Applied Soft Computing 24 (2014) 994–1012

Contents lists available at ScienceDirect

Applied Soft Computing

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

Improved churn prediction in telecommunication industry using data mining techniques

A. Keramati[a,*], R. Jafari-Marandi[a], M. Aliannejadi[b], I. Ahmadian[c], M. Mozaffari[a], U. Abbasi[a,d]

[a] School of Industrial and Systems Engineering, University of Tehran, Tehran, Iran
[b] Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran
[c] Department of Industrial Engineering, K.N. Toosi University of Technology, Tehran, Iran
[d] Design and Manufacturing Systems-Arts et Métiers ParisTech, Paris, France

ARTICLE INFO

Article history:
Received 20 November 2012
Received in revised form 29 July 2014
Accepted 18 August 2014
Available online 29 August 2014

Keywords:
Telecommunication
Churn prediction
ANN
KNN
SVM
Decision tree

ABSTRACT

To survive in today's telecommunication business it is imperative to distinguish customers who are not reluctant to move toward a competitor. Therefore, customer churn prediction has become an essential issue in telecommunication business. In such competitive business a reliable customer predictor will be regarded priceless. This paper has employed data mining classification techniques including Decision Tree, Artificial Neural Networks, K-Nearest Neighbors, and Support Vector Machine so as to compare their performances. Using the data of an Iranian mobile company, not only were these techniques experienced and compared to one another, but also we have drawn a parallel between some different prominent data mining software. Analyzing the techniques' behavior and coming to know their specialties, we proposed a hybrid methodology which made considerable improvements to the value of some of the evaluations metrics. The proposed methodology results showed that above 95% accuracy for Recall and Precision is easily achievable. Apart from that a new methodology for extracting influential features in dataset was introduced and experienced.

1. Motivation and significance

Nowadays business managers have started to appreciate the important role of churn prediction in their way of prosperity. In the literature, it has been repeatedly indicated that customer retention in comparison to absorbing new customers is significantly more achievable and less expensive. In today's competitive business environment, losing a customer should be considered as a real disaster. Loss of a customer can be contemplated in three different aspects. First, losing existing customers is figuratively equivalent to have a critical machine irreparably broken down due to the fact that they are any company's most precious assets. Furthermore, by the same imaginary assumption, losing a customer would mean passing our asset intentionally to our competitor. Finally it is too laborious a task to gain a new customer. To make the matters worse, even if a new customer were absorbed, they even would not be as loyal as the old customers. It may take some time for just a proportion of them to become slightly loyal. Therefore, the prevention strategy is absolutely worthwhile. Customer retention plays a major role in many enterprises, especially matured ones, including telecommunications and finances [1]. Acquiring it requires and rears churn prediction, which is another term and keyword in customer retention. It can be explained as predicting customers' probable tendency to switch to our competitor.

In today's telecommunication business environment, competition is tremendously fierce. The services and customers' options also have become more comparable and more competitive. This is the reason why customer loyalty tends to erode. It costs customers figuratively nothing to switch from a service provider to another. They are customers after all and they have freewill to switch to a better and probably more inexpensive service in a competitive market. We, as company managers, ought to take every necessary step so as to get in their way of leaving. It is imperative to distinguish customers who are not reluctant to move toward another competitor before they actually consider so. Therefore, dealing with the probability of customers' churn has become an inevitable issue in telecommunication industry [2]. The telecommunication service companies are annually facing with loss of valuable customers to competitors. Due to the changes and improvements in

* Corresponding author. Tel.: +98 9122388846.
E-mail addresses: keramati@ut.ac.ir (A. Keramati), ruholla.jafari@ut.ac.ir (R. Jafari-Marandi).

http://dx.doi.org/10.1016/j.asoc.2014.08.041
1568-4946/

telecommunications' services technologies in the last few years, customers' churn has resulted in magnificent losses and it has proven itself as a real issue [3].

2. Introduction

The number of mobile phone users during past years has perceptibly multiplied. Statistical report for the end of 2010 indicated that the number of mobile phone users in Iran has surpassed 70 million, more than 90% of the country population. Therefore, telecommunication market is in the point of being saturated, especially for big cities. Mobile phone penetration rates for certain city have gone above 100%, which means there are more subscriptions than inhabitants. As a result, heat of the competition in today's telecommunication market is distinguishably high. Proposed products and service offerings are becoming more and more similar and replaceable. The fact that customers, in most of the cases, are able to self-centrically prefer a service provider better brings about the eroding of customer loyalty. Then, Iran telecom operators in near future need to start giving a great amount of attention to customer churn prediction and customer retention strategies and should they fail to do so they would not survive. Furthermore, it has been repeatedly shown that taking customer retention strategy can be profitable for a company [4].

Not so long ago data mining techniques have been in use to tackle the challenging customer churn problems in telecommunication service field [3]. Due to the aforementioned heat of the competition in telecommunication market, these data mining techniques are mainly employed to cope with the churn prediction issues which has been receiving an unprecedented attention in the telecommunication industry and research. They are mostly and prevalently applied using the customer log-files or questionnaires so as to come up with some knowledge helping to determine the customers who are likely to churn. Although most of these data mining techniques use essentially different techniques to achieve approximately same result, the knowledge is totally worth spending and beneficial for almost any telecommunication company. The following paragraph is devoted to illustrate the importance of exact churn prediction.

Applying data mining techniques itself can be time-consuming and extravagantly expensive. Data Gathering, Data Cleaning, and Data Preprocessing in Iran rather immature telecommunication companies also can be painstakingly hard. It is important to understand that the knowledge extracted for data mining probably will be employed to make important decisions about the customers who are repeatedly known as the most important asset of any business. Therefore, the more accurate and reliable the extracted knowledge is, the more proper and appropriate decisions can be made. Making appropriate decision about such a vulnerable matter needs tremendous certainty and that is achievable by reliable knowledge extracted by reliable techniques. On the other hand, improper decision based on falsified, bias or flawed knowledge may lead to devastating situations where no one can make restitution.

Contemplating previous researches, they can be categorized by two main aspects. First is the studies which were centrally concerned about churn determinants. They analyzed well-known churn determinates and verified them by using customer behaviors in telecommunication. Some attributes, such as customer satisfaction, switching costs, customer demographics, tendency to change behavior, and service usage, have been prevalent among churn determinants [1]. Second, there are other studies such as this paper whose authors' interest has been to improve the outcome of churn prediction by state of the art computational methods. Having said that, although this paper's main part is to propose a novel methodology for more accurate prediction, a feature extraction method for dimensionality reduction is also presented.

A rather distinguished and important part of this paper is the idea of using intelligent algorithms for prediction of future matters. Intelligent algorithm has proven to be versatile and applicable for different tasks. They have been in use in variety of problems, the subject matter be Internet such as [5,6], ATM management [7], or the most basic statistical problem estimating missing value [8]. There have been many research works which employ neuro-fuzzy inference system to combine the advantages of fuzzy logic and neural network for classification and regression problems. There are many articles as an example for this adaptation: [9–14]. The combination of neural network and fuzzy logic in these studies benefits mostly from the human-like logic behind fuzzy systems and the connections of the neural networks. There are, moreover, approaches in which the use of meta-heuristic algorithms such as Genetic Algorithm (GA) for the parameter estimation of neural network or fuzzy system had been employed: [15,16]. Taking this strategy has many advantages because these types of problems are not closed form and more than too often the researcher opts to use an iterative algorithm to estimate their parameters. Meta-heuristic algorithms converge at a local minimum/maximum which is proven to be often sufficient. All the aforementioned works try to utilize intelligent algorithm to improve their estimates. These estimates in many works have come to natural phenomena. Such approaches have made great advances in different discipline such as production and inventory, or even natural hazards and risk management: [17–22].

However this paper's most important contribution is the churn prediction improvement by hybridizing some well-known algorithm. After experimenting with four mentioned prominent algorithms and coming to know their special features, a hybrid algorithm using all four is introduced. Not only did we prove that the suggested methodology is competitive with the best of the four, but also it has a tuning parameter that can be manipulated to predict the way its users need. If a decision maker needs to exactly know which customers will be churned, or if they want to realize who are the customers that have even slightest tendency to leave, the algorithm can be tuned to return the desired outcomes. Computational achievement in this paper is really strong since it has been able to push recall and precision measures, two of the most important measures for evaluating binary prediction, not to return less than 95% accuracy. Moreover, as it is an essential part of any data mining task we have investigated into the influence of the columns in our data. In real world setting computational limitation can bring about many restrictions. Thus, knowing the most influential columns can save many times an effort. To that end, a novel feature extraction methodology is suggested. The suggested method is experienced and consequently compared to traditional statistical means.

The rest of this paper is organized by 5 other sections. Section 3 is a short review of other research papers concerning telecommunication churn prediction. Section 4 describes in details the four classifiers algorithm used in our experiments. Section 5 is to show our experiences and results, whereas in Section 6 a hybridized methodology is explained and illustrated. Last but not least, Section 7 is the final discussion and conclusion of this paper experiments.

3. Literature review

Accurate and reliable prediction of churn customer is important in the development of appropriate retention strategies. Huang et al. [3] proposed a method based on ordinal regression to predict time of churn and tenure of customer in mobile telecommunication industry. They treated customer tenure as an ordinal outcome variable and take advantage of ordinal regression to form

a model. Their results showed that ordinal regression could be an alternative technique for survival analysis for the prediction of mobile customers' churn time. This paper has been the first study, its authors claimed, to use ordinal regression as a potential technique for modeling customer tenure. Zhang et al. [1] investigated the effects of interpersonal influence on the accuracy of customer churn predictions and proposed a novel prediction model that is based on interpersonal influence and combines the propagation process and customers' personalized characters. The comparison between results of traditional attributes-based models, network attributes-based models and combined attributes models proved that incorporating network attributes into predicting models can greatly improve prediction accuracy. Kim et al. [23] examined the communication patterns among subscribers in a mobile telecommunication company and introduced network analysis. They accomplished further improvements in churn prediction compared with the traditional machine learning approach that handles personal information stored in companies considering a propagation process in a network based on call detail records which transfers churning information from churners to non-churners. In a rather novel study, Verbeke et al. [24] used social network information to predict churn. They present an alternative modeling approach that integrates social network effects within a customer churn prediction setting using relational learning algorithm. Non-Markovian network effects are incorporated within relational classifiers. In this study, significant impact of social network effects on the performance of a customer churn prediction model is discussed and a novel parallel modeling setup is introduced to boost the profits generated by a retention campaign.

Hung et al. [25] intended to illustrate how to apply IT technology in order to facilitate telecom churn management. Their main goal, unlike to the most of other churn analysis research, was not to predict customer's churning attitude so as to decide what measures should be taken about the retention. However, authors opted for decision tree, neural network and K-means cluster among data mining techniques to come up with a churn predictive model. They had categorized their churn attributes into three segments according to their billing amount (to assess 'customer value'), tenure month (to appraise 'customer loyalty'), and payment behaviors (to engage 'customer credit risks'). Normally in churn analysis research the effective churn deterministic variables are chosen by running a regression model or other means; however, they simply used expertise's outlook in order to discern effective churn deterministic variables.

The purpose of using association rules is to reduce a large amount of information to a small and more understandable set of statistically supported statements [26]. Few studies have considered or illustrated the pre-processing step during data mining whose aim is to filter out unrepresentative data or information. All the same, the aim of Tsai and Chen [27] was to examine whether association rules can be adapted in the data pre-processing stage to reduce a large amount of information to a small and more understandable data variables in order to improve the prediction performance of neural networks and decision trees. They, after discussing some preprocessing steps, presented the important processes of developing MOD customer churn prediction models by data mining techniques. Their study contains a preprocessing stage for selecting important variables by association rules and also it consisted of a model construction stage by neural networks (NN) and decision trees (DT). Their experimental results showed that using association rules allows the DT and NN models to provide better prediction performances over a chosen validated dataset. Authors also investigated the combination of the data reduction and model development steps using data mining techniques for the problem of MOD customer churn prediction.

Lee et al. [28] focused on building an accurate and concise predictive model with the purpose of churn prediction by utilizing a partial least squares (PLS)-based methodology on highly correlated data sets among variables. Not only did they present a prediction model to accurately predict customers' churning behavior but also a simple but implementable churn marketing program was employed. Their proposed methodology allows a marketing manager to maintain an optimal (at least a near optimal) level of churners effectively and efficiently through their marketing programs. In this research PLS is employed as the prediction modeling method. Ning et al. [29] conducted an experimental investigation of customer churn prediction in telecom industry and proposed the use of boosting to enhance a customer churn prediction model. Unlike other boosting methods that improve the accuracy of a given basis learner, the authors suggested to separate customers into two clusters based on the weight assigned by the boosting algorithm. The proposed model provides an opportunity to an "Implementation Zone" where customers with the highest churn propensity can be addressed for retention actions.

