Expert Systems with Applications 39 (2012) 1414–1425

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Customer churn prediction in telecommunications

Bingquan Huang *, Mohand Tahar Kechadi, Brian Buckley

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

ARTICLE INFO

Keywords:
Churn prediction
Imbalanced datasets
ROC and AUC techniques
Logistic Regressions
Linear Classifications
Naive Bayes
Decision Trees
Multilayer Perceptron Neural Networks
Support Vector Machines
Evolutionary Data Mining Algorithms

ABSTRACT

This paper presents a new set of features for land-line customer churn prediction, including 2 six-month Henley segmentation, precise 4-month call details, line information, bill and payment information, account information, demographic profiles, service orders, complain information, etc. Then the seven prediction techniques (Logistic Regressions, Linear Classifications, Naive Bayes, Decision Trees, Multilayer Perceptron Neural Networks, Support Vector Machines and the Evolutionary Data Mining Algorithm) are applied in customer churn as predictors, based on the new features. Finally, the comparative experiments were carried out to evaluate the new feature set and the seven modelling techniques for customer churn prediction. The experimental results show that the new features with the six modelling techniques are more effective than the existing ones for customer churn prediction in the telecommunication service field.

1. Introduction

The service companies of telecommunication service businesses in particular suffer from a loss of valuable customers to competitors; this is known as customer churn. In the last few years, there have been many changes in the telecommunications industry, such as, the liberalisation of the market opening up competition in the market, new services and new technologies. The churn of customers causes a huge loss of telecommunication services and it becomes a very serious problem.

Recently, data mining techniques have emerged to tackle the challenging problems of customer churn in telecommunication service field (Au, Chan, & Yao, 2003; Coussement & den Poe, 2008; Hung, Yen, & Wang, 2006; John, Ashutosh, Rajkumar, & Dymitr, 2007; Luo, Shao, & Liu, 2007; Wei & Chiu, 2002; Yan, Wolniewicz, & Dodier, 2004). As one of the important measures to retain customers, churn prediction has been a concern in the telecommunication industry and research (Luo et al., 2007). Over the last decade, the majority of churn prediction has been focused on voice services available over mobile and fixed-line networks. As mentioned in Luo et al. (2007) and Zhang et al. (2006), in contrast to the mobiles services, there are less researchers to investigate the churn prediction for the land-line telecommunication services.

Generally, the features used for churn prediction in mobile telecommunication industry includes customer demographics, contractual data, customer service logs, call details, complaint data, bill and payment information (Hadden, Tiwari, Roy, & Ruta, 2006; Hung et al., 2006; John et al., 2007; Luo et al., 2007; Wei & Chiu, 2002). In contrast to the mobiles services, there are less amounts of qualified information for land-line services providers (Luo et al., 2007; Zhang, Qi, Shu, & Li, 2006). The data of land-line communication services is different to mobile services. Some of this data is missing, less reliable or incomplete in land-line communication service providers. For instances, customer ages and complaint data, fault reports are unavailable and only the call details of a few months are available. Due to business confidentiality and privacy, there are no public datasets for churn prediction. For churn prediction in the land-line telecommunication service field, Luo et al. (Luo et al., 2007; Zhang et al., 2006) presented a set of features, which are the duration of service use, payment type, the amount and structure of monthly service fees, Proportions variables, consumption level rates variables and the growth rates of the second three months. Recently, Huang, Kechadi, and Buckley (2010) presented a set of features, including one six-month Henley segmentation, line-information, bill and payment information, account information, call details and service log data, etc. In addition, most of the literature (Hadden et al., 2006; Hung et al., 2006, 2010; John et al., 2007; Wei & Chiu, 2002) shows the features that are the aggregated call-details are important for the customer churn prediction. These features are obtained by aggregating the duration, fees and the number of calls for any types of calls for each period. However, the call details can be further divided into more precise information, according to different types of calls (e.g. international, local, mobile phone, and national calls). This more precise information might be more useful than the existing features of call details for churn prediction.

In order to improve the accuracy of customer churn prediction in telecommunication service field, we present a new set of features with seven modelling techniques in this paper. The new

* Corresponding author.
E-mail address: bquanhuang@gmail.com (B. Huang).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.024

features are the 2 six-month Henley segmentation, precise 4-month call details, information of grants, line information, bill and payment information, account information, Demographic profiles and service orders that are extracted from existing limited information. The modelling techniques are Logict Regression, Naive Bayes, Linear classifiers, Decision Tree C4.5 (C4.5), Multilayer perceptrons artificial neural networks, Support Vector Machines and the Evolutionary Data Mining Algorithm. Finally, the comparative experiments are carried out. The experimental results show that the presented features and six modelling techniques are more effective than the existing features for the customer churn prediction in land-line communicational services.

The rest of this paper is organised as following: next section introduces the evaluation criteria of churn prediction systems. Section 3 describes our methodology which includes the techniques of feature extraction, normalisation and prediction. Experimental results with discussion are provided in Section 4, and the conclusion of this paper and future works are made in Section 5.

2. Evaluation criteria

After a classifier/predictor is available, it will be used to predict the further behaviour of customers. As one of the important steps to ensure the model generalises well, the performance of the predictive churn model has to be evaluated. That is, the prediction rates of a predictor are needed to be considered. In this work, the prediction rates refer to true churn rate (TP) and false churn rate (FP). The objective of the application is to get high TP with low FP. Table 1 shows a confusion matrix (Japkowicz, 2006), where a_{11} is the number of the correct predicted churners, a_{12} is the number of the incorrect predicted churners, a_{21} is the number of the incorrect predicted nonchurners, and a_{22} is the number of the correct predicted nonchurners. From the confusion matrix, TP is defined as the proportion of churn cases that were classified correctly, calculated by

$$\mathrm{TP} = \frac{a_{11}}{a_{11} + a_{12}}. \tag{1}$$

and FP is the proportion of nonchurn cases that were incorrectly classified as churn, written as

$$\mathrm{FP} = \frac{a_{21}}{a_{21} + a_{22}}. \tag{2}$$

From these pairs of TP and FP, the Receive Operating Curves (ROC) technique (Bradley, 1997) can be used to find the expected pair of prediction rates (TP and FP).

However, usually, the ROC technique is hard used to evaluate the sets of the pairs (TP and FP) that are from different prediction modelling techniques or different feature subsets of data. To overcome to the problem, the technique of calculating area under a ROC curve (AUC) (Bradley, 1997) is used to evaluate models and feature sets for the churn prediction in this paper. The area under a ROC curve can be calculated by the following equation:

$$\mathrm{AUC} = \frac{S_0 - n_0 \times (n_0 + 1) \times 0.5}{n_0 n_1} \tag{3}$$

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

Table 1
Confusion matrix.

Actual	Predicted	
	CHUN	NONCHU
CHU	a_{11}	a_{12}
NONCHU	a_{21}	a_{22}

3. Methodology

Generally, our churn prediction system consists of sampling data, data preparation, and classification/prediction phases. Data sampling randomly selects a set of customers with the relative information, according the definition of churn. The data preparation (also called data preprocessing) phase includes data cleaning, feature extraction and normalisation steps. Data cleaning removes the irrelevant information which includes wrong spelling words caused by human errors, special mathematical symbols, missing values, strings ‘‘NULL’’, duplicated information, and so on. The feature extraction extracts a set of features to represent customers. The normalisation step normalises the values of features into a range (e.g. in between 0 and 1). The prediction phase predicts the potential behaviour of customers in the nearest future. In this paper, the feature extraction is mainly discussed. The feature/variables extraction, normalisation and prediction/classification steps are described in the following subsections.

