all these potential variables is difficult and time consuming. Focused analysis may come up with a result that explains the change but an unfocussed analysis can be a fruitless and frustrating exercise.

As a part this resource pressure, hospitals are often unable to have teams of analysts looking across all their data, searching for useful information such as

trends and anomalies. For example, typically the team charged with managing the patient costing system, which incorporates a large data repository, is small. These staff may not have strong statistical backgrounds or the time or tools to undertake complex multi-dimensional analysis or data mining. Much of their work is in presenting and analysing a set of standard reports, often related to the financial signals that the hospital responds to (such as cost, revenue, length of stay or casemix). Even with OLAP tools and report suites it is difficult for the users to look at more than a small percentage of the available dimensions (usually related to the known areas of interest) and to undertake some *ad hoc* analysis in specific areas, often as a result of a targeted request, e.g. what are the cost drivers for liver transplants?

Even disregarding the trauma of an adverse patient outcome, adverse events can be expensive in that they increase the clinical intervention required, resulting in higher-than-average treatment costs and length-of-stay, and can also result in expensive litigation. Unfortunately, adverse outcomes are not rare. A study by Wolff et. al. [1] focusing on rural hospitals estimated that .77% of patients experienced an adverse event while another by Ehsani et.al., which included metropolitan hospitals, estimated a figure of 6.88% [2]. The latter study states that *the total cost of adverse events ... [represented] 15.7% of the total expenditure on direct hospital costs, or an additional 18.6% of the total inpatient hospital budget.* Given these indicators, it is important that the usefulness of data mining techniques in reducing artefacts such as adverse effects is explored.

A seminal example of data mining use within the hospital domain occurred during the Bristol Royal Infirmary inquiry of 2001 [3] in which data mining algorithms were used to create hypotheses regarding the excessive number of infant deaths at the Bristol Royal Infirmary that underwent open-heart surgery. In a recent speech, Sir Ian Kennedy (who lead the original inquiry) said, with respect to improving patient safety, that *The [current] picture is one of pockets of activity but poor overall coordination and limited analysis and dissemination of any lessons. Every month that goes by in which bad, unsafe practice is not identified and rooted out and good practice shared, is a month in which more patients die or are harmed unnecessarily.* The roll of data mining within hospital analysis is important given the complexity and scale of the analysis to be undertaken. Data mining can provide solutions that can facilitate the benchmarking of patient safety provision, which will help eliminate variations in clinical practice, thus improving patient safety.

The *Power Knowledge Builder* (PKB) project provides a suite of data mining capabilities, tailored to this domain. The system aims to alert management to events or items of interest in a timely manner either through automated exception reporting, or through explicit exploratory analysis. The initial suite of algorithms (trend analysis, resource analysis, outlier detection and clustering) were selected as forming the core set of tools that could be used to perform data mining in a way that would be usable to educated users, but without the requirement for sophisticated statistical knowledge.

To our knowledge, PKB's goal is unique – it is industry specific and does not require specialised data mining skills, but aims to leverage the data and skills that hospitals already have in place. There are other current data mining solutions, but they are typically part of a more generic reporting solutions (i.e. Business Objects, Cognos) or sub-sets of data management suites such as SAS or SQL server. These tools are frequently powerful and flexible, but are not targeted to an industry, and to use them effectively requires a greater understanding of statistics and data mining methods than our target market generally has available. This paper introduces the PKB suite and its components in Section 2. Section 3 discusses some of the important lessons learnt, while Section 4 presents the current state of the project and the way forward.

2 The PKB Suite

The PKB suite is a core set of data mining tools that have been adapted to the patient costing domain. The initial algorithm set (anomaly detection, trend analysis and resource analysis), was derived through discussion with practitioners, focusing upon potential application and functional variation. Subsequently, clustering and characterisation algorithms were appended to enhance usefulness.

Each algorithmic component has an interface wrapper, which is subsequently incorporated within the PKB prototype. The interface wrapper provides effective textual and graphical elements, with respect to pre-processing, analysis and presentation stages, that simplifies both the use of PKB components and the interpretation of their results. This is important as the intended users are hospital administrators, not data mining practitioners and hence the tools must be usable by educated users, without requiring sophisticated statistical knowledge.

2.1 Outlier Analysis

Outlier (or anomaly) detection is a mature field of research with its origins in statistics [4]. Current techniques typically incorporate an explicit distance metric, which determines the degree to which an object is classified as an outlier. A more contemporary approach incorporates an implied distance metric, which alleviates the need for the pairwise comparison of objects [5, 6] by using domain space quantisation to enable distance comparisons to be made at a higher level of abstraction and, as a result, obviates the need to recall raw data for comparison.

The PKB outlier detection algorithm LION contributes to the state of the art in outlier detection, through novel quantisation and object allocation, that enables the discovery of outliers in large disk resident datasets in two sequential scans. Furthermore LION addresses the need realised during this project of the algorithm to discover not only outliers but also outlier clusters. By clustering similar (close) outliers and presenting cluster characteristics it becomes easier for users to understand the common traits of similar outliers, assisting the identification of outlier causality. An outlier analysis instance is presented in Figure 1, showing the interactive scatterplot matrix and cluster summarisation tables.

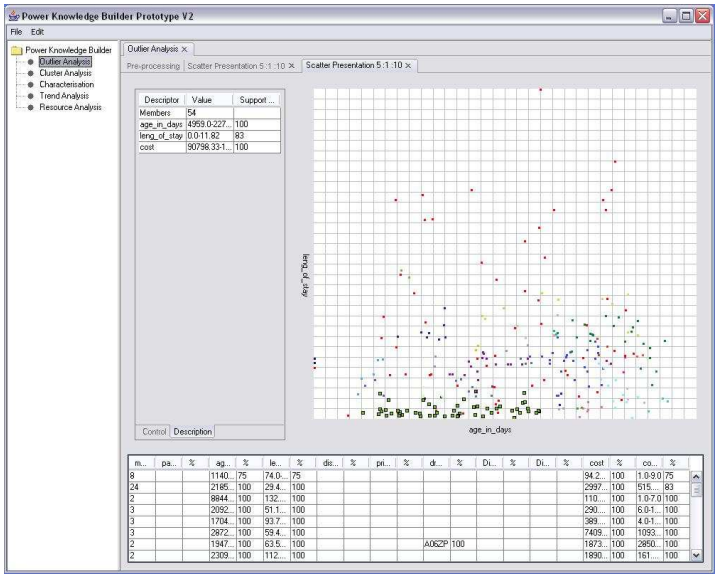


Fig. 1. Outlier Analysis with Characterisation Tables and Selection

Outlier detection has the potential to find anomalous information that is otherwise lost in the noise of multiple variables. Hospitals are used to finding (and in fact expect to see) outliers in terms of cost of care, length of stay etc. for a given patient cohort. What they are not so used to finding are outliers over more than two dimensions, which can provide new insights into the hospital activities. The outlier component presents pre-processing and result interfaces, incorporating effective interactive visualisations that enable the user to explore the result set, and see common traits of outlier clusters through characterisation.

2.2 Cluster Analysis

Given LION's cluster based foundations, clustering (a secondary component) is a variation of LION that effectively finds the common clusters rather than anomalous clusters. Given their common basis, both the outlier and clustering components require the same parameters and use the same type of result presentations. Once again the clusters are characterised to indicate common intra-cluster traits that can assist in identifying causality.

2.3 Characterisation

Characterisation (also a secondary component) was initially developed as a subsidiary for outlier and clustering analysis in order to present descriptive summaries of the clusters to the users. However it is also present as an independent tool within the suite. The characterisation algorithm provides this descriptive

cluster summary by finding the sets of commonly co-occurring attribute values within the set of cluster objects. To achieve this, a partial inferencing engine, similar to those used in association mining [7] is used. The engine uses the extent of an attribute value's (elements) occurrence within the dataset to determine its significance and subsequently its usefulness for summarisation purposes. Once the valid elements have been identified, the algorithm deepens finding progressively larger, frequently co-occurring elements sets from within the dataset.

Given the target of presenting summarised information about a cluster, the valid elements are those that occur often within the source dataset. While this works well for non-ordinal data, ordinal data requires partitioning into ranges, allowing the significant mass to be achieved. This is accomplished by progressively reducing the number of partitions, until at least one achieves a significant volume. Given the range 1 to 100, and initial set of 2^6 partitions are formed, if no partition is valid, each pair of partitions are merged, by removing the lowest significant bit, (2^5 partitions). This process continues until a significant mass is reached. This functionality is illustrated in Figure 1 through the presentation of a summarisation table with ordinal-ranges.

2.4 Resource analysis

Resource usage analysis is a domain specific application that provides a tool that analyses patterns of resource use for patient episodes (hospital stays). This novel algorithm is backed by an extended inferencing engine [7], that provides dynamic numeric range partitioning, temporal semantics and affiliated attribute quantisation, to provide a rich analysis tool. Furthermore the tool enables the association of extraneous variables with the resource patterns such as average cost and frequency. The resource usage results are presented as a set of sortable tables, where each table relates to a specified dataset partition. For example, the user can specify the derivation of daily resource usage patterns for all customers with a particular *Diagnosis Related Group*, partitioned by *consulting doctor*. By associating average cost and frequency with these patterns, useful information regarding the comparative cost effectiveness of various doctors may be forthcoming. A screenshot of the resource analysis presentation is provided in Figure 2 illustrating the clustering of daily resource use for consulting doctor id *1884*. Further questions such as *is it significant that a chest x-ray for one patient took place on day one, while for another patient in the same cohort it took place on day two?* can also be addressed. The resource analysis component automates and simplifies what would have previously been very complex tasks for costing analysts to perform.

2.5 Trend Analysis

Trend or time-series analysis is a comparative analysis of collections of observations made sequentially in time. Based upon previous research [8, 9], the underlying analysis engine undertakes similarity based time series analysis using *Minowski* Metrics and removing distortion through offset translation, amplitude

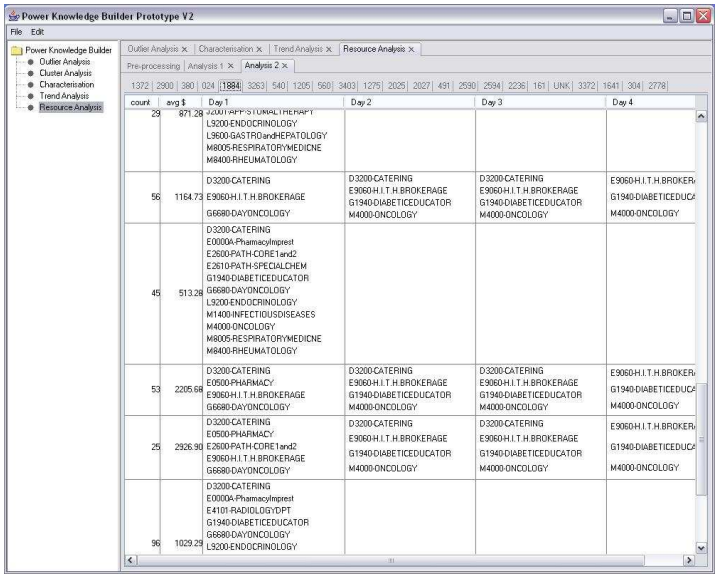


Fig. 2. Resource Analysis

scaling and linear trend removal. The component then provides a rich higher level functional set incorporating, temporal variation, the identification of offset and subset trends, and the identification of dissimilar as well as similar trends, as illustrated in Figure 3 . The component provides a comprehensive interface including ordered graphical pair-wise representations of results for comparison purposes. For example, the evidence of an offset trend between two wards with respect to admissions can indicate a causality that requires further investigation.

3 Lessons Learned

The PKB project began as a general investigation into the application of data mining techniques to patient costing software, between Flinders University and PowerHealth Solutions, providing both an academic and industry perspective. Now, 18 months on from its inception, many lessons have been learnt that will hopefully aid both parties in future interaction with each other and with other partners. From an academic viewpoint, issues relating to the establishment of beta test sites and the bullet proofing of code is unusual. While from industry the meandering nature of research and the potential for somewhat tangential results can be frustrating. Overall, three main lessons have been learnt.

Solution looking for a problem? It is clear that understanding data and deriving usable information and insights from it is a problem in hospitals, but how best to use the research and tools is not always clear. In particular, the initial project specification was unclear as to how this would be achieved.

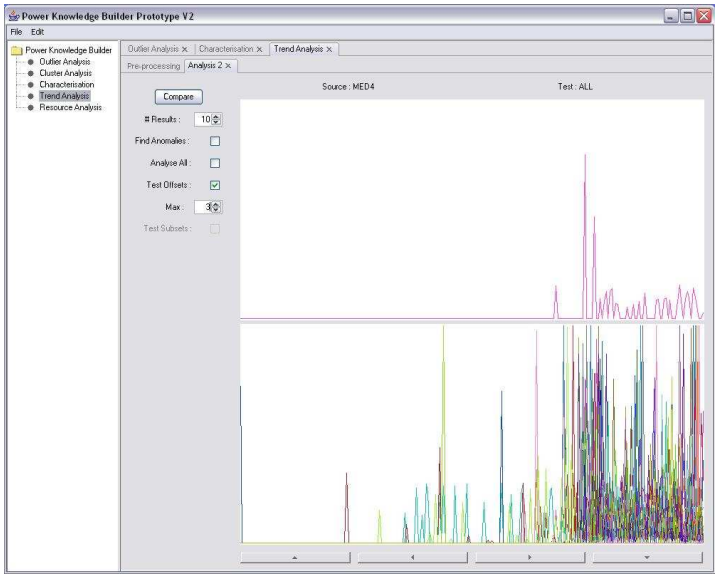


Fig. 3. Trend Analysis Presentation: parameter setting

As the project evolves it is crystallising into a tool suite that complements PowerHealth Solution's current reporting solution. More focus upon the application of the PKB suite from with outset would have sped up research but may also have constrained the solutions found.