De Bock and Poel [30] evaluated rotation-based ensemble classifiers for the prediction of customer defection. In rotation-based ensembles, feature extraction algorithms are applied to rotate the training data that is presented for training member classifiers in the ensemble. They compared two ensemble-based ensemble algorithms, i.e., Rotation Forest and RotBoost, to a set of often used benchmark algorithms, in terms of accuracy, AUC and top-decile lift on four real-life customer churn prediction applications. Their other interesting contribution was that they compared the influence of the use of three alternative feature extraction algorithms, i.e., principal component analysis (PCA), independent component analysis (ICA) and sparse random projections (SRP) on classification performance of both RotBoost and Rotation Forest. Idris et al. [31] proposed an intelligent churn prediction system for telecom by employing ensemble classification technique. The performance evaluation conducted on standard telecom datasets shows the effectiveness of proposed approach, which is based on RotBoost in combination with minimum redundancy and maximum relevance features, in handling high dimensionality of the telecom datasets and its high accuracy in predicting churners. In order to be able to better evaluate classification techniques, a cost-benefit analysis framework to the customer churn problem was applied by Verbraken et al. [32]. They proposed a new performance measure, the expected maximum profit criterion, which is aligned with the main objectives of the end users. Not only does the proposed framework assist companies with selecting the classifier which maximizes the profit, but also provides guidance about the fraction of the customer base to be included in the retention campaign.

Lin et al. [2] proposed three measurements including Modification Measures, Execution Effectiveness Measures and Cost-Benefit Analysis to evaluate the effectiveness of the churn models after retaining potential churner. They showed to be appropriate effectiveness measures so as to assist in evaluation of execution effectiveness and give a proper feedback to adjust data mining model or redesign marketing activities. Authors also designed a real experimental architecture for telecom churn management application to use it as a pilot project for real adopting data mining model. Pendharkar [33] proposed two GA-based neural network (NN) models to predict customer churn in subscription of wireless services. Their first GA-based NN model uses a cross entropy based criterion to predict customer churn, and their second GA based NN model attempts to directly maximize the prediction accuracy of customer churn. Using real-world cellular wireless services dataset and three different sizes of NNs, they compared the two GA-based NN models with a statistical z-score model using several model evaluation criteria, which include prediction accuracy, top 10% docile lift and area under receiver operating

characteristics (ROC) curve. The results of experiments indicate that both GA-based NN models outperform the statistical *z*-score model on all performance criteria.

Briefly, in other studies, Kim and Yoon [34] identified the determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Owczarczuk [35] tested the usefulness of the popular data mining models to predict the clients churn of the Polish cellular telecommunication company. Author dealt with prepaid clients who are far more likely to churn, less stable and much less we known about them. The results of his research showed that linear models, especially logistic regression, are a very responsive choice to mode churn of the prepaid clients. Kisioglu and Topcu [36] constructed a model by Bayesian Belief Network to identify the behaviors of customers propensity to churn. They analyzed three different scenarios that examine the characteristics of the churners and also suggested promotions to reduce the churn rate. Coussement and Van den Poel [37] developed a DSS for churn prediction. It tries to integrate free-formatted, textual information from customer emails with information derived from the marketing database. They investigated the beneficial effect of adding the voice of customers through call center emails – i.e. textual information – to a churn-prediction system that only uses traditional marketing information. Their findings proved having unstructured, textual information added into a conventional churn-prediction model will result in a significant increase in the prediction performance. Usero Sánchez and Asimakopoulos [38] examined the effects of mobile number portability (MNP) implementation on competition in the European Mobile communications industry. They concluded that subscriber churn rates are negatively affected by both the level of charges levied on subscribers wishing to maintain their current number (porting) when switching mobile providers and the length of time required switching. Sweeney and Swait [39] investigated the important additional role of the brand in managing the churn of current customers of relational services. Their research leaded to the enhanced understanding that the brand has a significant role to play in managing long-term customer relationships, and details how the usual tools of customer relationship management, satisfaction and service quality, relate to brand credibility. Their results from samples of retail bank and long distance telephone company customers indicated that brand credibility serves a defensive role. Huang and Kechadi [40] proposed a hybrid model-based learning system, which integrates the supervised and unsupervised techniques for predicting customer behavior. The system combines a modified k-means clustering algorithm and a classic rule inductive technique (FOIL). Authors concluded that their hybrid model is very promising and outperform the existing models as the result of a comparative study on a set of benchmarks and use of real telecom datasets.

Contemplating Table 1, which is the summary for literature review, one can see that data mining techniques, such as Decision Tree, Neural Network, Support vector machine, Bayesian Belief Networks, and Regression, have been prevalently employed in telecommunication customer churn prediction studies. Nevertheless, just a few of them have gone through the process of tuning these techniques parameters in order to compare them to each other. In this paper we have experienced, tuned, and compared four prominent data mining classification techniques namely Decision Tree, Artificial Neural Network, K-Nearest Neighbor and support Vector Machine. Apart from that, the average value of 98 and 97% accuracy respectively for Precision and Recall measures proved that our new hybrid methodology could predict test records with the least inaccuracies. To the best of our knowledge the process of identifying influential features, in other word significant features, is often performed by statistical tools and method. This paper, however, has introduced a hubristic methodology which extracts features with the most influential role in prediction accuracy.

4. Data mining techniques

4.1. Decision tree (DT)

One of the most popular and prevalently in use classification techniques among others is decision tree. Its exceptional flexibility and understandability are its greatest advantage which has resulted in its popularity. It is a great means to predict different categories (classes) by taking into account the values of predictor attributes. The decision tree's very informative visualization and its flexibility make it practically advantageous. As an experimental exploratory technique, especially when other techniques have failed, decision trees may successfully be employed. Decision trees can be seen as an answer to many problems caused in era of data and information. It has gained popularity because of its conceptual transparency. Decision trees are also useful to be used as knowledge-based experts systems [43].

Hunt et al. [44] was one of the first study aimed at constructing a decision tree. Their proposed algorithm is known as the general way of inducting decision trees in current software. Their proposed methodology is an understandable 4-stepts procedure. The following is Hunt's algorithm inducting decision trees:

1. Denote by *Dt* the set of training objects (data) that reach node *t*.
2. If *Dt* is an empty set, then *t* is a terminal node (a leaf node), labeled by the class Φt.
3. If *Dt* contains objects that belong to the same class *Ct*, then t is also a leaf node, labeled as *Ct*.
4. If *Dt* contains objects that belong to more than one class, then we use an attribute test to split the objects into smaller subsets.

The algorithm is effective and simple, but there is one important matter for our notice. It seems, in step 4, there is a methodology missing to choose an attribute among not-split ones to make use in time of splitting. In the literature, there is a concept called impurity, which is also known as goodness-of-fit. There are famous measures addressed to estimate the impurity of a node, such as GINI index, entropy, misclassification, Chi-square, and G-square measures. In the process of choosing which attribute is most appropriate to split (step 4), they are employed to estimate the impurity of each possible attribute. There are some well-known advantages for decision trees and they can be useful to decide whether employing decision tree is a viable choice [45]:

- Easy to understand and interpret.
- Inexpensive to be built. They require a small amount of training data compared with other classification techniques.
- Use of both numerical and categorical data without any restriction.
- "'Transparent-box" type, the classification rules can be understood 'at first sight'. Other classification techniques, such as artificial neural networks, act as "black-box" models, do not directly provide the user with the classification rules.

4.2. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is an information processing mechanism which is inspired from biological nervous systems. ANN comprises some interconnected elements, or neurons, working together as one system to solve specific problems. The structure of an ANN is determined by both the inter-neuron connections' arrangement and the nature of these connections. The way of training or adjusting the strengths of these connections so as to achieve a desirable overall behavior is known as learning process [46]. ANNs are well-known for their well-established empirical modeling power. Their most outstanding feature is their ability to learn

Table 1
Literature review summary.

Gopal and Meher [41] Filed Mobile telecommunication industry Data A major wireless telecommunications company Model Classical tenure modeling approach – survival analysis Methods Ordinal regression (OR) Multi-class classification (MC) Decision tree Cox proportional hazards (PH)	Churn deterministic Number of months in service Churn time Tenure Cumulative satisfaction Perceived losses transaction Failures Bad service quality
Zhang et al. [1] Filed Telecommunication Service Data Telecommunication service database of a mobile telecommunication company Model A novel prediction model based on interpersonal influence Classification model Propagation model Traditional attributes-based models Methods Lift curve AUC Hit rate	Churn deterministic Traditional attributes Customer satisfaction (*Prices, brand, key accounts level, function of the handset, complaints)* Switching barrier (*Loyalty points, Feedback gift, Discount level Lucky number)* Service usage (*Monthly call counts, Monthly payment, Short message, Diversity of service usage, Activeness of service usage,Duration of subscription)* Price sensitivity (*Chang billing suite, Free call duration, Free call duration, Frequency of paying)* Previous anomaly behavior (*Change of fees, Change of call counts, Anomaly status, Calling behavior)* Network attributes Neighbor composition/Tie strength/Homophily/Similarity/Structural cohesion/Structural cohesion/Structural cohesion/Influence of churn neighbors
Tsai and Chen [27] Filed Telecommunication companies that provides MOD services Data One of the telecommunication companies that provides MOD services in Taiwan Model NN and decision tree along with association rules	Churn deterministic The branches/Customers who either accept product promotion or not/The method of paying the MOD services/Whether the services are recommended by the employee or not/The speed of uploading and downloading data/discount method/Broadband services for either ADSL or FTTB/The broadband distance/from the MOD server to the customer/The times of logging on MOD/The basic fee for the MOD services/Extra fees for customer's preferences
Lee et al. [28] Filed Mobile telecommunication business environment Data Mobile phone service provider (data Center for CRM at Duke University) Model Linear and non-linear PLS models Logit regression models	Churn deterministic Number of days of current equipment/Handset (refurbished or new)/Range of overage minutes of use/Mean of unrounded minutes of use of completed voice calls/Mean of unrounded minutes of use of completed voice calls/Total number of months in service/Account spending limit/Billing adjusted total minutes over the life of the customer/Mean of number of minutes of use/Range of revenue of voice overage/Range of revenue of overage/Percentage change in monthly minutes of use vs. previous 3 month average/Average monthly number of calls over the life of the customer
Lin et al. [2] Filed Mobile telecommunication industry Model Confusion matrix/lift chart/cumulative gain Methods Data mining (Decision Tree, SVM, Regression, Clustering)/Modification Measures/Execution Effectiveness Measures/Cost-Benefit Analysis	Churn deterministic Billing amount Contract status Call behavior

Table 1 (*Continued*)

Pendharkar [33] Filed Wireless network Data Customer log files Methods GA based NN Statistical z-score	Churn deterministic The actual value of variable The maximum peak minutes allowed in the plan The monthly cost for using the peak minute service for less than or equal to peak minutes The cost per minutes that the customer has to pay for the ongoing
Ahn et al. [42] Filed Korean telecommunication Data Customer log files Methods Logistic regression	Churn deterministic Customer satisfaction: *Call quality/Tariff level/Billing/Value-added service/Customer services/Handset/Brand image/Switching cost/Age/Sex/Income/Monthly payment/Duration of subscription/Handset usage/Switching experience*
Hung et al. [25] Filed Taiwan wireless telecommunication Data Billing data/Call detail records (CDR)/Customer care Model Self-defined predictive model Method Data Mining Techniques: *Decision tree/Neural network/K-means cluster*	Churn deterministic Customer Demographic *Age/Tenure/Gender* Bill and Payment *Monthly Fee/Billing Amount/Count of overdue payment* Call detail record *In-net call Duration/Call Type* Customer Care *Count of Change of number/Count of barred or suspended*
Kim and Yoon [34] Filed Korean telecommunication Data Customer transaction Billing data Model Econometric model Binomial logistic model Method Logistic regression	Churn deterministic Customer dissatisfaction *Call drop rate/Call failure rate/Number of complaints* Switching cost *Loyalty points/Membership card program* Service usage *Monthly billed amount/Unpaid balances/Number of unpaid bills* Customer-related variables *Calling plans/Handset capability/Handset manufacture*
Coussement and Van den Poel [37] Field Belgian newspaper publishing company	Churn deterministic Client/company-interaction variables: *The number of complaints./Elapsed time since the last complaint./The average cost of a complaint (in terms of compensation newspapers)./The average positioning of the complaints in the current subscription./The purchase motivator of the subscription./How the newspaper is delivered./The number of conversions made in distribution channel, payment method and edition./Elapsed time since last conversion in distribution channel, payment method and edition./The number of responses on direct marketing actions./The number of suspensions./The average suspension length (in number of days)./Elapsed time since last suspension./Elapsed time since last response on a direct marketing action./The number of free newspapers.* Subscription-describing variables: group of variables describing the subscription *Elapsed time since last renewal./Monetary value./The number of renewal points./The length of the current subscription./The number of days a week the newspaper is delivered (intensity indication)./Which edition the subscriber has (X1, X2, X3)./The month of contract expiration./bad service quality*