3.1. Feature/variable extraction

As one of the most important factors, feature extraction can influence the performance of predictive models in the terms of prediction rates (high TP and low FP). If a robust set of features can be extracted by this phase, the prediction rates of TP and FP can be significantly improved and reduced, respectively. However, it is not easy to obtain such a good set of features. Until now, most of the feature sets have been introduced for churn prediction in mobile telecom industry (Hadden et al., 2006; Hung et al., 2006; John et al., 2007; Luo et al., 2007; Wei & Chiu, 2002) and fixed-line telecommunication (Huang et al., 2010; Luo et al., 2007; Zhang et al., 2006). However, these existing feature sets still can be improved. In this paper, we presents a new set of features for customer churn prediction in telecommunication service fields, which are described as follows:

- **Demographic profiles**: describe a demographic grouping or a market segment and the demographic information contains likely behaviour of customers. Usually, this information includes age, social class bands, gender, etc. The information of gender and counties are available and selected as two new features.
- **Information of grants**: some customers have obtained some special grants resulting in their bills being paid fully or partly by a third party (note: only one new feature is selected). For example, customers with a disability or are over 80 are more like to continue the services.
- **Customer account information**: This information mainly contains the types of service packages, credit controller indicators, junk mail indicators, the first date of using the services, creation date, the bill frequency, the account balance, equipment rents, payment types, and the summarised attributes which are call duration, the number of calls and standard prices and fees, current outstanding charges and charges paid. This information very straightly describes the customer accounts (we consider that one account number correspond to one customer). In addition, the account information concludes the indicators of broadband, bitstream access networks and local loop unbundling (LLU). These indicators present whether a customer has the broadband service, whether a customer is a bitstream user and whether a customer uses the LLU services. Unlike other account information, these three indicators may not be

straightly found in one file (e.g. a table in the database). Generally, they can be found by joining a number of relative files (e.g. tables) in database. For instance, the broadband indicator can be found by joining some files (tables) of broadband services with the account file (or table). Therefore, all of this account information might discriminate between churners and nonchurners for a classification model.

Different customers might have different bill frequencies. The features of call duration, the number of calls, standard fees and paid fees have to be re-calculated for a specified period. Because most customers ordered the bills every month, the selected period is 60 days. Let "*Dur*", "*Ncall*", "*Standfees*" and "*Paidfees*" be the call duration, the number of calls, the standard fees and the actual fees paid, respectively. They can be re-calculated by Eq. (4):

$$\begin{aligned} Ncall' &= \frac{Ncall\_M}{nDays} * 60 \\ Dur' &= \frac{Dur\_M}{nDays} * 60 \\ Standfees' &= \frac{Standfees\_M}{nDays} * 60 \\ Paidfees' &= \frac{Paidfees\_M}{nDays} * 60 \end{aligned} \tag{4}$$

where "*Ncall_M*" is the number of calls in the most recent bill, "*Dur_M*" is the duration of the most recent bill, "*Standfees_M*" is the fees of the most recent bill, "*Paidfees_M*" is the fees from customers, and "*nDays*" is the number of day of the bill, which can be obtained by Eq. (5).

$$nDays = endDate - startDate \tag{5}$$

where "*endDate*" and "*startDate*" are the dates of bill starting and ending.

- **Service orders**: describe the services ordered by the customer. Because a order list might be is very large, we only select the information of the last 3 service orders. This information is the quantity of the ordered services, the rental charges, the due date, the date of approving services as new features.
- **Henley segments**: the algorithm of Henley segmentation (Henley, 2009) divides customers and potential customers into different groups or levels according to characteristics, needs, and commercial value. There are two types of Henley segments: the individual and discriminant segments. The individual segment includes ambitious Techno Enthusiast (ATE) and Comms Intense Families (CIF) Henley segments. The discriminant segments are the mutually exclusive segments (DS) and can represent the loyalty of customers. The Henley segments ("DS", "ATE" and "CIF") of the most recent 2 six-months are selected as new input features. Similarly, the missing information of the Henley segments are replaced by neutral data.
- **Telephone line information**: this is the information of voice mail service (provided or not), the number of telephone lines, line types, district codes, and so on. The customers who have more telephone lines might prefer the services more and they might be more willing to continues using the services. This information cant be useful for a prediction model. Therefore, the number of telephone lines, district codes, and the voice mail service indicator are selected as a part of new features.
- **Complaint Information**: this complain information is from customers who complained the telecommunication services. Customers complain the services by email, posted-letter, and more generally by telephone to different departments. The details of content of complain information mostly is missing in the database. However, normally the indicator that indicates who made the complaint in a specified date can be found. Therefore, this paper only select this indicator from complaint information as a new features.
- **The historical information of bills and payments**: this concerns the billing information for each customer and service for a certain number of years. The last 4-month call details are available and only used in our prediction system (see the features of call details which are described below). In order to consist with the call details of the last 4 months, the bill information of these months are selected.

As mentioned above, different customers may have different bill frequencies. Therefore, the durations of customer bills might be different. In order to make bill durations be consistent, we segment the bill whose duration is more than 1 month into a number of monthly bills, each of which cover 1 month. For example, if a customer has one two-month bill, we need to segment this bill into two monthly bills. Each monthly bill describes the total standard charge fees, total rental charges, total value added tax (VAT), total call duration and total fees paid. Each bill also provides the details of the 8 types of calls, which are local calls, mobile phone calls (085, 086, 087 mobile phone calls, and mobile phone UK-calls), International calls and other types of calls. The call duration, standard fees, actual charged fees and fees paid are summarised for each type of calls in a monthly bill.

Beside the total standard charge fees, total rental charges, total value added tax (or total VAT), total call duration, total fees paid, the call duration of each type of calls, the standard fees of each type of calls, actual charged fees of each type of calls and the fees paid of each type of calls, this paper extracts the sum of the total standard charge fees, the sum of total rental charges, the sum of the total VAT, the sum of the total call duration and the sum of the total fees paid for the last 4 months as new features. Consider that the total standard charge fees, total rental charges, total VAT, total call duration and total fees paid, respectively are "$TotSF_i$", "$TotRC_i$", "$TotVAT_i$", "$TotD_i$" and "$TotFP_i$" for the *i*th month. The sum of total standard charge fees "*SSF*", the sum of total rental charges "*SRC*", the sum of the total VAT "*SVAT*", the sum of the total call duration "*SD*" and the sum of the total fees paid "*SFP*" can be calculated by Eq. (6).

$$\begin{aligned} SSF &= \sum_{i=1}^{4} TotSF_i \\ SRC &= \sum_{i=1}^{4} TotRC_i \\ SVAT &= \sum_{i=1}^{4} TotVAT_i \\ SD &= \sum_{i=1}^{4} TotD_i \\ SFP &= \sum_{i=1}^{4} TotFP_i \end{aligned} \tag{6}$$

Let the type of calls be *k*. Consider that the call duration, the standard fees, the actual charged fees and the fees paid for the *i*th month are MD_k^i, MSF_k^i, MAF_k^i and MPF_k^i, respectively. Let the change of the call duration, and the change of the standard fees, the change of the actual charged fees and the change of the fees paid between the two consecutive months *i*th and $i-1$th for the type of calls *k* be $ch\_MD_k^{i,i-1}$, $ch\_MSF_k^{i,i-1}$, $ch\_MAF_k^{i,i-1}$ and $ch\_MPF_k^{i,i-1}$, respectively. This subset of information is extracted as new features by Eq. (7).

$$
\begin{aligned}
ch\_MD_k^{i,i-1} &= \frac{\left|MD_k^i - MD_k^{i-1}\right|}{\sum_{i=2}^{4}\left|MD_k^i - MD_k^{i-1}\right|}\\
ch\_MSF_k^{i,i-1} &= \frac{\left|MSF_k^i - MSF_k^{i-1}\right|}{\sum_{i=2}^{4}\left|MSF_k^i - MSF_k^{i-1}\right|}\\
ch\_MAF_k^{i,i-1} &= \frac{\left|MAF_k^i - MAF_k^{i-1}\right|}{\sum_{i=2}^{4}\left|MAF_k^i - MAF_k^{i-1}\right|}\\
ch\_MPF_k^{i,i-1} &= \frac{\left|MPF_k^i - MPF_k^{i-1}\right|}{\sum_{i=2}^{4}\left|MPF_k^i - MPF_k^{i-1}\right|}
\end{aligned}
\tag{7}
$$