Educating practitioners. The practical barriers to data mining reside more in the structuring and understanding of the source data than in the algorithms themselves. A significant difficulty in providing data mining capabilities to non-experts is the requirement for the users to be able to collect and format source data into a usable format. Given knowledge of the source data, scripts can easily be established to accomplish the collection process. However where the user requires novel analysis, an understanding of the required source data is required. It is possible to abstract the algorithmic issues away from the user, providing user-friendly GUI's for result interpretation and parameter specification, however this is difficult to achieve for source data specification, as the user must have a level of understanding with respect to the nature of the required analysis in order to adequately specify it.

Pragmatics. The evaluation of the developed tools requires considerable analysis, from both in-house analysts and analysts from third parties who have an interest in the PKB project. The suite is theoretically of benefit, with many envisaged scenarios (based upon experience) where it can deliver useful results, but it is difficult to find beta sites with available resources.

4 Current State and Further Work

The second version of the PKB suite is now at beta-test stage, with validation and further functional refinement required from industry partners. The suite currently consists of a set of fast algorithms with relevant interfaces that do not require special knowledge to use. Of importance in this stage is feedback regarding the collection and pre-processing stages of analysis and how the suite can be further refined to facilitate practitioners in undertaking this.

The economic benefits of the suite are yet to be quantified. Expected areas of benefit are in the domain of quality of care and resource management. Focusing upon critical indicators, such as death rates and morbidity codes, in combination with multiple other dimensions (e.g. location, carer, casemix and demographic dimensions) has the potential to identify unrealised quality issues.

Three immediate areas of further work are evident: the inclusion of extraneous repositories, knowledge base construction and textual data mining. The incorporation of extraneous repositories such as meteorological and socio-economic within some analysis routines can provide useful information regarding causality. While the incorporation of an evolving knowledge base will facilitate analysis by either eliminating known information from result sets or the flagging of critical artefacts. As most hospital data is not structured, but contained in notes, descriptions and narrative the mining of textual information will also be valuable.

References

1. Wolff, A.M., Bourke, J., Campbell, I.A., Leembruggen, D.W.: Detecting and reducing hospital adverse events: outcomes of the Wimmera clinical risk management program. Medical Journal of Australia **174** (2001) 621–625
2. Ehsani, J.P., Jackson, T., Duckett, S.J.: The incidence and cost of adverse events in Victorian hospitals 2003-04. Medical Journal of Australia **184** (2006) 551–555
3. Kennedy, I.: Learning from Bristol: The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995. Final report, COI Communications (2001)
4. Markou, M., Singh, S.: Novelty detection: a review - part 1: statistical approaches. Signal Processing **83** (2003) 2481–2497
5. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In Gupta, A., Shmueli, O., Widom, J., eds.: 24th International Conference on Very Large Data Bases, VLDB'98, New York, NY, USA, Morgan Kaufmann (1998) 392–403
6. Papadimitriou, S., Kitagawa, H., Gibbons, P., Faloutsos, C.: Loci: Fast outlier detection using the local correlation integral. In Dayal, U., Ramamritham, K., Vijayaraman, T., eds.: 19th International Conference on Data Engineering (ICDE), Bangalore, India (2003) 315–326
7. Ceglar, A., Roddick, J.F.: Association mining. ACM Computing Surveys **38** (2006)
8. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods. Springer Verlag, New York (1987)
9. Keogh, E.K., Chakrabarti, K., Mehorta, S., Pazzani, M.: Locally adaptive dimensionality reduction for indexing large time series databases. In: ACM SIGMOD International Conference on Management of Data, Santa Barbara, CA, ACM (2001) 151–162

Combining text mining strategies to classify requests about involuntary childlessness to an internet medical expert forum

Wolfgang Himmel[1], Ulrich Reincke[2], Hans Wilhelm Michelmann[3]

[1]Department of General Practice/Family Medicine, University of Göttingen, Humboldtallee 38, 37073 Göttingen, Germany, whimmel@gwdg.de
[2]Competence Center Enterprise Intelligence, SAS Institute Heidelberg, In der Neckarhelle 162, 69118 Heidelberg, Germany, ulrich.reincke@ger.sas.com
[3]Department of Obstetrics and Gynaecology, University of Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany, hwmichel@med.uni-goettingen.de

Abstract. Text mining has been successfully applied, for example, in ascertaining and classifying consumer complaints. The classification of requests of laymen to medical expert forums is more difficult, because such requests can be very long and unstructured by mixing personal experiences with laboratory data. A sample of 988 questions to an Internet expert forum was first personally classified into 38 categories (32 subject matters and 6 different types of senders' expectations). We then trained several logistic regression models in order to build appropriate models for automatic classification of the senders' requests. These models were based on techniques to reduce the amount of information such as principal component analysis or singular value decomposition. Compared to our own classification, a 100% precision and 100% recall could be realized in 10 out of 38 categories of the validation sample. Even in the worst cases, precision and recall rates were above 60%. A combination of different text mining strategies is highly recommended for the classification of complex texts.

Keywords. text mining, natural language processing, consumer health informatics, internet, infertility

1 Introduction

Both healthy and sick people increasingly use electronic media to get medical information and advice. Internet users actively exchange information with others about subjects of interest or send requests to web-based expert forums ("ask-the-doctor" services) [1-3]. An automatic classification of these requests could be helpful for several reasons: (1) a request can be forwarded to the respective expert, (2) an automatic answer can be prepared by localising a new request in a cluster of similar

requests which have already been answered, and (3) changes in information needs of the population at risk can be detected in due time.

Text mining has been successfully applied, for example, in ascertaining and classifying consumer complaints or to handle changes of address in emails sent to companies. The classification of requests to medical expert forums is more difficult because these requests can be very long and unstructured by mixing, for example, personal experiences with laboratory data. Moreover, requests can be classified (1) according to their subject matter or (2) with regard to the sender's expectation. While the first aspect is of high importance for the medical experts to understand the contents of requests, the latter is of interest for public health experts to analyse information needs within the population. Therefore, the pressure to extract as much useful information as possible from health requests and to classify them appropriately is very strong.

To make full use of text mining in the case of complex data, different strategies and a combination of those strategies may refine automatic classification and clustering [4]. The aim of this paper is to present a highly flexible, author-controlled method for an automatic classification of requests to a medical expert forum and to assess its performance quality. This method was applied to a sample of requests collected form the section "Wish for a Child" on the German website www.rund-ums-baby.de.

2 Methods

A sample of 988 questions to the expert forum www.rund-ums-baby.de was first personally classified into 2 dimensions: (1) 32 subject matters, e.g. assessment of pregnancy symptoms or information about artificial insemination and (2) 6 different types of expectations, e.g. emotional reassurance or a recommendation about treatment options. In a second step, a word count of all words appearing in all requests of each classification category was performed. Then we applied 3 techniques to reduce the huge amount of information:

(1) Calculation of the average Cramer's V statistic for the association of each word with each category and the subsequent generation of indicator variables that sum for each subject and document all Cramer's V coefficients over the significant words. The selection criterion for including a term's Cramer's V statistic was the error probability of the corresponding chi-square test. Its significance level was set alternatively at 1%, 2%, 5%, 10%, 20%, 30% and 40%.

(2) Principle component analysis to reduce the 7 indicator variables per subject into 5 orthogonal dimensions.

(3) A 240 dimensional singular value decomposition (SVD), on the basis of the standard settings of the SAS Text Miner™ Software [5].

The sample was then split into 75% training data and 25% validation data. On the basis of the 38*7 Cramer's V indicators per category, the 38*5 principle components per category and the 240 SVDs, we trained several logistic regressions models with stepwise forward selection in order to build appropriate models for automatic classification of the senders' requests.

To assess the most appropriate model for a classification, we used the following selection methods: (1) Akaike Information Criterion, (2) Schwarz Bayesian Criterion, (3) cross validation misclassification of the training data (leave one out), (4) cross validation error of the training data (leave one out) and (5) variable significance based on an individually adjusted variable significance level for the number of positive cases.

By means of logical combinations of input variables with the different selection criteria 1,761 models were trained. The statistical quality of the subject classification was determined, among others, by recall and precision (Table).

Table. Classification Results on Validation Data

Class	N	Selection Criterion*	P (%)	Input Variables**	Precision	Recall
Subject Matters (examples)***						
Abortion	40	AVSL	40	pc	83	100
Tubal examination	19	XMISC	1	k	100	100
Hormones	36	AVSL	30	svd	67	89
Expenses	25	XERROR	1	k	100	100
Worries during pregnancy	49	XMISC	40	pc	87	100
Pregnancy symptoms	36	XERROR	40	k	90	100
Semen analysis	57	XMISC	20	k	87	87
Hormonal stimulation	40	AVSL	40	pc	77	100
Cycle	79	AIC	30	k	88	75
Expectations (complete)						
General information	533	AIC	40	k	83	89
Current treatment	331	AVSL	40	k	77	84
Interpretation of results	310	AVSL	40	k	79	86
Emotional reassurance	90	SBC	1	k	68	65
Interpretation	242	XERROR	30	k	60	82
Treatment opportunities	351	SBC	30	k	77	82

* AIC = Akaike Information Criterion, SBC = Schwarz Bayesian Criterion, XMISC = cross validation misclassification, XERROR = cross validation error, AVSL = adjusted variable significance level

** k = Cramer's V-variables, pc = principal components, svd = singular value dimensions

*** Out of the 988 requests sent to the expert forum, some examples of how the requests were classified are given. The complete list comprised 32 subject matters.

3 Results

A 100% precision and 100% recall could be realized in 10 out of 38 categories on the validation sample; some examples are given in the Table. The lowest rate for precision was 67%, for recall 75% in the subject classification. The last six categories ("Expectations"), however, performed considerably below their subject matter peers.

The Cramer's V-variables (k) and the principal components (pc) proved to be superior to the 240 singular value dimensions (svd) for automatic classification. At the workshop, we will also present how the combination of different predictive models improved recall and precision for both types of classification problems. Finally, we will sketch out the approach for automatic answering of health requests on the basis of text mining.

4 Discussion

A combination of different text mining strategies is highly recommended for complex texts. This combination should include automatic routines for text mining and self-made strategies to make full use of expert knowledge in the classification process. The classification of requests according to the senders' expectation was sub-optimal. This may be due to the somehow vague definition of what exactly defines a certain patients expectation and requires improvement if health experts try to make conclusions about health needs from these expectations. The overall performance of the subject classification, however, seems to be sufficient so that a semi-automatic answering of senders' requests in this medical area may be a realistic option for the future.

References

1. Eysenbach G., Diepgen T.L.: Patients looking for information on the Internet and seeking teleadvice: motivation, expectations, and misconceptions as expressed in e-mails sent to physicians. Arch. Dermatol. 135 (1999) 151-156
2. Himmel W., Meyer J., Kochen M.M., Michelmann H.W.: Information needs and visitors' experience of an Internet expert forum on infertility. J. Med. Internet Res. 7 (2005) e20
3. Umefjord, G., Hamberg, K., Malker H, Petersson, G.: The use of an Internet-based Ask the Doctor Service involving family physicians: evaluation by a web survey. Fam. Pract. 23 (2006) 159-166
4. Jeske D., Liu R.: Mining Massive Text Data and Developing Tracking Statistics. In: Banks D., House L., McMorris F.R., Arabie P., Gaul W. (eds.): Classification, Clustering and Data Mining Application. Springer-Verlag, Berlin (2004) 495-510
5. Reincke, U. Profiling and classification of scientific documents with SAS Text Miner. Available at http://km.aifb.uni-karlsruhe.de/ws/LLWA/akkd/8.pdf

Mining in Health Data by GUHA method

Jan Rauch

Center of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 *
University of Economics, Prague, nám W. Churchilla 4, 130 67
Prague, Czech Republic, rauch@vse.cz

1 Introduction

GUHA is an original Czech method of exploratory data analysis developed since 1960s [3]. Its principle is to offer all interesting facts following from the given data to the given problem. It is realized by GUHA procedures. The GUHA-procedure is a computer program the input of which consists of the analysed data and of several parameters defining a very large set of potentially interesting patterns. The output is a list of all prime patterns. The pattern is prime if both it is true in the analysed data and if it does not immediately follow from other more simple output patterns.

The most used GUHA procedure is the procedure ASSOC [3] that mines for association rules describing various relations of two Boolean attributes including rules corresponding to statistical hypotheses tests. The "classical" association rules with confidence and support [1] are also mined however under the name *founded implication* [4]. The procedure ASSOC was several time implemented, see e.g. [5]. Its last and contemporary most used implementation is the procedure 4ft-Miner [6] that has some new features. Implementation of procedure ASSOC does not use the well known a-priori algorithm. It is based on representation of analysed data by suitable strings of bits [6]. This approach proved to be very efficient to compute various contingency tables [6, 8]. It led to implementation of five additional GUHA procedures. All procedures are included in the system LISp-Miner [7] (http://lispminer.vse.cz).

These GUHA procedures were many time applied to various health data. The goal of the paper is to present experience with these applications. We use data set STULONG[1] (http://euromise.vse.cz/challenge2004/data/index.html) to demonstrate some important features of two GUHA procedures, see next section.

* The work described here has been supported by the grant 1M06014 of Ministry of Education, Youth and Sports of the Czech Republic and by the grant 25/05 of University of Economics, Prague

[1] The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomečková, MD, PhD. and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the EuroMISE Centre of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvárová, DrSc).

2 Applying 4ft-Miner and SD4ft-Miner

There are six GUHA procedures in the LISp-Miner system that mine various patterns verified using one or two contingency tables, see `http://lispminer.vse.cz/procedures/`. The contingency tables make possible to formulate patterns that are good understandable for users - nonspecialists in data mining. The bit string approach used in implementation also brings a possibility to easy deal with (automatically) derived Boolean attributes of the form $A(\alpha)$. Here α is a subset of the set $a_1, \ldots, a_k$ of all possible values of attribute A. Boolean attribute $A(\alpha)$ is true in a row o of analysed data matrix $\mathcal{M}$ if it is $a \in \alpha$ where a is the value of attribute A in the row o.

