Customer Churn Prediction for Broadband Internet Services

B.Q. Huang[1], M-T. Kechadi[1], and B. Buckley[2]

[1] School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
[2] Eircom Limited, 1 Heuston South Quarter, Dublin 8, Ireland

Abstract. Although churn prediction has been an area of research in the voice branch of telecommunications services, more focused studies on the huge growth area of Broadband Internet services are limited. Therefore, this paper presents a new set of features for broadband Internet customer churn prediction, based on Henley segments, the broadband usage, dial types, the spend of dial-up, line-information, bill and payment information, account information. Then the four prediction techniques (Logistic Regressions, Decision Trees, Multilayer Perceptron Neural Networks and Support Vector Machines) are applied in customer churn, based on the new features. Finally, the evaluation of new features and a comparative analysis of the predictors are made for broadband customer churn prediction. The experimental results show that the new features with these four modelling techniques are efficient for customer churn prediction in the broadband service field.

1 Introduction

Services 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 service 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 [4,16,15,3,11,7,17]. As one of the important measures to retain customers, churn prediction has been a concern in the telecommunication industry and research [3]. Until now the majority of churn prediction has been focused on voice services available over mobile and fixed-line networks. Most of the literature introduces the usage of variables/features (which are customer demographics, contractual data, customer service logs, call details, complain data, bill and payment, structure of monthly service fees, as so on)[3,8,11,15,16,18], and the common modelling techniques (which are Logistic Regressions (LR) Decision Trees (DT), Artificial Neural Networks (ANN) and Random Forest (RF)) [7,8,11,16,19].

T.B. Pedersen, M.K. Mohania, and A M. Tjoa (Eds.): DaWaK 2009, LNCS 5691, pp. 229–243, 2009.

Broadband Internet services are potentially one of the greatest sources of revenue for providers and consequently feature highly in their marketing campaigns. However, the above techniques have not been applied to the specific area of churn prediction in broadband service field. Until now either very little churn prediction has been carried out on the broadband Internet services over fixed-line networks, or the literature of churn prediction in telecommunication does not provide the details of methodologies for churn prediction using broadband information [4,16,15,3,11,7,17]. Therefore, it is necessary to investigate the churn prediction in Broadband Internet service field.

This paper presents a new set of features with four modelling techniques for customer churn prediction in one telecommunication service field – broadband Internet. The new set of features are extracted from Henley segmentation, broadband usage, dial types, the spend of dial-up, line-information, bill and payment information, account information, call details and service log data. The modelling techniques used to predict churns are LR, DT, ANN and Support Vector Machines (SVM). Finally, based on the proposed features and the modelling techniques, experiments are carried out. The experimental results show that the presented features with the modelling techniques are efficient for broadband customer churn prediction.

The rest of this paper is organised as following: next section introduces the evaluation criterias 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 Criterias

After a classifier or predictor is available, it will be used to predict the further behaviour of customers. As one of important step to ensure the model generalise well, the performance of the predictive churn model have to be evaluated. Table 1 shows a confusion matrix [10], 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, the most common evaluation criterias for a predictive model are introduced as follows:

– The overall accuracy (AC) is the proportion of the total number of predictions that were correct, calculated by $\frac{a_{11}+a_{22}}{a_{11}+a_{12}+a_{21}+a_{22}}$.

Table 1. Confusion Matrix

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

- The accuracy of true nonchurn (TN) is the proportion of nonchurn cases that were correctly identified, written as $\frac{a_{22}}{a_{21}+a_{22}}$.
- The accuracy of true churn (TP) is defined as the proportion of churn cases that were classified correctly, calculated by $\frac{a_{11}}{a_{11}+a_{12}}$.
- The false churn rate (FP) is the proportion of nonchurn cases that were incorrectly classified as churn, written as $\frac{a_{21}}{a_{21}+a_{22}}$.
- The false nonchurn rate (FN) is the proportion of churn cases that were incorrectly classified as nonchurn, written as $\frac{a_{12}}{a_{12}+a_{11}}$.

There are other evaluation criterias and the details of them can be found in [10]. In this paper, we are more interested in the high accuracy of true churn and the low false churn rate.