- **Call details**: the call details refer to call duration, price and types of call (e.g. International or local call) for every call. Therefore, it is difficult to store the call details for every call for every month for every customer. Most of the telecommunication companies keep the call details of the last few months. Thus the limited call details can only be used for customer churn prediction. But they can still reflect how often customer have used the services by relative fees, call duration and so on. For example, the low call duration presents that the customers did not often use the services; the customers might cease the services in the future. If the fees of services are suddenly increased or decreased, the customer might cease the services sooner. The use of call details in churn prediction is reported in the literature (Wei & Chiu, 2002; Zhang et al., 2006). However, most of the researchers (Wei & Chiu, 2002; Zhang et al., 2006) only did select the aggregated number of calls, duration and fees from all types of phone calls as features. The authors did not further divide these aggregated the number of the calls, duration and fees into the ones of international, local, national, mobile phone, free and whole sale line calls (note: free calls are the calls without charge fees, whole sale line calls are that the customers call the whole sale line customers). However, these divided aggregated number of calls, duration and fees may more accurately reflect the status of using services and might be more effective for the churn prediction.

  This research extracts the number of calls, duration and fees from the call details of the last 4 months as new features. Based on these features, the change number of calls, change duration and change fees are also considered as new features. The procedure of extracting these new features can be described as follows: (1) segment the call details into a number of defined periods, (2) aggregate the duration, fees and the number of calls for each period for the local calls, national calls, international calls, mobile phone calls and other calls of every customer, (3) the total of the number of calls, duration and fees of the last 4 months are the sum of these aggregated number of calls, duration and fees, respectively, (4) the change number of calls, duration and fees can be obtained (explain in the next paragraph). Literature (Wei & Chiu, 2002; Zhang et al., 2006) reports that the call-details of every 15 or 20 days are effective. So the defined period contains 15 days in this paper.

  Let the numbers of calls, duration and fees be "*NCALL*", "*DUR*" and "*FEES*", respectively. Also let x represent one type of calls (local calls, national calls, international calls, mobile phone calls or other type of calls). Consider that the aggregated number of calls, duration and fees on the segment i for the type of calls x are "$NCALL_x^i$", "DUR_x^i" and "$FEES_x^i$", respectively. The changed number of calls, changed duration and changed fees between two consecutive segments i and $i-1$ of the call type x can be obtained by Eq. (8).

$$
\begin{aligned}
ch\_DUR_x^{i,i-1} &= \frac{\left|DUR_x^i - DUR_x^{i-1}\right|}{\sum_{j=2}^{M'}\left|DUR_x^j - DUR_x^{j-1}\right|}\\
ch\_NCALL_x^{i,i-1} &= \frac{\left|NCALL_x^i - NCALL_x^{i-1}\right|}{\sum_{j=2}^{M'}\left|NCALL_x^j - NCALL_x^{j-1}\right|}\\
ch\_FEES_x^{i,i-1} &= \frac{\left|FEES_x^i - FEES_x^{i-1}\right|}{\sum_{j=2}^{M'}\left|FEES_x^j - FEES_x^{j-1}\right|}
\end{aligned}
\tag{8}
$$

  where "$ch\_DUR_x^{i,i-1}$", "$ch\_NCALL_x^{i,i-1}$" and "$ch\_FEES_x^{i,i-1}$" are the changed number of calls, changed duration and changed fees between two consecutive segments i and $i-1$ of the call type x, respectively (note: the features of the free calls exclude the features "*ch_FEES*" and $FEES^i$).
- **Incoming calls details**: the incoming calls refer to the received calls. Incoming calls details are the durations of received calls, a number of received calls and the respected fees. However, there is no charge for the most of received called. The fees of incoming calls are not selected as new features. The new features from the incoming call details include the call duration and number of received calls, the change number of calls, the change of received call duration for every 15 days. Similarly, the change number of calls, the change of received call duration can be obtained by Eq. (8).

3.2. Normalisation

In the extracted features (see Section 3.1), some predictors or classifiers (e.g. artificial neural networks) have difficulties in accepting the string values of features (e.g. genders, county names). In addition, although the values of the features which are obtained by Eq. (8) or (7) (e.g. change duration, change fees, change fees paid) are in the range between 0 and 1 and these features will not be normalised, the values of some numerical features (e.g. the number of lines, the duration, the number of calls, fees) lie in different dynamical ranges. The large values of these features have larger influence over the cost functions than the small ones. However, it cannot reflect that the large values are more important in classifier or predictor design.

To solve above problems, the values of features have to be normalised for some predictors (e.g. artificial neural networks). The procedure of normalisation can be described as follows: (1) the values of string features need to be rewritten into binary strings, then (2) the values of numerical features can be normalised into a similar range by Eq. (9).

$$
\begin{aligned}
\bar{x}_j &= \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad j = 1, 2, \ldots, \imath\\
\sigma_j^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)\\
y &= \frac{x_{ij} - \bar{x}_j}{r\sigma_j}\\
\tilde{x}_{ij} &= \frac{1}{1 + e^{-y}}
\end{aligned}
\tag{9}
$$

where x_j is the feature jth, i is the number of features, N is the number of instances or patterns and r is a constant parameter which is defined by a user. In this study, r is set to one.

3.3. Prediction/classification

Many techniques have been proposed for churn prediction in telecommunication. This paper only selects seven modelling techniques as predictors for the churn prediction. These seven modelling techniques are outlined as follows:

3.3.1. Logistic Regressions (LR)

Logistic regression (Hosmer & Lemeshow, 1989) is a widely used statistical modelling technique for discriminative probabilistic classification. Logistic regression estimates the probability of a certain event taking places. The model can be written as:

$$prob(y=1) = \frac{e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_k}}{1 - e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_k}} \tag{10}$$

where Y is a binary dependent variable which presents whether the event occurs (e.g. $y = 1$ if event takes place, $y = 0$ otherwise), $x_1, x_2, \ldots, x_K$ are the independent inputs. $\beta_0, \beta_1, \ldots, \beta_K$ are the regression coefficients that can be estimated by the maximum likelihood method, based on the provided training data. The details of the logistic regression models can be found in Hosmer and Lemeshow (1989).

3.3.2. Decision Trees

A method known as "divide and conquer" is applied to construct a binary tree. Initially, the method starts to search an attribute with best information gain at root node and divide the tree into sub-trees. Similarly, the sub-tree is further separated recursively following the same rule. The partitioning stops if the leaf node is reached or there is no information gain. Once the tree is created, rules can be obtained by traversing each branch of the tree. The details of Decision Trees based on C4.5 algorithm are in literature (Quinlan, 1993, 1996).

3.3.3. Naive Bayes (NB)

a Naive Bayes classifier calculates the probability that a given input sample belongs to a certain class. Given an sample X which consists fo a feature/variable vector $\{x_1, \ldots, x_n\}$, the probability for the class y_j can be obtained by Eq. (11):

$$p(y_j|X) = p(X|y_j)p(y_j) = p(x_1, x_2, \ldots, X_n|y_j)p(y_j) \tag{11}$$

where $p(y_j)$ is the prior probability of y_j. However, Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent, the likelihood can written by:

$$p(X|y_j) = \prod_{i=1}^{n} p(x_i|y_j) \tag{12}$$

Thus, the posterior can be written as

$$p(y_j|X) = p(y_j) \prod_{i=1}^{n} p(x_i|y_j) \tag{13}$$

Consider that there are a number of classes $Y = \{y_1, y_2, \ldots, y_k\}$. Thus, given a unkown sample X, a Naive Bayes classifier makes the classification decision by:

$$c = \operatorname*{argmax}_{y_j \in Y} p(y_j|X) \tag{14}$$

The details of Naive Bayes classifier be found in literature (Langley, Iba, & Thompson, 1992).