We use procedures *4ft-Miner* and *SD4ft-Miner* and data matrix *Entry* of the STULONG data set to demonstrate some of possibilities of GUHA procedures. Data matrix *Entry* concerns 1417 patients that are described by 64 attributes, see `http://euromise.vse.cz/challenge2004/data/entry/`.

The procedure 4ft-Miner mines for association rules $\varphi \approx \psi$ where φ and ψ are Boolean attributes. The rule $\varphi \approx \psi$ means that φ and ψ are associated in a way given by the symbol $\approx$ that is called *4ft-quantifier*. The 4ft-quantifier corresponds to a condition concerning a four-fold contingency table of φ and ψ. The Boolean attributes φ and ψ are automatically derived from columns of analyzed data matrix. The association rule $\varphi \approx \psi$ is true in data matrix $\mathcal{M}$ if the condition corresponding to $\approx$ is satisfied in contingency table of φ and ψ in $\mathcal{M}$. There are 14 types of 4ft-quantifiers implemented in 4ft-Miner including quantifiers corresponding to statistical hypotheses tests [6]. An example of the association rule $\varphi \approx \psi$ produced by 4ft-Miner is the rule

$$\text{Height}\langle 166, 175\rangle \wedge \text{Coffee(not)} \Rightarrow^{+}_{0.71,\ 51} \text{Triglicerides}(\leq 96)\ .$$

The corresponding contingency table is at Fig. 1 (patients with missing values are automatically omitted). It says that there are 51 patients satisfying both φ

$\mathcal{M}$	ψ	$\neg\psi$	
φ	51	122	173
$\neg\varphi$	144	812	956
	195	934	1129

Fig. 1. Contingency table of Height $\langle 166, 175\rangle \wedge$ Coffee(not) and Triglicerides(≤ 96)

and ψ, 122 patients satisfying φ and not satisfying ψ etc. The presented rule means that relative frequency of patients with triglycerides ≤ 96 mg% among patients 166–175 cm high that do not drink coffee (i.e. $\frac{51}{173} = 0.29$) is 71 per cent greater than average relative frequency of patients with triglycerides ≤ 96 mg in the whole data matrix (i.e. $\frac{195}{1129} = 0.17$). The support of the rule is 51.

The presented rule is one of results of a run of 4ft-Miner related to the analytical question *What combinations of patients characteristics lead to at least*

20 per cent greater relative frequency of extreme values of triglycerides?. We used parameters defining derived Boolean attributes like Height$\langle ?, ?+10\rangle$ (i.e. sliding window of the length 10) or Triglicerides$\langle \min, ?\rangle$ and Triglicerides$\langle ?, \max\rangle$ (i.e. extreme values of Triglicerides). There were 281 rules found among 918 300 rules tested in 59 seconds on PC with 1.58 GHz. For more details concerning 4ft-Miner see e.g. [6].

The procedure SD4ft-Miner mines for SD4ft-patterns – i.e. expressions like $\alpha \bowtie \beta : \varphi \approx \psi/\gamma$. An example is the pattern

$$\text{normal} \bowtie \text{risk} : \ \text{Skinfold\_Triceps}(5, 15\rangle \wedge \text{Educ(Univ)} \rightarrow^{\bowtie}_{0.4} \ \text{Diast}(60, 90\rangle/\text{Beer}(\leq 1) \ .$$

This SD4ft-pattern is verified using two contingency tables see Fig. 2. The

$\mathcal{T}_N$	ψ	$\neg\psi$	
φ	40	4	44
$\neg\varphi$	89	22	111
	129	26	155

$\mathcal{T}_R$	ψ	$\neg\psi$	
φ	47	46	93
$\neg\varphi$	231	155	386
	278	201	479

Fig. 2. Contingency tables $\mathcal{T}_N$ and $\mathcal{T}_R$

table $\mathcal{T}_N$ concerns normal patients drinking maximally one liter of beer/day. The second table $\mathcal{T}_R$ concerns risk patients drinking maximally one liter of beer/day. Informally speaking the pattern means that the set of normal patients differs from the set of risk patients what concerns the confidence of association rule *Skinfold_Triceps* $(5, 15\rangle \wedge$ *Educ(Univ)* $\rightarrow$ *Diast*$(60, 90\rangle$ (abbreviated by $\varphi \rightarrow \psi$) if we consider patients drinking maximally one liter of beer/day. The difference of confidences is 0.4. The confidence of $\varphi \rightarrow \psi$ on $\mathcal{T}_N$ is $\frac{40}{44} = 0.91$, the confidence of $\varphi \rightarrow \psi$ on $\mathcal{T}_R$ is $\frac{47}{93} = 0.51$. Generation and verification of $31.6 * 10^6$ of SD4ft-patterns took 23 minutes at the PC with 1.58 GHz and 13 patterns with difference of confidences ≥ 0.4 were found including the presented one.

The additional GUHA procedures implemented in the LISp-Miner system are KL-Miner, CF-Miner, SDKL-Miner, and SDCF-Miner [7, 8].

3 Conclusions

Tens of runs of GUHA procedures of LISP-Miner system were done on various health data. Some of them were very successful e.g. the analysis of data describing thousands of catheterisations in General Faculty Hospital in Prague [9]. We can summarize:

– The principle of GUHA method - to offer *all true patterns relevant to the given analytical question* - proved to be very useful. All the implemented procedures are able to generate and verify huge number of relevant patterns in reasonable time. The input parameters of GUHA procedures makes possible both to fit the generated patterns to a solved data mining problem and to tune the number of output patterns in a reasonable way.

– Simple conditions concerning frequencies from contingency tables are used that makes results good understandable even for non-specialists. However there are also patterns corresponding to statistical hypotheses tests see e.g. [3, 6–8] that are intended for specialists.
– Namely the procedure 4ft-Miner gives fast orientation in large data and it is useful to combine it with additional analytical software.
– It is important that the complexity of used algorithms is linearly dependent on number of rows of analysed data matrix [6, 8]

However there are lot of related open problems. The first one is how to efficiently use large possibilities of fine tuning of the definition of the set of relevant patterns to be generated and verified. An other problem is how to combine particular procedures when solving a given complex data mining task. The next step of solution chain depends both on the goal of analysis and on results of previous runs of procedures. Big challenge is automated chaining of particular procedures of LISp-Miner to solve given problem.

Thus one of our next research goals is to build an open system called *EverMiner* of tools to facilitate solving real problems using all procedures implemented in the LISp-Miner system. The *EverMiner* system will consists of typical tasks, scenarios, repositories of definitions of sets of relevant patterns etc. [7].

References

1. Aggraval, R. et al: Fast Discovery of Association Rules. In Fayyad, U. M. et al.: Advances in Knowledge Discovery and Data Mining. AAAI Press, 1996. pp 307–328
2. Burian, J. Rauch, J.: Analysis of Death Causes in the STULONG Data Set. In: Berka, Petr (ed.). Discovery Challenge. Zagreb : IRB, 2003, pp. 47–58
3. Hájek, P., Havránek, T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer–Verlag 1978.
4. P. Hájek (guest editor), International Journal of Man-Machine Studies, special issue on GUHA, vol. 10, January 1978
5. P. Hájek and A. Sochorová and J. Zvárová: GUHA for personal computers. Computational Statistics & Data Analysis, vol 19, pp. 149 - 153, February 1995
6. Rauch J, Šimůnek M (2005) An Alternative Approach to Mining Association Rules. In: Lin T Y, Ohsuga S, Liau C J, and Tsumoto S (eds) Data Mining: Foundations, Methods, and Applications, Springer-Verlag, 2005, pp. 219 – 238
7. Rauch J, Šimůnek M: GUHA Method and Granular Computing. In: Hu, X et al (ed.). Proceedings of IEEE conference Granular Computing. 2005, pp. 630–635.
8. Rauch J, Šimůnek Milan, Lín Václav (2005) Mining for Patterns Based on Contingency Tables by KL-Miner First Experience. In: Lin T Y et all (Eds.) Foundations and Novel Approaches in Data Mining. Berlin. Springer-Verlag, pp. 155–167.
9. Štochl J, Rauch J, Mrázek V.: Data mining in Medical Databases. In: Zvárová J et all (ed.). Proceedings of the International Joint Meeting EuroMISE 2004. Praha: EuroMISE, 2004, p. 36

An Investigation into a Beta-Carotene/Retinol Dataset Using Rough Sets

Kenneth Revett[1], Florin Gorunescu[2], and Marina Gorunescu[2]

[1] University of Westminster,
Harrow School of Computer Science
Harrow, Middlesex,
England HA1 3TP
revettk@westminster.ac.uk
[2] University of Medicine and Pharmacology
Department of Mathematics, Biostatistics, and
Computer Science
Craiova, Romania
{fgorun, mgorun}@umfcv.ro

Abstract. Numerous reports have implicated diets and/or conditions where levels of carotene/retinol are below minimal daily requirements may pre-dispose individuals to an increased susceptibility to various types of cancer. In this study, we investigate dietary and other factors that may influence plasma levels of these anti-oxidants. We use a rough sets approach to generate a series of rules that map values of clinically derived attributes to a discrete range of carotene/retinol levels. The resulting classifier produced an accuracy of approximately 90% for both beta-carotene and retinol The results indicate that age, smoking, and dietary intake of these endogenous anti-oxidants are predictive of plasma levels.

1 Introduction

Clinical studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. Anti-oxidants have long been known to reduce the risks of a variety of cancers, and substantial efforts have been made in the research and clinical communities to find ways of elevating anti-oxidants to levels that are therapeutic. [1]. In addition to well-known functions such as dark adaptation and growth, retinoids have an important role in the regulation of cell differentiation and tissue morphogenesis. Following numerous experimental studies on the effects of retinoids on carcinogenesis, clinical use of retinoids has already been introduced in the treatment of cancer (acute promyelocytic leukemia) as well as in the chemoprevention of carcinogenesis of the head and neck region, breast, liver and uterine cervix [2]. Given the importance of this class of chemicals in their role as

anti-carcinogens, we investigated a small dataset containing 315 patient records recorded over a 1-year period. The purpose of this study was to determine which attribute(s) that were collected had a positive correlation with plasma levels of these anti-oxidants.

1.1 Dataset Description

This dataset contains 315 observations on 14 variables. Two of the variables (attributes) consist of plasma levels of beta-carotene and retinol levels. The dataset was split into two – one containing the beta-carotene levels and all other attributes (except the retinol levels), resulting in 12 attributes and one decision class. The same technique was applied, leaving out the beta-carotene levels (retaining the retinol levels as the decision class.) There were no missing values and the attributes consisted of both categorical and continuous data. Since the decision attribute was continuous, it was discretised such that a binary decision was obtained. Briefly, the distribution of the decision class was generated and the median value was used as a cut point for the two classes. In order to determine the effectiveness of this cut-off point, several trials with a range of +/-2% of the median (up to 100%) were tried in an exhaustive manner to find (empirically) the best threshold value. The best threshold value was determined by the resulting classification accuracy. The attributes were discretised whenever possible in order to reduce the number of rules that would be generated using rough sets. Once the dataset had been transformed into a decision table, we applied the rough set algorithm, described formally in the next section.

2 Rough Sets Algorithm Description

Rough set theory is a relatively new data-mining technique used in the discovery of patterns within data first formally introduced by Pawlak in 1982 [3,4]. Since its inception, the rough sets approach has been successfully applied to deal with vague or imprecise concepts, extract knowledge from data, and to reason about knowledge derived from the data [5,6]. We demonstrate that rough sets has the capacity to evaluate the importance (information content) of attributes, discovers patterns within data, eliminates redundant attributes, and yields the minimum subset of attributes for the purpose of knowledge extraction.

3 Methods

The structure of the dataset consisted of 14 attributes, including the decision attribute (labelled ‘result’) which was displayed for convenience in table 1 above. There were 4,410 entries in the table with 0 missing values. The attributes contained a mixture of categorical (e.g. Sex) and continuous (e.g. age) values, both of which can be used by rough sets without difficulty. The principal issue with rough sets is to discretise the attribute values – otherwise an inordinately large number of rules are generated. We

employed an entropy preserving minimal description length (MDL) algorithm to discretise the data into ranges. This resulted in a compact description of the attribute values which preserved information while keeping the number of rules to a reasonable number (see the results section for details). We determined the Pearson's Correlation Coefficient of each attribute with respect to the decision class. Next, we partitioned the dataset into training/test cases. We selected a 75/25% training testing scheme (236/79 cases respectively) and repeated this process with replacement 50 times. The results reported in this paper are the average of these 50 trials. We used dynamic reducts, as experience with other rough sets based reduct generating algorithms (cf. Rosetta) has indicated this provides the most accurate result [9]. Lastly, we created the rules that were to be used for the classification purpose. The results of this process are presented in the next section.

4 Results

After separating the beta-carotene decision from the retinol decision attribute, the rough set algorithm was applied as described above from an implementation available from the Internet (see reference [2]). In table 3 displayed below, samples of the resulting confusion matrices are displayed. The confusion matrix provides data on the reliability of the results, indicating true positives/negatives and false positives/negatives. From these values, one can compute the accuracy, positive predictive

TABLE 3: Sample confusion matrices randomly selected from a series of 10 classifications.

	Low	High	
Low	32	6	**0.84**
High	3	38	**0.93**
	0.91	**0.86**	**0.89**

Table 4. A sample of the rules produced by the rough sets classifier. The rules combine attributes in conjunctive normal form and map each to a specific decision class. The '*' corresponds to an end point in the discretised range – the lowest value if it appears on the left hand side of a sub-range or the maximum value if it appears on the right hand side of a sub-range.

