Expert Systems with Applications 38 (2011) 2354–2364

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Building comprehensible customer churn prediction models with advanced rule induction techniques

Wouter Verbeke [a,*], David Martens [a,b], Christophe Mues [c], Bart Baesens [a,c]

[a] Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b] Department of Business Administration and Public Management, Hogeschool Gent, Universiteit Gent, Voskenslaan 270, B-9000 Ghent, Belgium
[c] School of Management, University of Southampton, Highfield Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Churn prediction
Data mining
Classification
Comprehensible rule induction
Ant Colony Optimization
ALBA

ABSTRACT

Customer churn prediction models aim to detect customers with a high propensity to attrite. Predictive accuracy, comprehensibility, and justifiability are three key aspects of a churn prediction model. An accurate model permits to correctly target future churners in a retention marketing campaign, while a comprehensible and intuitive rule-set allows to identify the main drivers for customers to churn, and to develop an effective retention strategy in accordance with domain knowledge. This paper provides an extended overview of the literature on the use of data mining in customer churn prediction modeling. It is shown that only limited attention has been paid to the comprehensibility and the intuitiveness of churn prediction models. Therefore, two novel data mining techniques are applied to churn prediction modeling, and benchmarked to traditional rule induction techniques such as C4.5 and RIPPER. Both AntMiner+ and ALBA are shown to induce accurate as well as comprehensible classification rule-sets. AntMiner+ is a high performing data mining technique based on the principles of Ant Colony Optimization that allows to include domain knowledge by imposing monotonicity constraints on the final rule-set. ALBA on the other hand combines the high predictive accuracy of a non-linear support vector machine model with the comprehensibility of the rule-set format. The results of the benchmarking experiments show that ALBA improves learning of classification techniques, resulting in comprehensible models with increased performance. AntMiner+ results in accurate, comprehensible, but most importantly justifiable models, unlike the other modeling techniques included in this study.

1. Introduction

In recent decades we have witnessed an explosion of data. Valuable information is contained in this data, but is hidden in the vast collection of raw data. Data mining entails the overall process of extracting knowledge from this data. Data mining techniques have been successfully applied in many different domains. Well-known examples are breast-cancer detection in the biomedical sector, market basket analysis in the retail sector (Berry & Linoff, 2004), and credit scoring in the financial sector (Baesens et al., 2003). This paper however focuses on the use of data mining to predict customer churn.

Customer churn prediction models aim to detect customers with a high propensity to attrite. An accurate segmentation of the customer base allows a company to target the customers that are most likely to churn in a retention marketing campaign, which improves the efficient use of the limited resources for such a campaign. Customer retention is profitable to a company, because: (1) Attracting new clients costs five to six times more than customer retention (Athanassopoulos, 2000; Bhattacharya, 1998; Colgate & Danaher, 2000; Rasmusson, 1999). (2) Long-term customers generate higher profits, tend to be less sensitive to competitive marketing activities, become less costly to serve, and may provide new referrals through positive word-of-mouth, while dissatisfied customers might spread negative word-of-mouth (Colgate, Stewart, & Kinsella, 1996; Ganesh, Arnold, & Reynolds, 2000; Mizerski, 1982; Paulin, Perrien, Ferguson, Salazar, & Seruya, 1998; Reichheld, 1996; Stum & Thiry, 1991; Zeithaml, Berry, & Parasuraman, 1996). (3) Losing customers leads to opportunity costs because of reduced sales (Rust & Zahorik, 1993). A small improvement in customer retention hence can lead to a significant increase in profit (Van den Poel & Larivière, 2004). That is why both accurate and comprehensible churn prediction models are needed, in order to identify respectively the customers who are about to churn and their reasons to do so. As will be discussed in Section 2, many data mining techniques have already been tested on their churn predictive power. Much less attention has been paid however to the comprehensibility and the justifiability of the developed models. Note that churn prediction is just one of the applications of data mining for

* Corresponding author. Tel.: +32 16 32 68 87; fax: +32 16 32 66 24.
E-mail addresses: wouter.verbeke@econ.kuleuven.be (W. Verbeke), david.martens@econ.kuleuven.be (D. Martens), C.Mues@soton.ac.uk (C. Mues), bart.baesens@econ.kuleuven.be (B. Baesens).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.08.023

marketing, others include customer lifetime value prediction (Glady, Baesens, & Croux, 2009), frequent itemset mining (Agrawal & Srikant, 1994) and sales forecasting (Thomassey & Happiette, 2007).

In this paper we introduce the application of two novel data mining techniques for customer churn prediction. The first technique, AntMiner+, uses Ant Colony Optimization (ACO) to infer rules from data, and explicitly seeks to induce accurate, comprehensible, and intuitive classification rule-sets (Martens et al., 2007). So far AntMiner+ has been successfully applied to credit scoring (Martens et al., 2006), software mining (Vandecruys et al., 2008), audit mining (Martens, Bruynseels, Baesens, Willekens, & Vanthienen, 2008), and business/ICT alignment prediction (Cumps et al., 2009). An advantage of AntMiner+ is the possibility to incorporate domain knowledge (Martens et al., 2006), ensuring intuitive decision support models.

The second technique is an Active Learning Based Approach (ALBA) for support vector machine (SVM) rule extraction (Martens, Van Gestel, & Baesens, 2009). ALBA manipulates a dataset by changing the class labels of data instances by the SVM predicted labels, and by generating additional data instances close to the class boundaries. Applying simple rule induction techniques such as C4.5 or RIPPER on the manipulated dataset results in improved learning, and thus in a more accurate, but still comprehensible, rule-set.

The remainder of this paper is structured as follows. First, in Section 2, the domain of customer churn prediction modeling is introduced by means of a broad literature study. Then in Section 3, the workings of AntMiner+ and ALBA are briefly explained. In Section 4 both techniques are applied to predict customer churn, and the setup and results of a series of experiments are discussed. The final section concludes the paper.

2. Customer churn prediction modeling

Customer relationship management, and customer churn prediction in particular, have received a growing attention during the last decade. Table 1 provides an overview of the literature on the use of data mining techniques for customer churn prediction modeling. The table summarizes the applied modeling techniques, the characteristics of the assessed datasets, and the validation and evaluation of the results. Also included are preprocessing steps like sampling and variable selection.

In this paper we argue that both accurate and comprehensible churn prediction models are needed, in order to identify respectively the customers that are about to churn, and their reasons to do so. As can be seen from Table 1, a myriad of modeling techniques has been tested in a search for the most accurate modeling technique: logistic regression, decision trees, neural networks, support vector machines, random forests, regression forests, and many more. The comprehensibility of churn prediction models on the other hand has received much less attention in literature (Lima, Mues, & Baesens, 2009). However, several studies focus on the analysis of churn drivers (Buckinx & Van den Poel, 2005; Kumar & Ravi, 2008), which illustrates the need to gain insight in the causes of churn and therefore confirms the need to build comprehensible models.

Furthermore, also the justifiability of a model should be considered in the evaluation of a churn prediction model. In a data mining context, a model is justifiable when it is in line with existing domain knowledge. For a model to be justifiable, it needs to be validated by a domain expert, which in turn means that the model should be comprehensible (Martens et al., 2006). A modeling technique that allows to take into account domain knowledge and results in models that behave intuitively correct, is of much greater use for churn prediction modeling than a technique that produces counter-intuitive results. The most frequently encountered and researched aspect of knowledge fusion, i.e. incorporating the knowledge representing the experience of an expert into a data mining approach, is the monotonicity constraint. This constraint demands that an increase in a certain input cannot lead to a decrease in the output. For instance, an increased number of calls to the customer helpdesk, should yield an increasing probability of churn. Including domain knowledge in churn prediction modeling is to our knowledge thus far only discussed in Lima et al. (2009), and is one of the main contributions of this paper.