3.3.4. Linear Classifiers (LC)

A linear classifier maps a feature space X into a set of class labels Y by a linear combination function. Usually, a liner classifier $f(x)$ can be written as follows:

$$f(\vec{x}) = sgn\left(\sum_i w_i x_i + b\right) \tag{15}$$

where $w_i \in \Re$ are the weights of the classifiers and $b \in \Re$ is constant. The value of $f(\vec{x})$ for input vector $\vec{x}$ determines the predicted class label. For example, in binary classifications, the class label is +1 if $f(\vec{x}) \geqslant 0$. Otherwise, the class label is -1. The weights w_i and constant b can be learned from a set of labelled training samples. The details of linear classifiers can be found in literature (Vapnik, 1998).

3.3.5. Artificial neural networks

A Multilayer Perceptron Neural Networks (MLP) is a supervised feed-forward neural network and usually consists of input, hidden and output layers. Normally, the activation function of MLP is a sigmoid function. If an example of MLPs with one hidden layer, the network outputs can be obtained by transforming the activation functions of the hidden unit using a second layer of processing elements, written as follows:

$$Output_{net}(j) = f\left(\sum_{l=1}^{L} w_{jl} f\left(\sum_{i}^{D} w_{li} x_i\right)\right), \quad j = 1, \ldots, J \tag{16}$$

where D, L and J are total number of units in input, hidden and output layer, respectively, and f is a activation function. The back-propagation (BP) or quick back-propagation learning algorithms would be used to train MLP. More details of MLP can be found on Rumelhart, Hinton, and Williams (1986).

3.3.6. Support Vector Machines (SVM)

A SVM classifier can be trained by finding a maximal margin hyper-plane in terms of a linear combination of subsets (support vectors) of the training set. If the input feature vectors are nonlinearly separable, SVM firstly maps the data into a high (possibly infinite) dimensional feature space by using the kernel trick (Boser, Guyon, & Vapnik, 1992), and then classifies the data by the maximal margin hyper-plane as following:

$$f(\vec{x}) = sgn\left(\sum_{i}^{M} y_i \alpha_i \phi(\vec{x_i}, \vec{x}) + \delta\right) \tag{17}$$

where M is the number of samples in the training set, $\vec{x_i}$ is a support vector with $\alpha_i > 0$, ϕ is a kernel function, $\vec{x}$ is an unknown sample feature vector, and δ is a threshold.

The parameters $\{\alpha_i\}$ can be obtained by solving a convex quadratic programming problem subject to linear constraints (Burges, 1998). Polynomial kernels and Gaussian radial basis functions (RBF) are usually applied in practise for kernel functions. δ can be obtained by taking into account the Karush–Kuhn–Tucker condition (Burges, 1998), and choosing any i for which $\alpha_i > 0$ (i.e. support vectors). However, it is safer in practise to take the average value of δ over all support vectors.

3.3.7. Data Mining by Evolutionary Learning (DMEL)

DMEL is a genetic classification technique. A DMEL classifier consists of a set of labelled rules which were found by the genetic algorithm. Given a unknown sample, the DMEL classifier applies the rules to match the sample and gives classification decision. The details of DMEL techniques can be found in literature (Au et al., 2003).

4. Experiments

4.1. Data

The 827,124 customers were randomly selected from the real-world database provided by the telecoms of Ireland in our experiments. In the training dataset, there are 13,562 churners and 400,000 nonchurners. In the testing dataset, the numbers of churners, nonchurners and total customers are the same as in the training dataset, respectively. Each customer is represented by the 738 features which are described in Section 3.1.

4.2. Experiment set-up

In order to evaluate the effect of the new presented features, three sets of experiments were carried out independently in this paper. The feature subsets of different types of call details were used in the first set of experiments. These feature subsets include: the features of international calls (INT), the features of mobile phone calls (MOB), the features of national calls (NAT), the features of local calls (LOC) and the features of the summarised calls (CD). In addition, the feature subsets contain the following combined feature sets: the combination "LCD" of LOC and CD (LCD = LOC + CD), the combination set "CDLMN" of LCD, MOB and NAT (CDLMN = LCD + MOB + NAT), and the combination set "CDLMNI" of the sets CDLMN and INT (CDLMNI = CDLMN + INT).

For the second set of experiments, the new feature set which is presented in this paper were used for the churn prediction. In addition, the different subsets of these new features were evaluated. That is, these subsets of new features were independently used to carry out the experiments. These subsets of new features are: (1) the subset of the fixed part features that have no change in the a number of months (e.g. account information, customer genders, distinct code, exchange code, spc code, Herenly segmentations), (2) the features of free calls, (3) the features of call details, (4) the features of incoming calls, (5) the features of WHS calls, (6) the features of the current two month bills, and (7) the subset of features of the 6 bills and payment information.

In order to comparatively study the features, the third set of experiments were performed, based on the new feature fullset and the existing feature sets for the customer churn prediction in the telecommunication service field. The existing feature sets are the feature set of Zhang et al. (2006) and the set of Huang et al. (2010). These two sets of features were used to carry out independently the experiments.

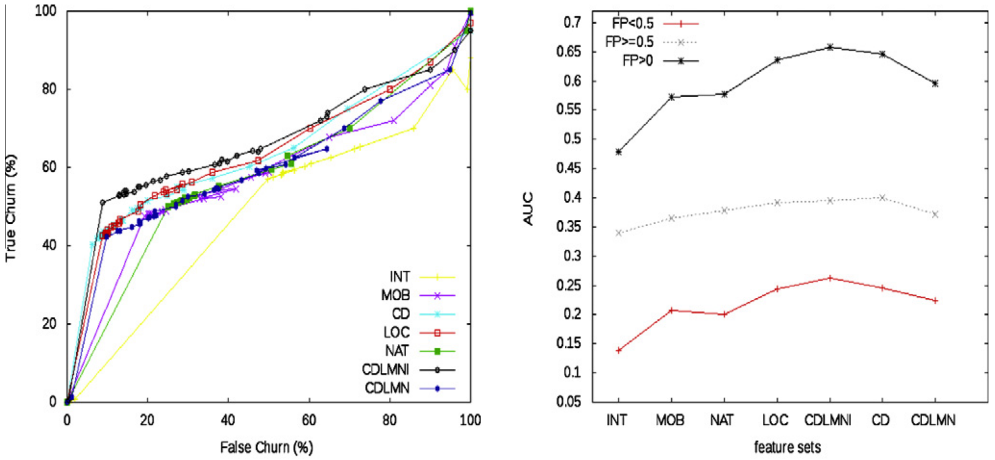
For the first and the second sets of experiments, six modelling techniques (LR, NB, LC, MLP, C4.5 and SVM) were used to make prediction for each set of features. However, for the third set of experiments, the above six modelling techniques and the DMEL techniques were used to predict or classify the behaviours of customers. In addition, Naive Bayes do not classify those instances for which conditional probabilities of any of the attribute value for every class is zero (Domingos & Pazzani, 1997). However, the feature sets used in the third set of experiments is high dimensional and might cause this problem. Therefore, the reduced or transformed features by Principal Component Analysis (PCA) (Jolliffe, 1986) are also used for Naive Bayes. For each subset of features, the training and testing datasets were not normalised when using the Decision Tree C4.5 or DMEL, but were normalised when using the NB, LC, LR, MLP or SVM. All the predictors were trained by 10 folds of cross-validations in each experiment. The same testing data were used for a trained model.

In addition, because there is usually a very large number of nonchurners and a very small numbers of churners, the class imbalance problem (Japkowicz, 2000) occurs in this application. The sampling technique was used to overcome this problem for each modelling technique. We sampled a subset of the training dataset by reducing the number of nonchurners. That is, for selecting data to train a model, the range of ratios ($\frac{n\_churners}{n\_nonchurners}$, where *n_churner* is fixed and *n_nonchurners* is variable) between churn and nonchurners were from 13.56 to 0.03. However, as above mentioned, the testing data for each trained model are the same (the numbers of churners and nonchurners were fixed in the testing phase).