Table 1 (*Continued*)

Data Marketing database and customer emails Method Logistic regression Churn deterministic Socio-demographic variables: variables describing the subscriber: *Age/Whether the age is known/Gender/Physical person (is the subscriber a company or a physical person)/Whether contact information (telephone, mobile number, email) is available* Renewal-related variable: variables containing renewal specific information *Whether the previous subscription was renewed before the expiry date./How many days before the expiry date, the previous subscription was renewed./The average number of days the previous subscriptions are renewed before expiry date./The variance in the number of days the previous subscriptions are renewed before expiry date./Elapsed time since last step in company retention procedure./The number of times the customer did not renew a subscription*	
Usero Sánchez and Asimakopoulos [38] Field European mobile communication industry Data Wireless Intelligence Database Model Econometric Model	Churn deterministic Independent variables *Portability period/Customer fee* Control variables *Industry growth/Penetration/(ARPU) Average Revenue per User of mobile services/UMTS/Number of operators*
Owczarczuk [35] Field Polish cellular telecommunication company Data Large data marts Method *Logistic regression/linear regression/Fisher linear discriminate analysis/decision trees*	Churn deterministic Demographical or personal data Clients' call direct records Average minutes of usage
Kisioglu and Topcu [36] Field Telecommunication industry of Turkey Data Large data set Method Bayesian Belief Networks	Churn deterministic Place of residence Age Tenure Tariff type Average billing amount Trend in billing amount Average minutes of usage Average frequency of usage
Sweeney and Swait [39] Field Retail banking and telecommunication in North American metropolitan Area Data Questionnaire Method Statistical methods	Churn deterministic Brand credibility Trust worthiness Expertise Customer satisfaction, loyalty commitment (LC), continuance commitment (CC), Word of-mouth (WOM) Switching propensity

Table 1 (*Continued*)

Huang and Kechadi [40]	Churn deterministic
Field	Demographic profiles
	Age/Gender/Social class bands/County code
Telecommunications industry	Account information
	Start time/account/number/type/fees/payment type/account balance/call information
Data	Call details
	Types of calls (e.g., international or local calls), number of calls/call duration/costs
Dataset provided by a telecom company	
Model	
Hybrid model-based learning system	
Method	
Data mining techniques:	
K-means clustering/First order inductive learning	
Kim et al. [23]	Churn deterministic
Field	Demographic profiles
	Age/gender/etc.
Mobile telecommunications industry	Product details
	Cell phone types and performance/duration from the last change of cell phone
Data	Service satisfaction factors
	Number of complaints/service quality score/loyalty score
A dataset obtained from a mobile telecommunication company	Propensity to telephone calls
	The proportion of non-voice calls/proportion of calls during the day time
Method	
A procedure using propagation process	
Ning et al. [29]	Churn deterministic
Field	Mobile plan and contract information
Mobile telecommunications industry	Billing
Data	Usage
Mobile customer database	Product holding information
Method	Customer care inbound/outbound information
Boosting method	
Logistic regression	

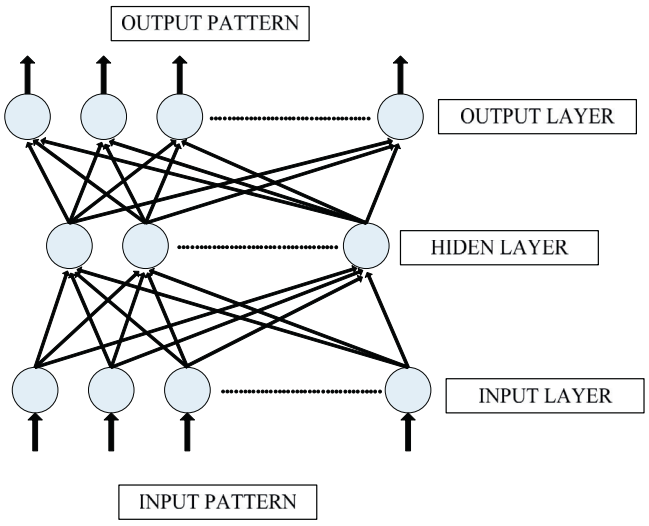

Fig. 1. Feed-forward network with one hidden layer.

automatically from available data in order to provide a means for predictions. They also are able to impose blind (hidden) insights into the hidden relationships [47].

ANNs can be categorized into two different groups considering their structures: Feed-Forwards and Recurrents. A Feed-Forward network's neurons are grouped into input, hidden and output layers. Feed-Forwards networks can be recognized by unidirectional arrows which flow from input layer through the output, not connecting neurons in a same layer but connecting them from previous layer just to the next one [46]. A simple Feed-Forward network with one hidden layer is presented in Fig. 1.

The most influential different between Feed-Forwards and Recurrents is that Recurrents unlike the others are allowed to point backwardly to previous layer. Having stunning ability to extract patterns which, by and large, are ambiguous for us, ANNs have introduced themselves as reliable means of abstruse classifications. However because of the well-known black-box nature, their explanation capability unlike the other methods is limited.

4.3. K-Nearest Neighbors (KNN)

Among various classification techniques, most of them rely on many assumptions made over data but in practical cases only few of them are applicable. Beyond such methods another type of learning algorithms named non-parametric learning has been introduced. These types of algorithms make no assumption on the data and therefore they are applicable on many real world problems. K-Nearest Neighbors (KNN) is one the most applicable and useful non-parametric learning algorithms. KNN may be known as a lazy algorithm, that is, all training data are used at testing phase. In fact there is no training phase and all data points are used directly in test phase, so these all points need to be used when it is to be tested [48].

KNN uses the distance between records so as to apply it for classification. In order to measure the distance between the points, KNN assumes that these points are scalar or multidimensional vectors in feature space. Euclidean distance is one of the most commonly used measuring methods used in KNN. All data points are vectors of feature space and the label will define their classes. The simplest case is when the class labels are binary but still it is applicable on arbitrary class numbers.

The classification problem for KNN can be stated as if we have some labeled data points (used for training) by which the model is trained and also some test unlabeled data by which the model's performance is measured. Assume that we are to classify the test data using only one neighbor. Let the test point be x and name y for the nearest train data (labeled data) to x. According to a rule (nearest neighbor rule), x has to be labeled after y's label. This may seem naive and unreliable but it has proved to work, especially when there are a large number of data points, its result improves. In cases where we talk about k neighbors (k is odd), the most straightforward way is to find the k nearest neighbors and assign x to the class in which most of k neighbors vote for it. For example let $k = 5$, three neighbors of x are of class C1 and two of them C2. According to KNN rule x is labeled as C1. This is obviously the simplest way and one extension as it is to do a weighted voting. One way to weigh the votes is to simply multiply it by inverse of its distance to x; this means the nearer the neighbor is the higher votes it has.

4.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) was first introduced by Vapnik [49]. The basic version of classifier would deal with two-class problems in which the training data are divided into two classes using a hyperplane which is defined by a number of support vectors. It was based on the Structural Risk Minimization (SRM) principle stated in computational learning theory. The basic idea behind supervised learning methods is to learn from observations. SVM endows us with a binary classifier by actually finding optimal separated hyperplanes through nonlinear mapping of the input vectors into a high-dimensional feature space. Based on support vectors, SVM constructs linear models to estimate the decision function by taking the advantage of the nonlinear class boundaries. In case the training data are linearly separable, SVM results in the optimal hyperplane with maximum distance between the hyperplane and those training sample which are closest to the hyperplane. These samples with minimum distance to the hyperplane are called support vectors. SVM only uses support vectors so as to find the hyperplane and so all other training samples are irrelevant. When the data are not linearly separable SVM uses nonlinear machines to find a hyperplane based on the minimum training error criterion [50].

In order to illustrate how SVM works we assume the simplest case where there are only two linearly separable classes. Label the training data by $\{x_i, y_i\}, \quad i = 1, \ldots l, \quad y_i \in \{-1, 1\}, \quad x_i \in \mathbb{R}^d$. Suppose we have some hyperplane which separates positive from negative examples. The points x which lie on the hyperplane satisfy the equation: $w \cdot x + b = 0$, w is normal to the hyperplane, $|b|/\|w\|$ is the perpendicular distance from the hyperplane to the origin, and $\|w\|$ is the Euclidean norm of w. Let d_+ (d_-) be the shortest distance from the separating hyperplane to the closest positive (negative) example. Define the "margin" of a separating hyperplane to be $d_+ + d_-$ [51]. Such case is shown in Fig. 2. For this particular case the algorithm simply looks for the separating hyperplane with maximum margin. It is formulated as follows, satisfying these constraints on all the training data:

$$x_i \cdot w + b \geq +1 \quad \text{for} \quad y_i = +1 \tag{1}$$

$$x_i \cdot w + b \leq -1 \quad \text{for} \quad y_i = -1 \tag{2}$$

which is restated such:

$$y_i(x_i.w + b) - 1 \geq 0 \, \forall i \tag{3}$$

The support vectors where the equality in Eq. (1) holds lie on the hyperplane $H_1 : \quad x_i \cdot w + b = +1$ with normal w and perpendicular distance from the origin $|1 - b|/\|w\|$. Similarly the support vectors, for which the equality in Eq. (2) holds, lie on the hyperplane $H_2 : \quad x_i \cdot w + b = -1$, with again normal w, and perpendicular distance from the origin $|-1 - b|/\|w\|$. Hence $d_+ = d_- = 1/\|w\|$ and the margin is simply $2/\|w\|$ [51]. Obviously H_1 and H_2 are parallel and no training points are between them. Thus minimizing $\|w\|^2$,

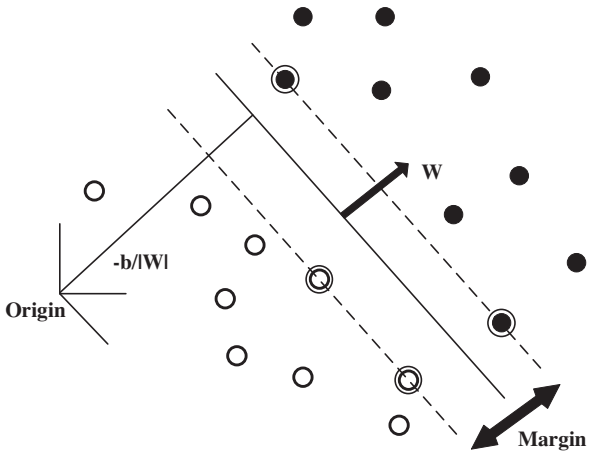

Fig. 2. Linear separating hyperplanes for the separable case [51].

subject to constraints Eq. (3) leads us to finding hyperplane pairs with maximum margin.

The minimization is done using Lagrangian multipliers mainly because the constraint Eq. (3) can be used easily to reformulate the problem and all the training data will only appear in the form of dot products between vectors [51]. What if the training points are not linearly separable? In such cases we introduce Kernel Functions which transform the problem to a high dimensional space in which the points in that are linearly separable. The decision function without using kernel is in the form stated in the Eq. (4) but in case the data are not linearly separable, it is transformed into a high dimensional feature space in the form stated in Eq. (5).

$$Y = sign\left(\sum_{i=1}^{N} y_i \alpha_i (x \cdot x_i) + b\right) \tag{4}$$

$$Y = sign\left(\sum_{i=1}^{N} y_i \alpha_i K(x, x_i) + b\right) \tag{5}$$

In Eq. (5) the function $K(x, x_i)$ is the kernel function which transforms the training data into a high dimensional feature space and only deals with dot product of the input samples. The vector $x = (x_1, x_2, x_3, \ldots, x_n)$ is the input and the vectors x_i, $i = 1, \ldots, N$ are support vectors. Also b and α_i are the parameters determining the hyperplane. Common types of Kernel functions are Polynomial, Radial Basis Function (RBF) and Neural Network kernel functions.