Although churn prediction modeling has been extensively researched, no general consensus exists on the performance of churn prediction modeling techniques. For instance, both Mozer, Wolniewicz, Grimes, Johnson, and Kaushansky (2000) Hwang, Jung, and Suh (2004) apply logistic regression and neural networks to predict churn. However, the first study finds neural networks to perform best and the second logistic regression. Broad benchmarking studies have not been published thus far, and widely varying methodologies and experimental setups impede to cross compare the results of different papers.

Although most studies summarized in Table 1 use private datasets, evaluating data mining techniques on publicly available datasets has many advantages (Vandecruys et al., 2008): (1) The creation of benchmarks is facilitated which makes it possible to compare and rank existing and new data mining techniques. (2) The impact of the characteristics of a dataset on the performance of a data mining technique is the same for all techniques. Comparing the results and rankings of techniques applied on a variety of datasets on the other hand allows to evaluate and study the effect of data characteristics on the performance of a technique. (3) Using publicly available datasets provides insight in the impact of each step of the followed methodology. Data preprocessing steps like input variable selection and sampling have a significant impact on the final result, possibly even to a larger extent than the choice of modeling technique. To summarize, using publicly available datasets improves the general comparability of results, techniques, and methodologies.

A final point of critique concerns the use of a single split up of the dataset in a training and a test set to validate the results of a model. It should be clear that the average result on multiple split ups provides a more reliable measure of performance than a single shot result. Furthermore, to draw valid conclusions about differences in performance of techniques, results should be tested whether they differ significantly or not. A common heuristic to test the significance of performance differences is for instance the Student's paired *t*-test (Dietterich, 1998), which will be applied in this paper.

3. Advanced rule induction techniques: AntMiner+ and ALBA

As churn prediction models should be both accurate and comprehensible, we will focus on the use of rule-based classification techniques. More specifically, we will induce rule-sets from a churn dataset using AntMiner+ and ALBA, as well as with more traditional rule induction techniques C4.5 and RIPPER. The workings of AntMiner+ and ALBA are explained briefly in the next two sections.

3.1. AntMiner+: classification based on Ant Colony Optimization

AntMiner+ is a classification technique that employs artificial ants to induce rules.[1] Previous benchmarking studies reveal that the models generated by AntMiner+ fulfill both the accuracy and comprehensibility requirements (Martens et al., 2007), and are also

[1] http://www.antminerplus.com.

Table 1
Overview of literature on churn prediction modeling.

Authors	Title & Journal	Year	What?	Techniques	Dataset-# cust. –# feat. – public (1) or private (2)	Metrics – sampling – feat. selection – validation
Eiben A.E., Koudijs A.E., Slisser F.	Genetic modeling of customer retention – *Lecture Notes in Computer Science*	1998	Comparison of modeling techniques, application on real-life dataset	Logistic regression, genetic programming, rough data analysis, CHAID	Financial services – 14.394 cust. – 213 feat. – (2)	PCC, Lift chart, CoC – no sampling – OMEGA software – hold-out
Madden G., Savage S.J., Coble-Neal G.	Subscriber churn in the Australian ISP market – *Information Economics and Policy*	1999	Development of a churn prediction model and analysis of churn drivers	Binomial probit model	Internet service provider – 592 cust. – 19 feat. – (2)	PCC, goodness-of-fit - no sampling – no feat. selection – likelihood ratio test
Datta P., Masand B., Mani D.R., Li B.	Automated cellular modeling and prediction on a large scale – *Artificial Intelligence Review*	2000	Description and application of automated cellular churn prediction modeling system	Neural network	Wireless telecom – 500.000 cust. – 200 feat. – (2)	Lift, payoff – undersampling – forward selection with decision tree, genetic alg. – hold-out
Mozer M.C., Wolniewicz R., Grimes D.B., Johnson E., Kaushansky H.	Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry – *IEEE Transactions on Neural Networks*	2000	Prediction of churn probability, retention incentive optimization, and profit gains estimation	Logistic regression, (boosted) decision tree, (boosted) neural network	Wireless telecom – 46.744 cust. – 134 feat. – (2)	Lift chart – no sampling – no feat. selection – 10-fold cross validation
Wei C.P., Chiu I.T.	Turning telecommunications call details to churn prediction: a data mining approach – *Expert Systems with Applications*	2002	Application of the C4.5 algorithm to create a churn prediction model, using a limited number of features	Decision tree (C4.5)	Wireless telecom – 114.000 cust. – over 12 feat. – (2)	Lift, lift chart, miss rate, false rate, detection error trade-off curve – oversampling – intuition – 3×10 fold cross validation
Au W.H., Chan K.C.C., Yao X.	A novel evolutionary data mining algorithm with applications to churn prediction – *IEEE Transactions on Evolutionary Computation*	2003	Application of a novel data mining technique to predict churn probabilities	Decision tree (C4.5), neural network, data mining by evolutionary learning	Wireless telecom – 100.000 cust. – 251 feat. – (2)	Top 5% lift, lift chart – undersampling – intuition – 10-fold cross validation
Hwang H., Jung T., Suh E.	An LTV model and customer segmentation based on customer value: a case-study on the wireless telecommunications industry – *Expert Systems with Applications*	2004	Churn prediction model as part of a customer lifetime value model	Logistic regression, decision tree, neural network	Wireless telecom – 16.384 cust. – 200 feat. – (2)	Error rate, lift chart – no sampling – R^2 method – hold-out
Buckinx W., Van den Poel D.	Customer base analysis: partial defection of behaviorally-loyal clients in a non-contractual FMCG retail setting - *European Journal of Operational Research*	2005	Comparison of techniques for partial defection prediction, focus on profitable customers in a non-contractual setting	Logistic regression, neural network, random forests	Grocery retail – 158.884 cust. – 61 feat. – (2)	PCC, AUC – no sampling – no feat. selection – hold-out
Larivière B., Van den Poel D.	Predicting customer retention and profitability by using random forests and regression forests techniques – *Expert Systems with Applications*	2005	Investigation of explanatory variables and modeling methods for customer churn prediction	Logistic regression, linear regression, random forests, regression forests	Financial services – 100.000 cust. – 30 feat. – (2)	AUC – no sampling – no feat. selection – hold-out, non-parametric test of De Long et al.
Hung S.Y., Yen D.C., Wang H.Y.	Applying data mining to telecom churn management – *Expert Systems with Applications*	2006	Comparative study and application of churn prediction modeling methods	Decision tree, neural network (on clustered segments)	Wireless telecom – 160.000 cust. – over 40 feat. – (2)	Hit ratio, top-decile lift – oversampling – intuition, EDA – hold-out, *t*-test
Lemmens A., Croux C.	Bagging and boosting classification trees to predict churn – *Journal of Marketing Research*	2006	Application of bagging and boosting techniques to improve predictive power of churn prediction models	Logistic regression, bagged & boosted decision trees	Wireless telecom – 100.000 cust. – 171 feat. – (1)	Error rate, top-decile lift, Gini – proportional and oversampling – principal components analysis – hold-out
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*	2006	Analysis of the results of a churn prediction modeling tournament with focus on method and shelve life	Logistic regression, decision tree, neural network, discriminant analysis, Bayes	Wireless telecom – 100.000 cust. – 171 feat. – (1)	Top-decile lift, Gini coefficient -
Burez J., Van den Poel D.	CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services – *Expert Systems with Applications*	2007	Development of churn prediction model, tested in real-life retention campaign	Logistic regression (with Markov chains), random forests	Pay-TV – 143.198 cust. – 81 feat. – (2)	PCC, cumulative lift, AUC – no sampling – no feat. selection – hold-out,
Burez J., Van den Poel D.	Handling class imbalance in customer churn prediction – *Expert Systems with Applications*	2008	Study on sampling methods, evaluation metrics and methods, and modeling techniques	Logistic regression, gradient boosting, (weighted) random forests	Banks, telecom, newspaper, pay-TV, supermarket – 32.371 to 143.198 cust. – 21 to 81 feat. – (2)	Error rate, AUC, lift – undersampling, CUBE – 5×2 fold cross validation, *t*-test
Coussement K., Van den Poel D.	Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques –	2008	Application of support vector machine in churn prediction in a newspaper subscription	Logistic regression, support vector machine, random forests	Newspaper subscription – 90.000 cust. – 82 feat. – (2)	PCC, AUC, top-decile lift – undersampling – no feat. selection – 10-fold cross validation, non-

