The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0368-492X.htm

Dimensionality and data reduction in telecom churn prediction

Wei-Chao Lin
Department of Computer Science and Information Engineering, Hwa Hsia Institute of Technology, Taipei, Taiwan
Chih-Fong Tsai
Department of Information Management, National Central University, Jhongli, Taiwan, and
Shih-Wen Ke
Department of Information and Computer Engineering, Chung Yuan Christian University, Jhongli, Taiwan

Received 13 March 2013
Revised 17 March 2014
Accepted 7 April 2014

Abstract

Purpose – Churn prediction is a very important task for successful customer relationship management. In general, churn prediction can be achieved by many data mining techniques. However, during data mining, dimensionality reduction (or feature selection) and data reduction are the two important data preprocessing steps. In particular, the aims of feature selection and data reduction are to filter out irrelevant features and noisy data samples, respectively. The purpose of this paper, performing these data preprocessing tasks, is to make the mining algorithm produce good quality mining results.

Design/methodology/approach – Based on a real telecom customer churn data set, seven different preprocessed data sets based on performing feature selection and data reduction by different priorities are used to train the artificial neural network as the churn prediction model.

Findings – The results show that performing data reduction first by self-organizing maps and feature selection second by principal component analysis can allow the prediction model to provide the highest prediction accuracy. In addition, this priority allows the prediction model for more efficient learning since 66 and 62 percent of the original features and data samples are reduced, respectively.

Originality/value – The contribution of this paper is to understand the better procedure of performing the two important data preprocessing steps for telecom churn prediction.

Keywords Data mining, Feature selection, Churn prediction, Data reduction, Dimensionality reduction

Paper type Case study

1. Introduction

It is the fact that retaining existing and valuable customers is the core managerial strategy of organizations to survive in industry. This leads to the importance of effective churn prediction. Customer churn means that customers are intending to move their custom to a competing service provider. Therefore, it is necessary to assess their customers' value in order to retain or even cultivate the profit potential of customers (Kim *et al.*, 2004).

This work was supported in part by the National Science Council of Taiwan under Grant No. NSC 99-2410-H-008-034-.

Kybernetes
Vol. 43 No. 5, 2014
pp. 737-749

0368-492X
DOI 10.1108/K-03-2013-0045

Regarding Berry and Linoff (2004), customer churn in the telecommunications industry can be divided into the voluntary and involuntary churners. Voluntary churn means that customers make a decision to terminate their service with the provider. On the other hand, involuntary churn means that the company (or service provider) withdraw the customers' service because of abuse of service, non-payment of service, etc. For the purpose of churn prediction, voluntary churn is the main focus for organizations.

In the recent literature, data mining techniques have been widely used for churn prediction, and they can provide better prediction performance than traditional statistical methods (Ngai *et al.*, 2009; Tsai and Lu, 2010). In general, the process of knowledge discovery in databases (KDD) or data mining contains data set selection, data preprocessing, data analysis, and result interpretation and evaluation (Bose and Mahapatra, 2001; Fayyad *et al.*, 1996).

Since the collected data sets for some specific domains tend to be very large, which usually contain some noisy information (or unrepresentative data), data preprocessing plays an important role in KDD. Particularly, its aim is to make the chosen data set as "clean" as possible for the later data analysis step. In other words, the data preprocessing techniques can be used to filter out certain numbers of noisy data from the chosen data sets, which are able to enhance the data quality as well as the mining results (Han and Kamber, 2000).

To perform data preprocessing, dimensionality reduction (or feature selection) and data reduction (or instance selection) are the two most active research problems in data mining. This is because the number of features and data samples selected for many problems is usually very large to date. Therefore, if too many instances are considered, it can result in large memory requirements and slow execution speed, and can cause over-sensitivity to noise (Wilson and Martinez, 2000). In addition, it is often the fact that they are not all equally informative and some data points will be further away from the sample mean than what is deemed reasonable. As a result, the mining result, such as the classification/prediction performance, without considering the data preprocessing step is very likely poorer than the one performing data preprocessing (Reinartz, 2002; Yang and Olafsson, 2006).

Specifically, the aim of feature selection is to select more representative features which have more discriminative power over a given data set. It is also called as dimensionality reduction (Guyon and Elisseeff, 2003). On the other hand, data reduction aims at discarding the faulty data (or outliers), in which outliers could be considered as noisy points lying outside a set of defined clusters and could lead to significant performance degradation (Aggarwal and Yu, 2001; Barnett and Lewis, 1994).

