Expert Systems with Applications 40 (2013) 5635–5647

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

An effective hybrid learning system for telecommunication churn prediction

Ying Huang [*], Tahar Kechadi

School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland

ARTICLE INFO

Keywords:
Classification
Telecommunication
Churn prediction
Hybrid model
Clustering
Rule induction

ABSTRACT

Customer churn has emerged as a critical issue for Customer Relationship Management and customer retention in the telecommunications industry, thus churn prediction is necessary and valuable to retain the customers and reduce the losses. Moreover, high predictive accuracy and good interpretability of the results are two key measures of a classification model. More studies have shown that single model-based classification methods may not be good enough to achieve a satisfactory result. To obtain more accurate predictive results, we present a novel hybrid model-based learning system, which integrates the supervised and unsupervised techniques for predicting customer behaviour. The system combines a modified k-means clustering algorithm and a classic rule inductive technique (FOIL).

Three sets of experiments were carried out on telecom datasets. One set of the experiments is for verifying that the weighted k-means clustering can lead to a better data partitioning results; the second set of experiments is for evaluating the classification results, and comparing it to other well-known modelling techniques; the last set of experiment compares the proposed hybrid-model system with several other recently proposed hybrid classification approaches. We also performed a comparative study on a set of benchmarks obtained from the UCI repository. All the results show that the hybrid model-based learning system is very promising and outperform the existing models.

1. Introduction

With recent evolution in the Information and Communication Technology (ICT) sector, numerous new and attractive services have been introduced, and they put huge pressure on traditional services. Customer churn has emerged as one of the major issues in Customer Relationship Management (CRM) in telecommunication services around the world, for both wireless providers and long-distance carriers. For instance, in the U.S., telecom providers of long-distance and international services have been bearing the churn rates from 45% to 70% percent for some years (Mattison, 2001). Under the fierce competitive environment, it becomes very important for the telecom operators to retain their existing customers as acquiring new customers is much more expensive. Consequently, predicting which customers are likely to stop their subscription and switch to competitors (churn) is critical. Predicting the potential churners and successfully retain them, especially the valuable ones, can substantially increase the profitability of a company.

In the telecommunications industry, operators usually capture the transactional data, which reflects the service usage, and some static data such as subscriber's personal information and contract details. Data mining (DM) methods have emerged as a good alternative to study the customer behaviour. We can find various DM techniques, such as decision tree, logistic regression, support vector machine, artificial neural networks, inductive rule learning, etc. They have been applied to predict customer behaviour (Huang, Huang, & Kechadi, 2011; Hwang, Jung, & Suh, 2004; Larivire & Poel, 2005; Wei & Chiu, 2002; Xia & dong Jin, 2008). Most of the existing predictive modelling techniques, applied to customer churn, are based on supervised learning; very few of them have been based on unsupervised learning. In addition, most of the classifiers use single model (i.e., only one data mining technique). Many of the single model-based classifiers can predict potential churners to a large extent. However, either the accuracy is not good enough for some of the techniques or there is a room for improving the prediction accuracy for some others, and a hybrid model is a good alternative for better classification performance. Moreover, usually the entire training data instances are all used to build prediction models. However, it may be more effective to predict a new data instance based on partial training instances that are more similar to the test data than other training instances.

The advantages of the proposed model over the other commonly used modelling techniques in the domain of churn prediction concern the following aspects: Firstly, the prediction model

[*] Corresponding author. Tel.: +353 873145218.
E-mail address: ying.huang.1@ucdconnect.ie (Y. Huang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.04.020

of the proposed system is a hybrid-model, which fully integrates supervised and unsupervised learning by employing a clustering method for dividing the training data and a rule induction method for generalising classification rules for each cluster. Secondly, the process for predicting a test data instance relies on the most appropriate sub-classifier, which is produced from the most similar cluster to the test instances rather than the entire training set. We argue that using only a sub-classifier may improve the prediction accuracy. Thirdly, the commonly used k-means algorithm is used for clustering the original training data; however, to enhance the clustering result, we apply a weighting technique, which makes full use of the causality between attributes and the target. The experimental results show that modified k-means enhances not only the clustering results but also the prediction results.

The hybrid-model based prediction system has four main phases, which are as follows:

(1) Data Discretisation: We apply a class-dependent discretization method on all the continuous attributes. The continuous data is transformed to the form of intervals, which is conducive to the rule induction.
(2) Weighted Clustering: We apply a weighted k-means clustering algorithm to divide the training data into a number of clusters. We use Path Analysis to calculate the weights of attributes.
(3) Rule Extraction: We apply a rule learning method (i.e., FOIL) to each cluster, extracted in the previous phase, to induce a set of classification rules, which constitutes a sub-classifier. Thus, each cluster corresponds to a sub-classifier.
(4) Prediction: We predict each test instance by choosing the most suitable sub-classifier, which corresponds to the *closest* cluster to the test data, according to a distance or similarity measure.

The remainder of the paper is organised as follows: Section 2 reviews some related work in the area. Section 3 describes the details of the proposed hybrid learning system. The experimental set-up and results are discussed in Section 4. We conclude and outline some future work in Section 5.

2. Related Work

Large number of machine learning and knowledge discovery techniques have been proposed and applied to the problem of customer retention in the domain of CRM. The techniques can be used in different phases of the data mining process, such as eliminating the noise and outliers, reducing the feature space by selecting most relevant attributes, predicting churn behaviour, etc.

In this section, we briefly introduce ways of building a classification model by reviewing several customer churn prediction models proposed in the literature. Hung, Yen, and Wang (2006) developed two classification methods by using three models, one is based on k-means clustering and decision tree (C5.0) and the other combines the back-propagation neural network (BPN) and decision tree. The models are evaluated using LIFT and hit ratio measures. Huang, Kechadi, and Buckley (2012) proposed a mining process that consists of feature extraction and classification. The authors implemented seven traditional classification modelling techniques (e.g., Decision Tree, Linear Regression, and Multilayer Perceptron Neural Networks, etc.) to build different predictive models, and for some models, they use different data processing methods. They have evaluated the performance of their approach using true positive and false positive measures.

Apart from the techniques described above, in fact, there is a large amount of literature reporting on the application of data mining techniques to study the behaviour of customers in telecoms, (Au, Chan, & Yao, 2003; Coussement & Den Poel, 2008; Huang et al., 2011; Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000). Ngai, Xiu, and Chau (2009) reviewed more than 80 papers about the application of data mining to Customer Relationship Management, and a lot of them concern the domain of customer churn prediction. Nonetheless, most of the proposed modelling techniques use a single model.

Recently, many researchers have started to study hybrid models to improve the classification effectiveness. Usually, hybrid models combine two or more techniques. For instance, the classification or clustering techniques can be sequentially combined (e.g., Khashei, Hamadani, & Bijari (2012) proposed a hybrid classification model by integrating artificial neural networks and multiple linear regression to yield more general and accurate classification result than single traditional method). Tsai and Lu (2009) suggested that the hybrid techniques can provide better performance than many single model-based approaches in numerous different domains. One can refer to Lee and Lee (2006) for some common types and structures of hybrid model-based classification methods. The authors have built a hybrid model called *SePI* (**S**egmentation by **P**erformance **I**nformation). The framework of *SePI* is based on three models, main model, discrimination model, and support model. Decision tree (C5.0), which found to be one of the best single model for a given data, is chosen as the main model; the discrimination model uses the performance information of the main model on the training dataset; the support model uses the data for which the main model predicted incorrectly, and ANN is employed as the support model. The key idea of *SePI* is that if the test examples cannot be predicted correctly by the main model, then they will be predicted by the support model.

For the customer churn prediction problem, most of the related work focuses on using only one data mining method. Table 1 shows the related implementations that employ hybrid modelling techniques for customer churn prediction. Most of the hybrid model-based applications normally follow a common pattern, which consists of two stages: the first stage deals with pre-processing the input data (e.g., reduce the dataset, detect the outliers, etc.), the second stage is dedicated to the mining of the preprocessed data to extract useful patterns. In our study, we also build a hybrid learning system by integrating two stages. However, the difference is that the first stage is no longer considered as the step for data pre-processing. We re-design the system by firstly segmenting the training data into different groups; we build a set of sub-classifiers by applying rule induction method on each group of the data; the second stage predicts the categories of the test examples, each example (e.g., *test_i*) is predicted by a sub-classifier that was produced by the closest sub-cluster to *test_i*.

3. Methodology

Our main motivation is the design of a hybrid learning system for predicting the customer future behaviour. The main idea behind our hybrid learning system is to predict a customer instance according to the training examples that are more similar to it. We assume that customers having similar behaviour patterns (characteristics) are more likely to behave the same in the future. Thus, it might be more accurate if an unlabelled instance is predicted using partial training instance, which has similar characteristics with the tested instance, rather than the whole data. This can be achieved by dividing the training data into clusters, and the test instance is assigned to the closest cluster to it.

In this paper, the main concern is the effectiveness of classification, thus, we do not discuss some data pre-process steps, such as data cleaning, normalisation, and feature selection, as they have

Table 1
Related work that applied Hybrid model for customer churn prediction. (DT: Decision Tree, NN: Neural Networks, KNN: K Nearest Neighbours, LR: Logistic Regression, SOM: Self Organising Maps, GA: Genetic Algorithms).