	Expert Systems with Applications		environment			parametric test of De Long et al.
Kumar D.A., Ravi V.	Predicting credit card customer churn in banks using data mining – *International Journal of Data Analysis Techniques and Strategies*	2008	Extensive study to compare the results of different sampling and modeling techniques to predict credit card customer churn	Logistic regression, decision tree, neural network, svm, random forest, rbf network & ensemble with majority voting	Credit card – 14.814 cust. – 22 feat. – (1)	PCC, specificity, sensitivity, AUC – under- & oversampling (combined) & SMOTE – CART – hold-out, 10-fold cross validation
Lima E., Mues C., Baesens B.	Domain knowledge integration in data mining using decision tables: case studies in churn prediction – *Journal of the Operations Research Society*	2008	Incorporation of domain knowledge in churn prediction models	Logistic regression, decision tree	Wireless telecom (2) – 5.000 and 100.000 cust. – 21 and 171 feat. – (1)	PCC, specificity, sensitivity, AUC – oversampling – Cramer's V-statistic, *t*-tests – hold-out, non-parametric test of De Long

intuitively correct (Martens et al., 2006). In this section the main workings of this technique are explained, starting with a short introduction to the basis of AntMiner+: Ant Colony Optimization.

3.1.1. Ant Colony Optimization

Ant Colony Optimization (ACO) is a metaheuristic inspired on the foraging behavior of real ant colonies (Dorigo & Stützle, 2004). A biological ant by itself is a simple insect with limited capabilities, and is guided by straightforward decision rules. However, these simple rules are sufficient for the overall ant colony to find short paths from the nest to the food source. ACO employs artificial ants that cooperate in a similar manner as their biological counterparts, in order to find good solutions for discrete optimization problems (Dorigo & Stützle, 2004). The first ACO algorithm developed was Ant System (Dorigo, Maniezzo, & Colorni, 1996), where ants iteratively construct solutions and add pheromone to the paths corresponding to these solutions. Path selection is a stochastic procedure based on not only a history-dependent pheromone value, but also a problem-dependent heuristic value. The pheromone value gives an indication of the number of ants that chose the trail recently, while the heuristic value is a problem-dependent quality measure. When an ant reaches a decision point, it is more likely to choose the trail with the higher pheromone and heuristic values. ACO has been applied to a wide variety of problems (Dorigo & Stützle, 2004), such as the vehicle routing problem (Bullnheimer, Hartl, & Strauss, 1999; Garcia, Montiel, Castillo, Sepúlveda, & Melin, 2009; Wade & Salhi, 2004), scheduling (Colorni, Dorigo, Maniezzo, & Trubian, 1994; Blum, 2005), and routing in packet-switched networks (Caro & Dorigo, 1998). Recently, ACO has also been applied in the data mining field, addressing both the clustering (see e.g. the work by Boryczka (2009), Abraham & Ramos (2003), Handl, Knowles, & Dorigo (2006)) and classification task (Liu, Abbass, & McKay, 2003; Martens et al., 2007; Parpinelli, Lopes, & Freitas, 2001), which is the topic of interest in this paper.

3.1.2. AntMiner+ Algorithm

ACO can be used to induce comprehensible and accurate rule-based classification models from data, as done in the AntMiner+ classification technique. This technique implements the $\mathcal{MAX}-\mathcal{MIN}$ Ant System (Stützle & Hoos, 2000) for classification. An environment is defined for the ants to walk through, such that each path corresponds to a classification rule. As such, the path chosen by each ant corresponds to a predictive rule. The principles of ACO drive the ants towards good predictive rules, as shown in the benchmarking study in Martens et al. (2007). The outline of the workings of the AntMiner+ algorithm is given in Algorithm 1. For more details on the workings of this technique, one may refer to Martens et al. (2007).

Algorithm 1 Pseudo-code of AntMiner+ algorithm

```
construct graph
while not early stopping or minimum percentage data covered do
    initialize heuristics, pheromones and probabilities of edges
    while not converged do
        create ants
        let ants run from source to sink
        evaporate pheromone on edges
        prune rule of best ant
        update path of best ant
        adjust pheromone levels if outside boundaries
        kill ants
```

(*continued on next page*)

Algorithm 1 (*continued*)

Algorithm 1 Pseudo-code of AntMiner+ algorithm

```
    update probabilities of edges
  end while
  extract rule corresponding to converged path
  flag data points covered by the extracted rule
end while
evaluate performance on test set
```

Advantages of AntMiner+ are not only the high accuracy and the comprehensibility of the generated models, but also the possibility to demand intuitive predictive models, which is crucial whenever comprehensibility is required. For example, when a classification rule is induced to predict whether or not a customer will churn, the rule "**if** *Helpdesk calls* > 5 **then** class = no churner", is an unintuitive rule, as we would expect that the more a customer calls the helpdesk the more probably he will churn, making the expected sign for this example " < ". The rule "**if** *Helpdesk calls* > 5 **then** class = churner" on the other hand, is intuitive. By imposing constraints on these inequality signs, such domain knowledge can be incorporated, resulting in intuitive, justifiable classification models. Note that these monotonicity constraints are imposed by the domain expert, who might disagree with such constraints and choose to impose no or other inequality constraints.

3.2. ALBA: Active Learning Based Approach for SVM rule extraction

The support vector machine (SVM) (Vapnik, 1995) is currently one of the state-of-the-art classification techniques. Benchmarking studies reveal that in general, the SVM performs best among current classification techniques (Baesens et al., 2003), due to its ability to capture non-linearities. However, its strength is also its main weakness, as the generated non-linear models are typically regarded as incomprehensible black-box models. The opaqueness of SVM models can be remedied through the use of rule extraction techniques, which induce rules that mimic the black-box SVM model as closely as possible. Through rule extraction, some insight is provided into the logics of the SVM model (Martens, Baesens, Van Gestel, & Vanthienen, 2007).