Related studies of feature selection and data reduction have shown some promising results that the performance of prediction models with one of the data preprocessing step is better than the one without data preprocessing (Feng *et al.*, 2014; Gunal and Edizkan, 2008; Leyva *et al.*, 2014; Orsenigo and Vercellis, 2013; Piramuthu, 2004; Tsai, 2009; Tsai and Chang, 2013; Wang and Chiang, 2008). However, they only focus on either selecting more representative features or reducing faulty data for better classification or prediction. This leads to the research question about which step (i.e. feature selection and data reduction) should be performed first since they both are very important to improve the mining performance.

In practice, it is inevitable to face this priority problem of performing these two data preprocessing steps. This is because in many domain problems there is usually no exact agreed number of variables. That is, the collected variables in a specific domain

would not be all informative. Furthermore, some data samples in a given large data set may be regarded as noisy data. Therefore, feature selection and data reduction should both be considered in order to develop a more effective model.

However, it is difficult to perform feature selection and data reduction at the same time over a given data set. Given a data set D containing m dimensional features and i data samples, when feature selection and data reduction are performed as the first and second preprocessing steps, respectively, it will lead to D_1 containing n dimensional features and j data samples (where $n<m$ and $j<i$). On the other hand, the data set D, which is preprocessed by data reduction and feature selection for the first and second preprocessing steps, respectively, will result in D_2 containing o dimensional features and k data samples (where $o<m$ and $k<i$). In addition, the numbers of features and data samples of D_1 and D_2 are usually not the same. Consequently, using the same data mining algorithm over D_1 and D_2 will produce different results.

Therefore, the aim of this paper is to perform feature selection and data reduction with different priorities to examine their performances for the domain of telecom customer churn prediction as the case study. In addition, three baselines will be compared, which are the data set without performing both data preprocessing steps, the data set preprocessed by feature selection only, and the data set preprocessed by data reduction only.

The rest of this paper is organized as follows. Section 2 describes the basic concept of feature selection and data reduction. The techniques used for these two preprocessing steps are also overviewed, which are principal component analysis (PCA), association rules, and self-organizing maps (SOMs). Section 3 presents the research design and process. Section 4 provides the experimental results and the conclusion is given in Section 5.

2. Literature review

2.1 Feature selection

It is usually the fact that the number of features (or variables) captured in a data set is relative large and not all of these features are informative or can provide high discrimination power. The aim of feature selection is to remove most irrelevant and redundant features from the chosen data set. This helps improve the performance of the prediction models. That is, the curse of dimensionality problem can be alleviated (Powell, 2007). In addition, given a data set feature selection can help people to understand which the important features are and how they are related with each other.

2.1.1 PCA. PCA is a multivariate statistical technique, which is able to find out important features for best describing the variance in a data set. That is, PCA is widely used to transform data samples into a new feature space and to use lower dimensional feature representation from the new feature space to denote the data sample. In other words, PCA performs a linear mapping of the data samples to a lower dimensional feature space in such a way that the variance of the data in the low dimensional feature representation is maximized.

The total variability of a data set produced by the complete set of m variables can often be accounted for primarily by a smaller set of k ($k<m$) components of these variables. Therefore, the new data set consists of n records on k components rather than n records on m variables as the original one. Specifically, by computing eigenvalues and eigenvectors of the principal components, we could find a combination of the original variables in linearity which makes the greatest variance. The first principal component accounts for as much of the variability in the data as possible, and

each succeeding component accounts for as much of the remaining variability as possible (Jolliffe, 1986).

2.1.2 Association rules. Association rules are used for discovering important and/or interesting relations between variables in large databases. That is, multiple independent elements that co-occur frequently and rules which relate to the co-occurred elements in a given data set can be identified (Agrawal *et al.*, 1993).

A well-known application of association rules is market basket analysis. A market basket contains purchasing transactions of customers. That is, it is a collection of items or itemsets which are purchased by a customer in a single transaction. As the number of customer transactions is usually very large and frequent itemsets is exponential to the number of different items, association rules can be used to examine as most frequent itemsets as possible. Questions like what products tend to be purchased together can be answered. For example, customers who bought product A will also buy product B for the 81.25 percent probability. On the other hand, customers who bought product B will also buy product A for the 65 percent probability.

The general concept of association rules is as follows. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items in a given data set *DB* containing a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Let X be a set of purchased items. A transaction T is said to contain X if and only if $X \subseteq T$. An association rule can be represented by the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \mathbf{\emptyset}$. The rule $X \Rightarrow Y$ holds in the transaction set *DB* with *confidence c* if c percent of the transactions that contain X as well as Y. The rule $X \Rightarrow Y$ has *support s* in the transaction set if s percent of the transactions in *DB* contain $X \cup Y$. Confidence means the strength of the rule and support indicates the frequency of the patterns occurring in the rule. As a result, rules with high confidence and strong support can be referred to as strong rules (Kantardzic, 2003).