In each set of experiments, each MLP with one hidden layer was trained. The number of input neurons of a MPL network is the same as the number of the dimensions of a feature vector. The number of output neurons of the network is the number of classes. Therefore, the number of output neurons is two in this application: one represents a nonchurner, the other represents a churner. If the numbers of input and output neurons are n and m, respectively, the number of hidden neurons of the MLP is $\frac{m+n}{2}$. The sigmoid function is selected as the activation function for all MLPs in the experiments. Each MLP was trained by 3 folds of cross-validation and BP learning algorithm with learning rate 0.1, maximum cycle 1800 and tolerant error 0.05, based on the training dataset. The number of training cycles to yield the highest accuracy is about 1100 for the MLPs.

Based on the extracted and normalised features, each SVM was trained to find the separating decision hyper-plane that maximises the margin of the classified training data. Two sets of values: the regularisation term $C \in \{2^8, 2^7, \ldots, 2^{-8}\}$ and $\sigma^2 \in \{2^{-8}, 2^{-7}, \ldots, 2^8\}$ of radial basis functions (RBF) were used to find the best parameters for the churn prediction. Altogether, 289 combinations of C and σ^2 with 3 folds of cross-validation were used for training each SVM. The optimal parameter sets (C, σ^2) yielding a maximal classification accuracy of SVMs were $(2^{-4}, 2^2)$ for these three sets of experiments.

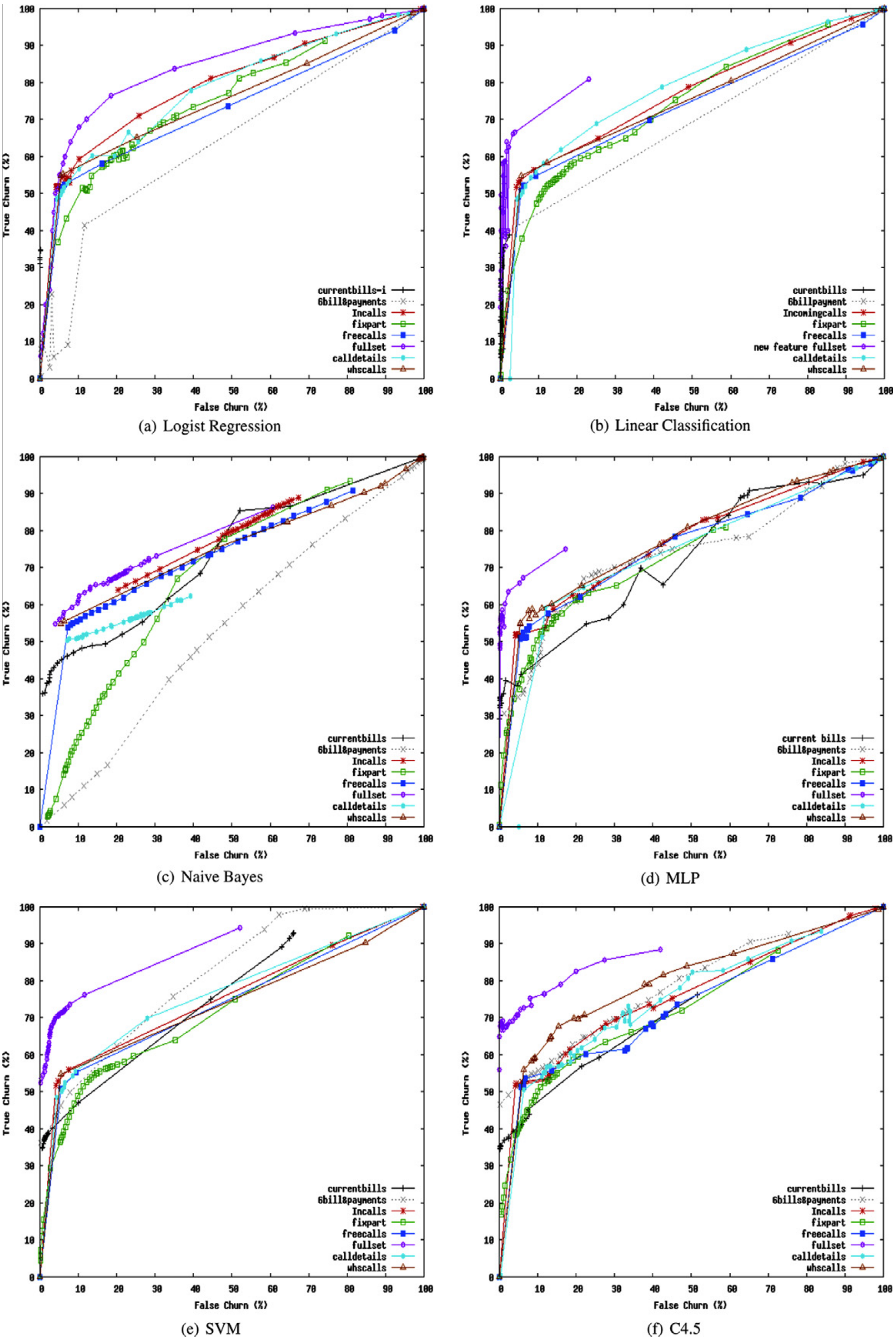
In the third set of experiments, each DMEL model was trained by the following parameters: population size is 30, the number of generations is 300, the Z-value is 1.96 for finding interested features, the probabilities of mutation and crossover are 0.1% and 60%, respectively.



(a) ROC

(b) AUC

Fig. 1. Prediction results when using call details.



(a) Logist Regression

(b) Linear Classification

(c) Naive Bayes

(d) MLP

(e) SVM

(f) C4.5

Fig. 2. The ROC from the different subsets of the new features VS. different modelling techniques.

4.3. Results and discussion

After the three sets of experiments were completed, we obtained the prediction results in terms of the prediction rates – false churn rate (FP) and true churn rate (TP). Based on these pairs of FP and TP, the ROC were plotted into Figs. 1(a), 2 and 4. The *x*-axis and *y*-axis of each subfigure of these figures present the prediction rates FP and TP, respectively. Each ROC curve of each subfigure consists of a sequence of points, each of which presents a pair of prediction rates (FP, TP) for a specified sampling rate. Thus, a decision maker can easily use a ROC to select the his/her interested prediction rates (FP and TP) from a subset of features and a modelling technique. Generally, in each ROC subfigure, the curve which is closer to the left-top corner presents the better prediction result. However, occasionally, it is difficult to decide which curve is better than others in most of the figures (e.g. the curve from incoming calls the one from call details in Fig. 2(f)). Therefore, the AUC technique was used to calculate the area under each ROC curve. The values of the AUC for each feature subset for each modelling technique were plotted into Figs. 1(b), 3 and 6. With the ROC figures, these AUC figures can show which modelling technique or which feature set is the best one. The *y*-axis and *x*-axis of each AUC figure/subfigure, respectively represent the AUC values and the different feature sets or modelling techniques.