ALBA is a rule extraction algorithm that uses specific concepts of the SVM, being the support vectors, and simple rule induction techniques such as C4.5 and RIPPER (Martens et al., 2009). Active learning entails the control of the learning algorithm over the input data on which it learns. More specifically, active learning focuses on the problem areas (Cohn, Atlas, & Ladner, 1994), which for rule extraction are those areas in the input space were the noise is the highest. These regions are found near the SVM decision boundary, which marks the transition of one class to another. First, we can change the labels of the data instances by the SVM predicted labels. In this manner the induced rules will mimic the SVM model and all noise is omitted from the data, removing any apparent conflicts in the data. Second, to incorporate the active learning approach additional data instances are generated close to the decision boundary. For this explicit use is made of the support vectors, which are typically close to the decision boundary. The support vectors are thus used as proxies for the decision boundary by generating additional data instances close to the support vectors. Since the distribution of the support vectors will follow the data distribution, more support vectors will be found in dense input areas, and less in more sparse ones. This implicit incorporation of the existing data distribution in the extra data generation step, eliminates the necessity to explicitly take into account density measures. The Active Learning Based Approach is described formally in Algorithm 2. A full discussion on ALBA can be found in Martens et al. (2009).

After improving and expanding the input data, simple rule induction techniques such as C4.5 and RIPPER are applied to induce rule-sets. C4.5 is a popular decision tree builder (Quinlan, 1993) where each leaf assigns a class label to observations. Each of these leaves can be represented by a rule and therefore C4.5 builds comprehensible classifiers. RIPPER is a rule induction technique, generating a list of ordered rules (Cohen, 1995; Tan, Steinbach, & Kumar, 2005; Witten & Frank, 2000a). The name RIPPER is an acronym for Repeated Incremental Pruning to Produce Error Reduction. A detailed overview of this technique can be found in Cohen (1995) and Witten and Frank (2000a).

Algorithm 2 Pseudo-code of ALBA algorithm

```
preprocess data D = {(x_i, y_i)}_{i=1}^{N}
split data in training data D_tr, and test data D_te in a 2/3, 1/3 ratio
tune SVM parameters with gridsearch on D_tr
train SVM on D_tr = {(x_i, y_i)}_{i=1}^{N_tr}, providing an oracle SVM mapping a data input to a class label
change the class labels of the training data to the SVM predicted class
% Calculate the average distance distance_k of training data to support vectors, in each dimension k
for k = 1 to n do
  distance_k = 0
  for all support vectors sv_j do
    for all training data instance d in D_tr do
      distance_k = distance_k + |d_k − sv_{j,k}|
    end for
  end for
  distance_k = distance_k / (#sv × N_tr)
end for
% Create 1000 extra data instances
for i = 1 to 1000 do
  randomly choose one of the support vectors sv_j
  % Randomly generate an extra data instance x_i close to sv_j
  k = 1 to n
    x_{i,k} = sv(j,k) + [(rand − 0.5) × distance_k / 2] with rand a random number in [0,1]
  end for
  provide a class label y_i using the trained SVM as oracle: y_i = SVM(x_i)
end for
run rule induction algorithm on the data set containing both the training data D_tr, and newly created data instances {(x_i, y_i)}_{i=1:1000}
evaluate performance in terms of accuracy, fidelity and number of rules, on D_te
```

4. Customer churn prediction with AntMiner+ and ALBA

4.1. Dataset

AntMiner+ and ALBA are applied on a publicly available dataset downloaded from the KDD library.[2] The dataset is obtained from a wireless telecom operator, and consists of 5000 observations. For each observation 21 features are available, with no missing values. 14.3% of the customers are indicated to churn in the coming three months. For a full description of the dataset, one may refer to Larose (2005).

[2] http://www.datalab.uci.edu/data/mldb-sgi/data/.

Table 2
Top eleven ranked features with chi-squared based filter and intuitive sign relations with churn.

Feature	Constraint	What?
Day_Mins		Daytime usage (minutes/month)
Day_Charge	+	Charge for daytime usage ($/month)
CustServ_Calls	+	Number of calls to customer service
Intl_Plan		International plan subscriber (0 = no, 1 = yes)
Vmail_Plan		Voicemail plan subscriber (0 = no, 1 = yes)
Vmail_Message		Number of voice mail messages
Intl_Charge	+	Charge for international calls ($/month)
Intl_Mins		International usage (minutes/month)
Eve_Mins		Evening usage (minutes/month)
Eve_Charge	+	Charge for evening usage ($/month)
Intl_Calls		Number of international calls

4.2. Data preprocessing

Data preprocessing was conducted in the form of discretization, input selection, and oversampling.

4.2.1. Discretization

All continuous variables are discretized following Fayyad and Irani (1993). Discretization and other data preprocessing procedures are performed using the open-source data mining workbench Weka[3] (Witten & Frank, 2000b).

4.2.2. Input Selection and monotonicity constraints

To make sure that only relevant variables are included in the dataset, and to decrease the computational burden, an input selection procedure is performed using a chi-squared based filter (Martens et al., 2006; Thomas, Edelman, & Crook, 2002). First, the observed frequencies of all possible combinations of values for class and variable are measured. Based on this, the theoretical frequencies, assuming complete independence between the variable and the class, are calculated. The hypothesis of equal odds provides a chi-squared test statistic; higher values allow one to more confidently reject the zero hypothesis of equal odds; hence, these values allow one to rank the variables according to predictive power. In this manner, the set of features was reduced from twenty to eleven as listed in Table 2. Also included in Table 2 are the signs of the expected relations between the explanatory variables and the class variable, expressing domain knowledge. For instance, the positive relation sign between *churn* and *day_charge* indicates that according to domain knowledge a customer is more likely to churn when he is charged more. As can be seen from the table, only for a few features an explicit relation is presumed, in accordance with Lima et al. (2009). The resulting AntMiner+ rule-set will be enforced to comply with domain knowledge by imposing monotonicity constraints.

4.2.3. Oversampling

The class variable of the dataset is heavily skewed: the number of churners (14.3%) is much smaller than the number of non-churners (85.7%). This causes the classification modeling techniques to experience difficulties in learning which customers are about to churn. Since predicting future churners is a principle objective of the model, oversampling is used to improve learning (Provost, Fawcett, & Kohavi, 1997). Fig. 1 illustrates the principle of oversampling. Observations of the minority class in the training set are copied and added to the training set. Only the training data is oversampled and the test set is not, in order to provide an unbiased indication of the performance of the model towards future predictions.

[3] http://www.cs.waikato.ac.nz/ml/weka.

Depending on the number of times churning class observations are repeated in the training set, the resulting accuracy (percentage of instances correctly classified as churners or non-churners, *PCC*), sensitivity (percentage of churners that is correctly predicted, *Sens*) and specificity (percentage of non-churners that is correctly classified, *Spec*) varies. Table 3 illustrates this problem, showing exploratory single shot results for AntMiner+ on the studied dataset for different degrees of oversampling. The results for the other reviewed classification techniques are similar. For the original dataset (0 oversampling) we observe that AntMiner+ reaches a reasonable accuracy and a high specificity, but a fairly low sensitivity. This reflects the difficulties encountered in learning. A higher degree of oversampling results in a higher sensitivity, but implies a declining specificity. Because the share of churners in the dataset increases with oversampling, the importance of sensitivity relative to specificity increases in the calculation of accuracy. This means that, as the degree of oversampling increases, the accuracy decreases, unless learning and thus classification improves. However, since the cost of not detecting a churner is likely to be higher than the cost of targeting a non-churner in a retention campaign, this is a trade-off that one is willing to make. A reasonable trade-off between sensitivity on the one hand and specificity and accuracy on the other hand, is reached at three times oversampling. At this oversampling rate, the distribution of churners versus non-churners is almost even, as there are about as much churners (57.2%) as non-churners in the dataset (42.8%). This is reflected in the results, which show good performances both in terms of specificity and sensitivity. At five times oversampling an even higher sensitivity is reached. This happens however at the cost of a decrease in the comprehensibility, since the number of rules (#R) increases from 12 to 17. Finally, also note that the bigger the dataset, the higher the computational requierments. This can become an issue for techniques such as SVM, which scales non-linearly with the number of observations. Therefore we decided to repeat each observation corresponding to a churner three times in the final training datasets.