Therefore, the task of using association rules is to reduce a large amount of information to a small and more understandable set of statistically supported statements (Kantardzic, 2003). For customer churn prediction, Tsai and Chen (2010) apply association rules to select important and representative variables and show their outperformance over non-feature selection.

2.2 Data reduction

According to Wilson and Martinez (2000), one problem with using the original data points is that there may not be any data points located at the precise points that would make for the most accurate and concise concept description. Therefore, the aim of data reduction is to reduce a data set, which results in a smaller data set, but the integrity of the original data set is closely maintains. That is, it keeps less data count and more information amount. In some cases generalization accuracy can increase when noisy instances are removed and when decision boundaries are smoothed to more closely match the true underlying function.

The data which are removed can be regarded as outliers (or bad data). Specifically, outliers are the data points which are highly unlikely to occur given a model of the data. One approach to perform this task is based on the distances to neighboring data points by some clustering algorithm (Ghosting *et al.*, 2008).

2.2.1 SOMs. A SOM (Kohonen, 1987) is the predominant clustering technique, and it is comparable to traditional clustering techniques like the k-means algorithm (Smith and Gupta, 2000). In particular, the learning process of SOM is based on an unsupervised competitive learning algorithm, a process of self-organization. It usually consists of an input layer and the Kohonen layer which is designed as two-dimensional arrangement of

neurons, and it is used to map a high-dimensional data into a two-dimensional representation space.

That is, SOM can explore the groupings and relations within such data by projecting the data on a two-dimensional image that clearly indicates regions of similarity. Figure 1 shows an example of a 5×5 SOM.

The main properties of SOM can be stated as:

(1) The map allocates different numbers of nodes to inputs based on their occurrence frequencies. If different input vectors appear with different frequencies, the more frequent one will be mapped to larger domains at the expense of the less frequent ones.

(2) The distance relationships between the input data are preserved by their images in the map as faithfully as possible. While some distortion is unavoidable, the mapping preserves the most important neighborhood relationships between the data items, i.e., the topology of their distribution.

2.3 Prediction techniques

To predict customer churn, some prediction models need to be developed. In particular, supervised machine learning techniques can be employed. Given a training set, which is composed of a number of training data samples and one specific class label (i.e. prediction output) is associated with each data sample, the learning task is to compute a model that approximates the mapping between the input-output examples and correctly labels the training set with some level of accuracy. According to Tsai and Lu (2010), three most popular and widely used techniques for churn prediction are artificial neural networks, decision trees, and logistic regression (LR).

2.3.1 Artificial neural networks. Neural networks (or artificial neural networks) are motivated by information-processing units as neurons in the human brain that a neural network is made up of artificial neurons (Haykin, 1999). Specifically, the multilayer perceptron (MLP) neural network is the most widely developed neural network model. It consists of an input layer including a set of sensory nodes as input nodes, one or more hidden layers of computation nodes, and an output layer of computation nodes.

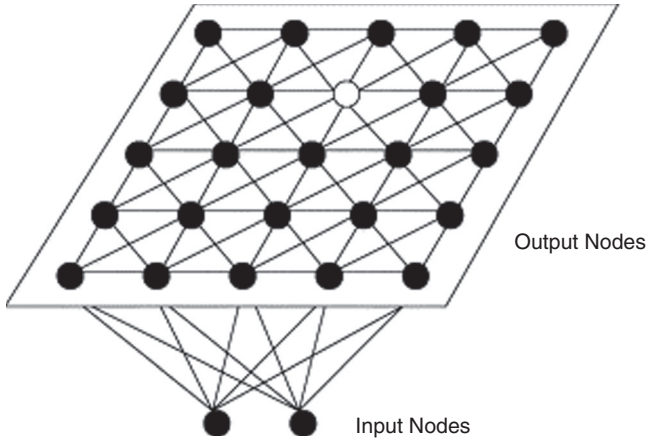

Figure 1.
A 5×5 SOM

A multilayer network is typically trained by a backpropagation learning algorithm. It performs weights tuning to define whatever hidden unit representation is most effective at minimizing the error of misclassification. That is, for each training example its inputs are fed into the input layer of the network and the predicted outputs are calculated. The difference between each predicted output and the corresponding target output is calculated. This error is then propagated back though the network and the weights between two layers are adjusted so that if the training example is presented to the network again, then the error would be less.

