Key-Attributes-Based Ensemble Classifier for Customer Churn Prediction

Yu Qian, Liang-Qiang Li, Jian-Rong Ran, and Pei-Ji Shao

Abstract—Recently, it has been seen that the ensemble classifier is an effective way to enhance the prediction performance. However, it usually suffers from the problem of how to construct an appropriate classifier based on a set of complex data, for example, the data with many dimensions or hierarchical attributes. This study proposes a method to construct an ensemble classifier based on the key attributes. In addition to its high-performance on precision shared by common ensemble classifiers, the calculation results are highly intelligible and thus easy for understanding. Furthermore, the experimental results based on the real data collected from China Mobile show that the key-attributes-based ensemble classifier has the good performance on both of the classifier construction and the customer churn prediction.

Index Terms—Customer churn, data mining, ensemble classifier, key attribute.

1. Introduction

Data mining has become increasingly important in management activities, especially in the support of decision making, most of which can be attributed to the task of classification. Therefore, classification analysis has been widely used in the study of management decision problems[1][2], for example, trend prediction and customer segmentation. Obviously, classification methods with high accuracy would reduce the decision loss of misclassification. However, with the increasing complexity of modern management and the diversity of related data, the results provided by a single classifier are suspected of having poor semantics and thus are hard to understand in management practice, especially for the prediction tasks with complex data and managerial scenarios[3].

In recent years, ensemble classifiers have been introduced into solving complicated classification problems[4], and they represent the new direction for the improvement of the performance of classifiers. These classifiers could be based on a variety of classification methodologies, and could achieve different rates of correct classified individuals. The goal of the classification result integration algorithms is to generate more certain, precise, and accurate results[5].

In literature, numerous methods have been suggested for the creation of ensemble classifiers[6][7]. Although the ensemble classifiers constructed by any of the general methods have archived a great number of applications in classification tasks[8][9], they have to face two challenges in performance under some real managerial scenarios. The first one is the expensive time cost for classifiers’ training/learning, and the second one is about the poor semantic understanding (management insights) of the classification results.

In this research, we propose a method which builds an ensemble classifier based on the key attributes (values) that are filtered out from the initial data. Experiment results with real data show that the proposed method not only has high relative precision in classification, but also has high comprehensibility of its calculation results.

2. Related Work

2.1 Classification Models for Churn Prediction

In most real applications, studies are mainly focused on improving the performance of a single algorithm in predicting activities, typically in predicting the customer churn in the service industry.

In this stream, Hu et al. analyzed and evaluated three implementations for decision trees in the churn prediction system with big data[9]. Kim et al. used a logistic regression to construct the customer churn prediction model[9]. Tian et al. adopted the Bayesian classifier to build a customer churn prediction model[9]. More complicatedly, artificial neural network (ANN) and random forest (RF) have been adopted to build the customer churn prediction model. Ultsch introduced a self-organizing map (SOM) to build the customer churn prediction model[10]. Rodan et al. used support vector machine (SVM) to predict customer churn. Au et al. built the customer churn prediction model based on evolutionary learning algorithms[10].

2.2 Ensemble Classifier

The main idea of the ensemble classifier is to build multiple classifiers on the collected original data set, and then gather the
results of these individual classifiers in the classification process. Here, individual classifiers are called base/weak classifiers. During the training, the base classifiers are trained separately on the data set. During the prediction, the base classifiers provide a decision on the test dataset. An ensemble method then combines the decisions produced by all the base classifiers into one final result. Accordingly, there are a lot of fusion methods in the literature including voting, the Borda count, algebraic combiners, and so on.

The theory and practices in literature have proved that an ensemble classifier can improve the performance of classification significantly, which might be better than the performance provided by any single classifier. Generally, there are two methods to construct an ensemble classifier: 1) algorithm-oriented method: Implementing different classification learning algorithms on the same data, for example, the neural network and decision tree; 2) data-oriented method: Separating the initial dataset into parts and using different subsets of training data with a single classification method.

Particularly, for the decline of the prediction precision caused by the complex data structure, processing the training data is a feasible way for ensemble classifier construction. Bagging and boosting are two typical ensemble methods of handling the datasets.

As mentioned before, the focuses of these researches have been put on the prediction accuracy of each single model. However, we could also address the problems of constructing an ensemble classifier based on the data distribution for better prediction results.