Using Kernel functions, the input data are transformed into a high dimensional feature space and in many complex cases it makes the feature space easy to classify and in the case of telecommunication churn prediction, number of features are numerous and it is a good idea to use Kernel functions to make them easy-to-handle.

5. Experiments

In the previous section, we introduce 4 different and prominent classification techniques. In order to evaluate and compare these techniques normally they have to undergo a performance evaluation procedure. As it is presented in Fig. 3, the whole data have been separated into two train and test sets. Usually train set comprises 70% of data set and consequently test set is 30% of it. A classifier which has been trained by train set will be used so as to predict test set records. Finally, using a lot of known evaluation measures, for instance misclassification, the predicted values will be compared to real value of test set.

5.1. Dataset

We used the dataset which was randomly collected from an operator call-center's database over a 12-month period. The dataset contains 3150 customer data such as number of Call Failure (CF), number of Complains (Co), Subscription Length (SL), Charge Amount (CA), Seconds of Use (SU), Frequency of use (FU), Frequency of SMS (FS), Distinct Calls Number (DCN), Age Group (AG), Type of Service (TS), Status (St), and Churn (Ch). Obviously the class attribute is Churn. Looking into data, we saw that there were 495 records with the class label churned and the rest, i.e. 2645 records, were non-churned. For the sake of keeping our experiments in proper randomization environment, for each single block we randomly select 70% of dataset for training set and the rest for the validation process, keeping the proportion the same. In another word, our every single experiment was done by a training set with 347 records labeled churned and 1858 non-churned ones (totally 2645 records). Also all of this paper test process has been performed by 148 churned and 347 not-churned classes (totally 495). By the aforementioned process we prepared 5 blocks in order to run and compare experiments.

5.2. Evaluation measures

In terms of binary classification there are known measures to check the accuracy or performance of a classifier. They range from as simple as accuracy or misclassification measures to precision and recall which can show more insightful information about the

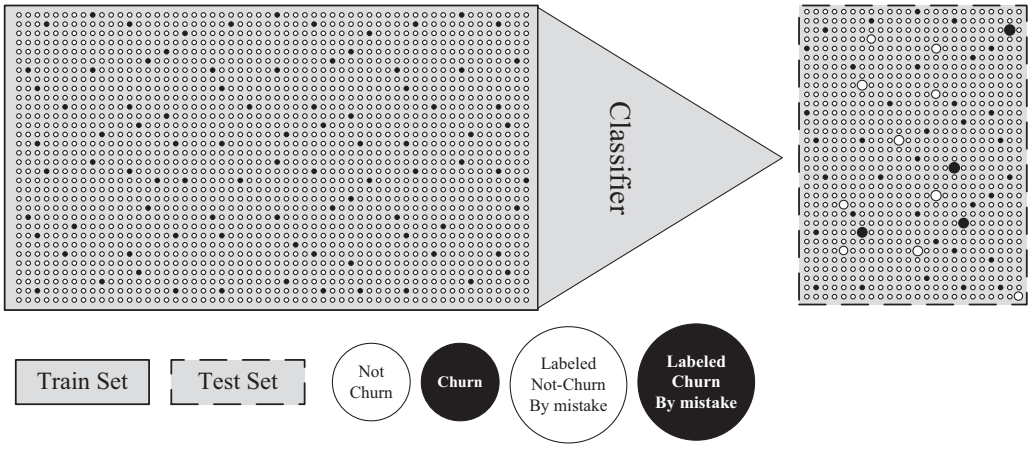

Fig. 3. Presentation of a classification experience.

Table 2
Confusion matrix, shown with totals for positive and negative tuples.

		Predicted class		Total
		Yes	No	
Actual class	Yes	*TS*	*FN*	P
	No	*FP*	*TN*	N
Total		$\tilde{P}$	$\tilde{N}$	

performance of classifiers. Obviously there is not a better or more eligible classifier for all the cases and a data mining practitioner normally chooses one of them based on his/her needs. That is to say, understanding different classifiers will make it easy to grasp the meaning of the various measures. The following is the short description for these known measures.

There are four terms better to be introduced in order to get familiar with each measure in terms of both calculating and preference. True positives (TS): refer to the positive tuples that were correctly labeled by the classifier. Let *TS* be the number of true positives. True negatives (TN): These are the negative tuples that were correctly labeled by the classifier. Let *TN* be the number of true negatives. False positives (FP): These are the negative tuples that were incorrectly labeled as positive. Let *FP* be the number of false positives. False negatives (FN): These are the positive tuples that were mislabeled as negative. Let *FN* be the number of false negatives. The terms are summarized in the confusion matrix of Table 2. Moreover, Eqs. (6)–(10) respectively show the definitions for accuracy, miscalculation, precision and recall measures.

$$\text{Accuracy} = \frac{TP + TN}{FP + FN} \tag{6}$$

$$\text{Missclasification} = \frac{FP + FN}{P + N} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$F\text{-Score} = \frac{\text{Precision} \times \text{Recall} \times 2}{\text{Precision} + \text{Recall}} \tag{10}$$

The accuracy and miscalculation of a classifier's result on a given test set are concerned with the percentage of test set tuples that are correctly or incorrectly classified, whereas precision and recall are known respectively as a measure of exactness and completeness. As it mentioned before it is highly related to the case of classification to opt for one of these measures. Since this paper classification is about the question of predicting whether a customer churn or not in order to make a decision about taking proper retention actions, in this case, we argue that precision and recall are more eligible measures. As it has been repeatedly mentioned in the literature, normally the cost of losing a valuable customer is far more than taking preventive or retention actions. And since important and valuable customers are normally taken into consideration of churn classifications, the exactness and completeness of classification performance is of our interest. In other words, if an actual non-churn is classified as churn by the classifier it is not favorable but do not cause a big loss on the other hand knowing an actual churn for a non-churn can bring about a steep loss by relying on the outcome of the classifier. All in all, to conform to this paper assumptions and case of study we have used precision and recall measures in order to evaluate the performance of every classifier in this paper.

5.3. DT

In order to examine the powerfulness or in other word responsiveness of decision trees dealing with churn classification problem we tried to construct decision trees with most of available and frequently in use software in data mining environment. We built the following completed block design ANOVA (Table 3), first to examine the responsiveness of decision trees handling churn prediction problem, and then to compare the current software. There are three values for each cell in the ANOVA table which are respectively, from the left, the value of recall (RE), precision (PR) measure and misclassification (MI) for each decision tree run.

ANOVA completed block design results (Appendix A) for both *F*-Score and Misclassification measures depicted that there is significant different in performance of all decision trees. It means that some of them have meaningfully done better job. WEKA (*Random Forest*) shows to be the most responsive decision tree for churn prediction. Although, MATLAB decision performance was not as good a classifier as a few of the WEKA decision trees were, its performance was acceptable and was among good ones.

5.4. ANN

In this paper, Artificial Neural Network is employed in quest of identifying the best fitted network for the special characteristics of churn prediction problem. This paper dataset comprises 12 attributes: 11 attributes are inputs and one of them (churn) is the output. In order to find the best network for churn prediction purpose we built both one hidden and two hidden layers networks and observed their performances. For each and every number of neurons in one hidden layer and also for each and every combination of the numbers for neurons in two-hidden-layers networks we randomly prepared 10 new experiments (training and test set).

Table 3
Prediction experiments by decision trees.

Experiments	1			2			3			4			5		
	RE	PR	MI	RE	PR	MI	RE	PR	MI	RE	PR	MI	RE	PR	MI
MATLAB	0.83	0.80	55	0.82	0.77	63	0.82	0.83	51	0.74	0.85	58	0.82	0.83	52
R (*Part*)	0.40	0.97	90	0.75	.78	69	0.51	0.93	78	0.62	0.86	71	0.49	0.88	86
R (*rparty*)	0.78	0.75	71	0.77	0.73	76	0.77	0.73	76	0.8	0.72	74	0.67	0.83	68
WEKA (*AD.Tree*)	0.61	0.73	92	0.36	0.92	101	0.41	0.82	110	0.38	0.97	95	0.61	0.72	93
WEKA (*BF.Tree*)	0.79	0.86	52	0.75	0.84	58	0.77	0.86	58	0.81	0.93	38	0.76	0.78	66
WEKA (*FT*)	0.78	0.86	52	0.83	0.79	59	0.75	0.86	61	0.75	0.86	55	0.78	0.81	60
WEKA (*J48*)	0.83	0.80	57	0.85	0.77	61	0.71	0.90	61	0.78	0.92	41	0.77	0.87	52
WEKA (*LADTree*)	0.69	0.70	90	0.66	0.76	84	0.69	0.79	81	0.72	0.69	90	0.66	0.80	76
WEKA (*LMT*)	0.81	0.88	46	0.84	0.81	52	0.78	0.89	53	0.78	0.92	43	0.83	0.87	44
WEKA (*NBTree*)	0.77	0.83	59	0.81	0.77	62	0.67	0.93	59	0.69	0.78	62	0.71	0.88	58
WEKA (*Random Forest*)	0.79	0.90	45	0.83	0.90	41	0.81	0.94	39	0.83	0.91	38	0.81	0.89	44
WEKA (*Random Tree*)	0.75	0.79	68	0.82	0.82	55	0.79	0.85	56	0.75	0.87	54	0.76	0.83	60
WEKA (*REPTree*)	0.78	0.81	60	0.78	0.71	73	0.81	0.84	55	0.67	0.89	62	0.76	0.83	59
WEKA (*Simple Cart*)	0.85	0.85	46	0.75	0.87	55	0.83	0.88	45	0.78	0.93	42	0.83	0.80	57

Table 4
F-Score results for the each number of neurons in one-hidden-layer networks.

N. neurons	1	2	3	4	5	6	7	8	9	10
F-Score	0.681	0.669	0.706	0.694	0.778	0.756	0.801	0.787	0.783	0.681
N. neurons	11	12	13	14	15	16	17	18	19	20
F-Score	0.829	0.811	0.799	0.834	0.835	0.833	0.788	0.839	**0.845**	0.835

Table 5
F-Score results for the each combination for the number of neurons in two-hidden-layer networks.

N. neurons	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.662	0.659	0.642	0.609	0.679	0.657	0.561	0.621	0.598	0.626	0.661	0.6	0.552	0.652	0.573
2	0.675	0.652	0.666	0.671	0.752	0.705	0.657	0.74	0.68	0.762	0.753	0.733	0.733	0.659	0.745
3	0.675	0.697	0.671	0.703	0.747	0.783	0.743	0.727	0.702	0.76	0.773	0.783	0.792	0.765	0.793
4	0.736	0.854	0.677	0.735	0.808	0.765	0.755	0.772	0.801	0.815	0.825	0.816	0.795	0.813	0.749
5	0.71	0.742	0.774	0.795	0.826	0.771	0.753	0.828	0.842	0.823	0.822	0.849	0.836	0.828	0.818
6	0.75	0.755	0.772	0.789	0.755	0.819	0.813	0.822	0.786	0.833	0.774	0.823	0.823	0.823	0.83
7	0.775	0.793	0.784	0.744	0.765	0.808	0.747	0.824	0.841	0.83	0.843	0.829	0.842	0.829	0.847
8	0.789	0.803	0.845	0.789	0.791	0.81	0.841	0.783	0.857	0.828	**0.877**	0.847	0.845	0.789	0.851
9	0.783	0.784	0.791	0.8	0.807	0.811	0.85	0.83	0.758	0.854	0.825	**0.87**	0.841	0.838	0.863
10	0.769	0.755	0.699	0.821	0.777	0.805	0.817	0.845	0.828	0.852	0.828	0.857	0.836	0.829	0.808
11	0.778	0.756	0.745	0.784	0.821	0.787	0.798	0.795	0.841	**0.871**	0.852	0.775	0.816	0.872	0.837
12	0.784	0.831	0.826	0.798	0.752	0.833	0.814	0.86	0.783	0.858	0.836	0.86	0.865	0.863	0.84
13	0.78	0.773	0.811	0.812	0.796	0.792	0.833	0.824	0.83	0.772	0.835	0.799	0.854	0.855	0.86
14	0.651	0.73	0.786	0.813	0.844	0.806	0.783	0.804	0.839	0.867	0.824	0.811	0.829	0.859	0.858
15	0.701	0.722	0.73	0.84	0.688	0.789	0.807	0.82	0.82	0.86	0.861	0.843	0.859	**0.867**	0.862

Table 6
KNN tuning.