2.3.2 Decision trees. A decision tree classifies an instance by sorting it through the tree to the appropriate leaf node, i.e. each leaf node represents a classification. Each node represents some attribute of the instance, and each branch corresponds to one of the possible values for this attribute.

To construct tree, it is based on splitting the training set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

In literature, the CART (Classification and Regressing Tree) decision tree is the popular technique for constructing a classification or regression tree according to its dependent variable type, which may be categorical or numerical (Breiman *et al.*, 1984). That is, a decision tree with a range of discrete (symbolic) class labels is called a classification tree, whereas a decision tree with a range of continuous (numeric) values is called a regression tree.

2.3.3 LR. LR is a type of probabilistic statistical classification model. LR measures the relationship between a categorical dependent variable and one or more independent variables, which are usually continuous, by using probability scores as the predicted values of the dependent variables. LR allows us to look at the fit of the model as well as at the significance of the relationships between dependent and independent variables that are modeled (Hosmer and Lemeshow, 2000).

The LR function can be written as:

$$P = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{1}$$

where P is the probability of 1, e is the base of the natural logarithm and α and β are the parameters of the model.

3. Research methodology

3.1 The case data set

The data set for the later experiments is based on customer churn prediction from Cell2Cell, a wireless telecom company[1]. This data set can be used to develop a model for churn prediction in order to retain potential churners to remain with the company.

The data set contains 51,306 subscribers with 173 different features (i.e. variables), including 34,761 churners and 16,545 non-churners, from July 2001 to January 2002. In addition, the subscribers have to be mature customers who were with the telecom company for at least six months. Churn was then calculated based on whether the subscriber left the company during the period 31-60 days after the subscriber was originally sampled.

3.2 The experimental process

There are eight different procedures to construct eight different prediction models, respectively, shown in Figure 2, which belong to five categories described as follows:

- Category 1: the baseline – this process is to develop a baseline prediction model based on the original data set without performing any data preprocessing steps. Therefore, the comparative result allows us to understand whether the preprocessed data sets can make the prediction model performs better than the model without data preprocessing.
- Category 2: feature selection – in this process, we consider PCA and AR as the two feature selection methods, respectively. That is, we would like to know which feature selection method can result in better prediction performance.
- Category 3: data reduction – in addition to examining the model performance by feature selection, data reduction based on SOM is also considered. The result can be used to compare if the model only followed by the data reduction step can provide better prediction performance than the one only followed by the feature selection step.
- Category 4: feature selection + data reduction – this procedure is based on performing feature selection first and then, the processed data set is further processed by SOM for the data reduction task.
- Category 5: data reduction + feature selection – apposed to Category 4, data reduction is performed first, and the reduced data set is further processed by feature selection. Therefore, we can examine which preprocessing step should be performed first in order to construct an effective prediction model.

3.3 Dimensionality and data reduction

For the dimensionality reduction method by PCA, we considered factor loadings $\geqslant 0.5$ as informative variables. In particular, the principal axis analysis to filter the variables

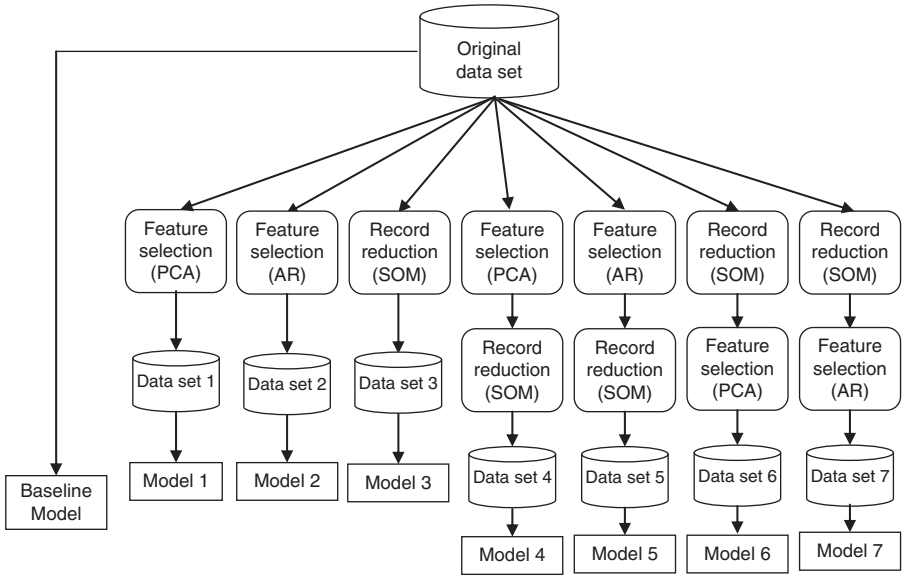

Figure 2.
The experimental process

is used. For association rules, 56 different support and confidence values are examined in order to identify the best selected features. In this paper, we found that 10 percent support and 80 percent confidence values can select the best features for prediction.