4.3. Experimental setup

To evaluate the results of AntMiner+ (with and without including domain knowledge) and ALBA (with rule induction using both C4.5 and RIPPER), a benchmarking study is performed that includes commonly used state-of-the-art classification techniques such as C4.5, RIPPER, SVM, logistic regression, and the simplistic majority vote. Because of the importance of comprehensibility in a churn prediction setting, both the results of C4.5 with standard pruning (*confidence factor* = 0.25) and with extensive pruning (*confidence factor* = 0.001) are included. Extensive pruning (indicated with XP in the results table) leads to fewer and smaller rules, and thus to a more comprehensible rule set. Also the results of ALBA combined with RIPPER are reported with standard and extensive pruning (minimum total weight of instances in a rule of respectively 2.0 and 12.0). While C4.5 induces an unordered rule-set, AntMiner+ and RIPPER induce ordered rule-sets. C4.5, RIPPER, logistic regression, and majority vote are all evaluated using the open-source Weka workbench (Witten & Frank, 2000b).

The parameters of AntMiner+ that need to be set are the total number of ants and the evaporation factor ρ, which are initialized to 1000 and 0.85 respectively, as suggested by Martens et al. (2006). The number of extra data instances generated by ALBA is set to 2000 to obtain a good balance between predictive performance and computational burden (Martens et al., 2009). For the SVM a RBF Kernel is chosen as it is shown to achieve a good overall performance (Baesens et al., 2003; Martens et al., 2009). The regularization parameter C, and bandwith parameter σ are chosen

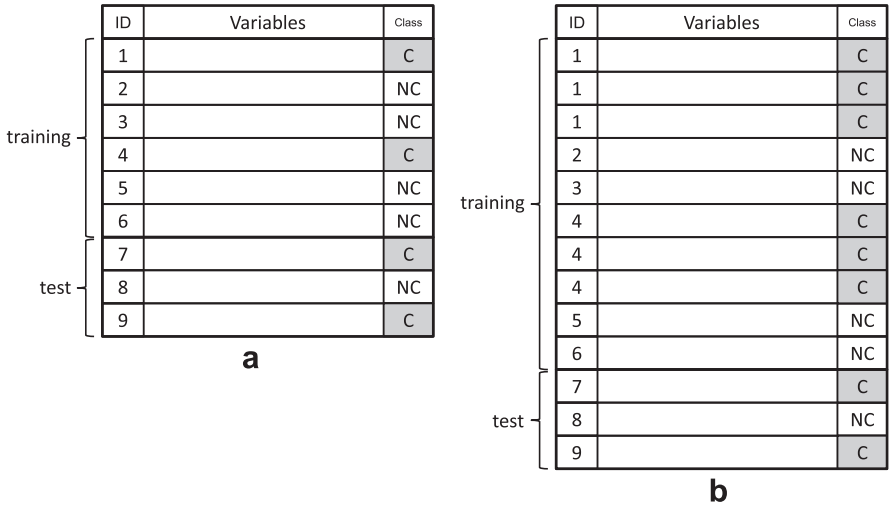

Fig. 1. Illustration of the principle of oversampling. A small dataset with 9 observations (a) is split into a training set of 6 observations and a test set of 3 observations. Training instances classified as churners (Class = C) are repeated twice in the oversampled dataset (b).

Table 3
Out-of-sample performance gain for AntMiner+ using oversampling.

# Oversampling	PCC	#R	Sens	Spec
0	87.66	1	16.34	99.14
1	90.33	12	75.34	92.95
2	90.33	12	75.78	92.87
3	**90.33**	**12**	**76.68**	**92.71**
4	89.73	12	72.20	92.80
5	90.07	17	84.98	90.91

using a gridsearch mechanism (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002).

The reported measures are the average out-of-sample performances on ten random 70/30 split ups of the dataset in training and test sets. Early stopping is applied since the dataset is relatively large. Therefore one third of the training datasets is set apart for validation. Four series of experiments were performed. In the first series the performance of each classification technique is tested on the original dataset. Then in the second series ALBA is applied as explained in Section 3.2, using the support vector machines trained on the original data. In the third series the original data are oversampled as explained in Section 4.2.3, and finally

Table 4
Average out-of-sample results of the churn prediction experiments.

Series	Technique	PCC	Sens	Spec	#R	St. Dev
Original	AntMiner+	90.85	37.09	**99.71**	7.7	0.54
	AntMiner + DK	90.73	36.26	**99.69**	8.0	0.44
	C4.5	**93.59**	64.93	**98.34**	21.1	0.49
	C4.5 XP	92.94	56.88	**98.90**	10.7	0.52
	RIPPER	92.92	62.31	**97.99**	5.8	0.83
	RIPPER XP	92.83	60.71	**98.15**	5.5	0.78
	SVM	92.51	78.29	94.85	–	0.52
	Logit	86.87	29.18	96.44	–	0.55
	Majority rule	85.79	0.00	**100.00**	–	0.56
ALBA	AntMiner+	92.83	53.95	**99.27**	13.9	1.24
	AntMiner + DK	91.29	41.32	**99.57**	15.0	0.44
	C4.5	**93.79**	65.53	**98.49**	73.3	0.43
	C4.5 XP	**93.70**	65.24	**98.44**	29.7	0.47
	RIPPER	**93.85**	65.95	**98.46**	20.5	0.39
	RIPPER XP	**93.87**	65.66	**98.54**	10.5	0.38
Oversampled	AntMiner+	93.15	65.76	**97.72**	14.8	0.51
	AntMiner + DK	92.62	67.98	**96.72**	16.3	0.81
	C4.5	91.66	**80.82**	**93.45**	23.6	0.21
	C4.5 XP	90.94	**82.29**	**92.31**	13.6	0.61
	RIPPER	91.73	**81.49**	**93.41**	7.7	0.57
	RIPPER XP	91.89	**81.28**	**93.64**	7.5	0.36
	Logit	88.31	73.42	**89.62**	–	3.41
ALBA Oversampled	AntMiner+	92.45	74.42	**95.44**	13.8	0.70
	AntMiner + DK	91.51	63.12	**96.20**	13.8	1.57
	C4.5	92.41	77.42	**94.89**	84.5	0.33
	C4.5 XP	92.32	77.90	**94.71**	40.8	0.33
	RIPPER	92.39	77.37	**94.88**	25.9	0.32
	RIPPER XP	92.41	77.46	**94.88**	14.9	0.30

the ALBA dataset of the second series is oversampled in the fourth series.

4.4. Results and discussion

The results of the churn prediction experiments are summarized in Table 4, with the best performances in terms of average percentage correctly classified (PCC), specificity, and sensitivity underlined. Also included in the table are the number of induced rules (#R), and the standard deviation (St. Dev) of the accuracy. As discussed in Section 2, a one-sided Student's paired *t*-test is used to test the performance differences. Performances that are not significantly different at the 5% level from the top performance with respect to a one-tailed paired *t*-test are tabulated in bold face. Statistically significant underperformances at the 1% level are emphasized in italics, and performances significantly different at the 5% level but not a the 1% level are reported in normal script. Since the observations of the randomizations are not independent, we remark that this standard *t*-test is used as a common heuristic to test the performance differences (Dietterich, 1998).