3 Methodology

The proposed churn prediction system for broadband Internet consists of sampling data, preprocessing, and classification/prediction phases. Data sampling randomly selects a set of customers and their relative information, according the definition of churn. The preprocessing (also called data preparation) includes data cleaning, feature extraction and normalisation steps. The main task of data cleaning is to remove 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 task of feature extraction is to select features to address customers. The process of normalisation is to normalise the values of features into a range. The task of prediction phase is to predict the further behaviour of customers. The following subsections describe the features/variable extraction, normalisation and prediction/classification steps.

3.1 Feature/Variable Extraction

The feature extraction plays the most important role which can directly influence the performance of predictive models in the term of prediction rates. If a robust set of features is extracted in this phase, a significant improvement will be yielded. However, it is not easy to obtain such a set of features. Until now, most of the feature sets have been introduced for churn prediction in mobile telecoms industry [8,11,16,3,15] and fixed-line telecommunication [3,18]. However, in these existing feature sets, the broadband Internet information is not included. Thus, it is difficult to use the existing feature set for churn prediction in broadband Internet service field. Based on broadband Internet service information, the following features are selected for broadband Internet churn prediction in telecommunication:

- **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 available demographic information for this research is gender and country. Therefore, these information may be useful for predicting the further behaviour of a customer.

- **Information of grants**: Some customers have obtained some special grants resulting in their bills being paid fully or partly by other parties. For example, a customer with a disability or over 80 are more unlikely to churn from that service.
- **Account information**: includes the account status, creation date, the bill frequency, the service packages, the account balance, payment types, dial-up types, dial-up cost, broadband opening date, download and upload capacity, total duration usage, average download and upload speeds, and general service usage information which includes the summarised call duration, the number of calls and standard prices, current outstanding charges and charges paid. Account information is very useful for predicting the customer behaviour for the next observation period.

Based on these new features, the average of call duration, the number of calls, the standard prices and the actual fees payed for 30 days (note the duration is a number of minutes) are also considered as new features. Let "$\overline{D_N}$", "$\overline{C_N}$" , "$\overline{SP_N}$" and "$\overline{FP_N}$" be the average of call duration, the number of calls, the standard prices and the fees payed in 30 days of the most recent bill, respectively. They are obtained by equation 1:

$$\begin{aligned} \overline{C_N} &= \frac{nCalls\_M}{nDays} * 30 \\ \overline{D_N} &= \frac{Duration\_M}{nDays} * 30 \\ \overline{SP_N} &= \frac{Fees\_M}{nDays} * 30 \\ \overline{FP_N} &= \frac{Fee\_C}{nDays} * 30 \end{aligned} \tag{1}$$

where "$nCalls\_M$" is the number of calls in the most recent bill, "$Duration\_M$" is the duration of the most recent bill, "$Fees\_M$" is the fees of the most recent bill, "Fee_C" is the fees from customers, and "$nDays$" is the number of day of the bill, which can be obtained by equation 2.

$$nDays = endDate - startDate \tag{2}$$

where "$endDate$" and "$startDate$" are the dates of bill starting and ending. In addition, the ratio between the actual fees that should be pay and the call-duration of the current bill is extracted as a new feature, which is written as equation 3.

$$R\_AMNT\_DUR = \frac{Fees\_M}{Duration\_M} \tag{3}$$

- **Service orders**: describe the services ordered by the customer. The quantity of the ordered services, the rental charges are selected as new features.
- **Henley segments**: The algorithm of Henley segmentation [2] splits 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.
- **Broadband Internet and telephone line information**: this includes information about voice mail service (provided or not), the number of broadband lines, the number of telephone lines, and so on. The customers who have more telephone or broadband Internet 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 and broadband lines, and the voice mail service indicator are selected as part of new features.
- **The historical information of payments and bills** : this concerns the billing information for each customer and service for a certain number of years. Each bill includes the total cost, prices, rental charges, call duration, charges paid so far, etc. Attributes monthly cost, rental charges, call duration and paid charges are extracted as new features. They are denoted by "mnCost", "mnRent_fees", "mnDur" and "paidfee", respectively. New other features are also created; the changed-cost, changed call-duration and rental changed-fees and are included in the set of new features. They are obtained by equation 4.