On the other hand, for the data reduction method by SOMs, 2×2, 3×1, 3×2, 3×3, 4×2, 4×3, 4×4, 5×1, 5×2, 5×3, 5×4, and 5×5 SOMs are constructed for comparisons. As there are churn and non-churn groups, two clusters corresponding to the two groups, respectively, of each SOM, which provide the highest rate of accuracy over the other clusters are selected as the clustering result. For the example of 2×2 SOM, given a training data set four clusters are produced represented by C_1, C_2, C_3, and C_4. Then, we can find out two out of the four clusters can be well "classified" into the churn and non-churn groups, respectively. The other two clusters whose data are not well classified or difficult to be classified by 2×2 SOM are discarded. In our experiments, we found that 4×4 SOM can provide the best clustering result for data reduction.

3.4 Artificial neural networks: the baseline prediction model

In this paper, artificial neural networks are used as the baseline prediction model in order to compare with the same models followed by different preprocessing steps shown in Figure 2. In particular, the MLP neural network with the back-propagation learning algorithm is constructed since approximately almost all business related studies utilize MLP (Smith and Gupta, 2000).

To construct a MLP neural network, there are two important parameters need to be setup in order to avoid the overfitting or overtraining problem. They are the number of hidden layer nodes and the training epoch. As Pendharkar (2002) point out that it is necessary to try at least three different parameters to obtain the optimal MLP, we consider four different hidden nodes, which are 8, 12, 16, and 24 and four different training epochs, which are 50,100,200, and 300. As a result, there are 16 prediction models constructed over one specific data set.

In addition, a tenfold cross-validation method is used. It is based on dividing ten equal and unduplicated parts of a data set. Any nine of the ten segments or subsets are selected to perform training. The remaining part is used for testing the model. As a result, there are ten testing results. Then, average prediction accuracy and errors can be obtained.

3.5 Evaluation methods

To evaluate the prediction performance of the eight prediction models, prediction accuracy and Types I and II errors are examined. They can be measured by a confusion matrix shown in Table I.

The rate of prediction accuracy can be obtained by:

$$\textit{Prediction accuracy} = \frac{a+b}{a+b+c+d}$$

Table I. Confusion matrix

↓Actual\predicted→	Churners	Non-churners
Churners	(a)	II (b)
Non-churners	I (c)	(d)

The Type I error shows the rate of prediction errors of a model, which incorrectly classifies the non-churners group into the churners group. Opposed to the Type I error, the Type II error presents the rate of prediction errors of a model to incorrectly classify the churners group into the non-churners group.

4. Experimental results

4.1 Prediction accuracy

Table II shows the average prediction accuracy rates of the eight models based on the 16 different parameter settings of MLP. As we can see, performing data reduction first and feature selection second is the best data preprocessing strategy, which can provide the highest prediction accuracy, i.e. 98.99 and 98.974 percent for SOM + PCA and SOM + AR, respectively. In particular, SOM + AR with MLP by the parameter of 50 learning epoch and eight hidden layer node can produce 99.01 percent prediction accuracy.

4.2 Types I and II errors

Tables III and IV show the Types I and II errors, respectively. Similar to prediction accuracy, performing data reduction first and feature selection second can provide the lowest Types I and II errors. More specifically, data preprocessing by SOM + PCA provides the lowest Type I error rate and SOM + AR for the lowest Type II error rate. However, there is no significant difference between them, i.e. <0.2 percent on average.

4.3 Reduction ratios vs prediction performances

Table V shows the reduction ratio and prediction performances of each method. Note that the number in each bracket means the reduction ratio. As we can see that AR produces a larger reduction ratio than PCA and make the MLP model slightly performs better in terms of accuracy and the Type I error. About performing either dimensionality reduction or data reduction, considering data reduction by SOM can allow MLP to provide better prediction performances than dimensionality reduction using PCA and AR.