Hybrid system	Data set	Techniques	Evaluation
SePI	Customer churn data of a Korean company; 11,587 subscribers' demographic and transaction data; 24 features.	DT (C5.0), NN, LR Lee and Lee (2006)	10-fold cross validation, hit ratio
KNN - LR	Customer data of a Chinese mobile carrier; training set (40,000 non-churners and 2,000 churners), testing set (10,000 non-churners and 495 churners); 93 features.	KNN, LR Zhang et al. (2007)	classification overall accuracy, ROC analysis
(NN - NN) and (SOM - NN)	dataset provided by American telecom company, including 34,761 churners and 16,545 non-churners from July 2001 to January 2002	NNs, SOMs Tsai and Lu (2009)	5-fold cross validation, overall prediction accuracy, two types of error assessment
Unsupervised techniques-Decision Tree with Boosting	Three sets of data are obtained from the Teradata center at Duke University, USA. Set1: 100,000 records, churn rate 50%; Set2: 50,000 records, churn rate 1.8%; Set3: 100,000 records, churn rate 1.8%	DT (C5.0) with boosting, clustering techniques (i.e., k-means, k-moid, SOM, FCM, and BIRCH) Bose and Chen (2009)	Lift Curve
DT - GA	1.2 million subscribers Call Details Records from the largest telecom operator in a developing country during 1st July and 1st September 2009.	DT (C4.5), GA Yeshwanth et al. (2011)	Accuracy, False Positive

been dealt with in previous studies (Huang, Huang, & Kechadi, 2010b; Huang, Buckley, & Kechadi, 2010). So, the learning system consists mainly of the following phases: data discretisation, data clustering, rule induction for each cluster, and classification/ prediction. The clustering results vary due to the nature of the k-means related algorithms, thus, we apply cross validation to objectively calculate the accuracy of the learning system. We randomly partition the dataset into five subsets with approximately equal size; one subset is used for test validation, while the remaining four subsets are used as the training data and they are part of the `Training Data`. The cross validation process iterates for five times, each subset data has equal chance to be selected as test data, and the final performance estimation result of the model is the average of the five validations. Fig. 1 illustrates the process of one validation. Basically, clustering and inductive rule leaning are considered as the two core steps of the hybrid system.

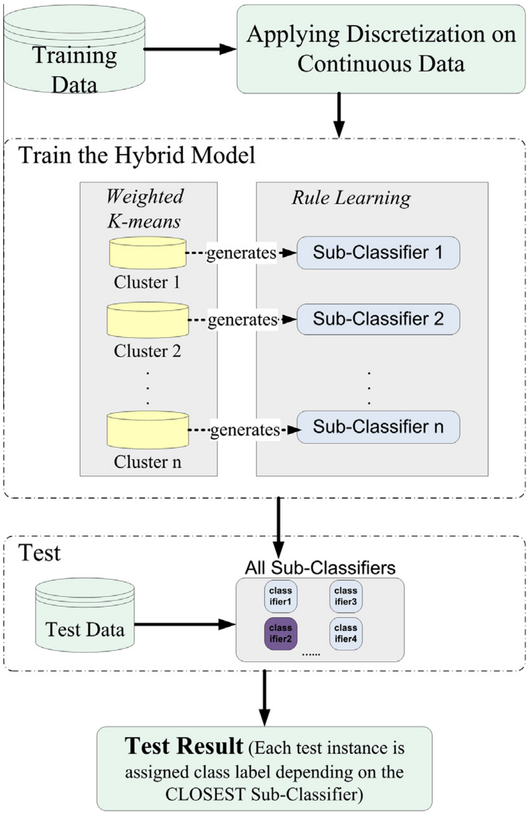


Fig. 1. The framework of the hybrid model-based prediction system.

3.1. Discretization

The discretisation process is widely used in data mining and statistical analysis problems. It partitions continuous attributes into discretised intervals, and it is usually carried out as the first step for implementing probability functions or inductive learning. Therefore, the effectiveness of discretization contributes to the performance of the entire analysis. Many discrete methods have been developed and have shown to be effective (Ching, Wong, & Chan, 1995; Fayyad & irani, 1993; Kurgan & Cios, 2004). The data used in this study contains a set of symbolic and continuous attributes. The continuous attributes are unsuitable for inductive rule learning. Therefore, we adopt the Class-Dependent discretisation technique for continuous attributes. This approach aims to maximize the dependency relationship (i.e., mutual dependence) between the class variables and continuous attributes, and automatically decide the number of intervals for an inductive learning application. More information about the process of discretisation can be found in Ching et al. (1995).

3.2. Clustering Algorithm

Clustering technique is widely used in many data mining applications. It groups data objects based on information that describes the objects and their relationships (Tan, Steinbach, & Kumar, 2005), called knowledge. In our approach we use clustering technique to segment the training data.

A number of different types of clustering methods, such as partitional clustering, hierarchical clustering, and fuzzy clustering, etc., have been studied in the last two decades or so. Among those clustering algorithms, k-means, which belongs to the family of partitional clustering, is very popular due to its simplicity. We adopted the core idea of this algorithm and proposed a weighted k-means clustering method to more effectively partition the data into groups with different characteristics, as the given application is very complex and its datasets are heterogeneous and containing a large number of attributes.

3.2.1. K-means

Algorithm 1 illustrates the main procedure of the classic k-means clustering. Consider a dataset D, let K be the number of clusters and o be an object of dataset D, C_i and c_i denote the i^{th} cluster and its centroid, respectively. The algorithm randomly selects K initial centroids; then it iterates through the following steps; each data object in D is assigned to its nearest cluster, this is followed by the recalculation of the cluster centroids. This algorithm stops when centroids are unchanged.

Algorithm 1. K-means Algorithm

```
Select $K$ objects as the initial centroids;
repeat
    Assign each object ($o$) to its closest centroid ($C_i$);
    Re-compute the centroids of all clusters;
until The value of objective function (SSE) does not change.
```

The main goal of clustering is that the objects within a group are similar and objects belonging to different group are dissimilar. Therefore, one needs to define similarity or a distance measure between data objects. In this analysis we use the Euclidian distance. We use the objective function SSE (Sum of the Squared Error) to measure the quality of the clustering, and SSE is calculated as follows:

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\|(c_i - o_{ij})\|^2 \tag{1}$$

where n_i is the number of data objects in cluster C_i and o_{ij} is the j^{th} data object in cluster C_i. The smaller the value of SSE, the better is the quality of the clustering. The algorithm stops when SSE is minimised. More details about the K-means algorithm can be found in Han and Kamber (2000).

3.2.2. Weighted K-means

The datasets we are using are provided by a telecom company. These datasets have very large number of attributes, and each attribute has different effect on the target (i.e., class). Thus, it may be more effective to deal with each attribute differently. We assign a weight for each attribute in order to reduce high influence of an attribute on the target. We believe that this can help to extract the knowledge from the datasets more objectively and accurately. In order to study the influence of each attribute, we consider the causal relationship between attributes and the target.

Vasconcelos, Almeida, and Nobre (1998) found that the simple correlation coefficient-based analysis does not reflect the direct effect of independent variables upon the dependent variable. Therefore, the correlation coefficient analysis usually tends to draw a wrong conclusion. Furthermore, multiple linear regression investigates the correlation relationship between multivariate variables and dependent variables, but it does not figure out the joint effect caused by multiple variables. Thus, to quantify the significance of attributes, we apply Path Analysis (PA) to discover the causality between variables. PA, which is a multivariate statistical analysis, provides not only the direct effect of an independent variable on a dependent variable but also the indirect effect of an independent variable on a dependent variable via other independent variables. PA has been applied in various fields of social science, such as public health psychology, agriculture, business, etc., one can refer to the applications in Duncan (1966), Scheiner, Mitchell, and Callahan (2000) and Vasconcelos et al. (1998). More details on PA can be found in Alwin and Hauser (1975) and Denis and Legerski (2006).

In order to analyse the causal effect of $x = \{x_1, x_2, \ldots, x_n\}$ (independent variables) on y (dependent variable), PA decomposes the correlation coefficient r_{yx_i} into the direct effect of x_i on y and the indirect effect of the x_i on y via other independent variables (say x_j). The direct effect of x_i on y is quantified by Path Coefficient p_{yx_i}. To calculate p_{yx_i}, we define the correlation coefficient matrices (2) and (3) that represent the correlation matrix among variables x $(x_1, x_2, \ldots, x_n)$ and the correlation matrix between x and y, respectively:

$$R = \begin{pmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_nx_1} & r_{x_nx_2} & \ldots & 1 \end{pmatrix} \tag{2}$$

$$r_{yx} = \begin{pmatrix} r_{yx_1} \\ r_{yx_2} \\ \vdots \\ r_{yx_n} \end{pmatrix} \tag{3}$$

The direct path coefficient can be calculated as follows:

$$p_{yx} = R^{-1} r_{yx} \tag{4}$$

In addition to domain experts, it is usually difficult to set appropriate paths, which reflect the way the independent variables affect the dependent variables. This is also why there is little literature about PA, as a multivariate analysis technique. In this paper, we started by considering that the indirect effect of x_i on y (called indirect path coefficient) by combining it with one variable x_j $(j \neq i)$ only. In order to set more effectively the weights, we use another concept; the determinant factor $(R^2_{(i)})$, which reflects the integrated determinant effects of x_i on y. $R^2_{(i)}$ is calculated as follows:

$$R^2_{(i)} = 2p_{yx_i} r_{yx_i} - p^2_{yx_i} \tag{5}$$

We define the weight of each attribute as the ratio of the determinant factor $\left(R^2_{(i)}\right)$ over the sum of all determinant factors, (see the Eq. (6)):

$$w_i = \frac{R^2_{(i)}}{\sum_{i=1}^{n} R^2_{(i)}} \tag{6}$$

Due to the high-dimensionality of the datasets used, we adopt the Minkowski distance of order p, and we define p to be the total number of attributes. To reflect the influence of attributes, we integrate the weights in the distance measure as follows:

$$dist = \left(\sum_{i=1}^{p} w_i \cdot (x_i - x'_i)^p\right)^{1/p} \tag{7}$$

where x_i and x'_i represent the attribute values of two different objects.