Antecedents	=>	Consequent
Age([*, 45)) AND SmokeStat(1) => Decision(0) (corresponds to low levels)		
Age([50,*)) AND SmokeStats(3) AND Cholesterol([100,*]=> Decision(1)		
Age ([*,45)) AND SmokeStats(1) AND Cholesterol([100,*))=> Decision(0)		
BMI ([*,25.1))) AND Cholesterol ([100,*)) => Decision(0)		
DailyFibre ([*,35.7)) AND Alcohol ([*, 1.3)) => Decision(1)		

value and the negative predictive value of the results. The results indicate an overall classification accuracy of approximately 90%. In table 4, we present a sample of the

resulting rules that were generated during the classification process. The rules generated are in an easy to read if attribute x = A then consequent = B.

5 Discussion

In this study, we examined a dataset containing information on factors that have been reported to influence plasma levels of the common anti-oxidants beta-carotene and retinol. The results show that many of the attributes, especially age, alcohol consumption, dietary fat and cholesterol intake correlate inversely with anti-oxidant levels. In this study, we have reduced the dimensionality of the dataset by 25% (3/12) without any appreciable reduction in classification accuracy. In addition, a set of readily interpreted rules such as those listed in table 4 means the results can be interpreted more readily than those generated by neural networks. In addition, , through standard validation techniques such as N-fold validation, our results produced results that are better than those published elsewhere in the literature. The area under the ROC was approximately 90% for beta-carotene and retinol. These promising results indicate that rough sets can be a useful machine learning tool in the automated discovery of knowledge, even from small and often sparse biomedical datasets.

References

1. Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology 1989;130:511-521
2. Krinsky, NI & Johnson, EJ, Department of Biochemistry, School of Medicine, Tufts University, 136 Harrison Avenue, Boston, MA 02111-1837, USA; Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University, 136 Harrison Avenue, 711 Washington St, Boston, MA 02111-1837,USA.
3. Z. Pawlak . Rough Sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.
4. Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer (1991).
5. J. Wroblewski.: "Theoretical Foundations of Order-Based Genetic Algorithms". Fundamenta Informaticae 28(3-4) pp. 423–430, 1996.
6. D. Slezak.: "Approximate Entropy Reducts". Fundamenta Informaticae, 2002.
7. K. Revett and A. Khan, "A Rough Sets Based Breast Cancer Decision Support System," METMBS, Las Vegas, Nevada, June, 2005
8. Gorunescu, F, Gorunescu, F,, El-Darzi, E, Gorunescu, S., & Revett K. A Cancer Diagnosis System Based on Rough Sets and Probabilistic Neural Networks, First European Conference on Health care Modelling and Computation, University of medicine and Pharmacy of Craiova, pp 149-159.
9. Rosetta: Rosetta: http://www.idi.ntnu.no/~aleks/rosetta

Data Mining Applications for Quality Analysis in Manufacturing

Roland Grund

IBM Information Management, Technical Sales, Altrottstraße 31, 69190 Walldorf
rgrund@de.ibm.com

Abstract

Besides the traditional application areas like CRM data mining also offers interesting capabilities in manufacturing. In many cases large amounts of data are generated during research or production processes. This data mostly contains useful information about quality improvement. Very often the question is how a set of technical parameters influences a specific quality measure - a typical classification problem. To know these interactions can be of great value because already small improvements in production can save a lot of costs.

Like in general, also in manufacturing industry there is a trend towards simplification and integration of data mining technology. In the past analysis was usually done on smaller isolated data sets with a lot of handwork. Meanwhile more and more data warehouses have become available containing well-organized and up-to-date information. Modern software technology allows to build customized analytical applications on top of these warehouses. Data Mining is integrated, simplified, automated and it can be combined with standard reporting, so that even an "end-user" can make use of it. In practice, however, the successful implementation of such applications still includes a variety of challenges.

The talk shows examples from the automotive industry (DaimlerChrysler and BMW) and provides a brief overview on IBM's recent data mining technology.

Towards Better Understanding of Circulating Fluidized Bed Boilers: A Data Mining Approach

Mykola Pechenizkiy[1], Antti Tourunen[2], Tommi Kärkkäinen[1],
Andriy Ivannikov[1], Heidi Nevalainen[2]

[1]Department of MIT, University of Jyväskylä, P.O. Box 35, 40351 Jyväskylä, Finland
mpechen@it.jyu.fi, tka@mit.jyu.fi, aivanni@cc.jyu.fi
[2]VTT Processes, P.O. Box 1603, 40101 Jyväskylä, Finland
{Antti.Tourunen, Heidi.Nevalainen}@vtt.fi

Abstract. Fuel feeding and inhomogeneity of fuel typically cause process fluctuations in the circulating fluidized bed (CFB) process. If control systems fail to compensate the fluctuations, the whole plant will suffer from fluctuations that are reinforced by the closed-loop controls. This phenomenon causes reducing efficiency and the lifetime of process components. Therefore, domain experts are interested in developing tools and techniques for getting better understanding of underlying processes and their mutual dependence in CFB boilers. In this paper we consider an application of data mining (DM) technology to the analysis of time series data from a pilot CFB reactor.

Keywords: CFB reactor, process monitoring and control, time-series mining.

1 Introduction

Self-sufficiency of energy is one of the most significant issues in EU policy. Enlargement of EU makes the challenge of ensuring the stable and reliable energy production even bigger, as the energy consumption in Eastern Europe will be increasing rapidly, ensuring their economic growth so as to reach the level of Central and Western Europe. This will lead to thermal exploitation of low quality fuels, which remains the main local energy source in European Union and especially in Eastern Europe. Cheap and stabile fuel guarantee lowest price for power production.

Continuous and growing increase of fluctuations in electricity consumption brings new challenges for the control systems of boilers. Conventional power generation will face high demands to ensure the security of energy supply because of increasing share of renewable energy sources like wind and solar power in power production. This can lead to frequent load changes which call for novel control concepts in order to minimize emissions and to sustain high efficiency during load changes.

From combustion point of view the main challenges for the existing boilers are caused by a wider fuel selection (increasing share of low quality fuels), increasing share of bio fuels, and co-combustion. In steady operation, combustion is affected by the disturbances in the feed-rate of the fuel and by the incomplete mixing of the fuel

in the bed, which may cause changes in the burning rate, oxygen level and increase CO emissions. This is especially important, when considering the new biomass based fuels, which have increasingly been used to replace coal. These new biofuels are often rather inhomogeneous, which can cause instabilities in the feeding. These fuels are usually also very reactive. Biomass fuels have much higher reactivity compared to coals and the knowledge of the factors affecting the combustion dynamics is important for optimum control. The knowledge of the dynamics of combustion is also important for optimizing load changes [2].

The development of a set of combined software tools intended for carrying out signal processing tasks with various types of signals has been undertaken at the University of Jyväskylä in cooperation with VTT Processes[1]. This paper presents the vision of further collaboration aimed to facilitate intelligent analysis of time series data from circulating fluidized bed (CFB) sensors measurements, which would lead to better understanding of underlying processes in the CFB reactor.

2 Addressing Business Needs with the Data Mining Approach

Currently there are three main topics in CFB combustion technology development; once through steam cycle, scale up (600 – 800 MWe), and oxyfuel combustion.

The supercritical CFB combustion utilizes more cleanly, efficiently, and sustainable way coal, biofuels, and multifuels, but need advanced automation and control systems because of their physical peculiarities (relatively small steam volume and absence of a steam drum). Also the fact that fuel, air, and water mass flows are directly proportional to the power output of the boiler sets tight demands for the control system especially in CFB operation where huge amount of solid material exist in the furnace.

When the CFB boilers are becoming larger, not only the mechanical designs but also the understanding of the process and the process conditions affecting heat transfer, flow dynamics, carbon burnout, hydraulic flows etc. have been important factors. Regarding the furnace performance, the larger size increases the horizontal dimensions in the CFB furnace causing concerns on ineffective mixing of combustion air, fuel, and sorbent. Consequently, new approaches and tools are needed in developing and optimizing the CFB technology considering emissions, combustion process, and furnace scale-up [3].

Fluidization phenomenon is the heart of CFB combustion and for that reason pressure fluctuations in fluidized beds have been widely studied during last decades. Other measurements have not been studied so widely. Underlying the challenging objectives laid down for the CFB boiler development it is important to extract as much as possible information on prevailing process conditions to apply optimization of boiler performance. Instead of individual measurements combination of information from different measurements and their interactions will provide a possibility to deepen the understanding of the process.

[1] Further information about the project, its progress and deliverables will be made available from http://www.cs.jyu.fi/~mpechen/CFB_DM/index.html.

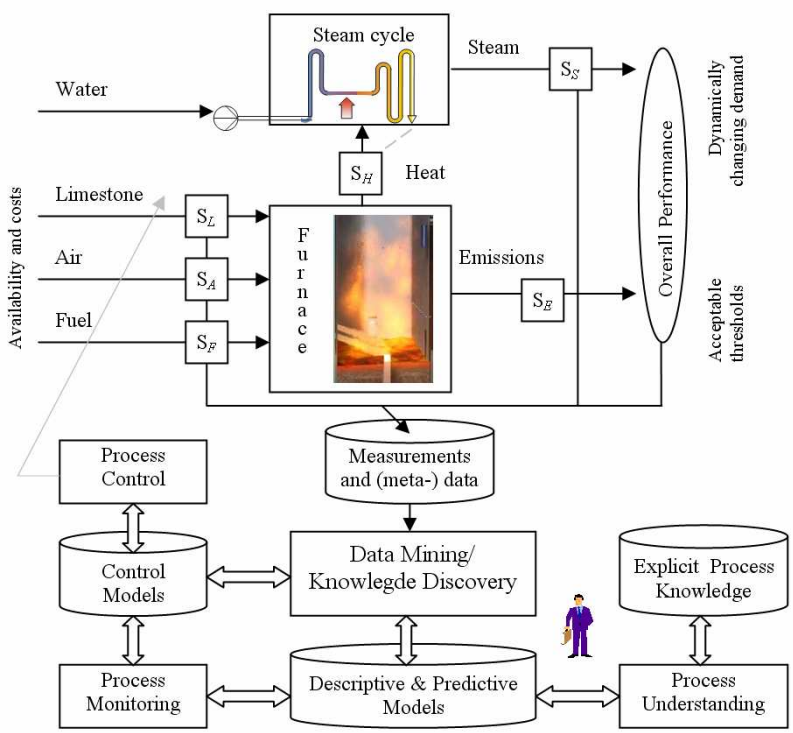


Fig. 1. A simplified view of a CFB boiler operation with the data mining approach

A very simplified view on how a CFB boiler operates is presented in the upper part of Fig. 1. Fuel (mixture of fuels), air, and limestone are the controlled inputs to the furnace. Fuel is utilized to heat production; air is added for enhancing the combustion process and limestone is aimed at reducing the sulfur dioxides (SO_2). The produced heat converts water into steam that can be utilized for different purposes. The measurements from sensors S_F, S_A, S_L, S_H, S_S and S_E that correspond to different input and output parameters are collected in database repository together with other meta-data describing process conditions for both offline and online analysis. Conducting experiments with pilot CFB reactor and collecting their results into database creates the necessary prerequisites for utilization of the vast amount of DM techniques aimed to identifying valid, novel, potentially useful, and ultimately understandable patterns in data that can be further utilized to facilitate process monitoring, process understanding, and process control.

The estimation of boiler's efficiency is not straightforward. The major estimates that can be taken into account include ratio of produced volumes of steam to the volumes of the consumed fuels, correspondence of volumes of emissions to environmental laws, amortization/damage of parts of boiler's equipment, reaction to fluctuations in power demand, costs and availability of fuels, and others. Correspondingly to these factors, a number of efficiency optimization problems can be defined.

However, developing a common understanding of what basic, enabling, and strategic needs are and how they can be addressed with the DM technology is essentially important. From the DM perspective, having input and output measurements for processes, first of all we are interested; (1) to find patterns of their relation to each other including estimation of process delays, level of gain, and dynamics, (2) to build predictive models of emissions and steam volumes, (3) to build predictive models of char load, having few measurements of it under different conditions during the experiments (in a commercial plant there is no way to measure char inventory on-line as it can be done with oxygen concentration and other output parameters). Currently, we are concentrated on estimation of the responses of the burning rate and fuel inventory to changes in fuel feeding. Different changes in the fuel feed, such as an impulse, step change, linear increase, and cyclic variation have been experimented with pilot CFB reactor. In [1] we focused on one particular task of estimating similarities in data streams from the pilot CFB reactor, estimating appropriateness of the most popular time-warping techniques to the particular domain.

3 Concluding Remarks and Future Work

We recognized basic, enabling, and strategic needs, defined what the current needs are and focused on them, yet continuing to define the most important direction of our further joint research efforts. Those include development of techniques and software tools, which would help to monitor, control, and better understand the individual processes (utilizing domain knowledge about physical dependence between processes and meta-information about changing conditions when searching for patterns, and extracting features at individual and structural levels). Further step is learning to predict char load from few supervised examples that likely lead to adoption of active learning paradigm to time series context. Progressing in these directions we, finally, will be able to address the strategic needs related to construction of intelligent CFB boiler, which would be able to optimize the overall efficiency of combustion processes.

Thus, a DM approach being applied for time series data being accumulated from CFB boilers can make critical advances addressing a varied set of the stated problems.

Acknowledgments. The work is carried out with a financial grant from the Research Fund for Coal and Steel of the European Community (Contract No. RFC-CR-03001).