	1	2	3	4	5	6	7	8	9	10
Euclidean	0.5119	0.5119	0.4170	0.4170	0.3968	0.3695	0.3592	0.3362	0.2819	0.2489
City block	0.5544	0.5544	0.4981	0.4747	0.4615	0.4219	0.4351	0.4309	0.3821	0.3739
Cosine	**0.7343**	0.7343	0.7168	0.7189	0.7007	0.7119	0.7079	0.7113	**0.7317**	0.7234
Correlation	0.7178	0.7178	0.7247	0.7128	0.7083	0.7148	0.6920	0.7067	0.7298	0.7298

Tables 4 and 5 are the average *F*-Score measure for each 10 experiences.

Contemplating the outcome of aforementioned experiences we can see that a network with two hidden layer can by far outperform a network with just one hidden layer since the best performance of two hidden layers by respectively, 8 and 11 neurons was 0.877 and the best performance of one hidden layer with 19 neurons was 0.845 (Tables 4 and 5).

5.5. KNN

The other data mining technique that has been applied to this paper dataset is KNN. It is among those techniques which radically needs beforehand tuning. The tuning parameters of KNN model using cross-validation are presented in Table 6. This table reports the *F*-Scores calculated using the number of neighbors and method of distance calculation. The results showed that using cosine distance method along with 1 neighbor has led to the best *F*-Score gained by KNN.

5.6. SVM

As it mentioned before SVM has several parameters to be tuned before final usage. So as to specify them, we applied cross-validation tuning for SVM parameters. The best *F*-Score achieved by SVM was 0.8383. The summary table for the final result of different kernel on test data is shown in Table 7. Our experience proved that the best tuning of parameter for this paper classification problem is applying polynomial kernel with order of 4. This has been used, from now on, whenever SVM is mentioned.

Table 7
SVM summary tuning.

Kernel type	Best *F*-Score	Kernel params
RBF	0.821	Sigma = 0.498
MLP	0.6618	$P1 = 0.11$, $P2 = -1.61$
Polynomial	0.8383	Order = 4

5.7. Comparison

So far we have introduced 4 different soft computing techniques for churn prediction. Except for decision tree, other soft-computing techniques' demeanors were observed and consequently tuned to be adapted to the spirit of churn prediction problem in their best ways. This part of paper's aim is to compare the performances and the demeanors of all of them. Table 8 shows a complete block design table in which all of the introduced soft computing techniques results (*F*-Scores) for the aforementioned blocks are presented. Except for block 3 in which SVM has slightly outperformed ANN, it has been ANN that has had the best results. Tables 9 and 10 are respectively summary and the ANOVA complete block design analysis result for the techniques. The ANOVA results (Table 10) as it could be predicted, have rejected the hypothesis that the

Table 8
Complete block design table for comparison between DT, ANN, KNN and SVM.

Experiment	1	2	3	4	5
Best DT	0.843	0.824	0.831	0.844	0.849
ANN	0.881	0.871	0.857	0.851	0.851
KNN	0.723	0.742	0.744	0.763	0.773
SVM	0.806	0.831	0.860	0.823	0.823

Table 9
Summary for Table 8.

Summary	Count	Sum	Average	Variance
Best DT	5	4.191	0.8382	0.000107
ANN	5	4.311	0.8622	0.000177
KNN	5	3.745	0.749	0.000381
SVM	5	4.143	0.8286	0.000391

performances of different techniques approaching churn prediction are similar (p-value $\ll 0.05$). Furthermore, if we are to compare the experience techniques so as to choose the best one, we simply see the Average and variance results of each technique. One may rank these techniques by the value of average (descending) and by the value of variance (ascending) respectively as follows: for Average: ANN, Best DT, SVM and KNN. For variance: Best DT, ANN, KNN, and SVM. Because of KNN the worse Validity (lower average) and its undesirable reliability (3rd variance) it is not being chosen as the best technique is beyond doubt. The value of Best DT's variance indicates its best reliability among the other techniques. Nevertheless because of its lower value of average comparing to ANN, which shows less validity, it simply cannot be chosen as the best technique. Although Best DT and SVM had performed roughly the same, the Best DT's value of variance is much more desirable. All in all, in view of the highest accuracy of ANN and its acceptable variance value among the others, ANN can be chosen as the best techniques among the others.

As it was implicitly mentioned, Best DT was the second best techniques. Due to the fact that this paper computation is mostly based on MATLAB software and the best DT (WEKA-*Random Forest*) was one of the decision trees in WEKA software, from now on in this paper DT's results are actually the ones calculated by MATLAB. Fig. 4 is another behavior comparison of DT and ANN. In this plot, although DT shows to have a less variable demeanor (lower variance – better reliability), ANN obviously has outperformed DT. There are two other points about the plot in Fig. 4. First, due to the little growth in classifiers' accuracy, it seems unnecessary for both ANN and DT to have their train set proportion raised more than 0.7. Furthermore, the value of 0.6 for F-Score with the 0.05 proportion of train set seems to be a little bit unreasonable. Our further experiences showed that this peculiarity is actually because of the special nature of churn data. Not only for this paper data set but also for nearly all the other churn data the proportion of the churned customer is considerably lower than non-churned ones (in this paper case, churn proportion is 19%). This paper proposed a methodology that has actually taken the advantage of this speciality of churn problems to come up with the more accurate methodology to approach churn prediction.

Table 11
Features abbreviations.

Number of fail calls	CF	Frequency of use	FU
Complaints	Co	Frequency of SMS	FS
Subscription length	SL	Distinct call numbers	DCN
Charge amount	CA	Age group	AG
Second of use	SU	Type of service	TS
Status	St		

5.8. Feature extraction

The matter of finding out which features in a database are the most influential is an important part of any data mining task. There are different techniques and approaches for such matter: [52,53]. However since such statistical approach was already tried on this dataset, see [54], we introduced a new and more practical method of dimension reduction. This paper data set consisted of 11 features. These features and their abbreviations are shown in Table 11. In order to extract more important features we used a new methodology which proved to be effective for distinguishing more influential features in such processes. We ran DT classifier with 2048 set of different features, that is, every single possible set of features was experienced. For example CF (Call Failure) features was used for classification process once just by itself, and 10 other times along with the other 10 features and other times along with every possible set of other features. In order to eliminate the random error we run the algorithm 10 times for each 2048 sets of features and then we calculated the average value of 10 F-Scores' value for further analysis. Note that for the feature extraction process, we opted to use DT because it had shown to be the swiftest classifier. Although we ourselves in this paper have shown that the kind of classifier we use can make a difference in the result, here we do not concern about how accurate the classifier is. Here, in feature extraction, the only thing that matters is how much a well-chosen set of features can lead to improvements. Table 12 shows the order of features ranked by their influences after applying the explained methodol-

Table 10
ANOVA complete block design result for Table 8.

Source of variation	*SS*	*df*	*MS*	*F*	*p*-Value	*F*
Treatments	0.03613	3	0.012043	36.96846	2.43E−06	3.490295
Blocks	0.000313	4	7.84E−05			
Error	0.003909	12	0.000326			
Total	0.033585	19				

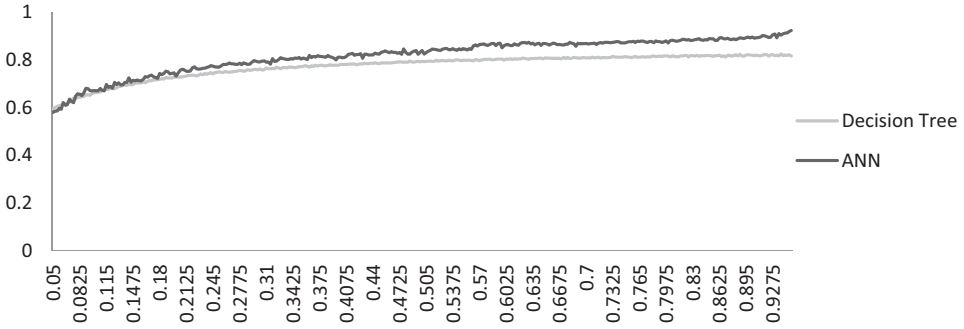

Fig. 4. Plot of F-score versus train set proportion (DT-MATLAB and ANN).

Table 12
Feature extraction – sets of one feature.

	Fe	Number	Sum	Average
1	FU	1024	725.8586	0.708846
2	Co	1024	722.621	0.705685
3	SU	1024	719.3543	0.702494
4	St	1024	707.559	0.690976
5	SL	1024	702.5501	0.686084
6	DCN	1024	699.0004	0.682618
7	AG	1024	693.7532	0.677493
8	CF	1024	690.6193	0.674433
9	FS	1024	688.2293	0.672099
10	CA	1024	682.9907	0.666983
11	TS	1024	671.0787	0.65535

Table 13
Feature extraction – top ten with maximum average among set of 1–6 features (total number of features set is 1485).

	Fe 1	Fe 2	Fe 3	Fe 4	Fe 5	Fe 6	Number	Sum	Average
1	Co	SL	SU	DCN	AG	St	32	25.99895	0.812467
2	Co	SL	SU	FU	AG	St	32	25.98944	0.81217
3	CF	Co	SL	SU	AG	St	32	25.85625	0.808008
4	CF	Co	SL	FU	AG	St	32	25.84574	0.807679
5	Co	SL	FU	DCN	AG	St	32	25.83967	0.80749
6	Co	SL	SU	FU	DCN	AG	32	25.76135	0.805042
7	CF	Co	SL	SU	FU	AG	32	25.72194	0.803811
8	Co	SL	SU	FU	DCN	St	32	25.69614	0.803004
9	CF	Co	SL	SU	FU	St	32	25.68352	0.80261
10	CF	Co	SL	SU	DCN	St	32	25.64151	0.801297

ogy. On the contrary of what we had anticipated Frequency of use, number of complains and second of use stood as the most influential features. On the other hand charge amount and total number of Call Failure were recognized as the least influential of all. In another study, [54], on the same data set, the hypothesis that subscription length, distinct call numbers, and type of service are mediators feature in the proposed model was rejected. Except for type of service which we have also recognized it as the least unimportant of all, the other two did not stand too unimportant a feature in our experience anyway.

In order to extract more interesting insights we calculated the average influences for set of 2–6 features. Table 13 represents the top ten most influential set of features. All of the top ten sets of features were from set of 6 features which indicates that using another feature might become influential. However, to check the results of most influential set of 2–6 features see Appendix B.

6. Proposed methodology

So far we have discussed and experienced 4 prominent data mining classification techniques and all of them showed different demeanors. It was mentioned earlier that the proportion of churn data is usually too small for a train-set system. As it was seen earlier in the plot presented in Fig. 4 DT or ANN performances did not drop dramatically by decreasing the train-set proportion of data. Therefore, here in the proposed methodology we have come up with the idea of using all of 4 experienced classifiers to make a better and more accurate hybrid classifier. One can see the general idea behind our proposed methodology by contemplating Fig. 5.

As it is shown in Fig. 5, the 4 classifiers, which we call interior classifier, are used to build the new one. Additionally, although in this figure the train set for the new classifier has been divided into 4 equal pieces, the proposed methodology itself justifies the proportion of these pieces automatically. Thus, the overall strategy behind our proposed methodology is clear. Apart from justifying the train-set division, and since there are going to be four different opinions about each of test-set records (4 interior classifiers), we needed to use a technique which could make the new classifier able to decide about its final decision by consolatory from the four interiors. We opt for a simple score-base technique. The technique is simple, after all four interior classifiers have drawn their decision about a customer (churn 0, not churn 1), each interior classifier score will be multiplied with their decisions and finally if the summation of these multiplications is bigger than the half of the overall scores, which had been assigned for each interior classifier, the classifier decision will be churn. So as to assign a score to every interior classification we used our experiences. Table 14 shows the procedure for score assignment. In this table except for average, variance and scores columns there are two other important columns named validity and reliability. The value of validity and reliability columns is calculated respectively to the average and variance columns. Unlike reliability and variance, for validity and average, the higher value of average leads to the higher value of validity. Finally, the scores column is calculated using both validity and reliability columns. The calculation performed for the results presented in Table 14 is done using Eqs. (11)–(13). However, in Eq. (13) a Ceiling Function with the significance of 0.01 has been used and that means the outcome of the fraction will be rounded up.

$$Va_i = \frac{A_i}{\sum_i Ai} \tag{11}$$

$$Re_i = \frac{V_i}{\sum_i Vi} \tag{12}$$

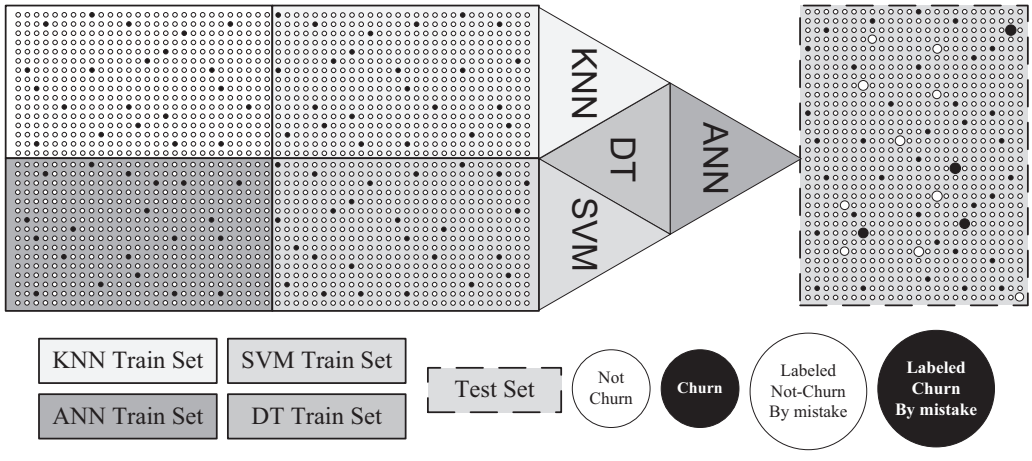

Fig. 5. Proposed methodology.

