Data Mining. New Trends, Applications and Challenges

Bart Baesens

University of Southampton

Data Mining: New Trends, Applications and Challenges

Bart Baesens[1]

ABSTRACT

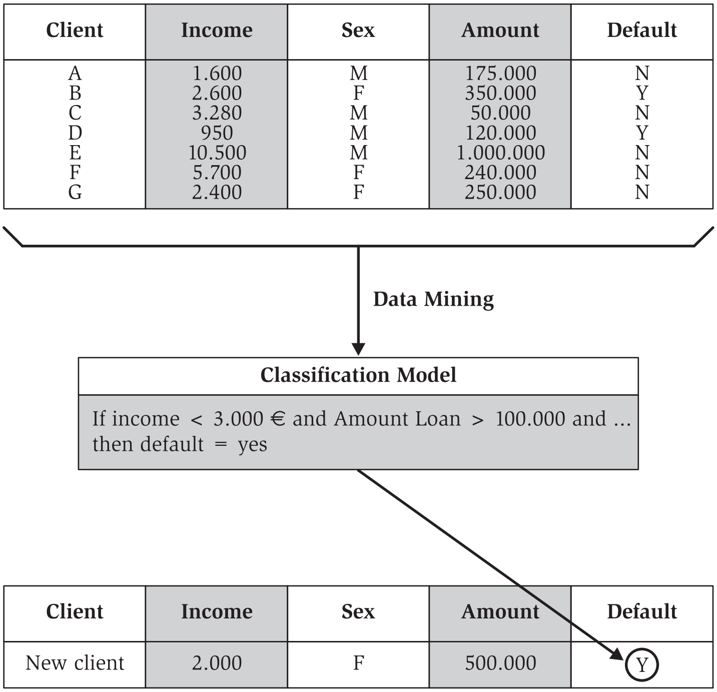
Data mining involves extracting interesting patterns from data to create and enhance decision support systems. Whereas in the early days of data mining, some tasks already relied on statistical and operations research methods such as linear programming and forecasting, data mining methods nowadays are based on a variety of methods including linear and quadratic optimisation, as well as on concepts such as genetic algorithms and artificial ant colonies. Their use has quickly become widespread, with applications in domains ranging from credit risk, marketing, or fraud detection to counter-terrorism. In all of these, data mining is increasingly forming a key part in the actual decision making. Nonetheless, many challenges still need to be tackled, ranging from data quality issues to e.g. the problem of how to include domain experts' knowledge, or how to monitor the performance of the obtained models. In this paper, we outline a series of upcoming trends and challenges within data mining.

JEL CODES C44, C49, C45, C53

KEYWORDS data mining, techniques, applications

I. Introduction: Data Mining

The research area of data mining refers to the process of extracting previously unknown patterns or models from (often very large) data sets. Data mining techniques and applications can be categorised as being either predictive or descriptive (Witten and Frank, 2000). Predictive data mining entails predicting the value for a certain target variable, based on historical data. When this target is discrete, we refer to the task at hand as classification. Applications include predicting the repayment behaviour of loan applications (known as credit scoring), predicting churn, classifying an insurance claim as fraudulent or not, and many more. The credit scoring example problem is illustrated in Figure 1. Regression on the other hand is the task of predicting the value of a continuous target variable. Typical examples are stock price, credit loss and sales amount prediction.



Client	Income	Sex	Amount	Default
A	1.600	M	175.000	N
B	2.600	F	350.000	Y
C	3.280	M	50.000	N
D	950	M	120.000	Y
E	10.500	M	1.000.000	N
F	5.700	F	240.000	N
G	2.400	F	250.000	N

Classification Model

If income < 3.000 € and Amount Loan > 100.000 and ...
then default = yes

Client	Income	Sex	Amount	Default
New client	2.000	F	500.000	Y

Figure 1. Predicting the creditworthiness of loan applications based on historical data.

Descriptive data mining aims at finding patterns that describe underlying relationships in the data. Examples are association rules and clustering. Association rule mining e.g. looks for frequently occurring patterns in the data and is often used for market basket analysis. A well-known example is the rule: if someone buys diapers then this person is also likely to buy beer. Although its actual truthfulness has been questioned, its use as a marketing vehicle for data mining has surely been effective.

In this paper, we will start out by giving a brief overview of the history of data mining. Next, we will discuss some recent upcoming trends and challenges facing data mining techniques. More specifically, we will address the issue of data quality, the need for interpretable white-box data mining models, the role of domain knowledge, the need for backtesting and stress testing data mining models, networked-based learning, and some exciting new application areas. Note that these trends and challenges have also been discussed in Baesens et al. (2009).