3.3. Inductive Rule Learning

Inductive rule learning is another core of this study. A set of rules can be generated from each cluster, and the rules can be used to represent the characteristics of the clusters. The main advantage of the rule-based learning is that the results can easily be understood and interpreted. In this research, we apply FOIL (First Order Inductive Learning), which is a classic sequential covering technique based algorithm, to learn about each group of the training data.

Sequential covering algorithm learns one rule at a time and removes any positive examples it covers, and then it iterates through this process until all positive examples are covered by the learned rules. FOIL is an extension of the sequential covering algorithm to the first-order representation. FOIL has been widely used by rule-based classification algorithms and techniques, such as Qin and Lawry (2005), Qin and Lawry (2008), Rai, Thoke, and Verma (2012) and Yin and Han (2003). These classification models implement pruning strategies or optimisation process on the basis of

FOIL's architecture. For the application of customer churn prediction, we are interested more on the behaviour of potential churners. Moreover, as FOIL is capable of sequentially generating a set of positive rules (i.e., churn rules in this application), we use it for rule-based learning. Algorithm 2 briefly outlines the FOIL algorithm, the input variable *Example* denotes the training set with labelled data examples.

Algorithm 2. FOIL (*Examples*)

```
Pos ← Positive Examples;
Neg ← Negative Examples;
$Learned\_rules \leftarrow \phi$;
while Pos is not empty do
    Rule← Learn-A-Rule (Examples, Neg);
    $Learned\_rules \leftarrow Learned\_rules \cup$ Rule;
    Pos ← Pos - {positive examples covered by Rule};
    Examples ← Examples - {any examples covered by Rule};
end while
return Learned_rules
```

Algorithm 3. Learn-A-Rule (*Examples*, Neg)

```
Rule ← the most general positive rule;
repeat
    Candidate_cond← generate candidate conditions for Rule;
    Best_cond← max (Foil_Gain);
    add Best_cond to the antecedent of Rule;
    covered_Neg← negative examples that are covered by Rule;
until There is no negative examples can be covered.
return Rule
```

More information of this algorithm, such as the calculation of information gain, can be found in Quinlan and Cameron-jones (1993).

3.4. Classification

To predict whether a customer is about to leave or not, we use clustering together with the inductive rule learning approach described in Section 3.3. We argue that customers having similar characteristics also behave similarly. In the case of churn problem, customers (instances) who have similar patterns, for instance, having very close calls behaviour, spending almost the same, etc., are more likely to belong to the same class (`churn` or `non-churn`). Thence, we aim to predict each test instance (tc_i) by referring to the instances with similar behaviour (we call them `Close Instances`). The goal is to have all `Close Instances` in the same cluster (`Close Cluster`) to tc_i. `closest` is quantified by the minimum distance measure between the instance and clusters, and the Minkowski distance (Eq. (7)) is used in this case. The final classifier (`Close Sub-classifier`) is produced by applying the rule learning method (FOIL) on the `Close Cluster`. Therefore, the classification of a tc_i is carried out by the following principles:

- If the `Close Cluster` is pure and all instances in `Close Cluster` are positive (`churn`), then tc_i is predicted as `churn`;
- If the `Close Cluster` is pure and all instances in `Close Cluster` are negative (`non-churn`), then tc_i is predicted as `non-churn`;
- If the `Close Cluster` is impure (having both `churn` and `non-churn`), then the class label of tc_i is assigned by the `Close Sub-classifier`, which consists of a set of positive rules. tc_i is predicted as `churn` if it was covered by any of the rules, otherwise, it will be assigned to the majority class (`non-churn` in this application).

In addition to the cases mentioned above, if the number of `Close Clusters` is more than one, that is, more than one clusters have the same minimal distance with tc_i, then tc_i is predicted by the majority prediction of all `Close Sub-classifier`. For example, three clusters (cl_1, cl_2, cl_3) are all `Close Clusters`, if the sub-classifiers generated by cl_1 and cl_2 predict tc_i to be `non-churn`, while the result `churn` is obtained from cl_3, then tc_i is labelled to `non-churn`.

4. Experiments and discussion

4.1. Data

Experiments were conducted on a telecom dataset. The whole dataset consists of 104,199 customer records: 6,056 churners and 98,143 non-churners. Each customer record is characterised by 121 attributes, including 11 symbolic attributes and 110 continuous attributes. The attributes mainly consist of the following information:

Demographic profiles:
describes the demographic information, including age, gender, social class bands, county code, etc.
Account information:
all information about customer accounts, e.g., start time, account number/type, fees, payment type, account balance, call information, etc.
Call details:
this group of information describes many aspects related to different types of calls (e.g., international or local calls), number of calls, call duration, costs, etc.

Apart from the attribute information mentioned above, one can refer to Huang, Kechad, and Buckly (2009) and Huang, Sato, Huang, Kechadi, and Buckley (2010a) for more other information about the data (e.g., Information of grants, service orders, historical information of payments, etc.).

4.2. Evaluation

In order to evaluate the performance of a predictive model objectively, we use 5-fold cross validation model. The data is randomly partitioned into five nearly equal size sets; during each iteration one set is used as the validation data, while the remaining four sets of data are used for training. In other words, the validation process is repeated five times; each of the five datasets is used exactly once as the validation data. The five validation results are then averaged to produce the final result.

Usually, a predictor is considered to be effective if a large number of instances can be correctly predicted. Thus, in much literature, the overall predictive accuracy (i.e., $\frac{num\_correct}{num\_total}$) is employed to evaluate the classification performance, such as Tsai and Lu (2009), Yeshwanth, Raj, and Saravanan (2011) and Zhang, Qi, Shu, and Cao (2007). However, the overall accuracy-based evaluation criterion does not suite imbalanced data (the telecom data in this study is one of them). For instance, for imbalanced data, a high overall inaccuracy might be caused by the fact that many data examples were

Table 2
Confusion matrix.

		Predict	
		CHURN	NONCHURN
Actual	CHURN	a_{11}	a_{12}
	NONCHURN	a_{21}	a_{22}

wrongly assigned to the majority class. Therefore, in this work, for better assessing the power of a classification technique, we adopt the rates of True Churn (*TC*), which indicates the proportion of churn examples that were correctly predicted as churn, and False Churn (*FC*), which is the proportion of non-churn examples that were wrongly predicted as churn. The objective of this application is to obtain high *TC* with low *FC*. Table 2 shows the confusion matrix, where the four parameters are defined:

a_{11}: denotes the number of churn cases that are predicted as churn.
a_{12}: represents the number of churn cases that are predicted as non-churn.
a_{21}: denotes the number of non-churn cases that are predicted as churn.
a_{22}: represents the number of non-churn cases that are predicted as non-churn.

The *TC* and *FC* are calculated as follows:

$$TC = \frac{a_{11}}{a_{11} + a_{12}} \quad \text{and} \quad FC = \frac{a_{21}}{a_{21} + a_{22}} \tag{8}$$

In this paper, we also use Receiver Operation Characteristic (ROC) curve, which is plotted by a set of TC-FC pairs (True Churn - False Churn), to demonstrate the predictive accuracy, Vuk and Curk (2006) gave the definition of ROC and how it can be applied for evaluating classification tasks. Obviously, the bigger the rate of True Positive and smaller the rate of False Churn are, the better are the predicted results. In some cases, it is difficult to use ROC to compare several prediction results. For instance, if one ROC curve (roc1) was plotted for predicted results with very high TCs and FCs, and another ROC curve (roc2) was plotted for a set of other predicted results with relatively smaller TCs and FCc simultaneously; only domain experts can identify which result is better than the other. To overcome this problem, the area under the ROC curve (AUC) is frequently used to measure the predictive accuracy. Bradley (1997) discusses the use of area under the ROC curve for the evaluation of machine learning algorithms. AUC can be calculated by the following equation:

$$AUC = \frac{S_0 - n_0 \times (n_0 + 1) \times 0.5}{n_0 n_1} \tag{9}$$

where S_0 is the sum of the ranks of the class 0 (churn) test patterns, n_0 is the number of patterns in the test set which belongs to class 0, and n_1 is the number of patterns which belong to class 1 (non-churn). The details of AUC can be found in Bradley (1997) and Vuk and Curk (2006).

4.3. Experiment set-up

In order to comprehensively investigate the proposed learning system, three sets of experiments were conducted. The first experiment is for evaluating the performance of the hybrid model and the second and third are for comparing it to well-known existing models.