References

1. Pechenizkiy M. et al. Estimation of Similarities between the Signals from the Pilot CFB-Reactor with Time Warping Techniques. Unpublished technical report (2006)
2. Saastamoinen, J., Modelling of Dynamics of Combustion of Biomass in Fluidized Beds, Thermal Science 8(2), (2004) 107-126
3. Tourunen, A. et al. Study of Operation of a Pilot CFB-Reactor in Dynamic Conditions, In: Proc. 17th Int. Conf. on Fluidized Bed Combustion, ASME, USA (2003) FBC 2003-073

Clustering of Psychological Personality Tests of Criminal Offenders

Markus Breitenbach[1], Tim Brennan[24], William Dieterich[34], and Gregory Z. Grudic[1]

[1] University of Colorado at Boulder, Boulder CO 80309, USA
[2] Institute for Cognitive Science, University of Colorado at Boulder, Boulder CO 80309, USA
[3] Graduate School of Social Work, University of Denver, Denver CO 80202, USA
[4] Northpointe Institute for Public Management Inc.,

Abstract. In Criminology research the question arises if certain types of delinquents can be identified from data, and while there are many cases that can not be clearly labeled, overlapping taxonomies have been proposed in [18], [20] and [21]. In a recent study Juvenile offenders (N = 1572) from three state systems were assessed on a battery of criminogenic risk and needs factors and their official criminal histories. Cluster analysis methods were applied. One problem we encountered is the large number of hybrid cases that have to belong to two or more classes. To eliminate these cases we propose a method that combines the results of Bagged K-Means and the consistency method[1], a semi-supervised learning technique. A manual interpretation of the results showed very interpretable patterns that were linked to existing criminologic research.

1 Introduction

Unsupervised clustering has been applied successfully in many applied disciplines to group cases on the basis of similarity across sets of domain specific features. A typical analytical sequence in the data mining process is to first identify clusters in the data, assess robustness, interpret them and later train a classifier to assign new cases to the respective clusters.

The present study applies several unsupervised clustering techniques to a highly disputed area in criminology i.e. the existence of criminal offender types. Many contemporary criminologist argue against the possibility of separate criminal types [22] while others strongly support their existence (see [20,23]). Relatively few studies in criminology have used Data Mining techniques to identify patterns from data and to examine the existence of criminal types. To date, the available studies have typically used inadequate cross verification techniques, small and inadequate samples and have produced inconsistent or incomplete findings, so that it is often difficult to reconcile the results across these studies. Often, the claims for the existence of "criminal types" have emerged from psychological or social theories that mostly lack empirical verification. Several

attempts have been made to integrate findings from available classification studies These efforts have suggested some potential replications of certain offender types but have been limited by their failure to provide clear classification rules e.g. psychopathic offenders have emerged from a large clinical literature but there remains much dispute over how to identify them, and what specific social and psychological causal factors are critical, and whether or not this type exists among female offenders or among adolescents and whether there are "sub-types" of psychopaths. Thus, a current major challenge in criminology is to address whether reliable patterns or types of criminal offenders can be identified using data mining techniques and whether these may replicate substantive criminal profiles as described in the prior criminological literature.

In a recent study Juvenile offenders ($N = 1572$) from three U.S. state systems were assessed using a battery of criminogenic risk and needs factors as well as official criminal histories. Data mining techniques were applied with the goal to identify intrinsic patterns in this data set and assess whether these replicate any of the main patterns previously proposed in the criminological literature [23]. The present study aimed to identify patterns from this data and to demonstrated that they relate strongly to only certain of the theorized patterns from the prior criminological literature. The implications of these findings for Criminology are manifold. The findings firstly suggest that certain offender pattern can be reliably identified data using a variety of data mining unsupervised clustering techniques. Secondly, the findings strong challenge those criminological theorists who hold that there is only one general "global explanation" of criminality as opposed to multiple pathways with different explanatory models (see [24]).

The present paper describes some difficult analytical problems encountered in applied criminological research that stem from the kind of data produced in this field. A first major problem is that the data is noisy and often unreliable. Second, the empirical clusters are not clear cut so that cases range from strongly classified to poorly classified boundary cases with only weak cluster affiliations. Certain cases may best be seen as hybrids (close to cluster boundaries) or outliers. The distortion of clusters may also be a problem since it is well known many clustering algorithms assign a label to every point in the data, including outliers. Such "forcing" of membership - for both hybrids and outliers - may distort the quality and interpretation of the clustering results. Standard methods such as K-Means will assign cases to the closest cluster center no matter how "far away" from the cluster centers the points are. Some algorithms like EM-Clustering [2] output probabilities of class-membership, but nonetheless eliminating outliers in a unsupervised setting was a hard problem in this area of applied research. In this context we acknowledge that much work has been done to make clustering more robust against outliers, such as using clustering ensembles [3,4] or combining the results of different clustering methods [5], but we are not aware of a method to eliminate points in an aggressive way to obtain a more refined clustering solution, i.e. removing points that are not "close enough" to the cluster center.

Thus, in this research we also demonstrate a methodology to identify well clustered cases. Specifically we combine a semi-supervised technique with an initial standard clustering solution. In this process we obtained highly replicated offender types with clear definitions of each of the reliable and core criminal patterns. These replicated clusters provide social and psychological profiles that bear a strong resemblance to several of the criminal types previously proposed by leading Criminologists [23,20]. However, the present findings firstly go beyond these prior typological proposals by grounding the type descriptions in clear empirical patterns. Secondly they provide explicit classification rules for offender classification that have been absent from this prior literature.

2 Method

We started with an initial solution obtained "manually" using standard K-means and Wards minimum variance method. These have been the preferred choice in numerous social and psychological studies to find hidden or latent typological structure in data [10,11].

However, despite its success standard K-means is vulnerable to data that do not conform to the minimum-variance assumption or expose a manifold structure, that is, regions (clusters) that may wind or straggle across a high-dimensional space. These initial K-means clusters are also vulnerable to remaining outliers or noise in the data. Thus, we proceeded with two additional methods designed to deal more effectively with these outlier and noise problems.

2.1 Bagged K-Means

Bagging has been used with success for many classification and regression tasks [9]. In the context of clustering, bagging generates multiple classification models from bootstrap replicates of the selected training set and then integrates these into one final aggregated model. By using only two-thirds of the training set to create each model, we aimed to achieve models that should be fairly uncorrelated so that the final aggregated model may be more robust to noise or any remaining outliers inherent in the training set.

In [6] a method combining Bagging and K-means clustering is introduced. In our analyses we used the K-means implementation in R [7]. We generated 1000 random bags from our initial sample of 1,572 cases with no outliers removed to obtain cluster solutions for each bag. The centers of these bags were then treated as data points and re-clustered with K-means. The final run of this K-means was first seeded with the centers from our initial solution, which was then tested against one obtained with randomly initialized centers. These resulted in the same solution, suggesting that initializing the centers in these ways did not unduly bias K-means convergence. The resulting stable labels were then used as our final centers for the total dataset and in the voting procedure outlined below.

2.2 Semi-Supervised Clustering

Zhou et.al. introduced the consistency method in [1], a semi-supervised learning technique. This method, given one labeled example per class, assigns all remaining unlabeled cases in accordance with the underlying intrinsic structure of the dataset. Thus, whereas K-means tends to favor (or impose) hyper-spherical clustering structure, the semi-supervised method is more sensitive to almost any arbitrary cluster structure or shape intrinsic to the data being analyzed as long as the point-clouds are connected. The method works by propagating labels from the labeled points to all other points over each iteration. However, the further the point is away from the labeled point, the fewer that label information is propagated. An unlabeled point is assigned to the class with the highest value of activation. This allows the method to follow the shape of arbitrarily shaped clusters as long as they are dense. This is illustrated in figure 1(a) and shows the label assignment for different steps of the propagation. The method has demonstrated good performance on high-dimensional domains such as various image classification tasks.

2.3 Obtaining a Refined Solution: Consensus Cases and Voting Procedure

To tackle the problem of hybrid case elimination we use a voting methodology to eliminate cases in which different algorithms produce a disagreement similar to [5] that combines hierarchical and partitioning clusterings.

In this paper we propose the following solution: First, we use Bagged K-Means [6] to get a stable estimate of our cluster centers in the presence of outliers and hybrid cases. To eliminate cases that are far away from the cluster centers, we will use the obtained centers in a semi-supervised setting with the consistency method [1] to obtain a second set of labels. The labels from the semi-supervised method are obtained with a completely different similarity measure than the K-Means labels. K-Means assigns labels by using the distance to the cluster center (Nearest Neighbor) and works best given clusters that are Gaussian. The semi-supervised consistency method assigns labels with respect to the underlying intrinsic structure of the data and follows the shape of the cluster. These two fundamentally different methods of label assignments are more likely to disagree the further away the point is from the cluster center. We eliminate cases in which the labels do not agree. Note that the consistency method has been demonstrated to work well on high-dimensional data such as images. On the other hand it has been demonstrated that assignments of labels using Nearest Neighbor in high dimensional spaces are often unusable [8].

The process is illustrated in figure 1(b) with a toy example consisting of three Gaussians and a couple of hybrid cases placed in between. The labeling resulting from K-Means and the consistency method differ. The final voting solution consists of less hybrid cases (marked in blue in the bottom right figure).

Using the method outlined above results in roughly half the cases of our data being eliminated. The stability of these central core cases - as retained in the

consensus model - is shown by the almost identical matching of these core cases between the consensus model and the bagged K-means solution ($\kappa = .992, \eta = .994$) and also to the original K-means ($\kappa = 0.949, \eta = 0.947$).

3 Results

The core clusters obtained with this method were interpreted and relationships with types already identified in the criminology literature examined.

The clusters identified were Internalizing Youth A[20,13,16], Socialized Delinquents [12,14,15], Versatile Offenders[20], Normal Accidental Delinquents[18], Internalizing Youth B[20], Low-control Versatile Offenders[20,21] and Normative Delinquency [19]. All the clusters relate to types that have been previously identified various studies in the Criminology literature, but were never identified at the same time in one data set using clustering.

External validation requires finding significant differences between clusters on external (but relevant) variables that were not used in cluster development. By comparing the means and bootstrapped 95 percent confidence intervals of four external variables across the seven clusters from the core consensus solution we identified those variables. The external variables include three criminal history variables (total adjudications, age-at-first adjudication and total violent felony adjudications) and one demographic variable (age-at-assessment). These plots show a number of significant differences in expected directions. For example, clusters 4 and 7, which both match the low risk profile of Moffitt's AL type [18] have significantly later age-at-first adjudication compared to the higher risk cluster 6 that matches Moffitt's high risk LCP and Lykken's [20] Secondary Psychopath. This latter cluster has the earliest age-at-first arrest and significantly higher total adjudications - which is consistent with Moffitt's descriptions.

Finally, while our results indicate that boundary conditions of clusters are obviously unreliable and fuzzy, the central tendencies or core membership appear quite stable. This suggests that these high density regions contain sufficient taxonomic structure to support reliable identification of type membership for a substantial proportion of juvenile offenders.

Using the method in Section 2 we were able to remove most of the hybrid cases. In fact, the case removal was overly aggressive and removed roughly half the data set. However, the remaining cases were very interpretable on manual inspection. Our analyses also show that cluster boundaries are relatively unstable. Kappa's from 0.55 to 0.70, although indicating general overlap, also imply that boundaries between clusters may be imposed differently, and cases close to boundaries may be unreliably classified across adjacent clusters. Many of these cases may be regarded as hybrids with many co-occurring risk or needs and multiple causal influences. Lykken [20] recognized this by stating that many offenders will have mixed etiologies and will be borderline or hybrid cases (p. 21).

The presence of hybrids and outliers appears unavoidable given the multivariate complexity of delinquent behavior, the probabilistic nature of most risk factors and multiplicity of causal factors. Additionally, our findings on boundary

conditions and non-classifiable cases must remain provisional since refinements to our measurement space may reduce boundary problems. Specifically, it is known that the presence of noise and non-discriminating variables can blur category boundaries [10]. Further research may clarify the discriminating power of all classification variables (features) and gradually converge on a reduced space of only the most powerful features.

4 Conclusion

In this paper we report on our experiences with finding clusters in the Youth COMPAS data set which contains 32 scale scores used for criminogenic assessment.

Cluster analysis methods (Ward's method, standard k-means, bagged k-means and a semi-supervised pattern learning technique) were applied to the data. Cross-method verification and external validity were examined. Core or exemplar cases were identified by means of a voting (consensus) procedure. Seven recurrent clusters emerged across replications.

The clusters that were found using unsupervised learning techniques partially replicate several criminal types that have been proposed in previous criminological research. However, the present analyses provide more complete empirical descriptions than in most previous studies and allow. Additionally, the presence of certain sub-types among these major types is suggested by the present analysis. This is the first study in which most of the well replicated patterns were identified purely from the data. We stress that many prior studies provided only partial theoretical or clinical descriptions, omit operational type-identification procedures or provide very limited feature sets.

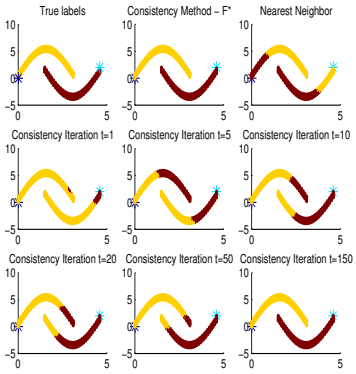
We introduced a novel way of hybrid-case elimination in an unsupervised setting and although we are still working on establishing a more theoretical foundation of the technique it has generally resulted in good results and very interpretable clusters. From the resulting clusters a classifier was build from the data in order to classify new cases.

It is noteworthy that the initial solution we obtained with an elaborate outlier removal process using Ward's linkage and regular K-Means was easily replicated using Bagged K-Means without outlier removal or other "manual" operations. In this instance Bagged K-Means appears to be very robust against noise and outliers.