II. History of Data Mining

Already in 1941, Durand was the first to use quadratic discriminant analysis for analysing a credit scoring data set of more than 7000 observations. Credit scoring, where the aim is to distinguish good payers from defaulters based on a set of customer characteristics, can undoubtedly be considered as one of the most popular application fields for both data mining and operations research (OR) techniques. Many of the specificities of the problem statement have led to the development of new data mining techniques, as will be illustrated in what follows.

In 1965, Mangasarian introduced the OR method of linear programming to perform linear and nonlinear separation of patterns. This OR based solution to a data mining problem has been validated in several empirical application settings, e.g. credit scoring (Thomas et al., 2002), customer relationship management (Padmanabhan and Tuzhilin, 2003), breast cancer detection (Mangasarian et al., 1995), etc. Furthermore, the more recently suggested Support Vector Machine (SVM) data mining algorithm, essentially builds upon the earlier OR theory of mathematical programming, to come up with powerful non-linear prediction models (Burges, 1998). The ideas of mathematical programming have also been successfully used in other data mining settings, e.g. for data clustering (Bradley et al., 1997), data visualisation (Abbiw-Jackson, 2006), decision tree construction (Bennett, 1992), attribute selection (Bredensteiner and Bennett, 1998; Yang and Olafsson, 2005), and classification cut-off setting (Thomas et al., 2002). Bradley et al. (1999) give an overview of mathematical programming formulations and challenges for data mining.

Earlier forecasting methods have also been introduced for data mining. Popular examples include linear regression and time series analysis. Both techniques have been successfully applied in data mining settings, and laid the foundation for developing new techniques, such as e.g. projection pursuit regression, multivariate adaptive regression splines (MARS), neural networks, regression trees, etc.

The OR based concepts of markov processes and monte carlo sampling have also been widely used in data mining. Examples are model selection and Bayesian network learning (Giudici and Castelo, 2003; Baesens et al., 2002). Social network analysis also often uses the markov property for learning structures from networked data (cf. infra).

Recently, ant colony optimisation (ACO) algorithms have been introduced as a new promising optimisation technique (Dorigo and Stützle, 2004). Initial successful applications include the travelling salesman problem, manufacturing control, and routing of packages through the Internet (Di Caro and Dorigo, 1998). However, the same optimisation method can also be applied for data mining. E.g., Martens et al (2007b) used ACO to develop AntMiner+, a system capable of extracting if-then classification rules from data. Genetic/evolutionary algorithms are another example of optimisation algorithms that can serve both traditional OR as well as data mining applications. As such, the latter have been successfully used to perform feature selection, train neural networks, infer classification rules, or optimise the weights of classifiers in a data mining ensemble (Freitas, 2007).

In what follows, we will outline a series of upcoming trends and challenges for data mining. It is obvious that these challenges will require new developments, e.g. by devising new (optimisation) algorithms, or re-using earlier introduced data mining methods in new settings.

III. Data Quality

Data mining relies on the use of data. Bad data yields bad models. This is often referred to as the GIGO principle: garbage data in, garbage models out. Hence, it is of crucial importance that data is of good quality in order to obtain good data mining models. Furthermore, it has been shown in many data mining studies that simple models typically perform well on most data sets. E.g., Baesens et al. (2003b) showed that logistic regression typically performs quite well for classification in their benchmarking study, whereas Holte (1993) obtained a similar finding using simple classification rules. Hence, in several areas, the best way to augment the performance of a data mining model often is to improve data quality. Measuring data quality is however not an easy endeavour and many challenges lie ahead in trying to improve data quality. In what follows, we will discuss five elements relating to data quality: data accuracy, data completeness, data bias and sampling, data transformations, and data definition.

Data accuracy determines to which extent the data accurately and consistently measures what it is intended to measure. Bad data accuracy can be caused by data entry errors and/or measurement errors. Also outliers might be a sign of bad data accuracy, although it needs to be noted that an outlier does not necessarily present an invalid observation. E.g., when considering ratio variables in a credit scoring data mining system, it is often observed that these variables have highly dispersed distributions (because the division of two Gaussian distribution is a Cauchy distribution with fat tails). Hence, adequate schemes should be adopted to deal with these valid, outlying observations in the most optimal way in order for the data mining system to better detect patterns in the data. One way is using robust winsorising schemes (Van Gestel et al., 2005). However, other interesting outlier handling schemes could be developed, taking into account the type of variable, why it is an outlier, and of course the data mining technique adopted.