	MLP	AR	PCA	SOM	SOM + AR	SOM + PCA	AR + SOM	PCA + SOM
50/8	92.878	93.021	92.346	98.835	99.01	98.993	98.703	97.46
50/12	92.795	92.885	92.435	98.788	98.999	98.993	98.683	97.46
50/16	92.655	92.907	92.321	98.726	98.999	98.993	98.809	97.459
50/24	93.39	92.723	92.394	98.752	98.994	98.993	98.77	97.454
100/8	92.67	92.9	92.412	98.851	98.998	98.993	98.77	97.454
100/12	92.593	92.795	92.314	98.763	98.992	98.993	98.61	97.432
100/16	92.544	92.688	92.251	98.554	98.993	98.993	98.757	97.483
100/24	92.351	92.449	92.188	98.705	99.004	98.993	98.717	97.459
200/8	92.672	92.894	92.341	98.58	98.81	98.982	98.278	97.398
200/12	92.591	92.709	92.281	98.543	98.93	98.998	98.418	97.399
200/16	92.206	92.476	92.153	98.29	98.998	98.993	98.605	97.421
200/24	92.19	92.242	91.954	98.437	99.014	98.977	98.497	97.387
300/8	92.441	92.714	92.382	98.595	98.858	98.982	98.125	97.415
300/12	92.534	92.569	92.21	98.556	98.973	98.993	98.518	97.382
300/16	92.226	92.362	92.202	98.318	99.008	98.987	98.316	97.376
300/24	91.919	92.025	91.868	98.433	99.008	98.982	98.478	97.365
Avg.	92.541 (7)	92.645 (6)	92.253 (8)	98.608 (3)	98.974 (2)	98.99 (1)	98.55 (4)	97.424 (5)

Table II. Prediction accuracy of the eight models (%)

	MLP	AR	PCA	SOM	SOM + AR	SOM + PCA	AR + SOM	PCA + SOM
50/8	16.428	16.462	18.433	16.07	1.602	1.418	1.753	3.474
50/12	16.096	16.936	18.15	1.603	1.602	1.418	1.771	3.494
50/16	16.239	16.454	18.486	1.583	1.623	1.418	1.825	3.491
50/24	16.161	16.613	18.185	1.601	1.631	1.418	1.784	3.493
100/8	16.472	16.748	18.726	1.58	1.624	1.418	1.72	3.555
100/12	16.526	16.561	18.427	1.587	1.625	1.418	1.784	3.519
100/16	16.752	16.793	18.156	1.583	1.636	1.418	1.805	3.472
100/24	16.569	17.223	18.372	1.523	1.619	1.418	1.805	3.516
200/8	16.659	16.513	18.401	1.661	1.605	1.432	1.719	3.627
200/12	16.421	17.265	17.868	1.591	1.613	1.418	1.742	3.588
200/16	16.185	16.445	18.108	1.561	1.536	1.418	1.794	3.589
200/24	16.435	16.972	18.023	1.575	1.61	1.432	1.796	3.59
300/8	16.232	16.623	18.586	1.589	1.568	1.432	1.732	3.543
300/12	16.209	16.452	17.989	1.573	1.625	1.418	1.751	3.61
300/16	16.157	16.705	18.147	1.599	1.62	1.432	1.771	3.544
300/24	16.788	16.84	18.04	1.61	1.624	1.432	1.765	3.626
Avg.	16.396	16.725	18.256	2.493	1.61	1.422	1.77	3.546
	(6)	(7)	(8)	(4)	(2)	(1)	(3)	(5)

Table III.
Type I errors of the eight models (%)

	MLP	AR	PCA	SOM	SOM + AR	SOM + PCA	AR + SOM	PCA + SOM
50/8	3.11	2.822	2.975	0.633	0.264	0.47	0.733	1.817
50/12	3.375	2.787	2.992	0.708	0.286	0.47	0.75	1.809
50/16	3.493	2.982	2.999	0.849	0.268	0.47	0.312	1.805
50/24	3.605	3.171	3.068	0.77	0.268	0.47	0.513	1.801
100/8	3.438	2.828	2.79	0.632	0.26	0.47	1.155	1.778
100/12	3.423	3.069	3.026	0.806	0.283	0.47	0.898	1.836
100/16	3.407	3.15	3.236	1.175	0.268	0.47	0.484	1.768
100/24	3.816	3.319	3.244	0.932	0.255	0.47	0.602	1.772
200/8	3.238	3.039	2.971	1.137	0.611	0.469	1.772	1.801
200/12	3.492	2.887	3.303	1.279	0.416	0.454	1.209	1.844
200/16	4.19	3.591	3.357	1.757	0.255	0.469	0.882	1.798
200/24	4.105	3.659	3.696	1.433	0.239	0.483	1.165	1.873
300/8	3.84	3.18	2.892	1.222	0.565	0.469	2.185	1.855
300/12	3.638	3.57	3.381	1.308	0.319	0.469	1.202	1.859
300/16	4.17	3.744	3.272	1.693	0.261	0.469	1.532	1.943
300/24	4.185	4.087	3.866	1.47	0.239	0.469	1.241	1.888
Avg.	3.658	3.243	3.192	1.113	0.316	0.469	1.04	1.828
	(8)	(7)	(6)	(4)	(1)	(2)	(3)	(5)