Table 14
Scores assignment.

	Count	Sum	Average (A)	Variance (V)	Validity (Va)	Reliability (Re)	Scores (S_i)
DT	5	4.050072	0.810014	0.000268	0.249249	0.259929	2.5
ANN	5	4.311	0.8622	0.000177	0.265307	0.284853	2.8
KNN	5	3.745	0.749	0.000381	0.230475	0.228978	2.3
SVM	5	4.143	0.8286	0.000391	0.254968	0.226239	2.4
Sum					1	1	10

Table 15
Proposed methodology outcome with ϕ variation.

ϕ	0.5			0.7			0.9			1		
	RE	PR	FS	RE	PR	FS	RE	PR	FS	RE	PR	FS
Block 1	0.317	1	0.482	0.662	0.837	0.739	0.709	0.847	0.772	0.797	0.837	0.817
Block 2	0.290	1	0.450	0.635	0.969	0.767	0.689	0.918	0.787	0.790	0.921	0.851
Block 3	0.378	1	0.549	0.736	0.893	0.807	0.770	0.877	0.82	0.898	0.887	0.893
Block 4	0.385	1	0.556	0.662	0.961	0.784	0.757	0.941	0.839	0.817	0.864	0.84
Block 5	0.263	0.928	0.41	0.641	0.896	0.748	0.655	0.851	0.740	0.790	0.847	0.818
ϕ	1.25			1.5			1.8			2.2		
	RE	PR	FS	RE	PR	FS	RE	PR	FS	RE	PR	FS
Block 1	0.824	0.797	0.811	0.824	0.797	0.811	0.844	0.801	0.822	0.966	0.821	0.888
Block 2	0.878	0.769	0.820	0.878	0.769	0.820	0.891	0.771	0.827	0.966	0.773	0.859
Block 3	0.912	0.808	0.857	0.912	0.808	0.857	0.932	0.811	0.868	0.98	0.784	0.871
Block 4	0.858	0.765	0.809	0.858	0.765	0.808	0.878	0.769	0.82	0.966	0.748	0.844
Block 5	0.905	0.779	0.837	0.905	0.779	0.837	0.926	0.783	0.848	0.979	0.784	0.871

$$S_i = Ceiling\left(\frac{Re_i + Va_i}{2}, 0.01\right) \tag{13}$$

$$D_{PM} = \begin{cases} \left(\sum_i S_i \times D_i\right) \times \phi \geq \dfrac{\sum_i S_i}{2} & 1 \\ \left(\sum_i S_i \times D_i\right) \times \phi < \dfrac{\sum_i S_i}{2} & 0 \end{cases} \tag{14}$$

Eq. (14) is the main equation of our proposed methodology. Except for the notation that was introduced in Table 14, there are some notations which needs clarification. First, i is the index for different methods and can have up to four different values: DT, ANN, KNN and SVM. Second, D_i is a binary variable and is the decision based on the technique i we would have made. Needless to say that D_{PM} is the decision that the proposed methodology will result in. Last but not least is ϕ which is the methodology behavior variable. It can alter between 1/2 and $\sum S_i/(2 \times Min(S_i))$. As one can see in the equation, it will be multiplied by the summation and will justify the results before the comparison with $\sum S_i/2$. If we choose 1/2 for ϕ, the classifier will decide in favor of churn only if all of the interior classifiers have voted for churn. On the other hand, if we choose $\sum S_i/2 \times Min(S_i)$ for ϕ, the classifier will distinguish the customer as churn only if at least one of the interior classifiers has known the case as churn. By manipulation of this variable one can decide how the algorithm approach would be. As explained, it can be strict and decide on being churn only after all of the four interiors have decided on being churn, or it can be less strict by deciding on churn if only one of them has decided on being churn. Both of these extreme ends can be used for different situations in real world. For instance, the former is usable for the cases that our retention scheme is expensive and we would only want to spend that kind of money on valuable customers who are definitely going to churn.

Table 15 shows the results for the proposed methodology. There are some noteworthy points about this table. First, as it was previously implied the Precision and recall Values showed to be considerably adjustable by ϕ alteration. By the value of 0.5 for ϕ all of the blocks, except for block 5, precision values are 1 which means every single customer who has been labeled churn by the classifier was actually churn. Moreover, by the value of 2.2 for ϕ, all of the blocks recall values are bigger than 0.96, which is a reasonably high accuracy, and the F-Scores' value is by far the best comparing to the other values assigned to ϕ. In order to compare the best interior classifier, ANN, and the best F-Score values in Table 15, in which $\phi = 2.2$, we performed a t-test. The p-value of 0.509 for this test showed that there is not a considerable difference between the performance of ANN and this proposed methodology. Nevertheless, as it was shown, the proposed methodology precision and recall values can be adjusted to be near perfect.

7. Conclusion and future research

In this paper we experienced four prominent classification techniques using an Iranian telecommunication company dataset. Artificial Neural Network (ANN) significantly outperformed the other three, namely K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM). Moreover, we have proposed a hybrid methodology in whose system all of the four aforementioned techniques are employed. It was shown that using the proposed technology a telecommunication company can gain a considerably higher than 95% accuracy for both Precision and Recall measures. Tables 16 and 17 enunciate the superiority of the hybrid methodology over the other four classifiers. Although it was seen that gaining a remarkably good Precision means a low Recall measure and vice versa, the proposed methodology has the capability

Table 16
Comparing recall value of hybrid methodology.

	1	2	3	4	5
Best DT	0.787	0.834	0.814	0.828	0.809
ANN	0.908	0.891	0.869	0.874	0.869
KNN	0.706	0.725	0.727	0.748	0.743
SVM	0.923	0.932	0.912	0.931	0.939
Best hybrid methodology	0.966	0.966	0.98	0.966	0.979

Table 17
Comparing precision value of hybrid methodology.

	1	2	3	4	5
Best DT	0.903	0.901	0.936	0.914	0.895
ANN	0.857	0.852	0.845	0.823	0.838
KNN	0.741	0.76	0.762	0.780	0.809
SVM	0.715	0.761	0.813	0.734	0.731
Best hybrid methodology	1	1	1	1	0.928

which makes the user able to decide whether better Precision, or better Recall, is desired. Apart from that, the proposed methodology performance, if it is measured by *F*-Score, was equivalent to ANN that was the best interior classifier.

Additionally, we introduced a new dimensionality reduction methodology so as to extract the most influential set of features. Frequency of use, total number of complaints, and seconds of use were shown to be the influential ones. Also on the contrary of the previous study on the same data set, we showed that frequency of SMS, charge amount and type of service were the least influential ones. Finally it can be concluded that the high value of Recall and Precision obtained by the hybrid methodology is due to the application of four different insights, four different classifiers. Not only this paper methodology for churn analysis can be used by telecommunication companies seeking for an accurate prediction, it can also be tested in other areas of business which deals with everyday customers. The analogy for this paper is the idea of four different experts (classifiers) expressing their opinions; so one researcher may approach this paper model by fuzzy logic for a decision-making procedure.

Appendix A.

See Tables 18–23.

Table 18
Complete block design ANOVA for the performance of different decision trees using *F*-Score measure.

Experiment	1	2	3	4	5
MATLAB	0.814	0.794	0.824	0.791	0.824
R (*Part*)	0.566	0.765	0.659	0.720	0.629
R (*rparty*)	0.765	0.749	0.749	0.758	0.741
WEKA (*AD.Tree*)	0.664	0.517	0.547	0.546	0.660
WEKA (*BF.Tree*)	0.823	0.792	0.812	0.865	0.769
WEKA (*FT*)	0.818	0.809	0.801	0.801	0.795
WEKA (*J48*)	0.815	0.808	0.794	0.844	0.816
WEKA (*LADTree*)	0.695	0.706	0.737	0.704	0.723
WEKA (*LMT*)	0.843	0.824	0.831	0.844	0.849
WEKA (*NBTree*)	0.798	0.789	0.779	0.732	0.785
WEKA (*Random Forest*)	0.841	0.863	0.870	0.868	0.848
WEKA (*Random Tree*)	0.769	0.82	0.819	0.805	0.793
WEKA (*REPTree*)	0.795	0.743	0.825	0.764	0.793
WEKA (*Simple Cart*)	0.85	0.805	0.854	0.848	0.815

Table 19
Completed block design ANOVA (summary) – decision trees – *F*-Score.

Summary	Count	Sum	Average	Variance
MATLAB	5	4.050072	0.810014	0.000268
R (*Part*)	5	3.339909	0.667982	0.006
R (*rparty*)	5	3.763001	0.7526	7.95E-05
WEKA (*AD.Tree*)	5	2.935319	0.587064	0.004888
WEKA (*BF.Tree*)	5	4.064216	0.812843	0.001297
WEKA (*FT*)	5	4.024756	0.804951	8.11E-05
WEKA (*J48*)	5	4.077724	0.815545	0.000339
WEKA (*LADTree*)	5	3.566033	0.713207	0.000275
WEKA (*LMT*)	5	4.19342	0.838684	0.000105
WEKA (*NBTree*)	5	3.885401	0.77708	0.00068
WEKA (*Random Forest*)	5	4.291454	0.858291	0.000164
WEKA (*Random Tree*)	5	4.007398	0.80148	0.000438
WEKA (*REPTree*)	5	3.920746	0.784149	0.000974
WEKA (*Simple Cart*)	5	4.17297	0.834594	0.000513

Table 20
Completed block design ANOVA – decision trees – *F*-Score.

Source of variation	*SS*	*df*	*MS*	*F*	*p*-Value	*F* crit
Decision Trees	0.358103	13	0.027546	22.44405	7.91E−17	1.913455
Blocks	0.000587	4	0.000147	0.11947	0.974968	
Error	0.063822	52	0.001227			
Total	0.422511	69				

Table 21
Complete block design ANOVA for the performance of different decision trees using Misclassification.

Experiment	1	2	3	4	5
MATLAB	55	63	51	58	52
R (*Part*)	90	69	78	71	86
R (*rparty*)	71	76	76	74	68
WEKA (*AD.Tree*)	92	101	110	95	93
WEKA (*BF.Tree*)	52	58	58	38	66
WEKA (*FT*)	52	59	61	55	60
WEKA (*J48*)	57	61	61	41	52
WEKA (*LADTree*)	90	84	81	90	76
WEKA (*LMT*)	46	52	53	43	44
WEKA (*NBTree*)	59	62	59	62	58
WEKA (*Random Forest*)	45	41	39	38	44
WEKA (*Random Tree*)	68	55	56	54	60
WEKA (*REPTree*)	60	73	55	62	59
WEKA (*Simple Cart*)	46	55	45	42	57

Table 22
Completed block design ANOVA (summary) – decision trees – misclassification.

Summary	Count	Sum	Average	Variance
MATLAB	5	279	55.8	23.7
R (*Part*)	5	394	78.8	83.7
R (*rparty*)	5	365	73	12
WEKA (*AD.Tree*)	5	491	98.2	55.7
WEKA (*BF.Tree*)	5	272	54.4	108.8
WEKA (*FT*)	5	287	57.4	14.3
WEKA (*J48*)	5	272	54.4	69.8
WEKA (*LADTree*)	5	421	84.2	36.2
WEKA (*LMT*)	5	238	47.6	21.3
WEKA (*NBTree*)	5	300	60	3.5
WEKA (*Random Forest*)	5	207	41.4	9.3
WEKA (*Random Tree*)	5	293	58.6	32.8
WEKA (*REPTree*)	5	309	61.8	45.7
WEKA (*Simple Cart*)	5	245	49	43.5

Table 23
Completed block design ANOVA – decision trees – misclassification.