Data completeness refers to the degree of missing values and/or observations in the data. Missing information should clearly be minimised, especially for variables that are retained in the data mining model; however, if present, appropriate procedures should be used to deal with it. Although many schemes have been suggested to deal with missing values (e.g. (multiple) imputation routines, introducing a separate category/variable, etc.), it is still not always clear which is the best scheme given the problem domain and the data mining technique adopted.

The next data quality criterion relates to data bias and sampling. When building a data mining model, one typically starts from a sample of data. A first important issue relates to the sampling strategy that will be adopted. Carefully selecting the observations using both active and adaptive learning strategies becomes more and more important. Also, selecting a (near-)optimal set of variables is far from straightforward, particularly in high-dimensional data sets, which often occur in bio-informatics applications. It needs to be noted that, in many cases, the sample that is taken is not representative for the population on which the data mining model built using that sample is going to be used. E.g., when building data mining models for credit scoring, one only has the target variable (good/bad) available for the sample of past accepts, whereas the future population may not be exactly the same, because of the past credit policy. This leads to the problem of reject inference, which, although studied already extensively, has to date not been solved in an unambiguous way (Thomas et al., 2002). Closely related to this is the problem of policy inference, where data mining models are being used to apply different policies to different customers, making it less evident to decide whether the actual outcome was due to the policy applied or to customer characteristics. Similar problems also often occur in other data mining settings.

Hence, studying the impact of data bias and its impact on data mining models is clearly an important issue of future research.

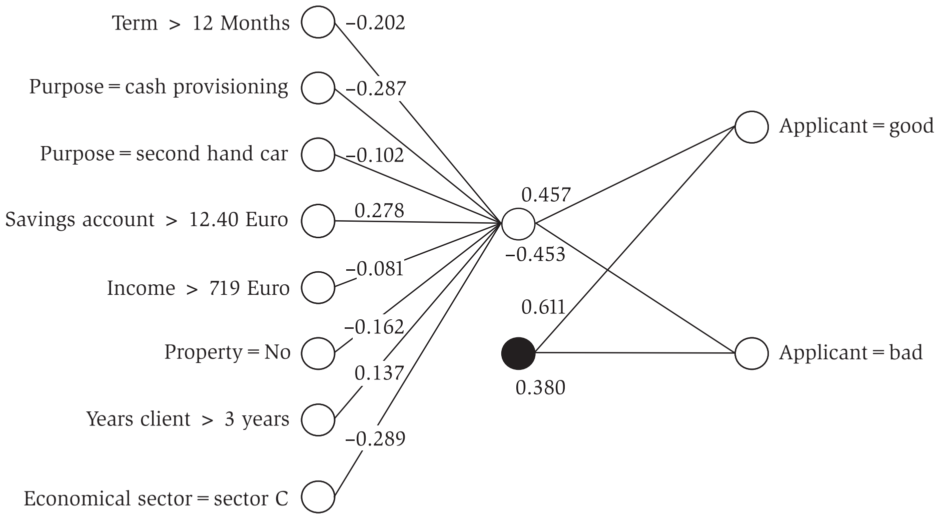
Data transformations are often applied to make attributes more informative for the data mining technique so as to more easily detect patterns. A popular example is coarse classification, which categorises categorical and/or continuous variables for more robust analysis or for introducing non-linear effects into linear models. Another example are Box-Cox transformations, which are a set of simple logarithmic transformations that may significantly improve the performance of classification models (Van Gestel et al., 2005). Principal component analysis is also often used to transform the variables to a set of uncorrelated principal components, typically at the cost of decreased interpretability. A crucial issue here is which transformations are most appropriate considering the data mining task at hand (i.e., classification, regression, clustering, etc.) and given model priorities (accuracy, interpretability, re-training efforts, etc.).

Data definition relates to the way data attributes have been defined. Consider e.g. a corporate credit risk rating system in which the ratio debt/earnings is used as a variable in a linear/logistic regression based data mining system. Because earnings can be both negative and positive, the impact of that variable on the default risk is not uniquely defined. By defining the variable in that way, one may already enforce the model to make a suboptimal decision, both in terms of performance and interpretability. Hence, it is of crucial important to think carefully about how variables have been defined and what might be their impact on the target variable of interest.

Summarising, in the preceding paragraphs we have outlined various issues related to data quality. We clearly stressed that these issues should be dealt with in the most optimal way in order to obtain a powerful data mining system. However, further research is needed on how to improve data quality by either redesigning data entry processes (using e.g. validation constraints, business rules, etc.), or by developing new procedures to encode, transform or redefine data in the most optimal way, hereby allowing the data mining system to better detect patterns in the data.