$$
\begin{aligned}
changed\_cost_{i,i-1} &= \frac{|mnCost_i - mnCost_{i-1}|}{\sum_{j=2}^{T} |mnCost_j - mnCost_{j-1}|} \\
changed\_Duration_{i,i-1} &= \frac{|mnDur_i - mnDur_{i-1}|}{\sum_{j=2}^{T} |mnDur_j - mnDur_{j-1}|} \\
changed\_rental\_Fees_{i,i-1} &= \frac{|mnRent\_fees_i - mnRent\_fees_i|}{\sum_{j=2}^{T} |mnRent\_fees_j - mnRent\_fees_j|}
\end{aligned}
\tag{4}
$$

  where mnCost_i, mnDuration_i and mnRent_fees_i are the cost, call-duration and rental fees of the bill for the month i^{th}.
- **Broadband monthly usage information**: This information is used to record the details of the broadband monthly usage for each customer. Monthly information can show frequency of broadband use, total upload/download, connection duration. The following types of customers may often churn: i) those who have short online sessions, ii)Those who have small upload/ download totals, iii) those who have greatly fluctuating monthly usage figures. Therefore features which capture this information must be included.

Therefore, some new features should be extracted from the usage information of broadband Internet for churn prediction in telecommunication service fields, especially in broadband Internet service field. Based on this information, the new extracted features are the sizes of the information downloaded and uploaded, the duration of using Internet every month, the changed sizes of the information downloaded and uploaded, and the changed the online duration of every consecutive two month, and the ratio between the total sizes of information downloaded/uploaded and the duration of online broadband Internet for a month.

Consider the sizes of the information downloaded and uploaded, the duration for month i are "DOW$_i$", "UP$_i$" and "ONT$_i$", respectively. If the change sizes of information downloaded and uploaded, and the duration of online on Internet are "CH_DOW$_{i,i-1}$", "CH_UP$_{i,i-1}$" and "CH_ONT$_{i,i-1}$" respectively, they can be calculated by equations (5), (6) and (7), respectively.

$$CH\_DOW_{i,i-1} = \frac{|DOW_i - DOW_{i-1}|}{\sum_{j=2}^{M'} |DOW_j - DOW_{j-1}|} \tag{5}$$

$$CH\_UP_{i,i-1} = \frac{|UP_i - UP_{i-1}|}{\sum_{j=2}^{M'} |UP_j - UP_{j-1}|} \tag{6}$$

$$CH\_ONT_{i,i-1} = \frac{|ONT_i - ONT_{i-1}|}{\sum_{j=2}^{M'} |ONT_j - ONT_{j-1}|} \tag{7}$$

Consider the ratio between the total sizes of information downloaded/ uploaded and the duration of online broadband Internet for month i is "R_GB_ONT$_i$". The ratio can be calculated by equation (8).

$$R\_GB\_ONT_i = \frac{DOW_i + UP_i}{ONT_i} \tag{8}$$

– **Call details**: If the customers did not use the services, he might cease the services in the future. If the fees of services are suddenly increased or decreased, the customer might cease the services sooner. Call-details can reflect this information – how often customer have used the services with relative payment, and so on. The use of call details in churn prediction is reported in [16,18]. The call-detail contain call duration, price and types of call (e.g. International or local call) of every call. It is difficult to store all call details of every call every month for every customer. Most of the telecommunication companies keep the call details of a few months. The limited call details can be used for churn prediction in telecommunication.

Based on these month call details, the aggregated number of calls, duration and fees are extracted as new features. The basic idea for extracting features is to segment the call details into a number of defined periods, then to aggregate the duration, fees and the number of calls for each period for every customer. In literature [16,18], it is reported that the call-details of every 15 or 20 days are efficient. In this paper, the six-month call details

are segmented into 15-day period, then number of calls, duration and fees of each 15-day period are aggregated for each customer. For a segment i of a customer's call details, let the aggregated number of calls, duration and fees be "CALL_N$_i$", "DUR$_i$" and "COST$_i$", respectively. The changed number of calls, changed-duration and changed-cost between two consecutive segment of call details can be obtained by Equation 9.