3. Research Method

3.1 Research Problem

Representing each user as an entity, then the dataset is composed by the values of user-attributes can be treated as an initial matrix as shown in Table 1. In which, the value of $x_i$ is typically vectors of the form $x_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots)$ and it denotes the whole values to the User. The value of $A_i$ is typically vectors of the form $A_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots)$ whose components are discrete or real value representing the values for attribute $A_i$ such as age, income, and location. The attributes are called the features of $x_i$.

<table>
<thead>
<tr>
<th>Table 1: Initial data matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_i$</td>
</tr>
<tr>
<td>User</td>
</tr>
<tr>
<td>$\cdots$</td>
</tr>
<tr>
<td>User</td>
</tr>
</tbody>
</table>

Given a vector of data $X = [x_1, \cdots, x_i, \cdots, x_n]^T$, a classifier learning program is given training examples of the form $\{(x_1, y_1), \cdots, (x_m, y_m)\}$ for some unknown function $y = f(x)$. An ensemble of classifiers, i.e., $F = \{f_1, f_2, \cdots, f_l, \cdots\}$, is a set of classifiers whose individual decisions are combined in some way to classify new examples. For the algorithm-oriented method, people try to train a set of classifiers by selecting some functions in $F$ and implement these functions on the data set $X$. Whereas, for the data-oriented methods, people train a set of classifiers by selecting one function in $F$ and implement it on each of the sub dataset $X_i$.

Since the basic research context of this study is about complicated training data and complex decision scenarios, both of the algorithm- and data-oriented methods would be taken into consideration in ensemble classifier construction by training a group of classifiers.

3.2 Key Attribute Selection

As the dimensionality of the data increases, many types of data classification problems become significantly harder, particularly at the high computational cost and memory usage. As a result, the reduction of the attribute space may lead to a better understandable model and simplifies the usage of different visualization techniques. Thus, a learning algorithm to induce a classifier must address the two issues: Selecting some key attributes (dimension reduction) and further splitting a dataset into parts according to the value distributions of these key attributes.

Key attribute selection used in this study is to select a lot of attributes from data sets and the selection is basically consistency with the goal of prediction.

The two ways of supervised and unsupervised methods can be used to select attributes. The supervised method is also called the management-oriented method. It is used to determine whether an attribute $A$ is a key attribute according to the management needs and prior knowledge. The typical method is asking some experts to label out the key attributes. The advantage of this method is that its calculation process is simple and its results have higher comprehensibility. To avoid the selection bias from the experts’ side, sometimes, the unsupervised method is used for data preprocessing by introducing some methods with the computational capacity of grouping or dimension reduction, for example, clustering or principal component analysis (PCA).

To simplify the calculation, we introduce the following “clustering-annotation” process to selecte the key attributes. Firstly, we use a clustering method to cluster the attributes of $(A_1, A_2, \cdots, A_n)$ into $l (l \leq n)$ groups, i.e., $\pi_1, \pi_2, \cdots, \pi_l$, according to their values’ similarity. In other words, if $A_{i1}, A_{i2}, \cdots, A_{ij}, \cdots$ are similar to each other, then

$$\pi_i = \{A_{i1}, A_{i2}, \cdots\}. (1)$$

Next, we associate one representative attribute for $\pi_i$ in accordance with the management semantics of attributes in $\pi_i$. The basic rule for the association is that the selected attribute should have strong potential correlation (management insights) with the decision-making problem.
3.3 Attribute Value Based Dataset Splitting

After the key attributes are selected, then the data set \( X \) would be split (clustered) into \( k \) parts by the value distributions of these key attributes.

The general method for such task is the binning method which can distribute sorted values into a number of bins. Assume that the maximum value of attribute \( A \) is \( max \) and its minimum is \( min \), and divide the original data set into \( k \) sub-dataset. The record \( x \), whose value of attribute \( A \) satisfies the following condition, will be classified as a member of the group \( C_i \):

\[
C_i = \left\{ x | x(A) \in \min + [(i-1), i] \times \frac{\max - \min}{k} \right\} \tag{2}
\]

where \( i = 1, 2, \cdots, k \).

In literature, researchers have introduced some efficient methods to split the initial dataset into sub-datasets automatically, for example, the maximum information gain method and Gini method\[8\][21]. The performance of such unsupervised methods is affected by the type, range, and distributions of the attribute values, and especially, they may suffer from the higher computational complexity.