IV. Interpretable Data Mining Models

During the past years, many data mining algorithms have been developed for a variety of different analysis problems. Examples are neural networks, support vector machines, Bayesian networks, genetic algorithms, fuzzy techniques, swarm intelligence, etc. An important focus of these developments was the detection of patterns in data, in the most optimal way, i.e. given a performance metric that was defined on beforehand (e.g. classification accuracy, R-squared, mean squared error, etc.). Although undoubtedly not unimportant, performance is however not the only criterion that matters for the successful deployment of data mining models in the business. More and more, interpretability and model readability are considered key critical success factors. A popular illustration of this is the recently suggested Basel II Capital Accord, which encourages financial institutions to develop data mining models estimating default risk, loss risk and exposure risk. Given the strategic impact of these models on e.g. capital levels maintained, financial regulators are reluctant to approve the use of complex, black-box models; hence the need for white-box data mining models that give a clear and transparent view of the patterns in the data. The often referred to Ockam's razor embodies this idea of model simplicity. However, aiming for model understandability often comes at the cost of decreased performance, so trade-offs between model readability and model performance need to be taken into account. During the past years, attempts have been made to balance this trade-off. Baesens et al. (2003a) used neural network rule extraction as a way of opening up the neural network black box and shed light on its



if Term > 12 months **and** Purpose = cash provisioning **and** Savings Account ≤ 12.40 Euro **and** Years client ≤ 3 **then** *Applicant = bad*
if Term > 12 months **and** Purpose = cash provisioning **and** Owns Property = No **and** Savings Account ≤ 12.40 Euro **then** *Applicant = bad*
if Purpose = cash provisioning **and** Income > 719 Euro **and** Owns Property = No **and** Savings Account ≤ 12.40 Euro **and** Years client ≤ 3 **then** *Applicant = bad*
if Purpose = second hand car **and** Income > 719 Euro **and** Owns Property = No **and** Savings Account ≤ 12.40 Euro **and** Years client ≤ 3 **then** *Applicant = bad*
if Savings Account ≤ 12.40 Euro **and** Economical sector = Sector C **then** *Applicant = bad*
Default class: *Applicant = good*

1. Savings Account	≤ 12.40 Euro													> 12.40 Euro
2. Economical sector	Sector C	other												–
3. Purpose	–	cash provisioning							second-hand car				other	–
4. Term	–	≤ 12 months				> 12 months			–				–	–
5. Years client	–	≤ 3			> 3	≤ 3	> 3		≤ 3			> 3	–	–
6. Owns Property	–	Yes	No		–	–	Yes	No	Yes	No		–	–	–
7. Income	–	–	≤ 719 Euro	> 719 Euro	–	–	–	–	–	≤ 719 Euro	> 719 Euro	–	–	–
1. applicant = good	–	x	x	–	x	–	x	–	x	x	–	x	x	x
2. applicant = bad	x	–	–	x	–	x	–	x	–	–	x	–	–	–
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Figure 2. Developing interpretable data mining models using neural network rule extraction and decision tables (Baesens et al. 2003a). The network above was used to extract propositional if-then rules presented in the middle, which were then transformed to a decision table given at the bottom of the figure. The white circles represent the neurons of a neural network. The black circle refers to the bias neuron (intercept). The numbers represent the weights connecting each of the nodes with the subsequent layer.

decision logic, by extracting propositional if-then rules mimicking its behaviour. Furthermore, in order to represent the rules in the most user-friendly way, decision tables were adopted. This entire process is illustrated in Figure 2. The translation of this work from neural networks to support vector machines is currently also being addressed (e.g. Martens et al., 2007a).

Martens et al. (2007b) used Ant-Colony Optimisation (ACO) algorithms to also extract propositional if-then rules from data. It needs to be noted that also other types of rules (e.g. M-of-N rules, oblique rules, or fuzzy rules) can be considered as viable alternatives. Further research is needed to study the most suitable rule representation in terms of simplicity and ease of use, so as to arrive at interpretable and user-friendly data mining models.

Van Gestel et al. (2005) introduced an alternative way of obtaining interpretable models by using an additive combination of linear (logistic regression) and non-linear (support vector machines) modelling techniques to optimally balance between model interpretability (obtained from the linear part) and model performance (obtained from the non-linear part). Note that also Bayesian networks are considered white-box probabilistic models with a high degree of model transparency and readability.