$$\begin{aligned} CH\_DUR_{i,i-1} &= \frac{|DUR_i - DUR_{i-1}|}{\sum_{j=2}^{M'} |DUR_j - DUR_{j-1}|} \\ CH\_N_{i,i-1} &= \frac{|CALL\_N_i - CALL\_N_{i-1}|}{\sum_{j=2}^{M'} |CALL\_N_j - CALL\_N_{j-1}|} \\ CH\_C_{i,i-1} &= \frac{|COST_i - COST_{i-1}|}{\sum_{j=2}^{M'} |COST_j - COST_{j-1}|} \end{aligned} \tag{9}$$

where M' is the number of call-detail segments; i and j are the indexes of call-detail segment, and $2 =< i,\ j <= M'$. In addition, the increment rates of the number of calls, duration and fees are calculated Equation 10.

$$\begin{aligned} R\_DUR_{i,i-1} &= \frac{CH\_DUR_{i,i-1}}{CH\_DUR_{i,i-1} + DUR_{i-1}} \\ R\_N_{i,i-1} &= \frac{CH\_N_{i,i-1}}{CH\_N_{i,i-1} + CALL\_N_{i-1}} \\ R\_C_{i,i-1} &= \frac{CH\_C_{i,i-1}}{CH\_C_{i,i-1} + COST_{i-1}} \end{aligned} \tag{10}$$

Thus, the new features includes the number of calls, duration, fees, the changed number of calls, changed-duration, changed-fees, the rates of the increased number of calls, the rates of the increased duration and the rates of the increased fees.

3.2 Normalisation

In the extracted features (See subsection 3.1), some predictors or classifiers (e.g. Artificial Neural Networks) have difficulties in accepting the string values of features, such as, genders, county names. The value of a feature was rewritten into binary strings.

In addition, the values of each of these features (e.g the number of lines, the sizes of information downloaded/uploaded, the duration of online on Internet, "R_GB4_ONT$_i$", "$\overline{C_N}$", "CALL_N$_1$", "DUR$_1$", "COST$_1$", "CALL_N$_M$", "mnDur$_1$" and "paidfee$_1$", etc.), lie in different dynamical ranges. The large values of 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

design. To solve this problem, the values of these features can be normalised into a similar range by Equation 11.

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij},\ j = 1, 2, \cdots, \iota$$

$$\sigma_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)$$

$$y = \frac{x_{ij} - \bar{x}_j}{r\sigma_j}$$

$$\tilde{x}_{ij} = \frac{1}{1+e^{-y}} \tag{11}$$

Where x_j is the feature j^{th}, ι 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 by one.

3.3 Prediction/Classification

Many techniques have been proposed for churn prediction in telecommunication. Three popular modelling techniques (Logistic Regression, Multilayer Perceptron neural networks and Decision Tree C 4.5) and one promising modelling technique (Support Vector Machines), are selected as predictors from the broadband churn prediction. These four modelling techniques are outlined as follows:

Logistic Regressions: Logistic regression [9] 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:

$$prob(y = 1) = \frac{e^{\beta_0 + \sum_{k=1}^{K}\beta_k x_k}}{1 - e^{\beta_0 + \sum_{k=1}^{K}\beta_k x_k}} \tag{12}$$

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,$\cdots$,x_K are the independent inputs. β_0,β_1, $\cdots$, β_K are the regression coefficients that can be estimated by the maximum likelihood method, based on the provided training data. The detail of the logistic regression models can be found in [9].

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-tree. Summarily, 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 [13,12].

Artificial neural networks: A 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:

$$Output_{net}(j) = f(\sum_{l=1}^{L} w_{jl} f(\sum_{i}^{D} w_{li} x_i)) \quad j = 1, \cdots, J \tag{13}$$

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. The more details with learning algorithm can be found on [14].