Table IV.
Type II errors of the eight models (%)

Method	Dimension	No. of data	Accuracy	Type I error	Type II error
Baseline	173 (0%)	50,355 (0%)	92.54% (7)	16.4% (6)	3.66% (8)
AR	25 (85%)	50,355 (0%)	92.65% (6)	16.73% (7)	3.24% (7)
PCA	102 (40%)	50,355 (0%)	92.25% (8)	18.26% (8)	3.19% (6)
SOM	170 (0%)	19,080 (62%)	98.61% (3)	2.49% (4)	1.11% (4)
AR + SOM	25 (85%)	15,025 (70%)	98.55% (4)	1.77% (3)	1.04% (3)
PCA + SOM	102 (40%)	17,991 (64%)	97.42% (5)	3.55% (5)	1.83% (5)
SOM + AR	22 (87%)	19,080 (62%)	98.97% (2)	1.61% (2)	0.32% (1)
SOM + PCA	57 (66%)	19,080 (62%)	98.99% (1)	1.42% (1)	0.47% (2)

Table V.
Reduction rations vs prediction performances

On the other hand, for the priority of performing dimensionality and data reduction the results show that employing data reduction first and dimensionality reduction second is the optimal way for churn prediction. Particularly, SOM + PCA performs best in terms of prediction accuracy and the Type I error.

5. Conclusion

Feature selection (or dimensionality reduction) and data reduction are the two important data preprocessing steps in the data mining process. The main goal of conducting each of the two steps is to make a given data set more "clean" and/or "representative" by filtering out irrelevant features and noisy data samples for obtaining good quality mining results. This paper is the first attempt to assess the performance of performing feature selection and data reduction steps by different priorities over the customer churn prediction domain problem.

Regarding the experimental results, for average prediction accuracy performing data reduction first by SOM and feature selection second by PCA can make the MLP neural network model provide the highest prediction accuracy rate and the lowest Type I error rate. In particular, on average this data preprocessing step transforms the training data set (i.e. 90 percent of the original data set) from a $173 \times 46{,}175$ matrix into $59 \times 17{,}547$, which allows a prediction model for more efficient learning.

Therefore, we can conclude that in customer churn prediction data preprocessing by performing data reduction first and feature selection second can produce a "better" data set to construct an optimal prediction model, where the training cost is largely reduced if compared with the model trained by the original training data set without data preprocessing.

It should be noted that although this paper considers three popular methods for data preprocessing, there are other methods available in the literature. However, from the practical standpoint, it is difficult to conduct a comprehensive study on all existing data reduction and feature selection methods. In addition, currently it is hard to define the most representative method in the churn prediction domain, and there is not a comparative study based on these methods, which can be regarded as one of the future research issues. In addition, other future work can focus on the development of more sophisticated models by combining multiple classifiers, i.e. classifier ensembles (Kittler *et al.*, 1998; Tsai *et al.*, 2011) and hybrid approaches (Lenar *et al.*, 1998; Tsai and Lu, 2009) to enhance the prediction performance. In particular, the classification techniques to be combined, the number of combined classifiers, and the combination methods should all be taken into account.

Note

1. www.fuqua-europe.duke.edu/centers/ccrm/index.html

References

Aggarwal, C.C. and Yu, P.S. (2001), "Outlier detection for high dimensional data", *Proceedings of the ACM SIGMOD International Conference on Management of Data, Santa Barbara, CA*, pp. 37-46.

Agrawal, R., Imielinski, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, DC, pp. 207-216.

Barnett, V. and Lewis, T. (1994), *Outliers in Statistical Data*, John Wiley & Son, New York, NY.

Berry, M.J.A. and Linoff, G. (2004), *Data Mining Techniques: for Marketing, Sales, and Customer Support*, John Wiley & Sons.

Bose, I. and Mahapatra, R.K. (2001), "Business data mining – a machine learning perspective", *Information & Management*, Vol. 39 No. 3, pp. 221-225.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, P.J. (1984), *Classification and Regressing Trees*, Wadsworth International Group.

Fayyad, U., Piatetsky, S.G. and Smyth, P. (1996), *Advances in Knowledge Discovery and Data Mining*, The MIT Press.

Feng, D., Chen, F. and Xu, W. (2014), "Supervised feature subset selection with ordinal optimization", *Knowledge-Based Systems*, Vol. 56, pp. 123-140.