Source of variation	*SS*	*df*	*MS*	*F*	*p*-Value	*F* crit
Decision trees	15,974.24	13	1228.788	32.66048	2.09E−20	1.913455
Blocks	284.8	4	71.2	1.892456	0.125681	
Error	1956.4	52	37.62308			
Total	18,215.44	69				

Appendix B.

See Tables 24–28.

Table 24
Feature extraction – sets of two features.

	Fe 1	Fe 2	Number	Sum	Average
1	Co	FU	512	380.9361	0.744016
2	Co	SU	512	377.0122	0.736352
3	FU	St	512	376.5365	0.735423
4	Co	St	512	375.0804	0.732579
5	SU	St	512	374.5311	0.731506
6	FU	AG	512	374.3681	0.731188
7	SL	FU	512	374.1209	0.730705
8	SU	FU	512	372.4886	0.727517
9	SL	SU	512	371.9856	0.726534
10	Co	SL	512	371.8025	0.726177
11	CF	FU	512	371.7391	0.726053
12	SU	AG	512	371.4803	0.725548
13	Co	AG	512	371.3476	0.725288
14	Co	DCN	512	370.167	0.722982
15	SL	St	512	369.1976	0.721089
16	FU	FS	512	368.6413	0.720003
17	CF	SU	512	368.5819	0.719886
18	CA	FU	512	368.3515	0.719436
19	FU	DCN	512	368.1364	0.719016
20	DCN	St	512	366.6653	0.716143
21	CF	Co	512	366.4885	0.715798
22	SU	DCN	512	366.2183	0.71527
23	Co	CA	512	365.2672	0.713413
24	Co	FS	512	364.5545	0.712021
25	CA	SU	512	364.0862	0.711106
26	SL	DCN	512	363.7678	0.710484
27	SL	AG	512	363.7069	0.710365
28	SU	FS	512	363.3403	0.709649
29	FU	TS	512	363.1088	0.709197
30	AG	St	512	362.2521	0.707524
31	CF	St	512	361.8828	0.706802
32	DCN	AG	512	361.582	0.706215
33	Co	TS	512	361.389	0.705838
34	CF	SL	512	360.2547	0.703622
35	FS	St	512	360.0698	0.703261
36	SU	TS	512	359.9179	0.702965
37	SL	FS	512	359.4622	0.702075
38	CF	AG	512	356.3965	0.696087
39	SL	CA	512	356.3791	0.696053
40	CF	DCN	512	356.2981	0.695895
41	FS	DCN	512	356.1689	0.695642
42	CA	St	512	355.2861	0.693918
43	CA	DCN	512	354.4789	0.692342
44	FS	AG	512	354.056	0.691516
45	TS	St	512	353.9038	0.691218
46	CA	AG	512	351.8454	0.687198
47	CF	CA	512	351.6518	0.68682
48	SL	TS	512	351.443	0.686412
49	CF	FS	512	351.2325	0.686001
50	CA	FS	512	349.9558	0.683507
51	DCN	TS	512	349.6551	0.68292
52	AG	TS	512	347.0557	0.677843
53	CF	TS	512	345.491	0.674787
54	FS	TS	512	344.1315	0.672132
55	CA	TS	512	341.6556	0.667296

Table 25
Feature extraction – sets of three features.

	Fe 1	Fe 2	Fe 3	Number	Sum	Average
1	CF	SU	FS	256	196.5769	0.767878
2	CF	CA	St	256	195.9482	0.765423
3	CF	CA	DCN	256	194.9184	0.7614
4	CF	Co	SU	256	194.7285	0.760658
5	CF	CA	FU	256	194.0358	0.757952
6	CF	SL	TS	256	193.7185	0.756713
7	CF	Co	CA	256	193.7131	0.756692
8	CA	FS	St	256	193.4997	0.755858
9	Co	SL	TS	256	193.2649	0.754941
10	Co	SL	FS	256	193.2363	0.754829
11	Co	CA	SU	256	192.9846	0.753846
12	SL	SU	FS	256	192.848	0.753312
13	CF	Co	TS	256	192.8413	0.753286
14	CF	FU	AG	256	192.6489	0.752535
15	CF	CA	TS	256	192.4778	0.751866
16	Co	SL	SU	256	192.4297	0.751678
17	SL	TS	St	256	192.4017	0.751569
18	FU	TS	St	256	192.3593	0.751403
19	CF	SL	CA	256	192.2544	0.750994
20	CF	CA	AG	256	191.8955	0.749592
21	CF	AG	St	256	191.8831	0.749543
22	SU	FU	DCN	256	191.5953	0.748419
23	CF	Co	DCN	256	191.4551	0.747872
24	CA	FS	TS	256	191.4283	0.747767
25	SL	CA	St	256	191.3618	0.747507
26	SL	FU	St	256	191.2842	0.747204
27	CF	CA	SU	256	191.2801	0.747188
28	FU	AG	TS	256	191.1021	0.746493
29	FU	DCN	AG	256	190.8747	0.745604
30	CF	Co	FS	256	190.7395	0.745076
31	SU	FU	FS	256	190.6273	0.744638
32	CF	SU	FU	256	190.5706	0.744417
33	CF	FU	St	256	190.5579	0.744367
34	CF	FU	FS	256	190.4609	0.743988
35	SL	AG	TS	256	190.453	0.743957
36	SL	FU	AG	256	190.4329	0.743879
37	Co	SU	FU	256	190.2399	0.743125
38	FU	FS	TS	256	190.0381	0.742336
39	CF	Co	St	256	190.0075	0.742217
40	SL	DCN	AG	256	189.7931	0.741379
41	CA	AG	St	256	189.7769	0.741316
42	SL	DCN	TS	256	189.7699	0.741289
43	Co	SL	DCN	256	189.5338	0.740366
44	Co	SL	CA	256	189.3992	0.73984
45	SL	FU	FS	256	189.3064	0.739478
46	SL	FS	TS	256	189.2802	0.739376
47	Co	SL	AG	256	189.2656	0.739319
48	CF	SL	St	256	189.1723	0.738954
49	Co	FS	St	256	189.1093	0.738708
50	Co	SU	DCN	256	189.0987	0.738667
51	Co	FS	AG	256	189.0562	0.738501
52	SU	TS	St	256	188.9362	0.738032
53	CF	FS	TS	256	188.8225	0.737588
54	CF	SU	St	256	188.7614	0.737349
55	CF	CA	FS	256	188.5879	0.736671
56	CA	FS	DCN	256	188.5861	0.736665
57	Co	SU	TS	256	188.5238	0.736421
58	SU	FS	St	256	188.5159	0.73639
59	CA	SU	FU	256	188.3462	0.735727
60	CA	DCN	St	256	188.181	0.735082
61	SL	CA	TS	256	188.1162	0.734829
62	SL	SU	St	256	188.0858	0.73471
63	Co	CA	TS	256	188.0424	0.734541
64	CF	SL	AG	256	188.0144	0.734431
65	Co	FU	TS	256	187.9779	0.734289
66	SU	FS	AG	256	187.8929	0.733957
67	CF	Co	SL	256	187.7269	0.733308
68	CA	FU	AG	256	187.6651	0.733067
69	CF	Co	FU	256	187.6551	0.733028
70	CF	FS	AG	256	187.5578	0.732648
71	CF	FS	DCN	256	187.5517	0.732624
72	SL	CA	AG	256	187.4301	0.732149
73	CF	TS	St	256	187.4222	0.732118
74	SL	FS	DCN	256	187.3917	0.731999
75	SL	SU	AG	256	187.3234	0.731732
76	SL	AG	St	256	187.2789	0.731558
77	Co	FU	DCN	256	187.2567	0.731471
78	Co	SL	St	256	187.1747	0.731151
79	CF	SL	FS	256	187.1683	0.731126
80	FU	DCN	St	256	186.9563	0.730298
81	Co	CA	AG	256	186.8406	0.729846
82	CF	SU	AG	256	186.785	0.729629
83	CA	DCN	TS	256	186.7809	0.729613
84	CF	SL	FU	256	186.7688	0.729566
85	FU	DCN	TS	256	186.7158	0.729359
86	CF	SU	DCN	256	186.6797	0.729217

Table 25 (*Continued*)

	Fe 1	Fe 2	Fe 3	Number	Sum	Average
87	Co	FU	St	256	186.6692	0.729177
88	FU	FS	AG	256	186.4875	0.728467
89	SL	SU	FU	256	186.3471	0.727918
90	Co	FS	DCN	256	186.2766	0.727643
91	SU	FS	TS	256	186.2331	0.727473
92	SL	FS	AG	256	186.1726	0.727237
93	Co	SL	FU	256	186.0784	0.726869
94	DCN	AG	TS	256	186.0059	0.726586
95	SU	FU	St	256	185.9973	0.726552
96	CF	Co	AG	256	185.9501	0.726368
97	FU	AG	St	256	185.94	0.726328
98	SL	FU	TS	256	185.8636	0.72603
99	CF	FU	TS	256	185.7027	0.725401
100	FU	FS	DCN	256	185.4229	0.724308
101	CA	FS	AG	256	185.2848	0.723769
102	CA	SU	AG	256	185.256	0.723656
103	DCN	TS	St	256	185.1138	0.723101
104	CF	FU	DCN	256	185.1113	0.723091
105	Co	FU	FS	256	185.0102	0.722696
106	SU	FU	TS	256	184.9801	0.722578
107	Co	CA	FS	256	184.8921	0.722235
108	CF	AG	TS	256	184.8915	0.722232
109	Co	SU	AG	256	184.6761	0.721391
110	Co	CA	FU	256	184.5976	0.721084
111	FU	FS	St	256	184.3899	0.720273
112	SL	FS	St	256	184.3349	0.720058
113	SL	SU	DCN	256	184.2673	0.719794
114	Co	FS	TS	256	184.2463	0.719712
115	CA	DCN	AG	256	184.1893	0.719489
116	CF	SL	SU	256	184.1844	0.71947
117	SL	DCN	St	256	184.1417	0.719304
118	CF	DCN	AG	256	183.9777	0.718663
119	SL	CA	SU	256	183.886	0.718305
120	CF	DCN	TS	256	183.8728	0.718253
121	FS	AG	St	256	183.8704	0.718244
122	Co	SU	St	256	183.6188	0.717261
123	CA	SU	St	256	183.5885	0.717142
124	CA	SU	FS	256	183.5821	0.717118
125	CA	FU	TS	256	183.3996	0.716405
126	CA	AG	TS	256	183.2571	0.715848
127	SL	FU	DCN	256	183.2148	0.715683
128	FS	AG	TS	256	183.028	0.714953
129	CA	TS	St	256	182.7659	0.713929
130	CF	SL	DCN	256	182.692	0.713641
131	Co	AG	St	256	182.6434	0.713451
132	SU	FU	AG	256	182.6375	0.713428
133	CF	SU	TS	256	182.2732	0.712005
134	Co	FU	AG	256	182.1831	0.711653
135	CF	FS	St	256	182.1336	0.711459
136	Co	CA	St	256	181.9683	0.710813
137	Co	SU	FS	256	181.9304	0.710666
138	SL	CA	FS	256	181.7523	0.70997
139	SL	SU	TS	256	181.6999	0.709765
140	SU	AG	St	256	181.4493	0.708786
141	CA	FU	St	256	181.1911	0.707778
142	AG	TS	St	256	181.0155	0.707092
143	CA	FU	DCN	256	180.8782	0.706556
144	Co	AG	TS	256	180.769	0.706129
145	FS	DCN	TS	256	180.7535	0.706068
146	SU	DCN	St	256	180.6309	0.705589
147	SU	DCN	TS	256	180.5589	0.705308
148	Co	DCN	AG	256	180.4659	0.704945
149	FS	DCN	AG	256	180.3913	0.704653
150	SU	FS	DCN	256	180.1884	0.703861
151	CA	FU	FS	256	180.062	0.703367
152	Co	CA	DCN	256	179.7506	0.702151
153	Co	DCN	TS	256	179.2518	0.700202
154	SU	DCN	AG	256	178.2996	0.696483
155	DCN	AG	St	256	178.2952	0.696466
156	CF	DCN	St	256	178.2535	0.696303
157	FS	TS	St	256	178.2299	0.696211
158	CA	SU	DCN	256	178.0927	0.695675
159	SL	CA	DCN	256	177.7278	0.694249
160	Co	TS	St	256	177.3118	0.692624
161	CA	SU	TS	256	177.0285	0.691518
162	SL	CA	FU	256	176.0162	0.687563
163	SU	AG	TS	256	175.8944	0.687087

Table 25 (*Continued*)

	Fe 1	Fe 2	Fe 3	Number	Sum	Average
164	FS	DCN	St	256	175.639	0.68609
165	Co	DCN	St	256	174.988	0.683547

Table 26
Feature extraction – top ten with maximum average influence sets of four features.