Despite recent advances, many challenges still lie ahead in obtaining interpretable data mining models. Examples of such challenges are:

- What is the preferred model representation in terms of simplicity (e.g. rule based models, linear models, graphical models, etc.)?
- How to decide upon the optimal balance between model interpretability and model performance?
- How to suitably quantify not just performance but both criteria?

V. Incorporating Domain Knowledge

Although data mining techniques were originally introduced to analyse large-scale data sets, the same techniques are also more and more being used to mine small data sets. Examples of these are in medical settings where only a limited number of observations may be available, or financial institutions with low default portfolios (e.g. exposures to sovereigns, banks, etc.). Detecting patterns from small volume data sets requires the use of special techniques. In this context, one often tries to include domain specific expert knowledge during the learning process. In this way, the purpose is to arrive at a model using an optimal combination of the limited data with the available business expert knowledge. How to consolidate the automatically generated data mining knowledge with the knowledge reflecting experts' domain expertise constitutes the so-called knowledge fusion problem (Martens et al., 2006).

Many types of constraints exist that a domain expert might want to incorporate. The most popular constraint is the monotonicity constraint, stating that an increase in a certain input variable cannot lead to a decrease in the output variable. A credit scoring example is that an increasing income, keeping all other variables equal, should yield a decreasing probability of loan default. Although almost all research into this field is focused on this constraint, the taxonomy of possible domain constraints, shown in Table 1, indicates that this is arguably a too limited view. It is, however, the most common type of constraint to be incorporated. Each constraint can be either mandatory, which we name a hard constraint, or simply preferred, which we name a soft constraint. An example of multivariate constraints is demanding that variable X is more important in the model than variable Y. A detailed overview of examples and existing research can be found in Martens and Baesens (2008). The monotonicity constraint typically corresponds to a univariate ordinal constraint, as depicted in Table 2.

Table 1. Taxonomy of possible domain constraints to incorporate into the data mining model.

	Univariate				Multivariate
	nominal	ordinal			
		linear	non-linear		
			piecewise linear	non-piecewise linear	
Soft					
Hard					

In early OR approaches to data mining, such as linear programming, it was very easy to include domain knowledge. E.g., an additional constraint specifying that the weight of one variable should be higher than a constant or the weight of some other variable, could often do the job. However, in recently suggested techniques (e.g., rule extraction techniques, neural networks, support vector machines, ACO systems, or Bayesian networks), it is less evident what is the optimal way to take into account domain knowledge. Hence, also in this area, many issues and challenges for further study remain. Examples are:

- What is the best way of formalising domain knowledge?
- How to elicit domain knowledge from business experts?
- How to include domain knowledge when learning patterns from data?
- How to resolve conflicts between patterns learnt from data and domain knowledge?

VI. Backtesting and Stress Testing Data Mining Models

Data mining models are typically constructed using a snapshot of historic data. Once deployed in the organisation, their performance may degrade over time. Three obvious reasons for this are: sampling variability, macro-economic influences and endogenous effects. Sampling variability is the uncertainty due to the fact that the sample only gives a limited view of the population. Macro-economic influences reflect the impact economic up- or downturns may have on the model. Endogenous effects represent changes induced by the firm itself, e.g. strategy changes, exploration of new market segments, etc. These three effects typically have a strong impact on the performance of a data mining model, and they typically occur in a mixed way. Clearly, this is a complicating factor when backtesting the models, i.e. tracking their performance over time. Hence, backtesting frameworks should be developed to constantly monitor the behaviour of data mining models, and perform diagnostic checks on its performance taking into account the influences mentioned above. In doing so, it is important to be aware about the philosophy that was adopted when designing data mining models. A distinction can sometimes be made between point-in-time (PIT) data mining models, which take into account both cyclical and non-cyclical information, and through-the-cycle (TTC) data mining models which only focus on non-cyclical inputs. E.g. a customer score can measure churn behaviour or default risk during the upcoming year (PIT), or over a longer term horizon (TTC). Knowing this helps to judge the significance of temporary fluctuations due to e.g. macro-economic volatility. Clearly, both PIT and TTC represent the extremes of a continuum and knowing towards which extreme the data mining system is situated, is key before any backtesting exercise can start.