4.3.1. Set-up I

The performance of this hybrid classification system highly depends on the result of clustering. Therefore, in our experiments, we compare the clustering results derived from the original and the modified k-means algorithm, respectively. In addition, to provide more accurate weight for each attribute, we also compare the weighted clustering approaches, which are based on the path analysis and other correlation analysis methods (i.e., Chi-Square and information gain).

Chi-Square can be used to evaluate the dependence between an input variable and the class, and it has been widely applied in many phases of data mining process. In this application, we calculate the Chi-Square statistic (χ^2) for all attributes, $\{\chi_1^2, \chi_2^2, \ldots, \chi_n^2\}$, and the weight of an attribute (say the i_{th} attribute), is defined by $\frac{\chi_i^2}{\sum_{i=1}^n \chi_i^2}$, where n is the total number of attributes.

In addition to Chi-Square, information gain is also used to measure the significance of attributes. Similar to the weighting methods that are based on path analysis and Chi-Square, we calculate the information gain of all attributes, $\{ig_1, ig_2, \ldots, ig_n\}$, and the final weight of each attribute is defined as the fraction of ig_i by the sum of the information gain of all attributes, which is $\sum_{i=1}^n ig_i$.

4.3.2. Set-up II

In the second set of experiments, we compare the classification performance of the hybrid learning system with six other different types of supervised classification techniques. These include tree-based methods, statistical analysis, and rule-based techniques. We briefly describe in the following the classification models that were used in this evaluation:

Decision Tree: C4.5 is one of the most well-known tree-based classifiers. A tree is constructed by using the strategy of divide-and-conquer. C4.5 starts to search an attribute with the highest information gain, and then the data is partitioned into classes according to the attributes' values. Each sub-tree is recursively further partitioned using the same policy until any of the stopping criteria is satisfied; e.g., a leaf node is reached. The decision rules are obtained by traversing the branches of the tree from the root to the leaves. The details of this algorithm can be found in Quinlan (1993).

Logistic Regression: Logistic regression (LR) is a type of regression analysis; it has been widely applied in applications of probability classification. The probability of an event may occur is estimated as follows:

$$P(y = 1|x_1, \ldots, x_k) = \frac{e^{b_0 + b_1 x_1 + \cdots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + \cdots + b_k x_k}} \tag{10}$$

where y is a binary independent variable representing the occurrence of an event (i.e., $y = 1$ an event occurred and $y = 0$ otherwise), $x_1, x_2, \ldots, x_k$ denote the independent variables and $b_1, b_2, \ldots, b_k$ are the regression coefficients that can be estimated by many methods (e.g., Maximum Likelihood Estimation).

k-NN: k-nearest neighbour is one of the most simplest classification technique. The aim of k-NN is to classify a test instance (x') by finding its neighbourhood, which consists of the k closest instances in the training set. The class label assignment of the test instance is based on the dominance of the classes in the neighbourhood, that is, the test instance should be assigned to the majority class of the k instances. Therefore, this modelling technique has two core elements: the value of k, and the distance metric of similarities between instances. It is not easy to set the most proper value of k. In this study, we set $k = 3$.

SVM: Support Vector Machine (SVM) is an effective and robust technique for classification applications. The SVM main objective is to find the best classification function to separate the data objects of the training set into multiple classes. In this study, we are concerned by binary classification application of customer churn prediction, we aim to find the best classification function (i.e., the maximum-margin hyperplane) to partition the training set into two classes. New instances are mapped into the model and are predicted as a category (class) based on the side of the hyper-plane they fall into.

Since the real dataset is not linearly separable, a non-linear classification function $f(x)$, corresponding to a hyperplane, is used to separate the two classes. A new data instance x_i is classified by the function $f(x)$ based on the following principles:

$$\begin{cases} \text{positive}, & \text{if } f(x) < 0 \\ \text{negative}, & \text{if } f(x) > 0 \end{cases} \tag{11}$$

The classification function can be written as follows:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(< x_i, c >) + b \tag{12}$$

where n is the number of instances in the training set, x_i is an instances in the training set, c is a new instance that does not have a class label, $k(.,.)$ is a kernel function and b is a threshold. The parameter α_i indicates whether x_i is a support vector or not, <.,.> indicates the inner production of vectors, therefore, the inner product computation for the new instance is only for the "support vectors" rather than the entire training instances. The obtained maximum margin hyperplane is the so called Support Vector Machine. The parameter α_i can be obtained by solving a convex quadratic programming problem. Burges (1998) proposed that polynomial kernels and Gaussian radial basis functions are often used as kernel functions (Boser, Guyon, & Vapnik, 1992).

OneR: OneR is a very efficient rule-based classification learning method. OneR can generate a set of simple rules named "1-rules", which can classify an object based on a single attribute. To avoid over-fitting, which is a common problem in the domain of classification, OneR requires all intervals to contain more than a user-specified number of examples in the same class. This technique has higher classification efficiency than many other classifiers.

PART: PART is an efficient rule-learning classification method; it infers rules by repeatedly building partial C4.5 decision trees; thus PART combines two paradigms for the rule generation; one is for creating rules from decision trees and the other is divide-and-conquer. Once a partial tree has been built, a single rule is extracted from it. The main advantage of PART over the other models lies in its simplicity, and it produces good rule sets that compare favourably with those generated by well-know modelling techniques (e.g., RIPPER).

The telecom data suffers from the imbalance problem since the data consists of a huge number of non-churner records and relatively a very small number of churner records. To overcome this issue we adopted under-sampling technique, which randomly removes some of the majority class samples to reduce its size. Therefore, in our experiment, we train each classifier on a set of samples, which have different churn rate $\left(\frac{num\_churners}{num\_total}\right)$ in the range [10%, 50%] with the step size of 5%, obtained by applying the under-sampling method to randomly remove the non-churn records. Table 3 shows all the used samples, the third and the fourth columns indicate the number of churners and non-churners in each training sample, respectively; and the last column denotes the size of testing data, which is one fold of the entire dataset.

Moreover, to get an ideal predictive result, it is necessary to have an appropriate parameter as the number of clusters (i.e., K). In the experiment, K is set in the range [1, 10] with the step size of 1. In order to find the optimal K, 5-fold cross validation is applied, so the optimal K would be the one that leads to the highest average AUC.

Table 3
Under-sampled training data and testing data that are used in the experiment.

Sample	Churn Rate %	#Train-churn	#Train-nonchurn	#Test
S1	10	4896	44064	20839
S2	15	4896	27744	20839
S3	20	4896	19584	20839
S4	25	4896	14688	20839
S5	30	4896	11424	20839
S6	35	4896	9092	20839
S7	40	4896	7344	20839
S8	45	4896	5980	20839
S9	50	4896	4896	20839

4.3.3. Set-up III

In addition to the traditional classification techniques, we also compare the classification performance of the proposed hybrid learning system with several other hybrid classification models. In the following we briefly introduced the main structures of these hybrid models and the necessary parameter setting:

SePI: This classification model consists of three main steps: (1) a Main Model is built by applying various single models on trainingsets; the model that has the best performance is selected as the Main Model; (2) a Discrimination Model is built to found out whether the Main Model performs well on the given application datasets; (3) a Support Model is built by applying various single models on the data that the Main Model does not predict correctly, and the model that has the best performance is selected. Based on these three models, the process of prediction is performed by firstly using the Discrimination Model to find out which model (Main Model or the Support Model) should be used to predict the final results. In this work, the Discrimination Model and the Support Model use NN (Neural Networks), and C5.0 is used for the main model.

k-NN-LR: This hybrid classifier combines k-NN and Logistic Regression. This method has two main phases: (1) split the training data X into m disjoint data sets, where m is the number of attributes; (2) transform X into a new data set $K(X)$ by employing the k-NN algorithm; then a Logistic Regression classifier is used to train the new data set $K(X)$. In this study, we set the number of the nearest neighbours to 7.

KM-BoostedC5.0: This modelling technique is established by combining clustering techniques and decision tree with boosting. This model is build specifically for a mobile service data set characterized by two types of attributes, which show the information of service usage and revenue contribution, respectively. Therefore, in the first stage, clustering techniques are used to segment the data, and the original data is transformed into data instances with cluster labels only; in the second stage, the obtained new data is used as the input of the boosted decision tree model (C5.0).

In the literature, five clustering techniques (i.e., K-means, K-medoid, SOM, FCM and BIRCH) were applied in the first stage to compare the performance. In this experiment, we use KM (K-means) for segmenting the data set since this literature states that KM generates the most number of models that could beat the benchmark models. In addition, based on attribute information, we segment the real telecom data into clusters by considering two groups of attributes, one group includes the information of demographic profiles and customer account information, the other group includes the information of call details. More information of this hybrid model is shown in Bose and Chen (2009). In order to treat this classification model fairly, in the first stage, instead of setting the number of clusters (K) arbitrary, we find the optimal K values for each data sample from candidate numbers in a range [1 : 10] with step size of 1.

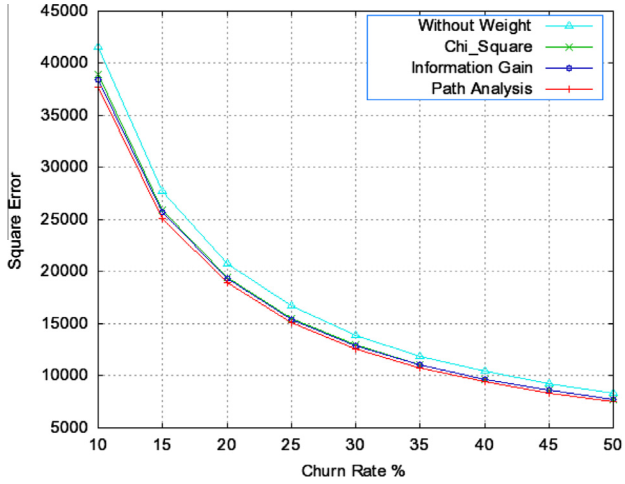


Fig. 2. Clustering results by 4 approaches.