Support Vector Machines: An 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 [5], and then classifies the data by the maximal margin hyper-plane as following:

$$f(\boldsymbol{x}) = sgn\left(\sum_{i}^{M} y_i \alpha_i \phi(\boldsymbol{x_i}, \boldsymbol{x}) + \delta\right) \tag{14}$$

where M is the number of samples in the training set, $\boldsymbol{x_i}$ is a support vector with $\alpha_i > 0$, ϕ is a kernel function, $\boldsymbol{x}$ 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 [6]. Polynomial kernels and Gaussian radial basis functions (RBF) are usually applied in practice for kernel functions. δ can be obtained by taking into account the Karush-Kuhn-Tucker condition [6], and choosing any i for which $\alpha_i > 0$ (i.e. support vectors). However, it is safer in practice to take the average value of δ over all support vectors.

4 Experimental Results and Discussion

The 139000 customers were randomly selected from the real-world database provided by Eircom[1] in our experiments. In the training dataset, there are 6000 churners, 94000 nonchurners and total 100000 customers. In the testing dataset, there are 39000 customers which includes 2000 churners and 37000 nonchurners. Each customer is represented by the features which are described in Section 3.1. Based on the datasets, three sets of experiments were carried out in this papers, independently.

In the first set of experiments, a number of different feature subsets were used. The features that describe demographic profiles, information of grants, account

Algorithm 1. The procedure of an experiment for a feature subset

1. Select the subset of features
2. Base on the selected feature subset, load the data from the training set.
3. When the predictive model is available, load the data from testing set, based on the selected subset of features.
4. Evaluate the the outputs from the predictive model.

information, henley segments, broadband Internet and telephone line information, 6-month call details and details of broadband usage information depending on the number of months selected, were used. The broadband monthly usage information for a number of months is formed using the current months data in addition to all previous months data e.g. the 3-month data subset contains the data for month 3, 2, and 1 and the 7-month data subset contains the information for month 7, 6, 5, 4, 3, 2 and 1. In the first set of experiments, the number of months is between 1 and 11. Thus, 11 different subsets of features were used. For each subset of features, the general procedure of an experiment which is described by Algorithm 1 was carried out. Four prediction modelling techniques LR, DT, MLP and SVM were used for each subset of features.

For the second set of experiments, the features that describe the details of broadband usage information depending on the number of months selected (including the summary information of broadband usage on the bills), were used. Similarly, the same procedure of selecting feature subsets and the same number of months (from 1 to 11) used in the first set of experiments, were used for the second set of experiments. Therefore, the second set of experiments also used 11 different subsets of features. For each subset of features, the general procedure described by Algorithm 1 was applied to each experiment. In addition, in the second set of experiments, the same modelling techniques were used for each subset of features.

For the third set of experiments, the features without broadband usage information were used. For this subset of features, four prediction modelling techniques (LR, DT, MLP and SVM) were used. Based on this subset of features, the procedure (see Algorithm 1) was carried out for each of these modelling techniques.

In each subset of features, LR, DT, MLP and SVM were trained and tested. The training and testing datasets were not normalised for the DT, but were normalised for the LR, MLP or SVM. All the predictors were trained by 10 folds of cross-validations in each experiment.

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

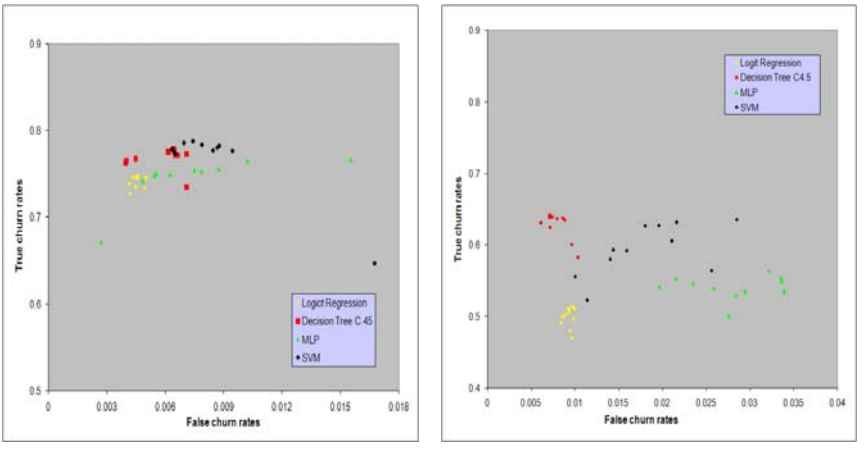

Fig. 1. ROC plot of false and true churn rates of the different number of months, where (a) and (b) presents the results of the first and second set of experiments, respectively