Ghosting, A., Parthasarathy, S. and Otey, M.E. (2008), "Fast mining of distance-based outliers in high-dimensional datasets", *Data Mining and Knowledge Discovery*, Vol. 16, pp. 349-364.

Gunal, S. and Edizkan, R. (2008), "Subspace based feature selection for pattern recognition", *Information Sciences*, Vol. 178, pp. 3716-3726.

Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection", *Journal of Machine Learning Research*, Vol. 3, pp. 1157-1182.

Han, J. and Kamber, M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

Haykin, S. (1999), *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall.

Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley.

Jolliffe, I.T. (1986), *Principal Component Analysis*, Springer Verlag.

Kantardzic, M. (2003), *Data mining – Concepts, Models, Methods, and Algorithms*, John Wiley & Sons.

Kim, M., Park, M. and Jeong, D. (2004), "The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunications services", *Telecommunications Policy*, Vol. 28, pp. 145-159.

Kittler, J., Hatef, M., Duin, R.P.W. and Matas, J. (1998), "On combining classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20 No. 3, pp. 226-239.

Kohonen, T. (1987), "Adaptive, associative, and self-organizing functions in neural computing", *Applied Optics*, Vol. 26, pp. 4910-4918.

Lenar, M.J., Madey, G.R. and Alam, P. (1998), "The design and validation of a hybrid information systems", *Journal of Management Information Systems*, Vol. 14 No. 4, pp. 219-237.

Leyva, E., Caises, Y., Gonzalez, A. and Perez, R. (2014), "On the use of meta-learning for instance selection: an architecture and an experimental study", *Information Sciences*, Vol. 266, pp. 16-30.

Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009), "Application of data mining techniques in customer relationship management: a literature review and classification", *Expert Systems with Applications*, Vol. 36, pp. 2592-2602.

Orsenigo, C. and Vercellis, C. (2013), "Linear versus nonlinear dimensionality reduction for banks' credit rating prediction", *Knowledge-Based Systems*, Vol. 47, pp. 14-22.

Pendharkar, P.C. (2002), "A computational study on the performance of ANNs under changing structural design and data distributions", *European Journal of Operational Research*, Vol. 138, pp. 155-177.

Piramuthu, S. (2004), "Evaluating feature selection methods for learning in data mining applications", *European Journal of Operational Research*, Vol. 156, pp. 483-494.

Powell, W.B. (2007), *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, Wiley-Interscience.

Reinartz, T. (2002), "A unifying view on instance selection", *Data Mining and Knowledge Discovery*, Vol. 6 No. 2, pp. 191-210.

Smith, K.A. and Gupta, J.N.D. (2000), "Neural networks in business: techniques and applications for the operations researcher", *Computers & Operations Research*, Vol. 27, pp. 1023-1044.

Tsai, C.-F. (2009), "Feature selection in bankruptcy prediction", *Knowledge-Based Systems*, Vol. 22 No. 2, pp. 120-127.

Tsai, C.-F. and Chang, C.-W. (2013), "SVOIS: support vector oriented instance selection for text classification", *Information Systems*, Vol. 38 No. 8, pp. 1070-1083.

Tsai, C.-F. and Chen, M.-Y. (2010), "Variable selection by association rules for customer churn prediction of multimedia on demand", *Expert Systems with Applications*, Vol. 37 No. 3, pp. 2006-2015.

Tsai, C.-F. and Lu, Y.-H. (2009), "Customer churn prediction by hybrid neural networks", *Expert Systems with Applications*, Vol. 36 No. 10, pp. 12547-12553.

Tsai, C.-F. and Lu, Y.-H. (2010), "Data mining techniques in customer churn prediction", *Recent Patents on Computer Science*, Vol. 3 No. 1, pp. 28-32.

Tsai, C.-F., Lin, Y.-C., Yen, D.C. and Chen, Y.-M. (2011), "Predicting stock returns by classifier ensembles", *Applied Soft Computing*, Vol. 11 No. 2, pp. 2452-2459.

Wang, J.-S. and Chiang, J.-C. (2008), "A cluster validity measure with outlier detection for support vector clustering", *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, Vol. 38 No. 1, pp. 78-89.

Wilson, D.R. and Martinez, T.R. (2000), "Reduction techniques for instance-based learning algorithms", *Machine Learning*, Vol. 38 No. 3, pp. 257-286.

Yang, J. and Olafsson, S. (2006), "Optimization-based feature selection with adaptive instance sampling", *Computers & Operations Research*, Vol. 33 No. 11, pp. 3088-3106.

Corresponding author
Dr Chih-Fong Tsai can be contacted at: cftsai@mgt.ncu.edu.tw