4.4. Results and Discussion

4.4.1. Experiment I

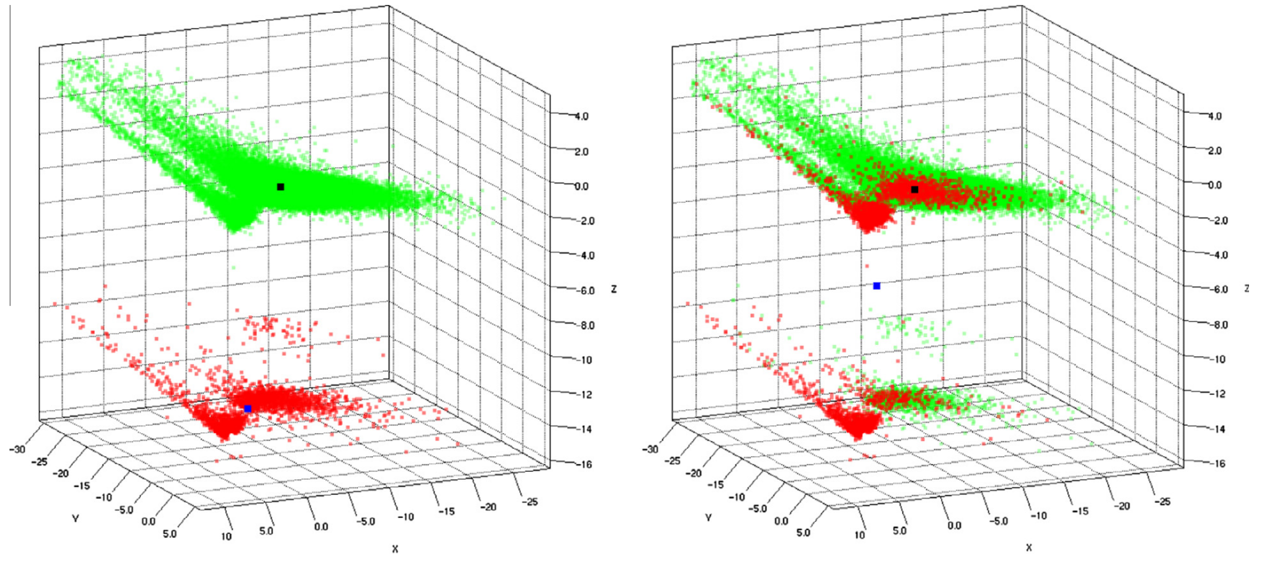
The original k-means clustering aims to minimize the Within-class Sum of Square Error (WSSE). Therefore, we use WSSE to evaluate the clustering performance. We use 9 samples to train the models and all the samples are used in performance evaluation and for comparing the clustering approaches considered in this study. Fig. 2 shows different clustering results produced by the original and weighted k-means. It also shows the clustering performance based on different statistical analysis. In this figure, the horizontal axis represents the churn rate while the vertical axis indicates the objective function (WSSE). Each plot is produced using different data samples. One can see that the weighted method produces better clustering results. In addition, the Figure also shows that the results generated by Path Analysis have lower square error than those of Chi-Square and Information Gain.

In order to demonstrate the performance of the modified clustering method, we display its results. The telecom data has a huge number of attributes, and to better illustrate the function of clustering, we used the PCA (Principal Component Analysis) technique to reduce its dimensionality. We plot the first, the second and the third principal components of all the observations, as illustrated in Fig. 3, X-axes, Y-axes and Z-axes represent the first, second, and third principal component, respectively. Since the customer churn prediction is a binary classification task, we assume that the data is separated into two clusters ($cluster_1$ and $cluster_2$). Fig. 3(a) and (b) show the clustering results produced by the weighted k-means and the original k-means, respectively. In the two figures, we show $cluster_1$ in red and $cluster_2$ in green. The center of each cluster is also shown, the blue square represents the center of $cluster_1$, while the black square represents the center of $cluster_2$. We can clearly see that the left figure has more separable clustering result than the right one, that is, the modified clustering method achieves a better clustering result than the one produced by original method.

In order to prove that this approach is suitable for other applications, we applied it on four other benchmark datasets (Ionosphere, SAheart, vehicle2 and Glass0), which are part of the UCI repository and KEEL (Knowledge Extraction based on Evolutionary Learning) datasets. All the four benchmark datasets require binary classification. Fig. 4 shows 3D clustering results for each dataset. We use the same symbols and colours as in Fig. 3, in order to better show the improvement of the weighted clustering, we display the result of each benchmark data in the most suitable view for observation. The left and right figures represent the modified and the original clustering results, respectively. The left figures separate the datasets more completely than the figures on the right side since the weighted method is able to find the better cluster centres.

4.4.2. Experiment II

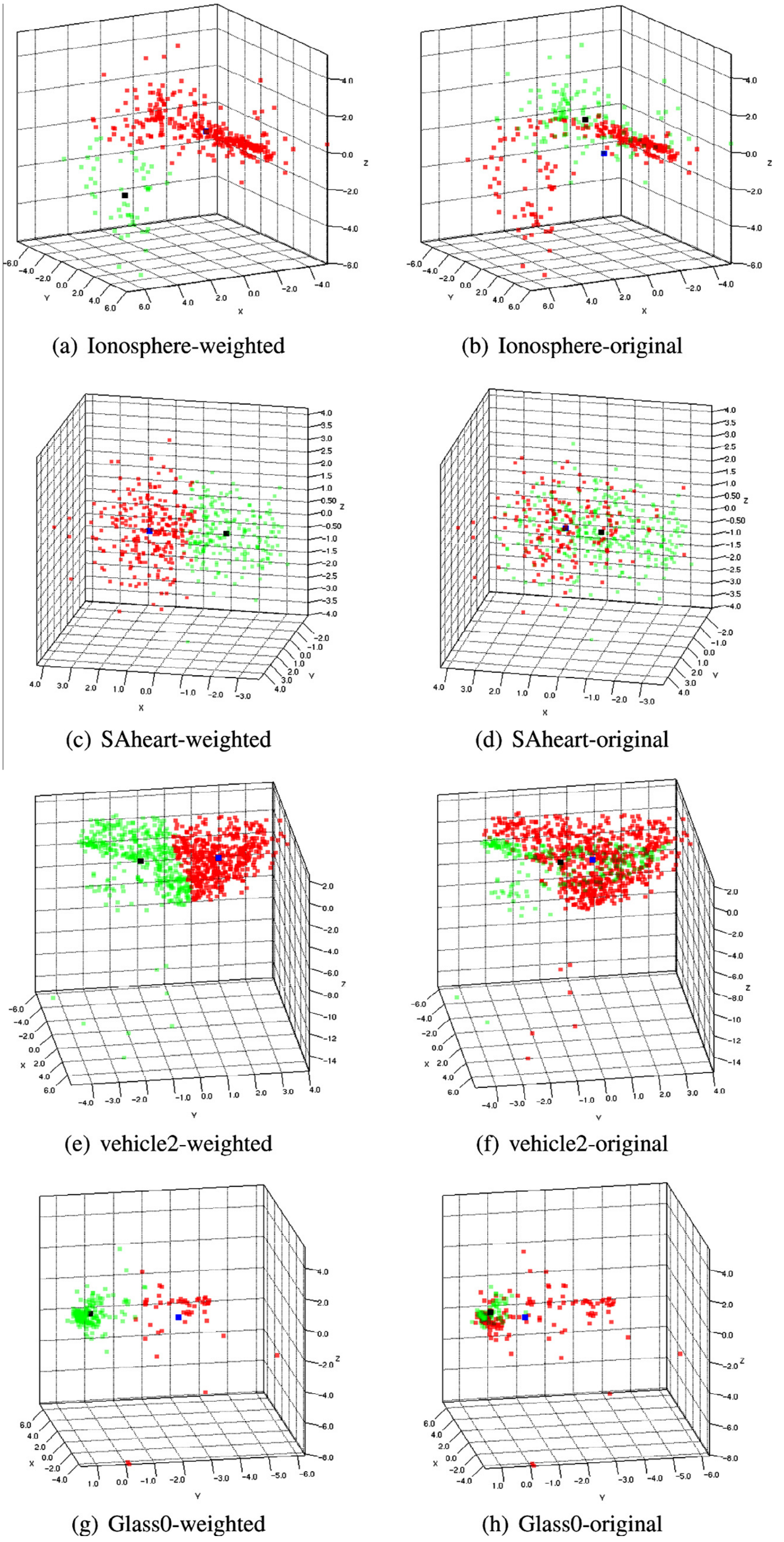
Based on the experimental set-up described in Section 4.3.2, we plot the ROC curve for each classification model. Fig. 5(a) compares the ROC curves of the proposed hybrid model and other well-established classification modelling techniques. The horizontal axis and the vertical axis represent the rates of False Churn and True Churn, respectively. We look for the predictive accuracy with high true churn and low false churn.