When working out a backtesting framework, a first critical question is which test statistics, metrics and/or measurements should be used for monitoring data mining model performance. E.g. in a clustering setting, one might be interested in monitoring cluster stability, whereas in a classification setting one might be interested in monitoring data stability, discriminatory power and/or probability calibration. When developing these backtesting diagnostics, one should clearly take into account the three influences measured above, combined with the fact that behaviour of customers might be correlated. Measuring and taking into account these correlation effects is a key challenge for a successful backtesting exercise. Furthermore, clear advice should be given regarding the significance levels to be adopted. Next, all the diagnostic checks need to be combined in a user-friendly intelligent dashboard environment allowing a constant, real-time monitoring of the data mining model. Finally, action plans should be available that specify what to do in case the data mining model starts degrading in performance, ranging from simple parameter adjustment and incremental learning to complete model re-estimation. Although preliminary work has been done in this area (see e.g. Castermans et al., 2007 for an overview on backtesting credit risk models), more research needs to be conducted to appropriately answer the above questions and come up with powerful backtesting frameworks that can be implemented in a user-friendly way.

Since data mining models are being more and more used as key inputs for strategic organisation processes (e.g. allocating marketing budgets, capital calculation in a Basel II setting), it is crucial to know how they behave under adverse economic circumstances. The aim of stress testing is to determine the impact of stress events (e.g. macro-economic downturns) on data mining models. This can range from simple sensitivity checks to advanced scenario based analysis, based on e.g. historical or hypothetical scenarios. It is clear that also in this area, much work lies ahead to come up with sound and solid methods for stress testing data mining models and/or making them less stress sensitive.

VII. Network-Based Learning

Typically, data mining is performed on a set of independent and identically distributed (iid) observations. However, in practice most observations relate to agents acting in a networked based structure. Examples of networked data are: web pages connected by hyperlinks, research papers linked by citations, customers linked by social networks (e.g. via telephone calls, or through social network sites such as LinkedIn or Facebook). It is believed that quantifying these social network structures and including them as part of the data mining learning process can have a significant impact on the type and power of patterns and/or models being learnt. For example, graph-based data mining algorithms can be used for community discovery and/or web structure mining, or information about an individual's social network may improve the performance of response or churn models in the marketing domain.

However, many challenges lie ahead. Since the class membership of one entity may influence the class membership of another, relational data mining algorithms are being developed that take into account the network dependencies. Furthermore, methods of collective inference are being devised to determine how the unknown nodes in a social network can be estimated together. Also, since a network functions as a complete structure, it is hard to separate it into a training and independent test set for performance calculation. Although preliminary work has been done in this area (see e.g. Macskassy and Provost 2007; Lu and Getoor, 2003), there is currently a strong need to see more empirical applications of network-based learning (in order to help us assess its added value over existing approaches), and potentially, for new algorithms to be developed. Again, this creates new opportunities for using earlier developed OR techniques in exciting novel settings.

VIII. New Application Areas

Having been originally introduced in well-known problem domains such as customer relationship management (CRM), or specific applications such as churn prediction, fraud detection, or credit scoring, data mining is currently being applied in many new exciting application areas. A first interesting novel application is web analytics, where data mining is being used for analysing all kinds of data collected from the World Wide Web (Liu, 2006). This involves web usage mining, which studies web logs for detecting web usage patterns, web content mining, which focuses on analysing the content of web pages, and web structure mining, which aims at mining communities from a set of hyperlinks.

Another popular application domain is bio-informatics, where data mining is being studied for analysing biological data, such as protein sequence databases (De Moor et al. 2003). Also text mining applications are becoming increasingly popular. Typical challenges in both are the relatively small number of observations compared to the high number of dimensions, which calls for special types of algorithms that are able to cope with such sparse data. Data mining is also becoming more and more popular for classical medical analysis problems (e.g. cancer detection, drug effect analysis, etc.). A key challenge in this field again concerns the merger between patterns learnt from data, and knowledge extracted from domain experts (e.g. medical practitioners).

Data mining techniques are currently also being applied to terrorism prevention. The application of social network mining techniques seems very promising in this area. However, major criticisms here relate to data confidentiality, and civil liberties and the protection of privacy. Consumer privacy is also a major concern for mining RFID (Radio Frequency Identification) data, particularly in a retail context. RFID is an increasingly popular supply-chain technology that overcomes several of the limitations of traditional bar codes and which can provide huge amounts of data on product identity, location and movement. RFID mining would be of particular interest for OR as it would allow one to track products as they move through the supply chain or are transported across e.g. large warehouse spaces, which can aid in the further optimisation of these processes.

Another exciting application is in software engineering, where data mining has been introduced to predict faults in software, or software development efforts (Lessmann et al. 2008). More recently, work has also begun on mining data in software repositories for tasks ranging from identifying code usage patterns to the analysis of change patterns to assist in future development or e.g. in assigning programmers to particular tasks. Research in this area is aided by the public availability of large amounts of data on open-source software development projects.