0.1, maximum cycle 1800 and tolerant error 0.05 were used to train the MLPs, based on the training dataset. The number of training cycles to yield the highest accuracy is about 600 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 $C \in \{2^8, 2^7, \cdots, 2^{-8}\}$ and $\sigma^2 \in \{2^{-8}, 2^{-7}, \cdots, 2^8\}$ of radial basis functions (RBF) were attempted to find the best parameters for the churn prediction. All together, 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 maximum classification accuracy of standard SVMs were $(2^{-6}, 2^8)$ for each set of experiments. The optimal parameter sets (C, σ^2) yielding a maximum classification accuracy of SVMs were $(2^{-6}, 2^7)$ for the first and second set of experiments. For the third set of experiments, the optimal parameter sets (C, σ^2) yielding a maximum classification accuracy of SVMs were $(2^{-5}, 2^8)$.

Table 3 shows the prediction rates (AC, TP, FP) for the third set of experiments, and Table 2 shows the prediction rates (AC, TP, FP) for the first and second sets of experiments. The prediction rates on the left hand side of Table 2 summarise the results of the four techniques (LR, DT, MLP and SVM) performed on the first set of experiments. The results on the right hand side of the Table 2 were obtained from the second set of experiments (that used broadband usage information only). Based on Table 2, Figures 1(a) and 1(b) plot the receiver operating characteristics (ROC) graphs, which are TP against FP for the first and second sets of experiments, respectively. A point in the plots presents a pair of prediction rates (FP, TP) from a modelling technique based on a subset of features. Figures 1(a) and Figures 1(b) present the prediction rates for the first and second sets of experiments, respectively. The results of LR, DT, MLP

Table 2. Prediction rates based on different datasets, where BU presents number months of broadband usage information

	All data/variables				Only Broadband usage info.			
	LR	DT	MLP	SVM	LR	DT	MLP	SVM
BU	1 month				1 month			
AC	98.27	98.42	98.19	98.26	96.59	97.39	95.18	96.77
TP	73.90	76.55	75.05	77.90	49.15	62.45	53.95	55.65
FP	0.42	0.40	0.55	0.64	0.84	0.72	2.59	1.01
BU	2 months				2 months			
AC	98.22	98.38	98.03	98.24	96.55	97.52	94.82	95.33
TP	73.55	76.75	75.45	77.60	51.15	63.10	53.50	56.45
FP	0.45	0.45	0.75	0.65	0.99	0.62	2.95	2.56
BU	3 months				3 months			
AC	98.29	98.41	98.03	98.22	96.62	97.46	95.78	96.06
TP	74.70	76.25	75.45	77.35	50.00	63.80	54.20	63.20
FP	0.43	0.40	0.75	0.65	0.86	0.72	1.97	2.16
BU	4 months				4 months			
AC	98.27	98.26	97.33	98.19	96.60	97.48	95.44	95.42
TP	74.80	77.50	76.70	77.70	50.30	64.10	54.60	63.55
FP	0.46	0.62	1.55	0.70	0.90	0.71	2.35	2.85
BU	5 months				5 months			
AC	98.22	98.24	97.83	98.17	96.61	97.47	95.66	96.37
TP	74.60	77.65	76.55	78.00	51.05	64.10	55.30	62.70
FP	0.50	0.65	1.02	0.74	0.93	0.73	2.16	1.81
BU	6 months				6 months			
AC	98.27	98.27	97.92	98.12	96.59	97.44	94.71	96.23
TP	74.60	77.55	75.60	77.90	51.40	63.90	56.40	62.75
FP	0.45	0.61	0.88	0.79	0.97	0.75	3.22	1.96
BU	7 months				7 months			
AC	98.26	98.26	97.98	98.05	96.60	97.38	94.53	95.98
TP	74.55	77.90	75.30	78.25	51.05	63.65	55.40	60.60
FP	0.46	0.64	0.79	0.88	0.94	0.80	3.35	2.11
BU	8 months				8 months			
AC	98.27	98.16	98.12	98.05	96.58	97.32	94.49	96.52
TP	74.80	77.25	74.95	78.05	50.60	63.75	54.90	58.05
FP	0.46	0.71	0.63	0.87	0.94	0.86	3.36	1.41
BU	9 months				9 months			
AC	98.22	98.21	98.22	98.05	96.49	97.28	94.40	96.55
TP	74.60	77.20	74.20	77.70	49.70	63.45	53.50	59.35
FP	0.51	0.65	0.49	0.85	0.98	0.89	3.39	1.44
BU	10 months				10 months			
AC	98.17	98.20	98.19	97.96	96.44	97.03	94.89	96.40
TP	73.40	77.15	74.85	77.65	48.05	60.05	53.00	59.25
FP	0.49	0.66	0.55	0.95	0.94	0.97	2.84	1.59
BU	11 months				11 months			
AC	98.21	97.96	98.06	96.60	96.37	96.87	94.82	96.47
TP	72.75	73.45	67.15	64.70	47.05	58.25	50.10	52.35
FP	0.42	0.71	0.27	1.68	0.96	1.04	2.76	1.14