(a) Weighted

(b) Original

Fig. 3. Original k-means vs. weighted k-means on Telecom dataset.



(a) Ionosphere-weighted

(b) Ionosphere-original

(c) SAheart-weighted

(d) SAheart-original

(e) vehicle2-weighted

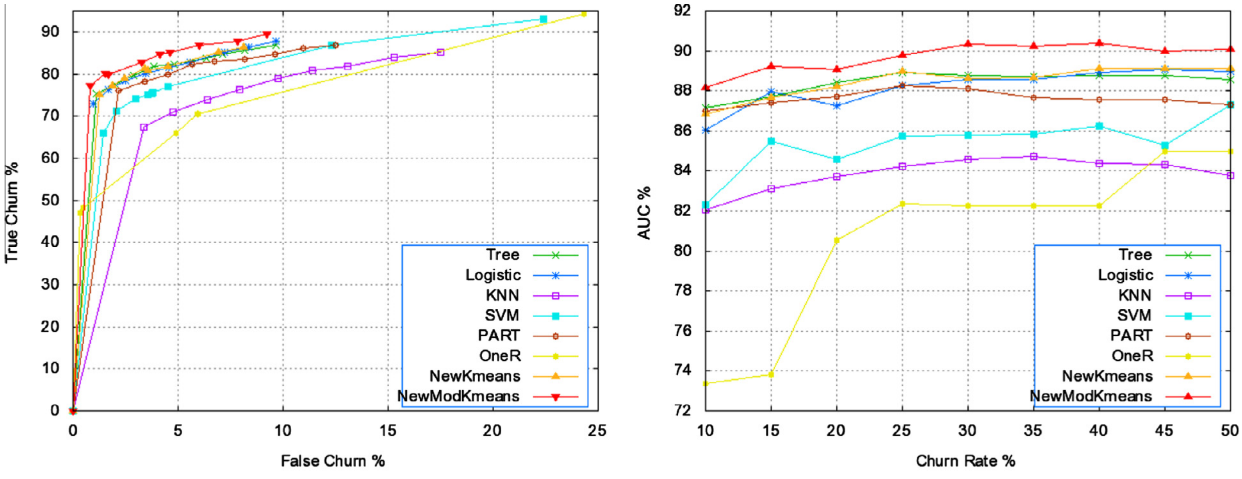
(f) vehicle2-original

(g) Glass0-weighted

(h) Glass0-original

Fig. 4. Original k-means vs. modified k-means on four benchmark datasets.

Each ROC curve is plotted by training the system with different churn rates (One can refer to Table 3 for more details about the samples). From this figure, we can observe that the ROC curve (labelled by *NewModKmeans*), which is obtained based on the



(a) ROC (b) AUC

Fig. 5. Compare the ROC curves and AUC of different classification techniques in customer churn prediction.

Table 4
An example of the prediction performance of a sample (S1, Churn Rate = 10%) based on different number of clusters. The maximum AUC value in bold.

Subset	Performance metric (%)	#Cluster (K)								
		2	3	4	5	6	7	8	9	10
*subset*1	AUC	84.86	84.98	84.88	85.43	86.71	86.24	87.36	86.78	86.67
*subset*2	AUC	87.01	88.71	86.41	88.51	85.12	84.02	86.41	89.02	84.36
*subset*3	AUC	87.74	85.27	86.64	86.58	85.76	83.51	86.98	87.83	86.53
*subset*4	AUC	87.85	88.16	87.75	86.65	87.83	88.29	87.87	88.15	88.21
*subset*5	AUC	86.13	87.2	87.15	85.99	86.02	85.7	86.54	88.27	87.45
Average	AUC	86.71 ± 1.24	86.86 ± 1.68	86.56 ± 1.07	86.62 ± 1.16	86.28 ± 1.03	85.55 ± 1.91	87.03 ± 0.59	***88.01*** ± ***0.81***	86.64 ± 1.44

Table 5
Prediction performance of samples based on different number of clusters. The maximum AUC value in bold.

Subset	Performance metric (%)	#Cluster (K)								
		2	3	4	5	6	7	8	9	10
*S*1	AUC	86.71 ± 1.24	86.86 ± 1.68	86.56 ± 1.07	86.62 ± 1.16	86.28 ± 1.03	85.55 ± 1.91	87.03 ± 0.59	***88.01*** ± ***0.81***	86.64 ± 1.44
*S*2	AUC	87.8 ± 1.38	87.08 ± 0.82	87.25 ± 0.23	87.66 ± 1.32	87.14 ± 0.6	86.89 ± 0.57	86.57 ± 2.13	87.1 ± 0.77	***88.49*** ± ***1.33***
*S*3	AUC	87.97 ± 0.73	87.48 ± 0.78	87.92 ± 0.69	87.39 ± 0.58	87.61 ± 1.55	88.15 ± 2.05	87.31 ± 1.27	88.07 ± 0.69	***88.35*** ± ***0.27***
*S*4	AUC	88.53 ± 1.46	88.54 ± 0.89	88.66 ± 0.71	88.44 ± 0.8	88.9 ± 1.23	88.26 ± 0.69	***89.2*** ± ***0.87***	88.55 ± 0.59	88.55 ± 0.38
*S*5	AUC	89.02 ± 1.01	88.93 ± 0.92	88.85 ± 0.97	88.6 ± 0.75	88.15 ± 0.85	88.59 ± 0.75	89.01 ± 0.86	***89.31*** ± ***0.97***	89.25 ± 0.91
*S*6	AUC	88.99 ± 1.39	89.13 ± 1.02	89.46 ± 1.04	***89.61*** ± ***1.04***	89.38 ± 1.04	89.21 ± 0.74	89.01 ± 0.53	88.75 ± 0.9	88.73 ± 0.81
*S*7	AUC	89.44 ± 0.79	89.15 ± 0.75	***89.51*** ± ***0.65***	88.88 ± 1.69	89.39 ± 1.26	88.7 ± 1.17	88.65 ± 1.36	88.03 ± 0.85	88.07 ± 1.06
*S*8	AUC	***89.32*** ± ***1.3***	88.51 ± 1.02	89.21 ± 0.64	89.03 ± 0.58	88.8 ± 1.37	88.94 ± 1.09	88.97 ± 0.91	89.21 ± 1.2	88.81 ± 1.66
*S*9	AUC	87.54 ± 1.46	***89.4*** ± ***0.71***	88.36 ± 1.43	88.32 ± 0.64	88.32 ± 0.93	88.36 ± 0.89	88.43 ± 0.89	87.91 ± 0.63	88.46 ± 0.99

Table 6
Compare the Overall Accuracy (OA) of the hybrid-model with the traditional classifiers based on all samples and Wilcoxon signed-rank test statistic values for testing the statistical significance when comparing the hybrid-model with the other classifiers. The maximum accuracy values in bold.

Sample data	Performance metric (%)	Classifiers						
		Hybrid-model	Tree (C4.5)	Logistic	knn	SVM	PART	OneR
*S*1	OA	***97.85***	97.53	97.54	94.31	96.68	96.57	96.61
*S*2	OA	***97.51***	96.79	97.03	92.94	96.39	95.52	96.52
*S*3	OA	***97.09***	96.11	96.4	91.46	95.67	94.5	93.43
*S*4	OA	***96.67***	95.32	95.56	89.89	95.17	93.618	92.72
*S*5	OA	***95.13***	94.46	94.65	88.27	95.03	92.68	92.65
*S*6	OA	93.92	93.09	93.72	86.63	***95.01***	91.35	92.66
*S*7	OA	***94.61***	92.41	92.35	84.84	94.4	90.04	92.67
*S*8	OA	***94.44***	91.46	91.31	82.98	87.63	88.87	76.74
*S*9	OA	90.01	***90.11***	90.18	81.04	78.51	87.45	76.74
Wilconxon Statistic Value (W)			W+=44, W-=1	W+=44, W-=1	W+=45, W-=0	W+=42, W-=3	W+=45, W-=0	W+=45, W-=0

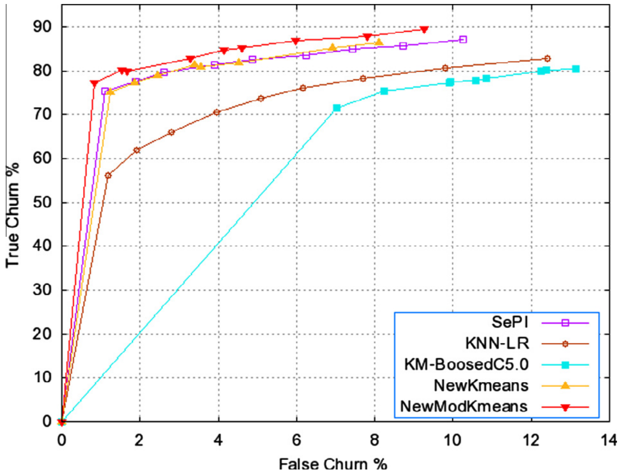


Fig. 6. Compare the prediction performance between the proposed hybrid-model and other hybrid classifiers based on ROC.

proposed hybrid model, is close to the left-upper corner most. Therefore, it compares favourably with the other six different classification models. We also plotted the ROC curve (labelled by *NewKmeans*) for the hybrid model by applying the original k-means. The figure illustrates that it has similar accuracy with Decision Tree for most of the points (with either higher rate of True Churn or lower rate of False Churn) and outperforms many other classifiers, but does not perform as good as the one based on weighted k-means. Fig. 5(b) compares the hybrid model-based learning system with the six classifiers by using Area Under ROC Curve as the evaluation metric. In this figure, the horizontal and vertical axes represent the churn rate and the AUC values, respectively. Each curve is plotted by considering all samples, and we can obviously see that the AUC values obtained by the hybrid system are all higher than the ones obtained by other classification models no matter what the churn rates of the samples are.