Closely related to this is the area of process mining, where data mining techniques are being used to mine process models (e.g. Petri nets) from event logs of an information system (Van der Aalst, 2004). This allows organisations to better understand their key operational processes and workflows by showing them in what order, by whom, etc. various activities are being performed in practice, and how this may differ from the processes designed “on paper”.

Also emerging are real-time data mining applications, which require that interesting patterns or anomalies are being discovered on a continuous basis. Data and customer behaviour is changing on a continuous basis. Hence, the idea of building a ‘static’ data mining model that is subsequently used for a fixed period of time is no longer appropriate. The data mining model has to be amenable to very quick, if not automatic updating to allow for real-time decision making. Applications are typically found in product configuration and pricing (Seow and Thomas, 2007), fraud prevention, or e.g. also in network intrusion detection systems (Lee et al., 2001).

The usage of data mining techniques for analysing videos (video mining) also offers many exciting applications. Examples are, e.g., image/video retrieval, video summarisation, visual surveillance, and real-time decision making during sports matches. Closely related, is the application of data mining to astronomy data, where the aim is to analyse data obtained by telescopes to find new phenomena, relationships and useful knowledge about the universe.

IX. Conclusions

The importance of data mining has steadily grown in recent decades, with an ever increasing range of techniques being developed. Whereas traditionally many applications have been in the finance and marketing domains, its use has also spread to several other domains. However, many challenges and opportunities for data mining are still ahead of us. In this paper we identified a number of outstanding issues and current trends relating to data quality, the interpretability of data mining models, the incorporation of domain knowledge into the data mining process, the need to backtest and stress test models, network-based learning, and a number of new application fields.

NOTE

[1] K.U.Leuven, Department of Decision Sciences and Information Management, Naamsestraat 69, B-3000 Leuven, Belgium, Bart.Baesens@econ.kuleuven.ac.be

REFERENCES

Baesens B., Setiono R., Mues C., Vanthienen J., Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation, *Management Science,* Volume 49, Number 3, pp. 312-329, March 2003a.

Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., Vanthienen J., Benchmarking State of the Art Classification Algorithms for Credit Scoring, *Journal of the Operational Research Society*, Volume 54, Number 6, pp. 627-635, 2003b.

Baesens B., Egmont-Petersen M., Castelo R., Vanthienen J., Learning Bayesian Network Classifiers for Credit Scoring using Markov Chain Monte Carlo Search, *Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR'2002),* IEEE Computer Society, Québec, Canada, pp. 49-52, August 2002.

Baesens B., Van Gestel T., Stepanova M., Van den Poel D., Vanthienen J., Neural Network Survival Analysis for Personal Loan Data, *Journal of the Operational Research Society, Special Issue on Credit Scoring,* Volume 59, Number 9, pp. 1089-1098, 2005

Baesens B., Mues C., Martens D., Vanthienen J., 50 years of Data Mining and OR: upcoming trends and challenges, *Journal of the Operational Research Society*, 2009, forthcoming.

Bennett K.P., Decision tree construction via linear programming, In M. Evants, editor, *Proceedings of the fourth Midwest Artificial Intelligence and Cognitive Science Conference*, pp. 97-101, Utica, Illinois, 1992.

Bradley P.S., Fayyad U.M., Mangasarian O.L., Mathematical programming for data mining: formulations and challenges, *Informs Journal on Computing*, Volume 11, Issue 3, pp. 217-238, 1999.

Bradley P.S., Mangasarian O.L., Street W.N., Clustering via concave minimization, M.C. Mozer, M.I. Jordan, T. Petsche, eds., *Advances in Neural Information Processing Systems*, MIT Press, Cambridge MA, 1997.

Bredensteiner E.J., Bennett, K.P., Feature minimization within decision trees, *Computational Optimizations and Applications*, 10, pp. 111-126, 1998.

Burges J.C., A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2, pp. 121-167, 1998.

Castermans G., Martens D., Van Gestel T., Hamers B., Baesens B., An Overview and Framework for PD Backtesting and Benchmarking, *Conference on Credit Scoring and Credit Control X*, Edinburgh (U.K.), July 2007

De Moor B., Marchal K., Mathys J., Moreau Y., Bioinformatics: Organisms from Venus, technology from Jupiter, algorithms from Mars, *European Journal of Control*, 9, 237-278, 2003.

Di Caro G., Dorigo M., Antnet: Distributed stigmergetic control for communications networks, *Journal of Artificial Intelligence Research*, 9, pp. 317-365, 1998.