Table 3. Prediction rates based on the data without broadband usage information

	LR	DT C4.5	MLP	SVM
AC	97.941	97.723	96.828	97.744
TP	69.600	72.300	69.200	72.950
FP	0.527	0.903	1.678	0.916

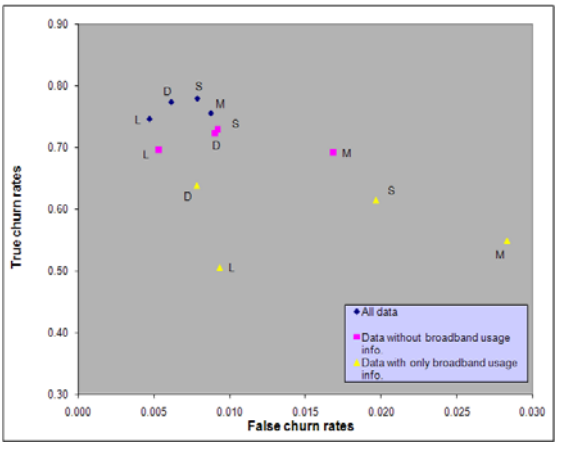

Fig. 2. ROC plot of false and true churn rates vs. the three different sets of data

and SVM models are shown in yellow, red, green and black, respectively, in these Figures. Table 2 Figures 1(a) and 1(b) show that:

1. The number of months of broadband usage information is between 3 to 9 to obtain better prediction rates
2. For the same subset of features, which type of modelling techniques would get higher prediction rates (TP) with lower prediction rates (FP) (e.g. the DT and SVM would get higher prediction rate (TP) than the LR and MLP; the DT and LR would get lower prediction rates (FP) than the SVM and MLP; the SVM would get slight high rates (TP) than than DT, etc.)
3. because the prediction rate (TP) is more significant when the FP is not very different, the SVM and DT outperform the LR and MLP.

As mentioned above, the number of months of broadband usage information is between 3 to 9 to obtain better prediction rates. In order to compare the efficiency of the feature subset that excludes broadband usage information with the feature subsets that include the broadband usage information, the average of prediction rates (FP, TP) for the months from 3 to 9 for each modelling technique were calculated. These pairs of the average prediction rates and the pairs of the prediction rates (FP,TP) from the third set of experiments are plotted in Figure 2: the yellow points are from the second set of experiment, the black points are from the first set of experiments, and the magenta points are from the third set of experiments. Figure 2 shows that:

1. For all the modeling techniques, the prediction rates for TP are reasonably high (above 73%) and the prediction rates for FP are very low (about max 1%)in the case of the first set of experiments.
2. The prediction rates (TP) are higher and the prediction rates (FP) are lower when all the information were used in the first set of experiments.
3. when the same data were used, the SVMs and DTs can obtain the highest prediction rates (TP) and lowest rates (FP); LR can get the lowest rates (TP), and MLP can provide the lower rates (TP) with highest rates (FP).
4. The prediction rates (TP, FP) obtained on the information without broadband information are also high (about TP of 71% and FP of 1.1%).
5. The prediction rates (TP, FP) obtained on the information with only broadband information are about 50% and 2%), which are quite good, considering the condition of the experiments, in which the number of customer that are churning is very low in comparison with the non-churners.