In addition, the original k-means, decision tree, logistic regression, PART, SVM, KNN, and OneR are successively followed by the proposed model. For several samples, the hybrid system achieves nearly two percentage points higher on AUC than decision tree. The k-means based hybrid model, in fact, has similar performance than decision tree, it achieves a little higher AUC only when the churn rate becomes relative high (i.e., 40%, 45%, 50%).

In terms of the number of clusters (K), as mentioned in Section 4.3.2, each sample might have different optimal K due to different churn rates and data objects distributions. As mentioned in previous section, we applied the cross validation method for all samples. Table 4 shows an example of obtaining the final prediction performance and deciding the optimal K for a training sample (here S1 is an example sample), and AUC is considered as the a metric measure. The last row indicates the average performance and the standard deviation of the 5 folds of validation, and the maximum value is obtained for $K = 9$. We did the same cross validation on each training samples. Table 5 shows the prediction performance based on all samples, and the maximum AUC values acquired by the optimal K are marked in bold. Moreover, it is good to see that the standard deviation for all the obtained accuracy metric values is very small, especially the acquired optimal performance metric in many cases. This reflects the consistency and stability of the proposed model.

ROC and AUC can objectively evaluate the classification accuracy by considering the ratios of both true and false positive. In a sense, they can better examine the skills of models than many other evaluation methods; therefore they are commonly used as a metric for the classifiers. Moreover, the overall accuracy (OA) is also important for evaluating a prediction model comprehensively. Table 6 compares the proposed hybrid model with the traditional classifiers on the overall accuracy rate. In this table, each row indicates the accuracy rate obtained from each sample, and we highlight the maximum accuracy rates in bold. Each OA of the hybrid-model is calculated for the corresponding optimal K (number of clusters). It is interesting to see that the new model compares favourably with the other ones in most of the cases; in addition, SVM achieves the best accuracy on $S6$, decision tree and the proposed one has more or less the same accuracy on $S9$.

Furthermore, in order to study the significance of this comparative study, which decides whether there is a significant difference between the proposed model and the traditional classifiers, it is necessary to use a statistical test method. We utilise the *Wilcoxon signed_rank* test on the overall accuracy rates since it is a non-parametric hypothesis test and does not require the data to be normally distributed. We simply point out the data and parameters related to the test rather than describe the *Wilcoxon* test steps in detail. In our case, the paired data would be the third column in Table 6 and one of the remaining columns (i.e., any column indicating the accuracy rates of a traditional classifier), the paired data size $N = 9$. We take the significance test between hybrid-model and SVM as an example. We firstly propose the null hypothesis (i.e., there is no significant difference between the hybrid-model and SVM), after ranking the absolute differences of the paired data, the positive total rank $W+ = 42$, while the negative total rank $W- = 3$. Thus, the *Wilcoxon* test statistic $W = 3$. With a data sample size $N = 9$, the critical value for a two-tailed test at $\alpha = 5\%$ and one-tailed test at $\alpha = 2.5\%$ is 6. The test statistic value is 3 (less than 6), therefore, we can reject the null hypothesis and conclude that there is a significant difference between the hybrid-model and SVM with 95% confidence, furthermore, it is likely that the hybrid-model is better than SVM with 97.5% confidence. To test the significant difference between the hybrid-model and other classifiers, the positive and negative total rank values are shown in Table 6, each with smaller total rank equal to the *Wilcoxon* test statistic W. Thus, we can conclude that the prediction accuracy of the hybrid-model is significantly better than the other six classifiers.

4.4.3. Experiment III

Fig. 6 compares the prediction performance between the proposed hybrid model and other hybrid classifiers described in Section 4.3.3. The figure illustrates that the proposed hybrid model obviously compares favourably well with the other three hybrid classification methods. NewKmeans (the new hybrid model based on original k-means) outperforms k-NN-LR and KM-Boosted C5.0, in addition, this figure shows that NewKmeans is as good as SePI in the first several points, which are the results of the samples having relative lower churn rates; for the samples with higher churn rate, NewKmeans produces lower ratio of False Churn than SePI.

In addition to the True Churn and False Churn illustrated in the ROC Curve, we also compare our model to the above mentioned hybrid classification methods. To do so, we use AUC and the overall accuracy and the results are shown in Table 7. In this table, each row depicts the performance (AUC and OA) achieved by each hybrid technique and the maximum AUC and accuracy are highlighted in bold. As one can notice, our proposed hybrid model shows the absolute advantages on both AUC and OA for all the samples used ($S_1, \ldots, S_9$).

In order to show that the proposed hybrid modelling technique is more general and can be applied to many other classification or prediction applications, we applied this system on 22 benchmark datasets, which are collected from the UCI machine learning repository and KEEL datasets. Table 8 shows the AUC values of all the used datasets, which are calculated by the hybrid model-based

Table 7
Comparison of prediction performance (AUC and OA) between the proposed hybrid-model and several other recent hybrid approaches based on a set of data samples. The maximum AUC and overal accuracy values, respectively in bold.

Sample data	Hybrid-Model		SePI		KNN-LR		KM-BoostedC5.0	
	AUC%	OA%	AUC%	OA%	AUC%	OA%	AUC%	OA%
*S*1	***88.15***	***97.85***	87.14	97.56	77.51	96.38	82.20	89.61
*S*2	***89.25***	***97.51***	87.82	96.95	80.01	96.01	83.74	89.36
*S*3	***89.08***	***97.09***	88.46	96.37	81.59	95.41	83.72	89.37
*S*4	***89.80***	***96.67***	88.67	95.25	83.26	94.59	83.56	90.82
*S*5	***90.33***	***95.13***	88.89	94.42	84.24	93.70	83.65	88.51
*S*6	***90.24***	***93.92***	88.59	93.18	84.91	92.82	83.88	87.20
*S*7	***90.39***	***94.61***	88.74	92.15	85.30	91.49	83.63	86.49
*S*8	***89.96***	***94.44***	88.50	90.96	85.42	87.32	83.87	87.29
*S*9	***90.10***	***90.01***	88.44	89.60	85.14	87.31	83.58	88.75
Avg	***89.70***	***95.24***	88.36	94.05	83.04	92.79	83.54	88.60

Table 8
Compare the AUC values (%) of the used classification modelling techniques based on the benchmark datasets. The maximum AUC values in bold.

Data	#Ins	#Atts	Class Distribution	Hybrid-model	Tree	Logistic	kNN	SVM	PART	OneR	SePI	KNN-LR
Australian	690	15	45%: 55%	***87.2***	85.3	88.4	77.6	83.7	82.3	86.3	83.2	85.2
breast	699	10	34%: 66%	95.5	93.9	93.1	93.5	93.9	93.2	88.9	95.2	***96.5***
credit	653	16	45%: 55%	***88.0***	85.4	85.2	84.7	85.3	84.8	86.1	83.9	85.0
eastWest	213	26	38%: 62%	91.2	***100.0***	***100.0***	92.3	98.1	93.9	***100.0***	***100.0***	95.0
German	1000	21	30%: 70%	63.4	64.8	***67.7***	61.1	64.7	63.6	50.0	65.0	64.3
*Glass*0	214	10	33%: 67%	***83.3***	77.1	69.8	81.3	50.0	77.1	69.8	77.1	65.6
*Glass*6	214	10	14%: 86%	***93.4***	82.9	84.2	85.0	70.0	82.6	65.0	82.6	89.2
Ionosphere	351	35	64%: 36%	***92.2***	87.0	86.3	82.2	80.2	89.3	77.0	91.7	85.8
heart	270	14	44%: 54%	83.8	82.0	***85.7***	81.0	83.0	82.0	73.0	81.8	72.5
hepatitis	80	20	17%: 83%	***77.8***	70.0	***77.8***	70.0	67.0	70.0	50.0	60.0	65.7
Japan	690	16	45%: 55%	87.8	***88.7***	84.1	83.0	86.1	83.8	86.7	85.8	83.2
atoms	1078	11	34%: 66%	71.4	64.7	51.8	69.3	52.6	65.8	72.9	64.7	***73.9***
page	5472	11	10%: 90%	***92.3***	90.4	81.4	88.0	69.8	89.0	83.9	90.4	75.2
parkinson	195	22	25%: 75%	82.3	84.5	81.4	***91.7***	50.0	74.0	71.9	80.4	85.5
pima	74	6	14%: 86%	***100.0***	97.9	72.9	97.9	75.0	97.9	97.9	97.9	95.8
teaching	151	6	35%: 65%	***70.7***	63.6	64.6	65.6	64.3	65.6	61.4	64.9	66.2
thyroid	215	6	30%: 70%	96.7	88.6	88.6	***99.7***	68.2	92.8	85.4	88.6	88.0
*thyroid*1	215	6	16%: 84%	***99.2***	96.7	99.2	***99.2***	79.2	98.4	92.6	94.2	98.4
*vehicle*0	846	19	23%: 77%	94.5	90.4	***95.6***	0.892	0.5	92.5	74.0	90.4	79.7
*vehicle*1	846	19	26%: 74%	***74.8***	66.7	70.1	69.8	50.0	56.6	57.1	66.7	65.2
wdbc	569	31	37%: 63%	***94.6***	91.1	91.9	92.4	92.7	92.2	88.9	91.1	94.1
westEast	213	26	62%: 38%	95.8	***100.0***	***100.0***	93.3	98.1	94.3	***100.0***	***100.0***	97.7
Average				***87.1***	84.1	82.7	83.9	73.3	83.3	78.1	83.4	82.2

system, the other six traditional models, and two of the used hybrid models (KM-BoostedC5.0 is not applied here since its modelling approach is only suitable for the telecom or mobile service data). In this table, the first column represents the dataset; the next three columns indicate the properties of datasets (i.e., the number of instances, the number of attributes and the class distribution). For each dataset, we partition the data into two sets, two-thirds of the data is used for training, and the remaining one-third is for testing. The AUC values are shown in each row for all models, and we highlight the maximum value in bold. Among all the benchmark datasets, the hybrid system is better 12 times. Moreover, we calculate the average AUC values for each classification method, and the hybrid model still has the maximum average value.