4.4.1. Predictive power

As can be seen from the table, the highest accuracy is reached using the combination of ALBA and RIPPER with extra pruning. C4.5 applied on the original dataset, and ALBA combined with C4.5 with standard and increased pruning and RIPPER with standard pruning do not perform significantly worse. Other techniques follow closely however, and except for logistic regression and majority vote all results lie in the interval between 90% and 94%.

Accuracy alone is not an adequate performance measure to evaluate the experimental results though, as it implicitly assumes a relatively balanced class distribution among the observations and equal misclassification costs (Baesens et al., 2003). The skewed distribution of the dataset, which is typical for churn prediction, was already mentioned. But also the assumption of equal misclassification costs cannot be sustained. Typically, a customer relationship manager who applies data mining techniques for customer churn prediction will mainly be interested in the correct detection of future churners. Even to the extent that it is preferred to include a certain number of customers that will not churn in the nearby future in a retention campaign. Indeed, the costs of including a number of non-churning customers do not weigh up to the costs a company incurs due to churn, at least to a certain extent.

As the costs associated with the incorrect classification of churners are clearly higher than the costs associated with the incorrect classification of a non-churner, it seems fair to us to assume unequal misclassification costs. Consequently, a high sensitivity is of more importance to a company than a high specificity. Of course this does not mean that specificity can be neglected. A classification technique that classifies all customers as churners might well result in including all churners in a retention campaign, but the retention marketing costs will be unjustifiably high. A trade-off has to be made in order to obtain a high specificity combined with a reasonable sensitivity. This allows the company to efficiently allocate its retention marketing budget, by focusing on the customers that are classified to have the highest propensity to attrite.

The highest sensitivity in the churn prediction experiments is reached with C4.5 XP on the oversampled dataset. C4.5, RIPPER and RIPPER XP do not perform significantly worse at the 5 percent level, while the result of AntMiner+ DK does not differ significantly at the 1% level. The highest specificity on the other hand is reached with AntMiner+ applied on the original dataset. However, only the SVM and the logit model perform significantly worse. Therefore, and because we are mainly interested in detecting future churners, we will no further asses the specificity in the evaluation of the results.

Oversampling the dataset appears to improve significantly the sensitivity of all data mining techniques. ALBA and ALBA Oversampled do as well, but only to a limited extent. This is a consequence of training the support vector machine on the orginal dataset. As can be seen from the table, the sensitivity of the SVM is remarkably high compared to the results of the other techniques applied on the original dataset, which illustrates the power of the non-linear SVM. However, this result imposes an upper bound to the results that can be achieved using ALBA. Even if C4.5, RIPPER, or any other technique classify the ALBA training dataset 100% correct, the resulting sensitivity of the model can not be higher than 78.29%. As can be seen from Table 4 the sensitivity of C4.5 (XP) and RIPPER (XP) applied on the oversampled ALBA dataset indeed lies around 78%. A possible solution which could lead to better performances exists in an adjustment of ALBA in order to take into account the class distribution of the dataset. The ALBA algorithm should strive towards class balance when adding datapoints that lie near the non-linear class boundaries. Or, if misclassification costs are unequal, even a distribution in favor of the minority class could be created. ALBA does not lead to the overall best results in this experimental setup, since oversampling leads to even higher sensitivities than the support vector machine trained on the original dataset achieves. ALBA however does lead to the highest accuracies and remains an interesting technique to improve the performance of rule induction techniques.

4.4.2. Comprehensibility

High accuracy, sensitivity, and specificity are not the only important aspects in evaluating a churn prediction model. As stressed in the literature review, also the reasons for customers to churn are very valuable information for a company. Such knowledge allows to develop a more effective retention strategy by focusing on the probable causes of churn. Therefore, comprehensibility of the classification model is an important requirement in churn prediction modeling. There is not really much to comprehend about a majority vote model. The only principle behind this technique is "majority wins". Therefore, majority vote adds almost no value. Logistic regression performs reasonably well as to comprehensibility, but its model structure is arguably more opaque than a rule-based representation. C4.5, RIPPER, and AntMiner+ on

Table 5
AntMiner+ rule-set for 3 oversampling with monotonicity constraints.

if *Intl_Plan* = 1 **and** *Intl_Calls* $\leqslant$ 2
if *Day_Mins* > 285.5 **and** *Day_Charge* > 48.53 **and** *Vmail_Message* $\leqslant$ 2
if *Intl_Plan* = 1 **and** *Intl_Charge* > 3.55 **and** *Intl_Mins* > 13.15
if *Intl_Plan* = 1 **and** *Vmail_Message* $\leqslant$ 2 **and** *Intl_Mins* > 13.15 **and** *Eve_Mins* > 248.15
if *CustServ_Calls* > 3 **and** *Intl_Plan* = 1 **and** *Intl_Mins* > 13.15
then class = churn
else class = non churn

Table 6
RIPPER rule-set for 3 oversampling without monotonicity constraints.

if *Day_Mins* > 248.65 **and** *VMail_Plan* = 0
if *CustServ_Calls* > 3 **and** *Day_Mins* $\leqslant$ 168.05
if *Intl_Plan* = 0 **and** *Intl_Calls* $\leqslant$ 2
if *Day_Mins* > 221.85 **and** *VMail_Plan* = 0 **and** *Eve_Mins* > 248.15
if *Intl_Plan* = 0 **and** *Intl_Charge* > 3.55
if *CustServ_Calls* > 3 **and** *Day_Mins* $\leqslant$ 221.85 **and** *Eve_Mins* $\leqslant$ 248.15 **and** *Intl_Charge* $\leqslant$ 3.55
if *VMail_Plan* = 0 **and** *Day_Mins* > 221.85 **and** *CustServ_Calls* > 3
then class = churn
else class = non churn

the other hand induce comprehensible rules from a dataset. The comprehensibility of the resulting model decreases however as the number of rules (#R) increases. As can be seen from Table 4, AntMiner+ and RIPPER clearly induce much less rules than C4.5, even with increased pruning. The issue faced by C4.5 is its greedy character, since every split made in the decision tree is irreversibly present in all leaves underneath. Hence AntMiner+ and RIPPER, which on average result in a comparable number of rules, are the most comprehensible classification techniques tested in the experiments. This confirms previous results (Martens et al., 2006; Vandecruys et al., 2008). Finally, comprehensibility is also important to check the justifiability of a model.

1. Intl_Plan	0														1			
2. Intl_Calls	≤2	>2													-			
3. Intl_Mins	-	≤13.15				>13.15									-			
4. CustServ_Calls	-	-				≤3							>3		-			
5. Vmail_Message	-	≤2			>2	≤2						>2		-	≤2			>2
6. Eve_Mins	-	-			-	≤248.15				>248.15		-		-	-			-
7. Day_Mins	-	≤285.5		>285.5	-	≤285.5		>285.5		-		-		-	≤285.5		>285.5	-
8. Day_Charge	-	-	≤48.53	>48.53	-	-		≤48.53		>48.53	-	-		-	-	≤48.53	>48.53	-
9. Intl_Charge	-	-	-	-	-	≤3.55	>3.55	≤3.55	>3.55	-	-	≤3.55	>3.55	-	-	-	-	-
1. class=churn	x	-	-	x	-	-	x	-	x	x	x	-	x	x	-	-	x	-
2. class=non churn	-	x	x	-	x	x	-	x	-	-	-	x	-	-	x	x	-	x
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Fig. 2. Decision table corresponding to the AntMiner+ rule-set in Table 5.