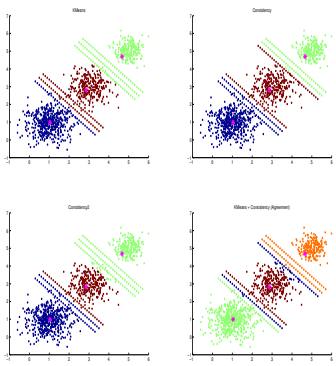
References

1. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. In S. Thrun, L.S., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16, Cambridge, Mass., MIT Press (2004)
2. Dempster, A., Laird, N., Rubin, D.: Maximum-likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society **39** (1977)
3. Dimitriadou, E., Weingessel, A., Hornik, K.: Voting-merging: An ensemble method for clustering. In: Lecture Notes in Computer Science. Volume 2130., Springer Verlag (2001) 217

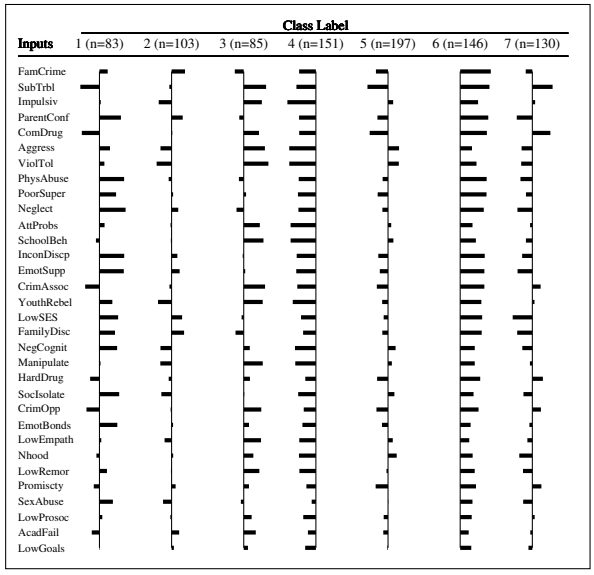
4. Topchy, A.P., Jain, A.K., Punch, W.F.: Combining multiple weak clusterings. In: Proceedings of the ICDM. (2003) 331–338
5. Lin, C.R., Chen, M.S.: Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging. In: IEEE Transactions on Knowledge and Data Engineering. Volume 17. (2005) 145 – 159
6. Dolnicar, S., Leisch, F.: Getting more out of binary data: Segmenting markets by bagged clustering. Working Paper 71, SFB 'Adaptive Information Systems and Modeling in Economics and Management Science" (2000)
7. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2004) 3-900051-07-0.
8. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is 'nearest neighbor" meaningful? Lecture Notes in Computer Science **1540** (1999) 217–235
9. Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2): 123–140.
10. Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In Arabie, P., Hubert, L., and De Soete, G. (eds.), *Clustering and Classification*, World Scientific Press, River Edge, NJ, pp. 345–379.
11. Han, J., and Kamber, M. (2000). *Data Mining - Concepts and Techniques*, Morgan Kauffman, San Francisco.
12. Miller, W. (1958). Lower-class culture as a generating milieu of gang delinquency. *Journal Of Social Issues* 14: 5–19.
13. Miller, M., Kaloupek, D. G., Dillon, A. L., and Keane, T. M. (2004). Externalizing and internalizing subtypes of combat-related PTSD: A replication and extension using the PSY-5 scales. *Journal of Abnormal Psychology* 113(4): 636–645.
14. Jesness, C. F. (1988). The Jesness Inventory Classification System. *Criminal Justice and Behavior* 15(1): 78–91.
15. Warren, M. Q. (1971). Classification of offenders as an aid to efficient management and effective treatment. *Journal of Criminal Law, Criminology, and Police Science* 62: 239–258.
16. Raine, A., Moffitt, T. E., and Caspi, A. (2005). Neurocognitive impairments in boys on the life-course persistent antisocial path. *Journal of Abnormal Psychology* 114(1): 38–49.
17. Brennan, T., and Dieterich, W. (2003). *Youth COMPAS Psychometrics: Reliability and Validity*, Northpointe Institute for Public Management, Traverse City, MI.
18. Moffitt, T. E. (1993). Adolescence-limited and life-course persistent antisocial behavior: A developmental taxonomy. *Psychological Review* 100(4): 674–701.
19. Moffitt, T. E., Caspi, A., Rutter, M., and Silva, P. A. (2001). *Sex Differences in Antisocial Behaviour*, Cambridge University Press, Cambridge,Mass.
20. Lykken, D. (1995). *The Antisocial Personalities*, Lawrence Erlbaum, Hillsdale, N.J.
21. Mealey, L. (1995). The sociobiology of sociopathy: An integrated evolutionary model. *Behavioral and Brain Sciences* 18(3): 523–599.
22. Farrington, D.P. (2005) *Integrated Developmental and Life-Course Theories of Offending*, Transaction Publishers,London (UK).
23. Piquero A.R. and Moffitt T.E. (2005) Explaining the facts of crime: How developmental taxonomy replies to Farrington's Invitation *Chapter in Farrington D.P (Ed) Integrated Developmental and Life-Course Theories of Offending*, Transaction Publishers,London (UK).
24. Osgood D. W. (2005). Making sense of crime and the life course *Annals of AAPSS, November 2005*, 602:196-211.



(a)



(b)



(c)

Fig. 1. (a) **Consistency Method**: two labeled points per class (big stars) are used to label the remaining unlabeled points with respect to the underlying cluster structure. F^* denotes the convergence of the series. (b) **Toy example**: Three Gaussians with hybrid cases in between them. Combining the labels assigned by K-Means (top, left) and the Consistency Method (top, right; bottom, left) with two different σ results in the removal of most of the hybrid cases (blue dots; bottom, right) by requiring consensus between all models build. The K-Means centers have been marked in magenta. (c) **Resulting Cluster Means**: Mean Plots of External Criminal History Measures Across Classes from the Core Consensus Solution with Bootstrapped 95% Confidence Limits.

Onto Clustering of Criminal Careers*

Jeroen S. de Bruin, Tim K. Cocx, Walter A. Kosters,
Jeroen F.J. Laros, and Joost N. Kok

LIACS, Leiden University, The Netherlands
tcocx@liacs.nl

Abstract. We analyze criminal careers through crime nature, frequency, duration and severity. We propose a tool that yields a visual clustering of these criminal careers, enabling the identification of classes of criminals.

1 Introduction

The Dutch national police annually extracts information from digital narrative reports stored throughout the individual departments. This data is compiled into a large and reasonably clean database that contains all criminal records from the last decade. This paper discusses a new tool that attempts to gain new insights into the concept of *criminal careers*, the criminal activities that a single individual exhibits, from this data.

The main contribution of this paper is in Section 4, where the criminal profiles are established and a distance measure is introduced.

2 Background

Background information on clustering techniques in the law enforcement arena can be found in [1, 4]. Our research aims to apply multi-dimensional clustering to criminal careers (rather than crimes or linking perpetrators) in order to constitute a visual representation of classes of these criminals. A theoretical background to criminal careers and the important factors can be found in [2].

3 Approach

We propose a *criminal career analyzer*, which is a multi-phase process visualized in Figure 1. Our tool normalizes all careers to "start" on the same point in time and assigns a profile to each offender. After this step, we compare all possible pairs of offenders on their profiles and profile severity. We then employ a specifically designed distance measure that incorporates crime frequency and the change over time, and finally cluster the result into a two-dimensional image using the method described in [3].

* This research is part of the DALE project (Data Assistance for Law Enforcement) as financed in the ToKeN program from the Netherlands Organization for Scientific Research (NWO) under grant number 634.000.430

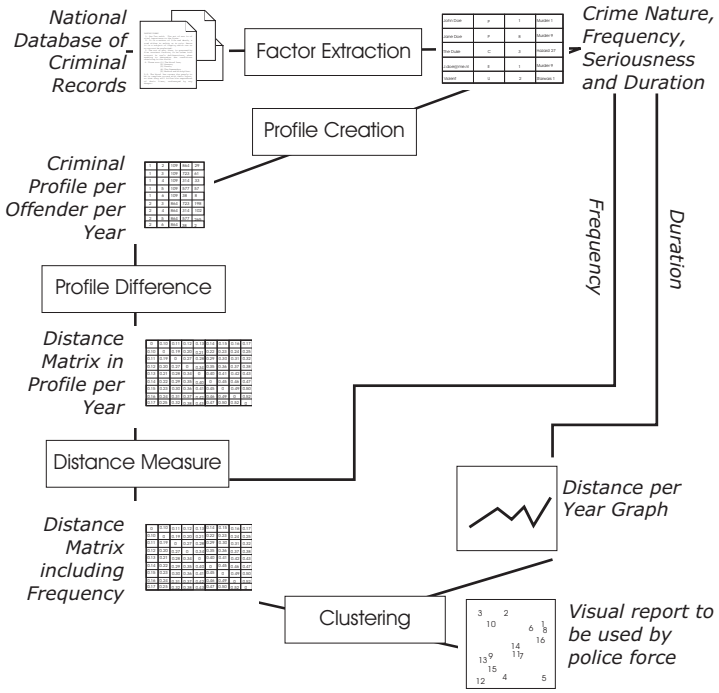


Fig. 1. Analysis of Criminal Careers

4 Method of Career Comparison

We distinguish eight different types of crime (vandalism, ..., sexual violence) divided into three severity classes (minor, intermediate, severe). Each offender's profile x is described by a table containing the percentages $Perc_{ix}$ of crimes that fall within each category i, summing to 1 (if at least one crime is committed). We also assign a severity class k to each crime type. Each severity class gets its own weighting factor $Fact_k$ (1,2,3, respectively). The total severity of crimes in class k for person x (the sum of the appropriate $Perc_{ix}$'s) are then described as Sev_{kx}.

After compilation of all individual profiles, we employ the method described in Formula 1 to create a *profile difference* matrix PD, where PD_{xy} denotes the profile difference between persons x and y:

$$PD_{xy} = \sum_{i=1}^{8} |Perc_{ix} - Perc_{iy}| + |\sum_{k=1}^{3} Fact_k \cdot Sev_{kx} - \sum_{k=1}^{3} Fact_k \cdot Sev_{ky}| \quad (1)$$

This intermediate distance matrix describes the profile difference per year for each possible pair of offenders. Its values all range between 0 and 4.

The crime frequency, or number of crimes, will be divided into categories (0, 1, 2–5, 5–10, >10 crimes per year) to make sure that the absolute difference shares

the range 0–4 with the calculated profile difference, instead of the unbounded number of crimes per year offenders can commit. The *frequency value difference* will be denoted by FVD_{xy}.

Criminal careers of one-time offenders are obviously reasonably similar, although their single crimes may differ largely in category or severity class. However, when looking into the careers of career criminals there are only minor differences to be observed in crime frequency and therefore the descriptive value of profile becomes more important. Consequently, the dependence of the profile difference on the crime frequency must become apparent in our distance measure. This ultimately results into a proposal for the *crime difference per year* distance $CPDY_{xy}$ between persons x and y:

$$CDPY_{xy} = \frac{\frac{1}{4} \cdot PD_{xy} \cdot FVD_{xy} + FVD_{xy}}{8} = FVD_{xy} \cdot \left(\frac{PD_{xy}}{32} + \frac{1}{8} \right) \qquad (2)$$

The factor 1/8 guarantees that $0 \leq CDPY_{xy} \leq 1$.

We have now calculated the career difference distance matrix containing all career comparison information between individual offenders. The clustering method that we incorporated in our tool was described by Broekens et al. [3] and allows data analysts, using node-coloring, to correct small mistakes made by naive clustering algorithms that result in local optima.

5 Experimental Results

Figure 2 gives an impression of the output produced by our tool when analyzing the beforementioned database.

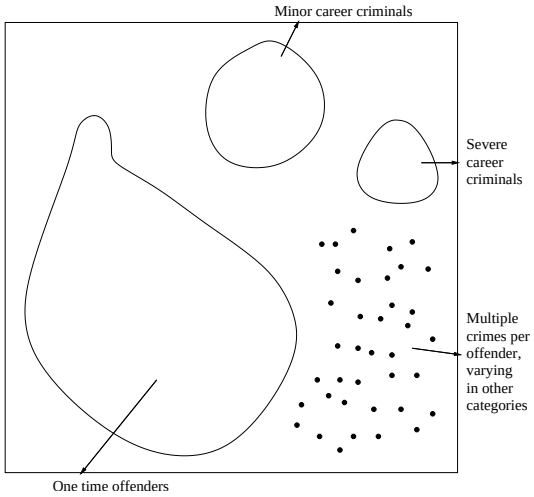


Fig. 2. Experimental results of tool usage

This image clearly shows what identification could easily be coupled to the appearing clusters after examination of its members. It appears to be describing reality very well. The large "cloud" in the left-middle of the image contains (most of the) one-time offenders. This seems to relate to the database very well since approximately 75% of the people it contains has only one felony or misdemeanour on his or her record. The other apparent clusters also represent clear subsets of offenders.

6 Conclusion and Future Directions

The tool we described compiled a criminal profile out of the four important factors describing a criminal career for each individual offender. We developed a specific distance measure to combine profile difference with crime frequency and the change of criminal behavior over time to create a visual two-dimensional clustering overview of criminal careers that is ready to be used by police experts.

The enormous "cloud" of one-time offenders gave a somewhat unclear situational sketch of our distance space. This problem, however, can not be easily addressed since a large part of the National Criminal Record Database simply consists of this type of offenders. Its existence shows, however, that our approach easily creates an identifiable cluster of this special type of criminal career, which is promising. One possible solution to this problem would be to simply not take these individuals into account when compiling our distance matrices.

The used method of clustering provided results that seem to represent reality well, and are clearly usable by police analysts, especially when the above is taken into account. The speed of the chosen approach, however was sub-optimal thus far. In the future, an approach like Progressive Multi Dimensional Scaling [5] could be more suited to the proposed task in a computative way, while maintaining the essence of career analysis. Future research will aim at solving both above mentioned concerns.

References

1. R. Adderley and P. B. Musgrove. Data mining case study: Modeling the behavior of offenders who commit serious sexual assaults. In *KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 215–220, New York, 2001.
2. A. Blumstein, J. Cohen, J. A. Roth, and C. A. Visher. *Criminal Careers and "Career Criminals"*. The National Academies Press, 1986.
3. J. Broekens, T. Cocx, and W.A. Kosters. Object-centered interactive multi-dimensional scaling: Let's ask the expert. To be published in the Proceedings of the 18th BeNeLux Conference on Artificial Intelligence (BNAIC 2006), 2006.
4. T.K. Cocx and W.A. Kosters. A distance measure for determining similarity between criminal investigations. In *Proceedings of the Industrial Conference on Data Mining 2006 (ICDM2006)*, LNCS. Springer, 2006.
5. M. Williams and T. Muzner. Steerable, progressive multidimensional scaling. In *IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 57–64. IEEE, 2004.