5. Conclusion and Future Work

In a competitive market, such as telecommunications, churn prediction is very important for operators to try to retain valuable customers and provide attractive services. Therefore, building an effective predictive model becomes a necessity and many researchers have been started looking at how to solve this problem. Moreover, most of well-established classification techniques are based on a single model. This paper presents a hybrid model-based classification learning system, which integrates a weighed k-means clustering and a classic inductive rule learning method (FOIL). The proposed technique combines the clustering and classification data mining approaches.

To evaluate the prediction accuracy of the proposed approach, we use ROC and AUC as the measure metric, and compare it with several well-known classifiers. The experimental results show that the proposed system compares favourably well with these well-known classifiers. In addition to the telecom data, we applied our techniques on 22 benchmark datasets, and compare it with other classifiers. The results show that a large number of the benchmark data gain the maximum accuracy when applying the hybrid model.

For future work, we will be committed to improving the proposed learning system by considering several issues. Firstly, as the pre-processing in data mining, outliers and redundant data examples can be detected, and for this hybrid model, eliminating outliers would greatly contribute to a better clustering result. Secondly, other clustering algorithms, which can decide about the number of clusters, might be applied in the first stage. Finally, many other single predictive models, such as decision tree, svm, neural networks, etc., can be applied as a second stage to build sub-classifiers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eswa.2013.04.020.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*, 37–47.

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

Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce, 19*(2), 133–151.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual ACM conference on computational learning theory, (COLT 1992)* (pp. 144–152). Pittsburgh, PA, USA: ACM Press. July 27-29..

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Journal of Pattern Recognition, 30*, 1145–1159.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Journal of Data Mining and Knowledge Discovery, 2*(2), 121–167.

Ching, J. Y., Wong, A. K. C., & Chan, K. C. C. (1995). Class-dependent discretization for inductive learning from continuous and mixed-mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*(7), 641–651.

Coussement, K., & Den Poel, D. V. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Journal of Expert Systems with Applications, 34*(1), 313–327.

Denis, D., & Legerski, J. (2006). Causal modeling and the origins of path analysis. *Journal of Theory and Science, 7*(1).

Duncan, O. D. (1966). Path analysis: Sociological examples. *The American Journal of Sociology, 72*(1), 1–16.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th international joint conference on artificial intelligence, (IJCAI-93), Chambéry, France, August 28–September 3* (pp. 1022–1027). Morgan Kaufmann.

Han, J. W., & Kamber, M. (2000). *Data mining: Concepts and techniques* (1st ed.). Morgan Kaufmann Publishers Inc..

Huang, B. Q., Buckley, B., & Kechadi, M-T. (2010). Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Journal of Expert Systems with Applications, 37*(5), 3638–3646.

Huang, B. Q., Kechad, M. T., & Buckly, B. (2009). Customer churn prediction for broadband internet services. In *Proceedings of the 11th international conference on data warehousing and knowledge discovery, (DaWaK '09), Linz, Austria, August 31–September 2. Lecture Notes in Computer Science* (Vol. 5691, pp. 229–243). Springer-Verlag.

Huang, B. Q., Sato, T., Huang, Y., Kechadi, M. T., & Buckley, B. (2010a). Using genetic K-means algorithm for PCA regression data in customer churn prediction. In *Proceedings of the advanced data mining and applications - 6th international conference, (ADMA 2010), Chongquing, China, November 19–21. Lecture Notes in Computer Science* (Vol. 6441, pp. 210–220). Springer.

Huang, B. Q., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Journal of Expert Systems with Applications, 39*(1), 1414–1425.

Huang, Y., Huang, B.Q., & Kechadi, M-T. (2010). A new filter feature selection approach for customer churn prediction in telecommunications. In *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM' 2010), Macao, China, 7-10 Dec. 2010* (pp. 338–342).

Huang, Y., Huang, B. Q., & Kechadi, M. T. (2011). A rule-based method for customer churn prediction in telecommunication services. In *Advances in knowledge discovery and data mining - 15th Pacific-Asia conference, (PAKDD 2011), Shenzhen, China, May 24–27. Lecture Notes in Computer Science* (Vol. 6634, pp. 411–422). Springer.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Journal of Expert Systems with Applications, 31*(3), 515–524.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Journal of Expert Systems with Applications, 26*(2), 181–188.

Khashei, M., Hamadani, A. Z., & Bijari, M. (2012). A novel hybrid classification model of artificial neural networks and multiple linear regression models. *Journal of Expert Systems with Applications, 39*(3), 2606–2620.

Kurgan, L., & Cios, K. (2004). CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering, 16*(2), 145–153.

Larivire, B., & Poel, D. V. D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Journal of Expert Systems With Applications, 29*, 472–484.

Lee, J. S., & Lee, J. C. (2006). Customer churn prediction by hybrid model. *Proceedings of the second international conference on advanced data mining and applications, (ADMA'06), Xi'an, China, August 14–16* (Vol. 4091, pp. 959–966). Berlin, Heidelberg: Springer-Verlag.

Mattison, R. (2001). *Telecom churn management: The golden opportunity*. Fuquay-Varina, N.C: APDG Publishing.

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

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Journal of Expert Systems with Applications, 36*(2, Part 2), 2592–2602.

Qin, Z., & Lawry, J. (2005). Linguistic rule induction based on a random set semantics. *Proceedings of the eleventh international fuzzy systems association world congress, (IFSA-05), Beijing, China, July 28-31* (Vol. 3, pp. 1398–1404). Tsinghua University Press and Springer.

Qin, Z., & Lawry, J. (2008). LFOIL: Linguistic rule induction in the label semantics framework. *Journal of Fuzzy Sets and Systems, 159*(4), 435–448.

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

Quinlan, J. R., & Cameron-jones, R. M. (1993). FOIL: A midterm report. In *Proceedings of the european conference on machine learning, (ECML-93), Vienna, Austria, April 5–7. Lecture Notes in Computer Science* (Vol. 667). Springer-Verlag.

Rai, D., Thoke, A. S., & Verma, K. (2012). Enhancement of associative rule based FOIL and PRM algorithms. In *Students conference on engineering and systems,(SCES), Allahabad, Uttar Pradesh, India, Mar 16-18* (pp. 1–4). Springer-Verlag.

Scheiner, S. M., Mitchell, R. J., & Callahan, H. S. (2000). Using path analysis to measure natural selection. *Journal of Evolutionary Biology, 13*(3), 423–433.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA, USA: Addison–Wesley Longman Publishing Co. Inc..

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Journal of Expert System with Applications, 36*(10), 12547–12553.

Vasconcelos, A. G., Almeida, V. R. M., & Nobre, F. F. (1998). The path analysis approach for the multivariate analysis of infant mortality data. *Journal of Expert System with Applications, 8*(4), 262–271.

Vuk, M., & Curk, T. (2006). ROC curve, lift chart and calibration plot. *Advances in methodology and Statistics, 3*(1), 89–108.

Wei, C. P., & Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Journal of Expert Systems with Applications, 23*(2), 103–112.

Xia, G. E., & dong Jin, W. D. (2008). Model of customer churn prediction on support vector machine. *Journal of Systems Engineering - Theory and Practice, 28*(1), 71–77.

Yeshwanth, V., Raj, V. V., & Saravanan, M. (2011). Evolutionary churn prediction in mobile networks using hybrid learning. In *Proceedings of the twenty-fourth international Florida artificial intelligence research society conference, (FLAIRS), Palm Beach, Florida, USA, May 18–20*. AAAI Press.

Yin, X. X., & Han, J. W. (2003). CPAR: Classification based on predictive association rules. In *Proceedings of the third SIAM international conference on data mining(SDM), San Francisco, CA, USA, May 1–3*. SIAM.

Zhang, Y. M., Qi, J. Y., Shu, H. Y., & Cao, J. T. (2007). A hybrid KNN-LR classifier and its application in customer churn prediction. In *Proceedings of the IEEE international conference on systems, man and cybernetics, (SMC), Montréal, Canada, 7–10 October* (pp. 3265–3269). IEEE.