1. Intl_Plan	0																	
2. Intl_Calls	≤2	>2																
3. Day_Mins	-	≤168.05			168.05 ≤ 221.85					221.85 ≤ 248.65						>248.65		
4. VMail_Plan	-	-			-					0				1		0	1	
5. CustServ_Calls	-	≤3		>3	≤3		>3			≤3			>3	-		-	-	
6. Eve_Mins	-	-		-	-		≤248.15		>248.15	≤248.15		>248.15	-	-		-	-	
7. Intl_Charge	-	≤3.55	>3.55	-	≤3.55	>3.55	-	≤3.55	>3.55	≤3.55	>3.55	-	-	≤3.55	>3.55	-	≤3.55	>3.55
1. class=churn	x	-	x	x	-	x	x	-	x	-	x	x	x	-	x	x	-	x
2. class=non churn	-	x	-	-	x	-	-	x	-	x	-	-	-	x	-	-	x	-
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

1. Intl_Plan	1																	
2. Intl_Calls	-																	
3. Day_Mins	≤168.05			168.05 ≤ 221.85						221.85 ≤ 248.65						>248.65		
4. VMail_Plan	-			-						0				1		0	1	
5. CustServ_Calls	≤3	>3	≤3	>3						≤3			>3	-		-	-	
6. Eve_Mins	-	-	-	≤248.15			>248.15	≤248.15		>248.15			-	-		-	-	
7. Intl_Charge	-	-	-	≤3.55		>3.55	-	-		-			-	-		-	-	
1. class=churn	-	x	-	x		-	-	-		x			x	-		x	-	
2. class=non churn	x	-	x	-		x	x	x		-			-	x		-	x	
	19	20	21	22		23	24	25		26			27	28		29	30	

Fig. 3. Decision table corresponding to the RIPPER rule-set in Table 6.

4.4.3. Justifiability

The rule-sets in Tables 5 and 6 are induced by respectively AntMiner+ DK (with inclusion of domain knowledge) and RIPPER. Figs. 2 and 3 show the decision tables, derived with the PROLOGA software (Vanthienen, Mues, Wets, & Delaere, 1998), corresponding to these rule-sets. A decision table is a more intuitive and user-friendly tabular representation of a rule-set, and consists of four quadrants which are separated by horizontal and vertical double-lines. The upper left quadrant contains the condition subjects, which represent the attributes in the rules of the rule-set. The action subjects in the lower left column describe the possible outcomes of the rules, which are in this case churn or non churn. Every column in the upper-right quadrant of the decision table comprises a classification rule, leading for a certain combination of condition states to the classification outcome in the lower-right quadrant marked with 'x'. A dash symbol ('-') in an entry column indicates that the value of the attribute is irrelevant within the context of that column.

It can be seen from Table 5 and Fig. 2 that all rules induced with AntMiner+ DK comply with the monotonicity constraints in Table 2. This is of great value for the practical use of the model since domain knowledge and prediction model are aligned and give complementary results. The rule-set in Table 6 induced by RIPPER on the other hands contains rules that do not comply with the constraint on the variable *Intl_Charge*. According to domain knowledge the more a client is charged, the more probably he will churn. The first two rules in the RIPPER rule-set violate this principle. Therefore this rule-set is not intuitive and will probably even be discarded by the user. Since *Intl_Charge* is included in the last row of the decision table in Fig. 3, one can easily see that the opposite relation is modeled by colums 22 and 23 which partly represent the first two rules induced by RIPPER. For instance, suppose two identical customers A and B with the same values for each feature, except for *Intl_Charge* which is equal to zero and ten for respectively customers A and B. The values of attributes *Intl_Plan*, *CustServ_Calls*, *Day_Mins*, and *Eve_Mins* are respectively equal to one, one, two hundred, and two hundred. Then the rule-set or decision table will classify customer A as a churner and customer B as a non-churner, which is not positively monotone with respect to *Intl_Charge*. It can be easily seen from the decision table that the AntMiner+ DK rule-set on the other hand results in a positive monotone relation between the class variable and *Intl_Charge*, in accordance with domain knowledge.

Although this small example might seem rather irrelevant, it illustrates perfectly that a data mining modeling technique should at least provide the possibility to impose constraints on a model, in order to include common domain knowledge and to improve the reliability of the model. This will enhance the comprehensibility of models and allow users to understand the workings of the model and the modeled relations between the variables inside the model.

To sum up, the results of the experiments show that ALBA improves learning by classification techniques and increases the accuracy, sensitivity, and specificity of the resulting models. The results are limited however by the performance of the support vector machine which depends on the dataset. AntMiner+ on the other hand results in accurate, comprehensible, but most important of all justifiable models. Since decision makers are reluctant to use unintuitive models, regardless of their accuracy, they will probably discard models that do not correspond with domain knowledge (Martens et al., 2007; Martens et al., 2006).

5. Conclusion

As discussed in the literature review, churn prediction models should be both accurate and comprehensible in order to improve the efficiency of retention marketing campaigns. This paper presents the application of AntMiner+ and ALBA on a publicly available churn prediction dataset. Both techniques explicitly seek to induce accurate as well as comprehensible rule-sets. The results are benchmarked to C4.5, RIPPER, SVM, and logistic regression. It is shown that ALBA, combined with RIPPER or C4.5, results in the highest accuracy, while sensitivity is the highest for C4.5 and RIPPER applied on an oversampled dataset. AntMiner+ results in less sensitive rule-sets, but allows to include domain knowledge, and results in comprehensible rule-sets which are much smaller than the rule-sets induced with C4.5. RIPPER also results in small and comprehensible rule-sets, but can lead to unintuitive models that violate domain knowledge. The comprehensibility of a churn prediction model is important since it facilitates the interpretation and the practical use of the model for marketing purposes. Comprehensibility also allows to check explicitly the concordance of a model with domain knowledge, which is of great importance since the intuitiveness of a model determines whether or not a model will be accepted by the end-users and effectively serve its purpose. AntMiner+ allows to include domain knowledge by imposing monotonicity constraints, leading to intuitive correct models that are still comprehensible and accurate, as proven by the results of the experiments.

Acknowledgements

We extend our gratitude to the Flemish Research Council for financial support (FWO postdoctoral research grant, Odysseus Grant B.0915.09), and the National Bank of Belgium (NBB/10/006).

References

Abraham, A., & Ramos, V. (2003). Web usage mining using artificial ant colony clustering. In *the congress on evolutionary computation* (pp. 1384–1391). IEEE Press.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *VLDB* (pp. 487–499).

Athanassopoulos, A. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research, 47*(3), 191–207.

Au, W., Chan, K., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation, 7*(6), 532–545.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

Berry, M., & Linoff, G. (2004). *Data mining techniques: For marketing, sales and customer relationship management*. New York, NY: John Wiley & Sons.

Bhattacharya, C. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of the Academy of Marketing Science, 26*(1), 31–44.

Blum, C. (2005). Beam-ACO – hybridizing ant colony optimization with beam search: An application to open shop scheduling. *Computers & Operations Research, 32*(6), 1565–1591.