4.3.1. Result I

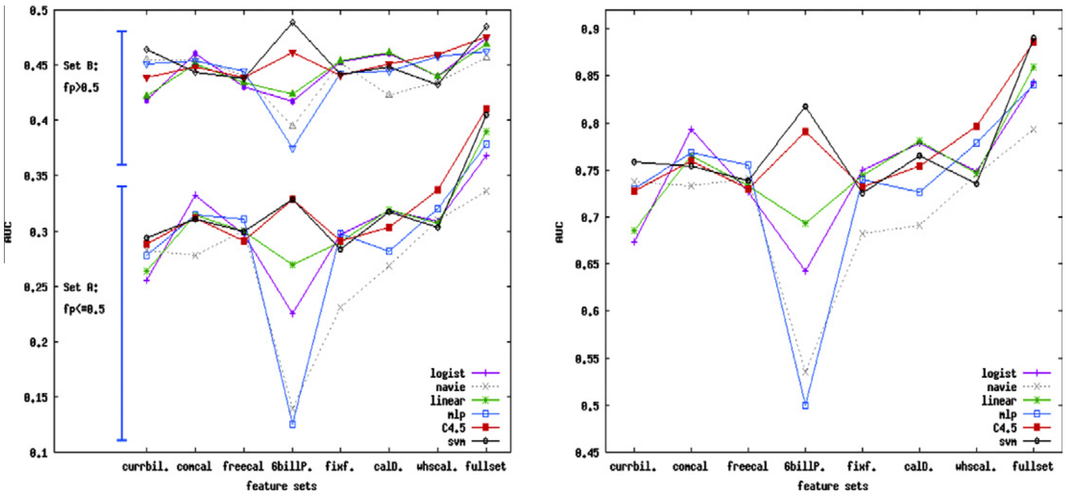
Fig. 1 shows the experimental results of the first set of experiments, which evaluate the feature subsets of the call detail information for the prediction. There are two subfigures in this figure: the one on the left hand side shows ROC curves, the other shows the relative AUC values. In Fig. 1(a), each point represents one pair of the average FP and the average TP from the six modelling techniques (LR, BN, LC, C4.5, MLP and SVM) for a certain sampling rate. There are 7 curves with different colours in this ROC figure, each of which represents a specified feature subset for the different sampling rates. The *y*-axis and *x*-axis of Fig. 1(b) represent the AUC values and different types of called details, respectively. There are three curves with different colours in this subfigure: the red, gray and black were plotted when $FP \leqslant 50\%$, $FP > 50\%$ and $FP \geqslant 0$, respectively. These two Fig. 1(a) and (b) together shows:

- Which feature subset of call details is more effective for churn prediction. For example, our presented feature subset of call details “CDLMNI” is the best one.
- What type of call detail is more significant for churn prediction (e.g. the most useful one is the information of local calls, and the information of international calls is less useful.)
- When the sampling rate $\frac{n\_churners}{n\_nonchurners}$ decreases (or the number of nonchurners increases), the false churn rate can reduce. But the true churn rate may reduce.
- When FP is less than about 5%, TP is very low for any feature subset (or call details) or any modelling technique. When $FP > 50\%$, the AUC values of any ROC curves are approximately equal. However, When $FP \leqslant 50\%$ or $FP \leqslant 100\%$, the AUC values are very different. Thus, most of the dicision makers might more consider he AUC values when $FP \leqslant 50\%$ or $FP \leqslant 100\%$.
- When the sampling rate $\frac{n\_churners}{n\_nonchurners}$ is very high (about 10), the prediction rates from all subsets of features are similar (high false churn rates and high true churn rates).

4.3.2. Result II

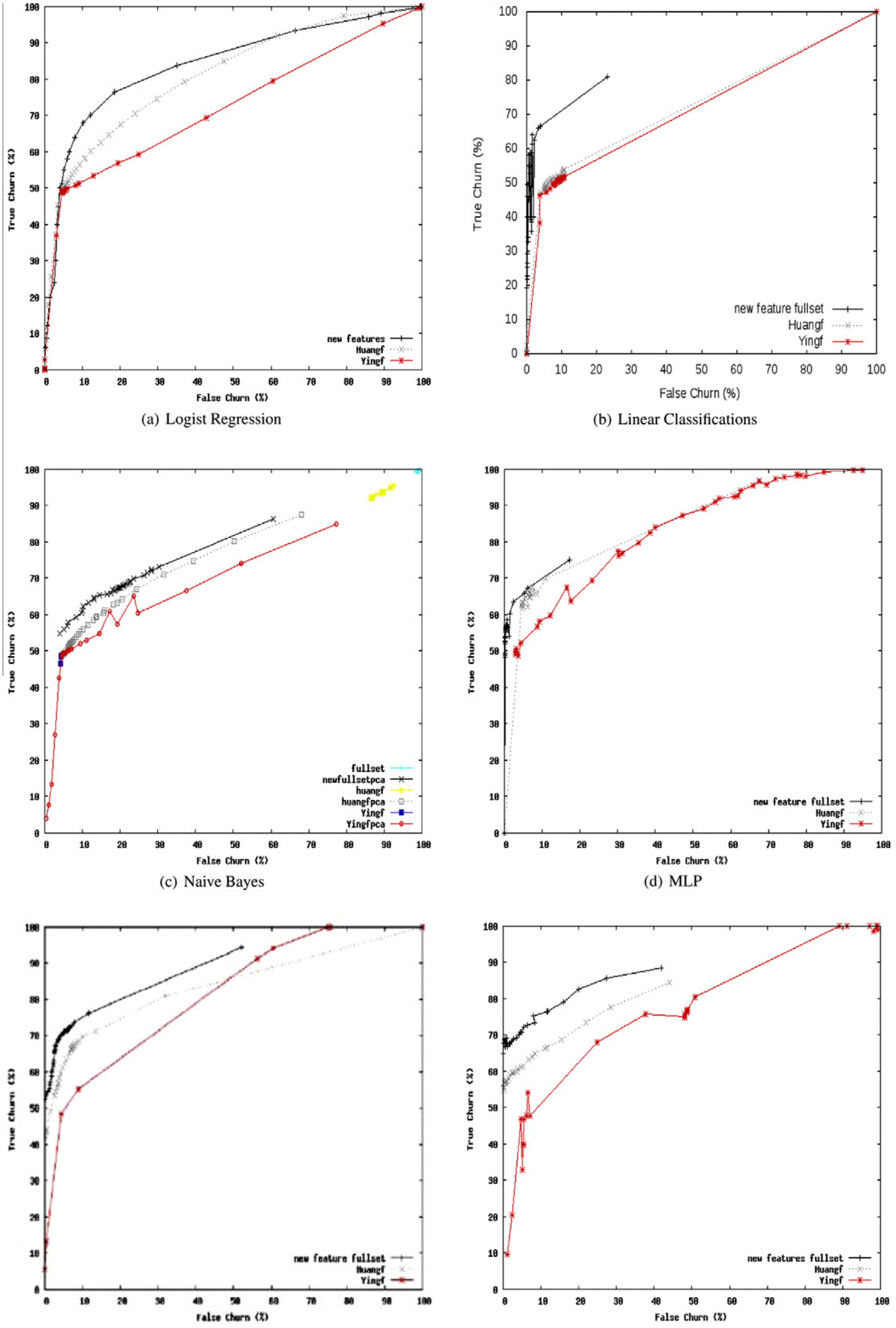
Fig. 2 shows the ROC curves which are based on the prediction rate pairs (FP and TP) when the six prediction modelling techniques and the different subsets of the new features were used in the second set of experiments. That is, Fig. 2(a)–(f) show the ROC curves when the prediction modelling techniques LR, LC, BN, MLP, SVM and C4.5 were, respectively used, based on the 8 subsets of the new features. These 8 subsets of features are the information of current bills, the information of the 6 bills with payments, the information of incoming calls (received calls), the fixed features that usually are not changed monthly or daily (e.g. account information, demographic profiles, telephone line information, Henley segments and grant information), the features of free calls, the features of call-details, the features of whole sale line calls, and the fullset of the new features. Based on these ROC curves, the AUC values were calculated and plotted into Fig. 3. Each subfigure of Fig. 2 shows the curve which is the closest to the left-top corner is from the fullset of the new features. This presents that the fullset is the best prediction results in term of the highest TP rates with the lowest FP when the same prediction modelling technique was used. Excepting to consider the fullset of the new features, the subfigures of Fig. 2 also show:

- If $FP < 5\%$, the FPs obtained by using the subsets of fixed features, current bills and the information of the 6-bills with payments are higher.
- Which feature subset would obtain the better results (the high TP and low FP) relies on the modelling techniques (for example, Fig. 2(e) shows the prediction results by using the incoming



(a) $FP > 0.5$ VS. $FP <= 0.5$ (b) $0 =< FP <= 1$

Fig. 3. The AUC from different subsets of the new features with different different modelling techniques.