	Fe 1	Fe 2	Fe 3	Fe 4	Number	Sum	Average
1	Co	SL	FU	AG	128	100.274	0.78339
2	Co	SL	FU	St	128	100.1265	0.782238
3	Co	SL	SU	St	128	99.97759	0.781075
4	Co	FU	AG	St	128	99.91599	0.780594
5	Co	SL	SU	AG	128	99.85633	0.780128
6	Co	SU	FU	St	128	99.81065	0.779771
7	CF	Co	FU	St	128	99.58711	0.778024
8	Co	SU	FU	AG	128	99.43807	0.77686
9	CF	Co	FU	AG	128	99.27647	0.775597
10	Co	FU	DCN	St	128	99.25348	0.775418

Table 27
Feature extraction – top ten with maximum average influence sets of five features.

	Fe 1	Fe 2	Fe 3	Fe 4	Fe 5	Number	Sum	Average
1	Co	SL	FU	AG	St	64	51.12456	0.798821
2	Co	SL	SU	AG	St	64	51.08254	0.798165
3	Co	SL	SU	FU	AG	64	51.01878	0.797168
4	Co	SL	SU	FU	St	64	50.95235	0.796131
5	Co	SL	SU	DCN	St	64	50.90555	0.795399
6	CF	Co	SL	FU	AG	64	50.79088	0.793608
7	CF	Co	SL	FU	St	64	50.74832	0.792943
8	Co	SL	SU	DCN	AG	64	50.70897	0.792328
9	CF	Co	SL	SU	AG	64	50.70111	0.792205
10	Co	SL	FU	DCN	AG	64	50.69537	0.792115

Table 28
Feature extraction – top ten with maximum average influence sets of six features.

	Fe 1	Fe 2	Fe 3	Fe 4	Fe 5	Fe 6	Number	Sum	Average
1	Co	SL	SU	DCN	AG	St	32	25.99895	0.812467
2	Co	SL	SU	FU	AG	St	32	25.98944	0.81217
3	CF	Co	SL	SU	AG	St	32	25.85625	0.808008
4	CF	Co	SL	FU	AG	St	32	25.84574	0.807679
5	Co	SL	FU	DCN	AG	St	32	25.83967	0.80749
6	Co	SL	SU	FU	DCN	AG	32	25.76135	0.805042
7	CF	Co	SL	SU	FU	AG	32	25.72194	0.803811
8	Co	SL	SU	FU	DCN	St	32	25.69614	0.803004
9	CF	Co	SL	SU	FU	St	32	25.68352	0.80261
10	CF	Co	SL	SU	DCN	St	32	25.64151	0.801297

References

[1] X. Zhang, J. Zhu, S. Xu, Y. Wan, Predicting customer churn through interpersonal influence, Knowl.-Based Syst. 28 (2012) 97–104.
[2] S.C. Lin, C.H. Tung, N.Y. Jan, D.A. Chiang, Evaluating churn model in CRM: a case study in Telecom, J. Converg. Inf. Technol. 6 (2011) 192–200.
[3] B. Huang, M.T. Kechadi, B. Buckley, Customer churn prediction in telecommunications, Expert Syst. Appl. 39 (2012) 1414–1425.
[4] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (2012) 211–229.
[5] H.M. Kuo, C.W. Chen, A novel viewpoint on information and interface design for auction web sites, Hum. Factors Ergon. Manuf. Serv. Ind. 22 (2012) 287–295.
[6] A. Keramati, R. Jafari-Marandi, Webpage clustering – taking the zero step: a case study of an Iranian website, J. Web Eng. 13 (2014).
[7] Y. Sekercioğlu, A. Pitsillides, A. Vasilakos, Computational intelligence in management of ATM networks, Soft Comput. 5 (2001) 257–263.
[8] A. Azadeh, S. Asadzadeh, R. Jafari-Marandi, S. Nazari-Shirkouhi, G. Baharian Khoshkhou, S. Talebi, A. Naghavi, Optimum estimation of missing values in randomized complete block design by genetic algorithm, Knowl.-Based Syst. 37 (2013) 37–47.
[9] C.-W. Chen, Stability analysis and robustness design of nonlinear systems: an NN-based approach, Appl. Soft Comput. 11 (2011) 2735–2742.
[10] C.-W. Chen, Application of fuzzy-model-based control to nonlinear structural systems with time delay: An LMI method, J. Vibr. Control 16 (11) (2010) 1651–1672.

[11] K. Subramanian, S. Suresh, A meta-cognitive sequential learning algorithm for neuro-fuzzy inference system, Appl. Soft Comput. 12 (2012) 3603–3614.
[12] A. Vasilakos, C. Ricudis, K. Anagnostakis, W. Pedryca, A. Pitsillides, Evolutionary-fuzzy prediction for strategic QoS routing in broadband networks, in: IEEE World Congress on Computational Intelligence, The 1998 IEEE International Conference on Fuzzy Systems Proceedings, IEEE, 1998, pp. 1488–1493.
[13] J.-W. Lin, C.-W. Chen, C.-Y. Peng, Potential hazard analysis and risk assessment of debris flow by fuzzy modeling, Nat. Hazards 64 (2012) 273–282.
[14] M.-L. Lin, C.-W. Chen, Application of fuzzy models for the monitoring of ecologically sensitive ecosystems in a dynamic semi-arid landscape from satellite imagery, Eng. Comput. 27 (2010) 5–19.
[15] P. Chen, C.-W. Chen, W. Chiang, D. Lo, GA-based decoupled adaptive FSMC for nonlinear systems by a singular perturbation scheme, Neural Comput. Appl. 20 (2011) 517–526.
[16] C.-W. Chen, P.-C. Chen, GA-based Adaptive Neural Network Controllers for Nonlinear Systems, 2010.
[17] M. Rabbani, M. Baghersad, R. Jafari, A new hybrid GA-PSO method for solving multi-period inventory routing problem with considering financial decisions, J. Ind. Eng. Manage. 6 (2013) 909–929.
[18] R. Tavakkoli-Moghaddam, R. Jafari-Marandi, A novel multi-objective genetic algorithm for cell formation problems, in: 9th International Industrial Engineering Conference, 2013.
[19] C.-W. Chen, K.F.-R. Liu, M.-L. Lin, C.-P. Tseng, A new viewpoint of hazard assessment and management for Taiwan's insurance issues, Nat. Hazards 65 (2013) 303–314.
[20] W.-K. Hsu, W.-L. Chiang, Q. Xue, D.-M. Hung, P.-C. Huang, C.-W. Chen, C.-H. Tsai, A probabilistic approach for earthquake risk assessment based on an engineering insurance portfolio, Nat. Hazards 65 (2013) 1559–1571.
[21] K.F.-R. Liu, H.-H. Liang, C.-W. Chen, J.-S. Chen, Y.-S. Shen, Combining scientific facts and significance criteria to predict the result of an environmental impact assessment review, J. Environ. Inform. 19 (2012) 93–107.
[22] C.-H. Tsai, C.-W. Chen, An earthquake disaster management mechanism based on risk assessment information for the tourism industry – a case study from the island of Taiwan, Tour. Manage. 31 (2010) 470–481.
[23] K. Kim, C.-H. Jun, J. Lee, Improved churn prediction in telecommunication industry by analyzing a large network, Expert Syst. Appl. (2014).
[24] W. Verbeke, D. Martens, B. Baesens, Social network analysis for customer churn prediction, Appl. Soft Comput. 14 (2014) 431–446.
[25] S.Y. Hung, D.C. Yen, H.Y. Wang, Applying data mining to telecom churn management, Expert Syst. Appl. 31 (2006) 515–524.
[26] M. Kantardzic, A. Kumar, Toward Autonomic Distributed Data Mining with Intelligent Web Services, 2003, pp. 544–552.
[27] C.F. Tsai, M.Y. Chen, Variable selection by association rules for customer churn prediction of multimedia on demand, Expert Syst. Appl. 37 (2010) 2006–2015.
[28] H. Lee, Y. Lee, H. Cho, K. Im, Y.S. Kim, Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model, Decis. Support Syst. 52 (2011) 207–216.
[29] L. Ning, L. Hua, L. Jie, Z. Guangquan, A customer churn prediction model in telecom industry using boosting, IEEE Trans. Ind. Inform. 10 (2014) 1659–1665.
[30] K.W. De Bock, D.V.D. Poel, An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction, Expert Syst. Appl. 38 (2011) 12293–12301.
[31] A. Idris, A. Khan, Y.S. Lee, Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification, Appl. Intell. 39 (2013) 659–672.
[32] T. Verbraken, W. Verbeke, B. Baesens, A novel profit maximizing metric for measuring classification performance of customer churn prediction models, IEEE Trans. Knowl. Data Eng. 25 (2013) 961–973.
[33] P.C. Pendharkar, Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services, Expert Syst. Appl. 36 (2009) 6714–6720.
[34] H.S. Kim, C.H. Yoon, Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market, Telecommun. Policy 28 (2004) 751–765.
[35] M. Owczarczuk, Churn models for prepaid customers in the cellular telecommunication industry using large data marts, Expert Syst. Appl. 37 (2010) 4710–4712.
[36] P. Kisioglu, Y.I. Topcu, Applying Bayesian Belief Network approach to customer churn analysis: a case study on the telecom industry of Turkey, Expert Syst. Appl. 38 (2011) 7151–7157.
[37] K. Coussement, D. Van den Poel, Integrating the voice of customers through call center emails into a decision support system for churn prediction, Inf. Manage. 45 (2008) 164–174.
[38] B. Usero Sánchez, G. Asimakopoulos, Regulation and competition in the European mobile communications industry: an examination of the implementation of mobile number portability, Telecommun. Policy 36 (2012) 187–196.
[39] J. Sweeney, J. Swait, The effects of brand credibility on customer loyalty, J. Retail. Consum. Serv. 15 (2008) 179–193.
[40] Y. Huang, T. Kechadi, An effective hybrid learning system for telecommunication churn prediction, Expert Syst. Appl. 40 (2013) 5635–5647.
[41] R.K. Gopal, S.K. Meher, Customer Churn Time Prediction in Mobile Telecommunication Industry Using Ordinal Regression, 2008, pp. 884–889.
[42] J.H. Ahn, S.P. Han, Y.S. Lee, Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry, Telecommun. Policy 30 (2006) 552–568.
[43] X. Liu, W. Pedrycz, The development of fuzzy decision trees in the framework of axiomatic fuzzy set logic, Appl. Soft Comput. J. 7 (2007) 325–342.
[44] E.B. Hunt, J. Marin, P.J. Stone, Experiments in Induction, 1966.
[45] F. Gorunescu, Data Mining: Concepts, Models, and Techniques, Springer, India, 2011.
[46] L. Özbakir, A. Baykasoğlu, S. Kulluk, A soft computing-based approach for integrated training and rule extraction from artificial neural networks: DIFACONN-miner, Appl. Soft Comput. J. 10 (2010) 304–317.
[47] S. Polak, B. Wiśniowska, M. Ahamadi, A. Mendyk, Prediction of the hERG potassium channel inhibition potential with use of artificial neural networks, Appl. Soft Comput. J. 11 (2011) 2611–2617.
[48] O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, USA, 2012.
[49] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[50] S. Kim, K.S. Shin, K. Park, An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case, 2005, pp. 636–647.
[51] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.
[52] Y. Jia, F. Nie, C. Zhang, Trace ratio problem revisited, IEEE Trans. Neural Netw. 20 (2009) 729–735.
[53] S. Xiang, F. Nie, G. Meng, C. Pan, C. Zhang, Discriminative least squares regression for multiclass classification and feature selection, IEEE Trans. Neural Netw. Learn. Syst. 23 (2012) 1738–1754.
[54] A. Keramati, S.M.S. Ardabili, Churn analysis for an Iranian mobile operator, Telecommun. Policy 35 (2011) 344–356.