Sequential patterns extraction in multitemporal satellite images

Andreea Julea[1,2,3] Nicolas Méger[2] and Emmanuel Trouvé[2]

[1]Universitatea Politehnica Bucuresti, LAPI
Bucharest, ROMANIA - Tel: +4021 402 4683
[2] Université de Savoie, ESIA/LISTIC - BP 806 - F-74016 Annecy Cedex, FRANCE
Tel: +33 450 096 548 Email: {nicolas.meger|emmanuel.trouve}@univ-savoie.fr
[3] Institutul de Stiinte Spatiale Bucuresti, ROMANIA - Tel/fax +4021 457 44 71
Email: {andreeamj}@venus.nipne.ro

Abstract. The frequency and the quality of the images acquired by remote sensing techniques are today so high that end-users can get high volumes of observation data for the same geographic area. In this paper, we propose to make use of sequential patterns to automatically extract evolutions that are contained in a satellite images series, which is considered to be a base of sequences. Experimental results using data originating from the METEOSAT satellite are detailed.

1 Introduction

Data mining techniques that aim at extracting local patterns can be successfully applied to process spatial data. For example, when considering geographical information systems, one can find association rules such as "if a town is large and if it is intersected by a highway then this town is located near to large surfaces of water" [4]. When dealing with satellites images, it is also possible to extract dependencies such as "if visible reflectance intensity ranges from 192 to 255 for the green band and if infrared reflectance intensity ranges from 0 to 63, then high yield is expected" [6]. In this paper, we propose an original approach based on the use of sequential patterns [2] for analyzing multitemporal remote sensing data [1, 3]. Indeed, as sequential patterns can include temporal order, they can be used for extracting frequent evolutions at the pixel level, i.e. frequent evolutions that are observed for geographical zones that are represented by pixels. Section 2 gives a brief introduction to sequential pattern mining while Section 3 details experiments on METEOSAT images.

2 Image series and base of sequences

Let us consider a series of remote sensing images that covers the same area during a period of time. Within each image, each pixel value gives the reflectance intensity of the geographical zone it represents. If the series contains for example 10 images, each image being acquired at a different date, then it is possible to

build for each pixel a sequence of 10 values that are ordered w.r.t. temporal dimension. This sequence of 10 values can be translated into a sequence of 10 symbols where each symbol is associated to a discretization interval. At the image level, this means that we can get a set of millions of short sequences of symbols, each sequence describing the evolution of a given pixel. This context has been identified in data mining as a base of sequences [2]. More precisely, a sequence is an ordered list of L events $\beta_1, \beta_2, \ldots, \beta_L$ which is denoted by $\beta_1 \rightarrow \beta_2 \rightarrow ... \rightarrow \beta_L$ and where each event is a non-empty set of symbols[1]. Let us consider a toy example of a base of sequences, $B = \{A \rightarrow K \rightarrow J \rightarrow B \rightarrow C \rightarrow C, C \rightarrow V \rightarrow T \rightarrow A \rightarrow B \rightarrow K, A \rightarrow G \rightarrow K \rightarrow B \rightarrow J \rightarrow K, A \rightarrow K \rightarrow J \rightarrow M \rightarrow V \rightarrow C\}$. This base describes the evolution of 4 pixels throughout 6 images. For example, values of one pixel are at level A in the first image, level K in the second, level J in the third, level B in the fourth, level C in the fith and level C in the sixth image. In this dataset, one can extract sequential patterns such as $A \rightarrow B \rightarrow K$. If such a pattern occurs in a sequence describing the evolution of a pixel, then the value of this pixel is A at a precise date, then this value changes to B sometime later before further changing to K. In more detail, a sequential pattern $\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_n$ is a sequence and it occurs in a sequence $\beta_1 \rightarrow \beta_2 \rightarrow \ldots \rightarrow \beta_m$ if there exist integers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$ such that $\alpha_1 = \beta_{i_1}$, $\alpha_2 = \beta_{i_2}$, ..., $\alpha_n = \beta_{i_n}$. A pattern is considered to be frequent if it occurs at least once in more than σ sequences, where σ is a user-defined threshold, namely the minimum support. Back to our toy example, if σ is set to 3/4, $A \rightarrow B$ turns to be a frequent sequential pattern. To sum up, if we consider an image series as a base of sequences where each sequence traces the evolution of a given pixel, it is possible to find frequent evolutions at the pixel level by extracting frequent sequential patterns. To do so, we can rely on the various complete algorithms that have been designed to extract frequent sequential patterns (e.g. [2, 7, 5]).

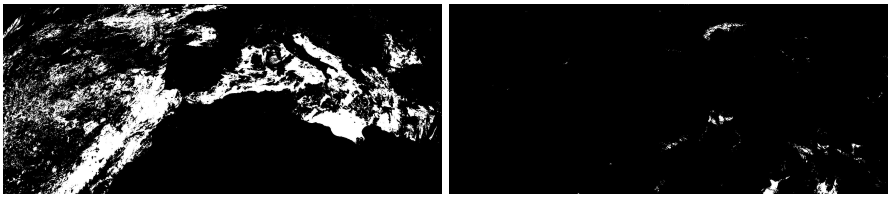
3 Experiments

We used M. J. Zaki's public prototype (http://www.cs.rpi.edu) that implements in C++ the cSPADE algorithm [7]. The images we processed are visible band images (0.5 - 0.9 μm) originating from European geostationary satellite METEOSAT that are encoded in a 256 gray scale format. They can be accessed for free in a slightly degraded JPEG format at http://www.sat.dundee.ac.uk [2]. We decided to use this data as the interpretation is quite straightforward when dealing with low definition and visible band images. The aim was to test this approach for further analyzing high definition images such as radar images from European Remote Sensing (ERS) satellites. Regarding METEOSAT data, we selected images that all cover the same geographic zone (North Atlantic, Europe, Mediterranean regions, North Africa) and that contains 2 262 500 pixels (905x2500). The 8 images we selected were acquired on the 7th, 8th, 9th, 10th,

[1] We refer the reader to [2] for more generic and formal definitions

[2] Website of the NERC satellite images receiving station of Dundee University.

11th, 13th, 14th and 15th of April 2006 at 12.00 GMT. We chose this time to get maximum exposure over the geographical zones covered by images. We then rediscretized the 256 grey levels into 4 intervals in order (1) to reduce the effects due to the acquisition process and to the JPEG degradation, (2) to facilitate results interpretation. Symbols $0, 1, 2, 3$ respectively relate to intervals $[0 - 50]$ (water or vegetation), $]50, 100]$ (soil or thin clouds), $]100, 200]$ (sand or relatively thick clouds), and $]200, 255]$ (thick clouds, bright sand or snow). First patterns appear at $\sigma = 77.5\%$ (execution time[3] = 2 s). It is worth noting that the number of extracted patterns does not exceed 110 until $\sigma = 10\%$ (execution time = 54 s) which facilitate interpretation of results. The first phenomena to be reported



a) b)

Fig. 1. Localization (white pixels) of pattern $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ (a) and pattern $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$ (b).

is cloud passing with patterns $2 \rightarrow 3$ and $3 \rightarrow 2$ ($\sigma = 45\%$). We then identified thick cloud passing over seas and oceans through the following patterns:

$0 \rightarrow 0 \rightarrow 3$ ($\sigma = 25\%$)
$3 \rightarrow 0 \rightarrow 0$, $0 \rightarrow 3 \rightarrow 0$ ($\sigma = 22.5\%$)
$0 \rightarrow 0 \rightarrow 0 \rightarrow 3$, $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ ($\sigma = 17.5\%$)
$0 \rightarrow 3 \rightarrow 0 \rightarrow 0$ ($\sigma = 15\%$) and $3 \rightarrow 0 \rightarrow 0 \rightarrow 0$ ($\sigma = 13\%$)
$0 \rightarrow 0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ ($\sigma = 12\%$) and $0 \rightarrow 0 \rightarrow 3 \rightarrow 0 \rightarrow 0$ ($\sigma = 11\%$)

As depicted in Figure 1, pattern $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ is mainly located in maritime zones. Thus, one can see the outline North Africa and Europe. Other patterns trace thin cloud passing over the oceans:

$0 \rightarrow 0 \rightarrow 2$, $2 \rightarrow 0 \rightarrow 0$, $0 \rightarrow 2 \rightarrow 0$ ($\sigma = 20\%$)
$0 \rightarrow 0 \rightarrow 1$, $1 \rightarrow 0 \rightarrow 0$, $0 \rightarrow 1 \rightarrow 0$ ($\sigma = 20\%$)
$0 \rightarrow 0 \rightarrow 1 \rightarrow 0$ ($\sigma = 14\%$)
$0 \rightarrow 0 \rightarrow 0 \rightarrow 1$, $0 \rightarrow 0 \rightarrow 2 \rightarrow 0$, $0 \rightarrow 2 \rightarrow 0 \rightarrow 0$, $1 \rightarrow 0 \rightarrow 0 \rightarrow 0$ ($\sigma = 13\%$)
$2 \rightarrow 0 \rightarrow 0 \rightarrow 0$ ($\sigma = 12\%$) and $0 \rightarrow 0 \rightarrow 0 \rightarrow 2$ ($\sigma = 11\%$)

The last discovered phenomena shows that some pixels did not change over the image series. For example, we found $0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0$ at $\sigma = 1.4\%$, which means that some ocean zones were not covered by clouds.

[3] All experiments were run on a AMD Athlon(tm) 64 3000+ (1800MHz) platform with 512 MB of RAM under SUSE Linux 10.0 operating system (kernel 2.6.13-15-default).

Another interesting pattern is $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$ at $\sigma = 0.7\%$. As Figure 1 shows, this pattern is located in the Alps (snowy zones, upper part of the image) and in North Africa (bright sand zones, lower part of the image). We here presented 24 patterns out of 110 whose minimum support is greater or equal to 10%. Thus, about 1/4 of extracted patterns can be considered of interest. Results can be refined by adding an infrared band. As an example, other experiments show that it permits to make distinction between snowy zones and bright sand zones. A final interesting result is that when localizing frequent sequential patterns, coherent spatial zones appear (e.g. maritime zones, snowy zones). We began to obtain similar results on radar images from ERS satellite covering the Alps by exhibiting geographical features such as glaciers.

4 Conclusion

We propose to consider a satellite images series as a base of sequence in which evolutions can be traced thanks to sequential patterns. First experiments confirm the potential of this approach by exhibiting well known phenomena. Future works include refined preprocessing such as scaling up from pixel level to region level by using both knowledge of the domain and signal processing.

References

1. IEEE transactions on geoscience and remote sensing : special issue on analysis of multitemporal remote sensing images, November 2003. Volume 41, Number 11, ISSN 0196-2892.
2. R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. S. P. Chen, editors, *Proc. of the 11th International Conference on Data Engineering (ICDE'95)*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.
3. F. Bujor, E. Trouvé, L. Valet, J.-M. Nicolas, and J.-P. Rudant. Application of log-cumulants to the detection of spatiotemporal discontinuities in multitemporal SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 42(10):2073–2084, 2004.
4. K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer and J. R. Herring, editors, *Proc. 4th Int. Symp. Advances in Spatial Databases, SSD*, volume 951, pages 47–66. Springer-Verlag, 6–9 1995.
5. J. Pei, B. Han, B. Mortazavi-Asl, and H. Pinto. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. of the 17th International Conference on Data Engineering (ICDE'01)*, pages 215–226, 2001.
6. W. Perrizo, Q. Ding, Q. Ding, and A. Roy. On mining satellite and other remotely sensed images. In *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 33–40, Santa Barbara, California, May 2001.
7. M. Zaki. Sequence mining in categorical domains: incorporating constraints. In *Proc. of the 9th International Conference on Information and Knowledge Management (CIKM'00)*, pages 422–429, Washington, DC, USA, November 2000.

Carancho - A Decision Support System for Customs

Norton Trevisan Roman[1], Everton Rufino Constantino[1], Helder Ribeiro[1], Jorge Jambeiro Filho[1,2], Antonella Lanna[2], Siome Klein Goldenstein[1], and Jacques Wainer[1]

[1] Institute of Computing – UNICAMP – Brazil
[2] Brazil's Federal Revenue

Abstract. This paper discusses Carancho, a decision support system for fraud detection in customs transactions. The system is based on the outlier detection paradigm as a form of detecting frauds. It has both an interactive and a batch mode. In the interactive mode, a customs expert simultaneously decides on whether a particular transaction is an outlier, and what is a "median" transaction against which the outlier is contrasted. In the batch mode, the system only selects potential outliers for further inspection. The novelty of the system is that instead of attacking the outlier problem in the many dimensions, we project the data into a set of specially formulated relevant dimensions, and detect the outliers in those dimensions.

1 Introduction

Foreign commerce has historically been of great importance as an economical and political instrument worldwide. Tax policy, among other instruments, are forms by which a government controls the trade of goods and services. The problem is that, as one might expect, whenever there is someone charging taxes, there is also someone else trying to avoid paying them.

Unfortunately, this is not only a simple matter of tax evasion. According to the U.S. Department of State [1], practises such as commodities overvaluation allow corporations to perform fund transfers transparently, possibly concealing money laundering schemes. These activities can be directly connected to drug traffic and smuggling. Under this light, an apparently minor tax-evasion offence could be just the tip of the iceberg of more severe crimes.

Some of the common problems found on customs transactions are:

- *Over/Undervaluation*: As mentioned before, incorrect price estimation can conceal illicit transfers of funds. Money laundering schemes in the case of overvaluation, and tax evasion offences in the case of undervaluation.
- *Classification errors*: When assigning a product to one of the predefined categories, the importer can make a honest mistake, misclassifying it. Nevertheless, such a mistake could also conceal an instance of tax evasion, should the misclassification lead to lower tax charges.