(a) Logist Regression

(b) Linear Classifications

(c) Naive Bayes

(d) MLP

(e) SVM

(f) C4.5

Fig. 4. The ROC from the new features and the different existing feature sets VS. different modelling techniques.

calls is light better than using WHS calls for SVM, whilest Fig. 2(f) shows the prediction results of using WHS calls is are better).

- When the number of non-churner training samples increased, the TP and FP rates generally reduced. This represents that the sampling techniques can change the TP and FP.

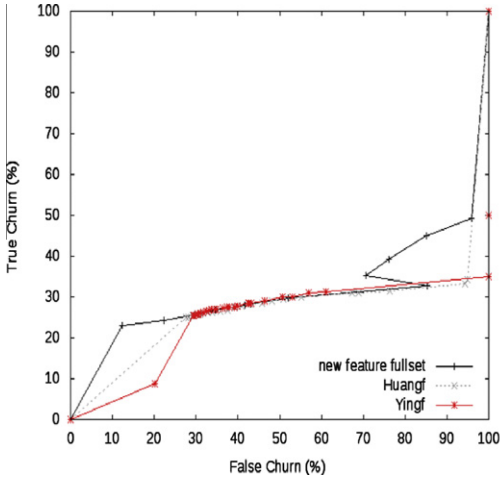


Fig. 5. The ROC from different feature subsets VS. the DMEL modelling techniques.

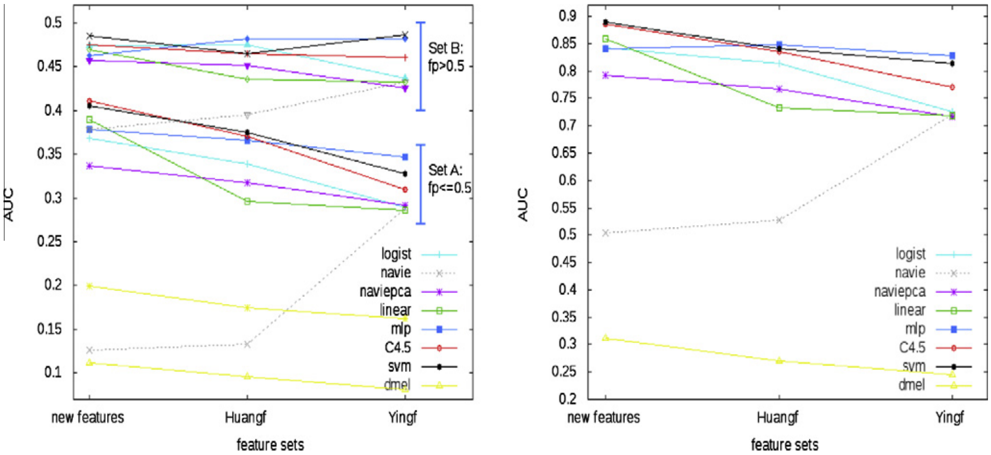
As mentioned above, the AUC of each ROC curve in Fig. 2 was calculated and plotted into Fig. 3. The *x*-axis and *y*-axis in each subfigure of Fig. 3 present the feature subsets and the AUC values, respectively. A point in Fig. 3 presents the AUC value of an relative feature subset and a prediction modelling technique. Usually, the higher values of AUC represent which feature subsets with the relative modelling techniques are better than others from this AUC figure. In Fig. 3(a), there are two sets of AUC curves: one set which is called "set A" was calculated when FP $\leqslant$50%, the other which is called "set B" was obtained when 50% < FP $\leqslant$ 100%. In addition, Fig. 3(b) shows the AUC when 0 $\leqslant$ FP $\leqslant$ 100%. The two subfigures of Fig. 3 shows:

- The AUC values by using the fullset of the new features are highest amongst the feature subsets. This indicates that the fullset are the best features for the churn prediction.
- The set B of AUC curves is smoother than the set A of AUC curves when the same prediction modelling technique was used. It implies that the prediction results (e.g. TP) from any two subsets of features is changed little when the FP > 50%. Usually, a decision maker may be interested in the low FP which is less than 50%. Therefore, the set A of AUC curves with the AUC curve set in Fig. 3(b) more accurately evaluates the feature subsets or the prediction modelling techniques.
- Which feature subset is the best to improve prediction results depends on which modelling technique is used in the prediction (excluding the fullset of features). For example, the features of the 6-bill with payment is the best for the SVM and C4.5. But it is not the best feature subset for the LC, LR, MLP and NB.
- Based on feature subset "6-bill & payments", the values (points) of AUC are very different for different modelling techniques. Especially, the values of AUC which are from NB and MLP are much lower. Two of the reasons for this may be that the size (the number of features) of this feature subset is highest and the data is less sensitive for NB and MLP. To solve this problem for the NB or MLP, the features should be transformed into low dimensions.
- In these prediction modelling, the results from SVM and C4.5 are the best. The SVM and C4.5 perform out other modelling techniques for the churn prediction.

4.3.3. Result III

When the third set of experiments was completed, which used three different feature sets (two existing feature sets and our new feature set) and seven prediction modelling techniques (LR, NB, LC, MLP, C4.5, SVM and DMEL), the prediction rates (FP and TP) were obtained. The ROC curves were plotted in Figs. 4 and 5, based on these FP and TP. Fig. 4(a)–(f) show the ROC curves when using LR, LC, NB, MLP and SVM, respectively. Fig. 5 shows the ROC curves when DMEL was used. In each subfigure of Fig. 4 or Fig. 5, there are at least three curves with different colours: the red colour curve represents the prediction result when using the Ying's feature set (Zhang et al., 2006), the gray curve represents the prediction results obtained by using the Huang's feature set (Huang et al., 2010), the black curve represents the prediction results obtained by using our new feature set. Each subfigure of Fig. 4 generally shows: (1) the black curve is the closet to the left-top corner in the figure. It implies the prediction results obtained by our new data set are the best (the highest TP with the lowest FP), (2) Similarly, the prediction result obtained by Huang's feature set is better than the Ying's feature set.

In Fig. 4(c), there are six ROC curves with different colours: (1) The black one was obtained when the data transformed from the new features by PCA was used for the modelling techniques. (2) The gray one was obtained when PCA was used to transform data, based on Huang's feature set. (3) The red one was obtained when PCA was used to transform data, based on Ying's feature set. (4) The green one was obtained by using the new feature set without any transformation. (5) The blue one was obtained by using Ying's feature set without any transformation. (6) The yellow one was ob-



(a) $FP > 0.5$ and $FP <= 0.5$

(b) $0 =< FP <= 1$

Fig. 6. The AUC from the new features and the existing feature sets with different modelling techniques.

tained by using Huang's feature set without any transformation. Fig. 4(c) shows that Naive Bayes obtained the bad prediction results when the features were not transformed. The reasons may be (1) high dimension features, (2) some of the attribute (feature) value for a class is zero (as above mentioned). This is why we used PCA to transform data for the Naive Bayes in our experiments. Therefore, we recommend that the input features or data with high dimensions to the Naive Bayes modelling techniques should be transformed for the churn prediction. The relative AUC values of these six ROC curves were calculated and plotted into Fig. 4, which will be discussed later.

Fig. 5 shows the ROC curves when the DMEL modelling technique and the three feature sets (new feature set, Huang's feature set and Ying's feature sets) were used. Contrasting to the ROC curves in Fig. 4, the FP and TP in the ROC curves in this figure are much more worst. That is, DMEL obtained the worst prediction rates in term of lowest TP with high FP. Therefore, DMEL techniques were not used to evaluate the subset of new features in the first and second sets of experiments.