The computation overhead is very different when the different modelling techniques were used for the churn prediction. The most expensive computational cost was spent on using the MLP, the computational cost of using the SVM is more expensive, the lower and lowest ones were spent on using the DT and LR respectively. In addition, the outputs of these modelling techniques are different. DT can provide churn reasons without likelihood. LR and MLP can give the likelihood/probability for customer behaviour. The SVM can provide only binary output which presents churn or nonchurn. Therefore, which types of modelling techniques should be used depends on the objectives of an application. For examples, if interested in churn reasons, the DT should be used; if the probabilities of churns and nonchurns is required, the MLP or LR might be more suitable to use.

5 Conclusions

This paper presented a new set of features, based on Henley segmentation, the broadband usage, dial types, the spend of dial-up, line-information, bill and payment information, account information, call details and service log data. Four modelling techniques (LR, DT, MLP and SVM) were used for customer churn prediction in telecommunication service field, especially broadband Internet. Finally, based on this new set of features, the comparative experiments of the four modelling techniques were carried out. The experimental results showed that the high true churn of 77% with the low false churn rate of 2% can be achieved using the proposed features. Experimental results also showed which modelling technique is more suitable for broadband churn prediction depends on the objective of the churn prediction. For examples, DT and SVM 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.

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

the principal components analysis methods and genetic algorithms. In addition, because the imbalance classification problem takes place in this application, the methods of imbalance classifications should be focused in the future.

Acknowledgements

This research was partly supported by Eircom of Ireland.

References

1. http://www.eircom.ie/cgi-bin/bvsm/bveircom/mainPage.jsp
2. http://www.henleymc.ac.uk/
3. Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service (June 2007)
4. Au, W., Chan, C., Yao, X.: A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Transactions on Evolutionary Computation 7, 532–545 (2003)
5. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Pro. the 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh,PA, July 1992, pp. 144–152. ACM Press, New York (1992)
6. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
7. Coussement, K., den Poe, D.V.: Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34, 313–327 (2008)
8. Hadden, J., Tiwari, A., Roy, R., Ruta, D.: Churn prediction: Does technology matter? International Journal of Intelligent Technology 1(2) (2006)
9. Hosmer, D., Lemeshow, S.: Wiley, New York (1989)
10. Japkowicz, N.: Why question machine learning evaluation methods? In: AAAI Workshop (2006)
11. John, H., Ashutosh, T., Rajkumar, R., Dymitr, R.: Computer assisted customer churn management: State-of-the-art and future trends (2007)
12. Quinlan, J.R.: C4.5: Programs for machine learning (1993)
13. Quinlan, J.R.: Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research 4, 77–90 (1996)
14. Rumelhart, D., Hinton, G., Williams, R.: Learning internal representations by error propagation, vol. 1. MIT Press, MA (1986)
15. Wang, H.-Y., Hung, S.-Y., Yen, D.C.: Applying data mining to telecom churn management. Expert Systems with Applications 31, 515–524 (2006)
16. Wei, C., Chiu, I.: Turning telecommunications call details to churn prediction: a data mining approach. Expert Systems with Applications 23, 103–112 (2002)
17. Yan, L., Wolniewicz, R., Dodier, R.: Customer behavior prediction - it's all in the timing. Potentials, IEEE 23(4), 20–25 (2004)
18. Zhang, Y., Qi, J., Shu, H., Li, Y.: Case study on crm: Detecting likely churners with limited information of fixed-line subscriber. In: 2006 International Conference on Service Systems and Service Management, October 2006, vol. 2, pp. 1495–1500 (2006)
19. Zhao, Y., Li, B., Li, X., Liu, W., Ren, S.: Customer churn prediction using improved one-class support vector machine. In: Li, X., Wang, S., Dong, Z.Y. (eds.) ADMA 2005. LNCS, vol. 3584, pp. 300–306. Springer, Heidelberg (2005)