– *Origin errors*: Such errors happen when an importer incorrectly declares the goods' country of origin. This is a direct effect of special customs restrictions regarding particular combinations of goods and their origin.
– *Smuggling*: Sometimes importers smuggle different materials into the country amongst the goods they are actually importing. This problem, as reported in [1], could also conceal money laundering schemes, specially when the smuggled material is a high-valued commodity, such as gold.

In this paper, we describe *Carancho*, a graphical decision support system designed to help customs officers decide what should be inspected taking into account all the past international operations.

2 Approaching the Problem

To address the four basic issues stated in the previous section, we follow some of the general ideas regarding the use of unsupervised learning for fraud detection presented in [2]. More specifically, we address such questions as outlier detection problems.

Our choice for the outlier detection framework is based on the assumption that the majority of international trading operations are in order, *i.e.*, that most importers have not only followed the laws properly, but also made no mistakes in the process of importing goods and/or services.

The central problem in applying standard outlier detection techniques to this problem is that the number of relevant attributes for an import declaration is extremely large. Although some of these attributes be continuous, such as weight and price, others are categorical with sometimes a rather large number of classes. Take as an example the merchandise code, that can assume one of approximately ten thousand different types, or the country of origin, which can be one of a set of over a hundred values.

Instead of tackling the outlier detection problem in a large dimensional space, we followed a different approach. Customs experts we interviewed declared that they felt that for each different type of fraud, a small set of single dimensions were enough to detect the outlier with some confidence. Regarding over and under evaluation, the experts defined a set of single dimensions[3] which they felt were the most significant. Their exact definitions are considered confidential, thus this strategic information cannot be made public in this paper.

It is important to notice that the officers' experience and expertise are codified into the system through these dimensions. We think that this approach, besides allowing outlier detection algorithms to be used, gives the system a higher adaptability to new situations than we could have achieve by hard coding a set of rules, as suggested in [3].

We followed a double approach in this research. The Carancho system can work as both a user directed decision support system, which helps the expert to select both outliers and medians, or as a more autonomous batch search for the outliers. Both approaches are described in more detail below.

[3] Currently we are dealing only with three of them.

2.1 Graphical-Assisted Manual Selection

The manual selection approach relies on the expertise and experience of the customs officer to determine what could be considered a suspicious import. Within this framework, the officer is responsible for deciding about the inspection of some goods based on a graphical representation of all operations which are similar to the one under evaluation.

Inspectors can define the level of similarity on the fly, applying filters to the database of transactions. The program retrieves the filtered relevant information, groups the data according to a set of predefined dimensions, and plots their histograms. This is a poor man's attempt to display the transactions' distribution according to these dimensions. The inspector can then, based on this overview of all the similar transactions, decide whether to inspect the cargo more carefully or to clear it right away.

2.2 Batch Search

The batch search is appropriate for inspection of all operations that have already been cleared, but that may deserve later investigation on documentation and financial statements. Basically, the system processes all the database, looking for outliers. This search is an attempt to automate the manual approach, using standard statistical estimation techniques.

Outliers detection takes place as follows. The system searches the whole database, grouping import documents according to their product code[4]. It then analyses each group according to the aforementioned dimensions.

For each of the strategic dimensions, the system finds its robust mean and standard deviations. Then, it looks for transactions that are more than three standard deviations away from the mean in any of these dimensions. That is, it defines a dimension outlier as any import operation whose declared value, for a specific dimension d, is

$$value(d) \geq (mean(d) + 3 \times \sigma(d))$$

An outlier is defined as any import that is an outlier in any of the dimensions. Or, in other words

$$outlier = outlier(d1) \vee outlier(d2) \vee outlier(d3)$$

The results of outlier detection are sent to a file, for later analysis.

3 Results

We ran the batch mode of the system on a 1.5 million transactions database. The outlier detection rate was around 0.9%. That is, from the total amount of import operations, around 14000 were considered outliers.

[4] Similar to the Harmonised System, developed by the World Customs Organisation.

This is a very interesting result, for it allows the customs department to pay special attention to these "strange" operations, as opposed to deciding which ones should be cleared and which should be inspected on a more subjective basis. This saves time and effort, not to mention financial resources.

Naturally, we have to bear in mind that being an outlier does not imply being an outlaw. Some outlier operations are in fact proper ones, they simply do not fit the pattern followed by the majority of similar operations. Nevertheless, this system will increase the customs officers' efficiency, making the decision process swifter and speeding the customs process.

We are still working on assessing the precision of our system. We choose not to use recall as an evaluation measure because there is no way of telling which irregular operations escaped detection.

4 Conclusion

In this paper we presented Carancho – a decision support system for the customs department, aimed at detecting some types of frauds.

We proposed a new approach to the problem, set in two fronts. The first front is intended to help customs officers to make their decisions more quickly, by showing them a graphical representation of the distribution of similar import operations in the past.

The second front is intended to detect those operations that are very unusual but, for some reason, were cleared from physical inspection. In sum, the first approach is meant to detect outliers as they happen, on the fly, whereas the second one is meant to detect past outliers.

Another novelty of our model is the fact that we broke up a multi-dimensional vector space (the dimensions analysed by the system) into a set of one-dimensional spaces, so that we could look for outliers in each of them separately, and then combine the results. Despite the fact that this measure simplified our task, we still need to determine how good a decision it was.

References

1. U.S. Department of State: International Narcotics Control Strategy Report. Volume II: Money Laundering and Financial Crimes. (2005)
2. Bolton, R.J., Hand, D.J.: Statistical fraud detection: A review. Statistical Science **17**(3) (2002) 235–255
3. Singh, A.K., Sahu, R.: Decision support system for hs classification of commodities. In: Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS 2004). (2004)

Bridging the gap between commercial and open-source data mining tools: a case study

Matjaž Kukar, Andrej Zega

University of Ljubljana, Faculty of Computer and Information Science,
Tržaška 25, SI-1001 Ljubljana, Slovenia,
matjaz.kukar@fri.uni-lj.si

Abstract. Most modern database systems now contain an option to include data mining modules. Such an option is very convenient for many business users, as it provides both customer support and data security. However, there arise situations where the vendor-provided data mining options are simply not enough. For such cases open source data mining tools can be used within database systems in order to provide the same security level and improved data mining abilities. We illustrate such an approach with Oracle database system and Weka data mining suite.

1 Introduction

In the last decade the information overload caused by the massive influx of raw data has caused traditional data analysis to become insufficient. This resulted in a new interdisciplinary field of data mining, encompassing both classical statistical, and modern machine learning tools to support the data analysis and knowledge discovery from databases. Every year more and more software companies are including elements of data mining into their products. Most notably, almost all major database vendors now offer data mining as an optional module within business intelligence. Data mining is therefore integrated into the product. While such an integration provides several advantages over the client-server data mining model, there are also some disadvantages. Among them, most important is the complete dependence on the data mining solution as provided by the vendor. On the other hand, most free, open source data mining tools offer unprecedented flexibility and are easily extendable.

We propose a server-side approach that has all the advantages (security and integration) of the commercial data mining solutions while retaining the flexibility of open-source tools. We illustrate our solution with a case study of coupling an open-source data mining suite Weka [5] into Oracle Database Server [2].

2 Methodology

There exist several different possibilities for coupling data mining into database systems [3]. Of those, experiments show that the *Cache-Mine* [3] approach (data read from SQL tables is cached; data mining algorithms are written in SQL as stored procedures) is the preferable in most cases.

Since Oracle Database Server includes a server-side Java virtual machine, extending it with Weka is relatively straightforward, once you get through the sufficient amount of user manuals. Our approach is rather similar to the *CacheMine* with data being cached in RAM memory due to Weka's requirements, and stored procedures being just interfaces to Java classes. Fig. 1 shows an architecture of the extended system. Compiled Java classes or whole JAR archives are uploaded to the server using Oracle-provided tools (loadjava). On the server, appropriate methods are published (in Oracle terminology) and thus made accessible from PL/SQL as stored procedures [1]. Naturally, all Weka classes are accessible from server-side Java programs. Connections to the database are made through JDBC using Oracle's server-side shortcut names. By using the shortcut names database connections remain local and data does not leave the server at any time.

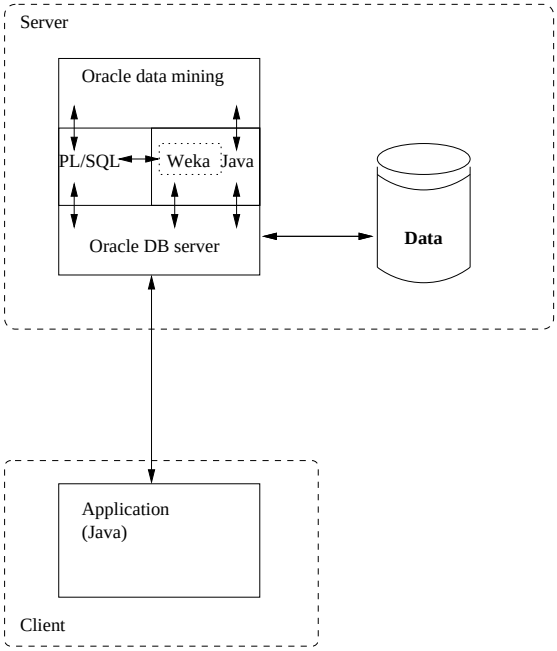


Fig. 1. The architecture of an Oracle Database Server extended with Weka.

3 Experimental evaluation

We performed several experiments on well-known UCI (for smaller problems) and KDD (for large problems) datasets. We compared predictive performance as well as model building time of server-side Weka and Oracle Data Mining[4]. In terms of predictive performance, comparable methods (decision trees, SVMs,

Table 1. A comparison of prediction performance between Weka and Oracle Data Mining on five domains.

	diab	mesh	car	nursery	ecoli
WEKA SMO	77.92	66.07	93.50	93.09	80.00
Oracle SVM	78.36	64.22	91.12	95.08	86.61
WEKA J48	73.05	66.96	89.02	96.41	82.22
Oracle DT	72.69	61.79	82.10	89.10	77.16
WEKA NB	75.65	60.71	88.58	90.57	82.96
Oracle NB	73.66	49.55	81.07	88.89	70.87

naive Bayes) in Weka often (although not always) perform slightly better than their Oracle equivalents (Tab. 1).

On the other side, comparing client- and server-side Weka shows that Oracle's server-side Java virtual machine is about ten times slower than Sun's client-side implementation, when compared on the same computer. Oracle are aware of this problem and claim that their virtual machine is intended for fast transaction processing and not heavy computation. For computationally-intensive Java applications they offer ahead-of-time (native) compilation of Java classes and JARs. By this approach Java bytecode is translated to C source and further compiled into dynamic link libraries (DLLs) that can be accessed from Java through Java Native Interface. Compiled bytecode can be up to ten times faster than interpreted one. Fig. 2 shows model building times for different configurations (Weka on client and server, natively compiled Weka on server – ncomp).

We also compared model building times between Weka and Oracle Data Mining (Fig. 3). For problem sizes of up to 100.000 rows (in this case about 10 megabytes), Weka algorithms perform quicker. For larger problems, Oracle data mining are faster since they are specialized to work with large collections of data.

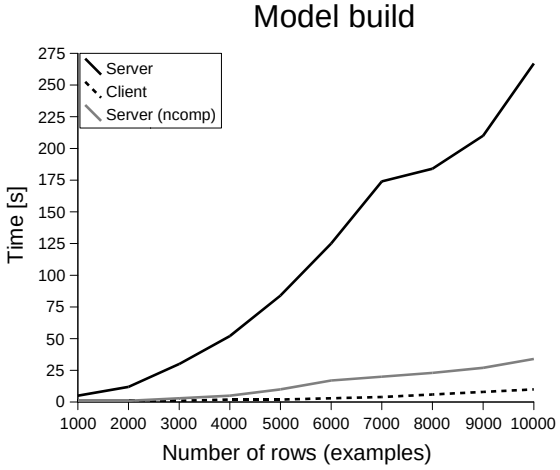


Fig. 2. Building decision tree models with Weka on large KDD datasets.

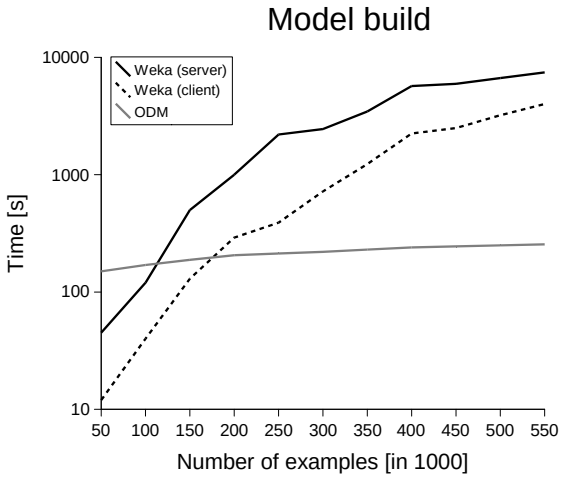


Fig. 3. Comparison of model building times (logarithmic scale) between server-side Weka and Oracle Data Mining.

4 Conclusion

In conclusion, Weka can be pretty successfully used within the Oracle Database Server. It is accessible both from Java and PL/SQL applications. Built models can be stored in a serialized form and later reused. Since data does not leave the server it is much more secure than in the usual client-server approach. One of our business customers (a large telecommunication company) who absolutely refused using a client-server approach and insisted exclusively on server-side data mining, now claim that they would be willing to use such a server-side data mining extension.

References

[1] *Oracle Database Java Developer's Guide 10g Release 1 (10.1).* Oracle, 2004.
[2] *Oracle Data Mining Application Developer's Guide.* Oracle, 2005.
[3] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: alternatives and implications. In *Proc. 1998 ACM-SIGMOD*, pages 343–354, 1998.
[4] M. Taft, R. Krishnan, M. Hornick, D. Muhkin, G. Tang, S. Thomas, and P. Stengard. *Oracle Data Mining Concepts, 10g Release 2 (10.2).* 2005.
[5] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2^{nd} edition, 2005.

Author Index