Boryczka, U. (2009). Finding groups in data: Cluster analysis with ants. *Applied Soft Computing, 9*(1), 61–70.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal Of Operational Research, 164*(1), 252–268.

Bullnheimer, B., Hartl, R., & Strauss, C. (1999). Applying the ant system to the vehicle routing problem. In Voss, S., Martello, S., Osman, I., Roucairol, C. (Eds.), *Meta-Heuristics: Advances and trends in local search paradigms for optimization*.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626–4636.

Caro, G. D., & Dorigo, M. (1998). Antnet: Distributed stigmergetic control for communications networks. *Journal of Artificial Intelligence Research, 9*, 317–365.

Cohen, W. W. (1995). Fast effective rule induction. In A. Prieditis & S. Russell (Eds.), *Proceedings of the 12th international conference on machine learning* (pp. 115–123). Tahoe City, CA: Morgan Kaufmann.

Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning, 15*(2), 201–221.

Colgate, M., & Danaher, P. (2000). Implementing a customer relationship strategy: The assymetric impact of poor versus excellent execution. *Journal of the Academy of Marketing Science, 28*(3), 375–387.

Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: A study of the student market in ireland. *International Journal of Bank Marketing, 14*(3), 23–29.

Colorni, A., Dorigo, M., Maniezzo, V., & Trubian, M. (1994). Ant system for jobÜshop scheduling. *Journal of Operations Research, Statistics and Computer Science, 34*(1), 39–53.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*, 313–327.

Cumps, B., Martens, D., De Backer, M., Viaene, S., Dedene, G., Haesen, R., et al. (2009). Inferring rules for business/ict alignment using ants. *Information and Management, 46*(2), 116–124.

Datta, P., Masand, B., Mani, D., & Li, B. (2000). Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review, 14*, 485–502.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, 26*(1), 29–41.

Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. Cambridge, MA: MIT Press.

Eiben, A., Koudijs, A., & Slisser, F. (1998). Genetic modeling of customer retention. *Lecture Notes in Computer Science, 1391*, 178–186.

Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th international joint conference on artificial intelligence (IJCAI)* (pp. 1022–1029). Chambéry, France: Morgan Kaufmann.

Ganesh, J., Arnold, M., & Reynolds, K. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing, 64*(3), 65–87.

Garcia, M. P., Montiel, O., Castillo, O., Sepúlveda, R., & Melin, P. (2009). Path planning for autonomous mobile robot navigation with ant colony optimization and fuzzy cost function evaluation. *Applied Soft Computing, 9*(3), 1102–1110.

Glady, N., Baesens, B., & Croux, C. (2009). A modified pareto/NBD approach for predicting customer lifetime value. *Expert Systems with Applications, 36*(2), 2062–2071.

Handl, J., Knowles, J., & Dorigo, M. (2006). Ant-based clustering and topographic mapping. *Artificial Life, 12*(1), 35–61.

Hung, S., Yen, D., & Wang, H. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*, 515–524.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications, 26*, 181–188.

Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies, 1*(1), 4–28.

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forest and regression forest techniques. *Expert Systems with Applications, 29*(2), 472–484.

Larose, D. (2005). *Discovering knowledge in data: An introduction to data mining*. New Jersey, USA: Wiley.

Lima, E., Mues, C., & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. *Journal of the Operational Research Society, 60*, 1096–1106.

Liu, B., Abbass, H. A., & McKay, B. (2003). Classification rule discovery with ant colony optimization. In *IAT* (pp. 83–88). IEEE Computer Society.

Madden, G., Savage, S., & Coble-Neal, G. (1999). Subscriber churn in the Australian ISP market. *Information Economics and Policy*.

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

Martens, D., Bruynseels, L., Baesens, B., Willekens, M., & Vanthienen, J. (2008). Predicting going concern opinion with data mining. *Decision Support Systems, 45*, 765–777.

Martens, D., De Backer, M., Haesen, R., Baesens, B., Mues, C., & Vanthienen, J. (2006). Ant-based approach to the knowledge fusion problem. In M. Dorigo, L. Gambardella, M. Birattari, A. Martinoli, R. Poli, & T. Stützle (Eds.), *Ant colony optimization and swarm intelligence, fifth international workshop. ANTS 2006* (Vol. 4150, pp. 84–95). Berlin, Germany: Springer-Verlag.

Martens, D., De Backer, M., Haesen, R., Snoeck, M., Vanthienen, J., & Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transaction on Evolutionary Computation, 11*(5), 651–665.

Martens, D., Van Gestel, T., & Baesens, B. (2009). Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 178–191.

Mizerski, R. (1982). An attribution explanation of the disproportionate influence of unfavourable information. *Journal of Consumer Research, 9*(December), 301–310.

Mozer, M., Wolniewicz, R., Grimes, D., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks, 11*(3), 690–696.

Neslin, S., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Detection defection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211.

Parpinelli, R. S., Lopes, H. S., & Freitas, A. A. (2001). An ant colony based system for data mining: Applications to medical data. In *Proceedings of the genetic and evolutionary computation conference (GECCO-2001)* (pp. 791–797). San Francisco, California, USA: Morgan Kaufmann.

Paulin, M., Perrien, J., Ferguson, R., Salazar, A., & Seruya, L. (1998). Relational norms and client retention: External effectiveness of commercial banking in canada and mexico. *International Journal of Bank Marketing, 16*(1), 24–31.

Provost, F., Fawcett, T., & Kohavi, R. (1997). The case against accuracy estimation for comparing induction algorithms. In *In proceedings of the 15th international conference on machine learning* (pp. 445–453). Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Rasmusson, E. (1999). Complaints can build relationships. *Sales and Marketing Management, 151*(9), 89–90.

Reichheld, F. (1996). Learning from customer defections. *Harvard Business Review, 74*(2), 56–69.

Rust, R., & Zahorik, A. (1993). Customer satisfaction, customer retention, and market share. *Journal of Retailing, 69*(2), 193–215.

Stum, D., & Thiry, A. (1991). Building customer loyalty. *Training and Development Journal, 45*(4), 34–36.

Stützle, T., & Hoos, H. H. (2000). $\mathcal{MAX}-\mathcal{MIN}$ ant system. *Future Generation Computer Systems, 16*(8), 889–914.

Suykens, J., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA: Addison Wesley.

Thomas, L., Edelman, D., & Crook, J. (Eds.). (2002). *Credit scoring and its Applications*. Philadelphia, PA: SIAM.

Thomassey, S., & Happiette, M. (2007). A neural clustering and classification system for sales forecasting of new apparel items. *Applied Soft Computing, 7*(4), 1177–1187.

Vandecruys, O., Martens, D., Baesens, B., Mues, C., De Backer, M., & Haesen, R. (2008). Mining software repositories for comprehensible software fault prediction models. *Journal of Systems and Software, 81*(5), 823–839.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.

Vanthienen, J., Mues, C., Wets, G., & Delaere, K. (1998). A tool-supported approach to inter-tabular verification. *Expert Systems with Applications, 15*(3–4), 277–285.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc..

Wade, A., & Salhi, S. (2004). An ant system algorithm for the mixed vehicle routing problem with backhauls. In *Metaheuristics: Computer decision-making* (pp. 699–719). Norwell, MA: Kluwer Academic Publishers..

Wei, C., & Chiu, I. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*, 103–112.

Witten, I. H., & Frank, E. (2000a). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

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

Zeithaml, V., Berry, L., & Parasuraman, A. (1996). The behavioural consequences of service quality. *Journal of Marketing, 60*(2), 31–46.