Dorigo M., Stützle T., *Ant Colony Optimization*, MIT Press, Cambridge MA, 2004.

Durand D., *Risk Elements in Consumer Instalment financing*, National Bureau of Economic Research, New York, 1941.

Fildes R., Nikolopoulos K., Crone S.F, Syntetos A.A, Forecasting and Operational Research: a review, Journal of the Operational Research Society, forthcoming, 2008.

Freitas A.A., A Review of Evolutionary Algorithms for Data Mining, In: O. Maimon and L. Rokach (Eds.), *Soft Computing for Knowledge Discovery and Data Mining*, pp. 61-93, Springer, 2007.

Giudici P., Castelo R., Improving Markov Chain Monte Carlo model search for data mining, *Machine Learning*, 50, 1-2, pp. 127-158, 2003.

Holte R.C., Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, *Machine Learning*, 11, 1, pp. 63-90, 1993.

Karasözen B., Rubinov A., Weber G.H., Optimization in Data Mining, *European Journal of Operational Research*, Volume 173, Issue 3, pp. 701-704, 2006.

Lessmann S., Baesens B., Mues C., Pietsch S., Benchmarking classification models for software defect prediction: A proposed framework and novel findings, *IEEE Transactions on Software Engineering*, 2008, forthcoming

Lee W., Stolfo S.J., Chan P.K., Eskin E., Wei F, Miller M.; Hershkop S., Junxin Z., Real time data mining-based intrusion detection, *DARPA Information Survivability Conference & Exposition II*, DISCEX '01, pp. 89-100, 2001.

Liu B., Web Data Mining: Exploring hyperlinks, content and usage data, *Springer*, 2006.

Lu Q., Getoor L., Link-based Classification, *Proceeding of the Twentieth Conference on Machine Learning (ICML-2003)*, Washington DC, 2003

Macskassy S.A., Provost F., Classification in Networked Data: A Toolkit and a univariate case study, *Journal of Machine Learning Research*, 8, pp. 935-983, 2007.

Mangasarian O.L., Linear and nonlinear separation of patterns by linear programming, *Operations Research*, 13, pp. 455-461, 1965.

Mangasarian O.L., Street W.N., Wolberg W.H., Breast cancer diagnosis and prognosis via linear programming, *Operations Research*, 43(4), pp. 570-577, July-August 1995.

Martens D., De Backer M., Haesen R., Baesens B., Mues C., Vanthienen J., Ant-Based Approach to the Knowledge Fusion Problem, *Lecture Notes in Computer Science*, ANTS Workshop 2006, Brussels (Belgium), pp. 84-95, September 2006.

Martens D., Baesens B., Van Gestel T., Vanthienen J., Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3), pp. 1466-1476, 2007a.

Martens D., De Backer M., Haesen R., Vanthienen J., Snoeck M., Baesens B., Classification with Ant Colony Optimization, *IEEE Transactions on Evolutionary Computation*, Volume 11, Number 5, pp. 651-665, 2007b.

Martens D., Baesens B., Building Acceptable Classification Models, *Annals of Information Systems*, 2008, forthcoming.

Olafsson S., Introduction to Operations research and data mining, *Computers and Operations research*, Volume 33, 11, 3067-3069, 2006.

Padmanabhan B., Tuzhilin A., On the use of optimization for data mining Theoretical interactions and eCRM opportunities, *Management Science;* Volume 49, Issue 10, pp. 1327-1343, 2003.

Rao M.R., Cluster analysis and mathematical programming, *Journal of the American Statistical Association*, Volume 66, pp. 622-626, 1971.

Seow H.Y., Thomas L.C., To ask or not to ask: that is the question, *European Journal of Operational Research*, 183, 1513-1520, 2007.

Thomas L.C., Edelman D.B., Crook J.N., *Credit Scoring and Its Applications*, Society for Industrial Mathematics, 2002.

Van Gestel T., Baesens B., Van Dijcke P., Suykens J., Garcia J. and Alderweireld T., Linear and nonlinear credit scoring by combining logistic regression and support vector machines, *Journal of Credit Risk*, Volume 1, Number 4, 2005.

Yang J., Olafsson S., Optimization based feature selection with adaptive instance sampling, *Computers and Operations Research*, Volume 33, Issue 11, pp. 3088-3106, 2005.

Van der Aalst W.M.P, Weijters A.J.M.M., Process Mining: A Research Agenda, *Computers in Industry*, 53(3), pp. 231-244, 2004.

Witten I.H., Frank E., *Data mining: practical machine learning tools and techniques with Java implementations*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2000.