In order to more preciously evaluate the effect of feature sets and modelling techniques, the AUC technique was used to calculate the AUC for the curves in each subfigure of Figs. 4 and 5. The relative AUC values for the ROC curves were plotted in Fig. 6. In each subfigure of Fig. 6, the x-axis and y-axis, respectively represent the different feature sets used in the third set of experiments (the feature sets are our new feature set, Huang's feature set and Ying's feature set) and the values of AUC. The AUC curves with different colours represent that different modelling techniques were used in the experiments. That is, in each subfigure of Fig. 6, different modelling techniques used in the experiments are labelled by the relative names. For example, the gray curves in Fig. 6 were obtained when the feature sets and Naive Bayes modelling technique were used for the churn prediction, and the relative label is "navie". For the label "naviepca", it implies Naive Bayes modelling techniques and PCA method were used for the prediction. In addition, the yellow curves in Fig. 6 were obtained from the ROC curves in Fig. 5. Two yellow curves are in Fig. 6(a): the one that consists of higher AUC values was obtained when FP > 50%, the other with lower AUC values was obtained when FP $\leqslant$ 50%. Therefore, a point in a AUC curve represents: (1) what AUC values is, (2) what feature set and what modelling technique were used in the experiments. Fig. 6(a) shows the AUC for each ROC curve when the following two intervals were selected: FP $\leqslant$ 50% and 50% < FP $\leqslant$ 100%. In this subfigure, there are two sets of AUC curves: set A was obtained when FP $\leqslant$ 50%, and set B was obtained when 50% < FP $\leqslant$ 100%. The AUC of each ROC curve is plotted in Fig. 6(b) when 0 $\leqslant$ FP $\leqslant$ 100%. From Fig. 6(a), the AUC curves of set B is smoother than the ones in set A. That is, values of AUC between using different feature set are less changed when FP > 50%. Thus, it implies the new features and the existing feature sets are approximately equally effective for the churn prediction when FP is very high. Actually, the expected FP is usually low for most decision makers (e.g. FP $\leqslant$ 50%). Therefore, set A of AUC and the ones of Fig. 6(b) may be more useful to evaluate feature sets or modelling techniques. The AUC curves of set A in Fig. 6(a) and the AUC curves in Fig. 6(b) mainly show:

- The AUC values from our new feature set are higher than the existing feature sets of other research when the same prediction modelling technique was used. This indicates that our new feature set is the best for the churn prediction.
- Which prediction modelling technique is more effective for churn prediction for a specified feature set. For examples, the C4.5 and SVM obtained higher values of AUC for the new feature set; that is, the C4.5 and SVM modelling techniques are more effective when using the new data set.
- The AUC values from DMEL are the lowest for any feature set. It implies that the DMEL modellig technique does not outperform most of the traditional modelling techniques. On the other hand, this technique is very sensitive with computation (see below). This is why the DMEL was not used to evaluate subset of features in the first and second set of experiments. We believe that DMEL is not suitable for this application.
- The prediction rates obtained by Naive Bayes modelling technique are very bad when the data with a large number of features were not transformed into low dimension. The prediction results obtained by Naive Bayes modelling technique are acceptable when the high dimensional data was transformed into low dimension.

The 7 modelling techniques have different computational complexity in the prediction. The computational cost of using the NB and DT is lowest. the computational overhead of using training the LC and LR is lower. The computational cost of using the SVM is expensive. However, the computational cost spent on the DMEL and MLP are much more expensive than others. As above mentioned, DMEL provided very poor prediction results. Therefore, the MLP and DMEL modelling techniques are very impractical for the churn prediction application on a very large dataset.

In addition, the outputs of these modelling techniques are different. DMEL can provide churn reasons and likelihood. DT can only provide churn reasons. LR, NB and MLP can give the likelihood/probability for customer behaviours. The SVM and LC can provide only binary output which presents churn or nonchurn. Therefore, which types of modelling techniques should be used depends on the objectives of the application. For example, if interested in churn reasons, the DT should be used; if the probabilities of churns and nonchurns is required, the NB, LR might be more suitable to use. If only interested in prediction rates, LC and SVM can be used.

5. Conclusions

This paper presented a new set of features for the customer churn prediction in the telecommunication, including the aggregated call details, Henley segmentation, account information, bill information, dial types, line-information, payment information, complain information, service information, and so on. Then seven modelling techniques (LR, LC, DT, MLP, SVM and DMEL) were used as predictors in this paper. Finally, based on the new feature set, the existing feature sets and the seven modelling techniques, the comparative experiments were carried out. In the experiments, each subset of the new feature were evaluated and analysed. In addition, the experiments provided: (1) the comparative effectiveness of the seven modelling techniques, and (2) the comparative effectiveness between the new feature set and the existing feature sets. The experimental results showed that: (1) the new the proposed feature set is more effective for the prediction than the existing feature sets, (2) which modelling technique is more suitable for customer churn prediction depends on the objectives of decision makers (e.g. DT and SVM with a low ratio should be used if interested in the true churn rate and false churn rate; the Logistic Regressions might be used if looking for the churn probability), (3) DMEL modelling techniques is impractical and ineffective for churn prediction on a large dataset with high dimension, (4) the high dimensional data for NB modelling technique is necessarily transformed into the low dimension, (5) what sampling ratio $\left(\frac{n\_churners}{n\_nonchurners}\right)$ is more suitable for the prediction depends on the objectives of decision makers to use a modelling technique (e.g. if more interested in true churn rates (or false churn rate), the ratio should become increase (or decrease)), and (6) the call detail combination feature set which consists of the feature sets of local, national, international, mobile phone and the sum of all of call details, can obtain higher true churn rates with lower false churn rates than any individual subset.

However, there are some limitations with our proposed techniques. In the future, other information (e.g. complain information, contract information, more fault reports, etc.) should be added into the new feature set in such a way to improve features. The dimensions of input features also should be reduced by using feature selection and extraction techniques which will be studied in the future. In addition, the imbalance classification problem takes place in this application and we only used the sampling technique to attempt to solve the problem. Therefore, more methods for imbalance classifications also should be focused in the future.

Acknowledgement

This research was partly supported by Telecom of Ireland.

References

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

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings the 5th annual ACM workshop on computational learning theory* (pp. 144–152). Pittsburgh, PA: ACM Press.

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

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

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

Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning, 29*(2–3), 103–130.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Churn prediction: Does technology matter? *International Journal of Intelligent Technology, 1*(2).

Henley, 2009. <http://www.henleymc.ac.uk/>.

Hosmer, D., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Huang, B., Kechadi, M.-T., & Buckley, B. (2010). A new feature set with new window techniques for customer churn prediction in land-line telecommunication. *Expert Systems with Applications, 37*(5), 3657–3665.

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

Japkowicz, N. (2000). *Learning from imbalanced data sets: A comparison of various strategies* (pp. 10–15). AAAI Press.

Japkowicz, N. (2006). Why question machine learning evaluation methods? In *AAAI Workshop. Boston*.

John, H., Ashutosh, T., Rajkumar, R., Dymitr, R. (2007). Computer assisted customer churn management: State-of-the-art and future trends.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the 10th national conference on ARTI CIAL intelligence* (pp. 223–228). MIT Press.

Luo, B., Shao, P., Liu, J. 2007. Customer churn prediction based on the decision tree in personal handyphone system service. In *International conference on service systems and service management* (pp. 1–5).

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufman Publishers.

Quinlan, J. R. (1996). Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research, 4*, 77–90.

Rumelhart, D., Hinton, G., & Williams, R. (1986). *Learning internal representations by error propagation* (Vol. 1). MA: MIT Press.

Vapnik, V.N. (1998). *The nature of statistical learning theory* (2nd ed., pp. 23–57).

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

Yan, L., Wolniewicz, R., & Dodier, R. (2004). Customer behavior prediction – it's all in the timing. *Potentials IEEE, 23*(4), 20–25.

Zhang, Y., Qi, J., Shu, H., Li, Y. (2006). Case study on crm: Detecting likely churners with limited information of fixed-line subscriber. In *2006 International conference on service systems and service management* (Vol. 2, pp. 1495–1500).