Equation (2) works well on data splitting with one attribute \( A \). Moreover, we could split data with a set of attributes as clustered in (1). To deal with a very large dataset, it is argued that the singular value decomposition (SVD) of matrices might provide an excellent tool\[22\].

Based on the values of selected key attributes of \( \pi_i \) in this study, the dataset \( X_n \) will be split as follows:

1) Extracting dataset \( X_n \) from \( X \) to form a \( m \times |\pi| \) matrix according to the key attributes of \( \pi_i \);

2) Computing the SVD of matrix \( X_n \) such that

\[
X_n = USV^T \tag{3}
\]

where \( U \) and \( V \) are orthonormal and \( S \) is diagonal. The column vectors of \( U \) are taken from the orthonormal eigenvectors of \( X_nX_n^T \), and ordered right to left from largest corresponding eigenvalue to the smallest.

3) The elements of \( S \) are only nonzero on the diagonal, and are called the singular values. By convention, the ordering of the singular values is determined by high-to-low sorting, so that we can choose the top-\( k \) eigenvalues of \( S \) and cluster the vectors \( x(\pi_i) \) in \( X_n \) into \( k \) clusters: \( C(\pi_i)_1, C(\pi_i)_2, \cdots, C(\pi_i)_k \).

Finally, the cluster information for \( X_n \) is further used to map each vector of \( x \) in \( X \) into the group \( C_j \) (where \( j = 1, 2, \cdots, k \)):

\[
C_j = \left\{ x | x(\pi_i) \in C(\pi_i)_j \right\}. \tag{4}
\]

3.4 Ensemble Classifier

To keep more managerial information, we can construct an ensemble classifier as following:

Firstly, given a decision-making goal, we cluster all the attributes into \( l \) groups and associate each group with a representative feature. Then, we introduce SVD to split the data matrix of \( X_n \) for the group \( \pi_i \) and the results are used to map all the vectors in \( X \) into \( k \) groups, each of which is a sub-dataset specially for the purpose of better prediction for the targeted decision-making goal. Next, based on the new generated sub-dataset, we can introduce the general algorithm or data oriented method to train a set of approximate classifiers and use them to perform the classification tasks for decision-making problem. At last, a fused result will be reported for the prediction.

Another important work is to select an appropriate classification algorithm for those aforementioned sub-datasets. Considering the cost of calculation and the precision of results, in this study, we choose three typical classification algorithms of neural net, logistic, and C5.0\[23] as the basic algorithms to build the hybrid model.

The classification of a new instance \( x \) is made by voting on all classifiers \( \{CF_t\} \), each has a weight of \( \alpha_t \) where \( t = \{\text{Neural net, Logistic, C5.0}\} \). The final prediction can be written as:

\[
CF(x) = \text{sign} \left( \sum_t \alpha_t CF_t(x) \right) \tag{5}
\]

where, \( \alpha_t \) is a value between \([0, 1]\) according to the performance of \( CF_t \). In order to simplify the calculation, \( \alpha_t \) can be set as 1 for the best classifier and 0 for the others.

3.5 Evaluation Method

In this paper, the precision\[21\] and receiver operating characteristic (ROC)\[24\] are used to evaluate the results.

Given a set of prediction results made by a classifier, the confusion matrix of two classes “true” and “false” are shown in Table 2. Here, variable the \( A, B, C, \text{ and } D \) are used to denote the number of true positive, true negative, false positive, and false negative results, respectively.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predict result</th>
<th>Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>True</td>
<td>( A )</td>
<td>( B )</td>
</tr>
<tr>
<td>False</td>
<td>( C )</td>
<td>( D )</td>
</tr>
<tr>
<td>Aggregate</td>
<td>( A+C )</td>
<td>( B+D )</td>
</tr>
</tbody>
</table>

Precision = \[
\frac{A}{A+C} \tag{6}
\]

True positive (Sensitive) = \[
\frac{A}{A+B} \tag{7}
\]

False positive = \[
1 - \frac{A}{A+C+D} \tag{8}
\]

The ROC is a graphical plot that illustrates the performance of a classifier system as its discrimination threshold is varied. The ROC analysis is originated from the statistical decision theory in the 1950s and has been widely used in the performance assessment of classification\[25\]. The ROC curve is plotted by treating the ratio of true positive as \( Y \)-axis and ratio of false positive as \( X \)-axis. The closer to the upper left corner of ROC curve, the higher the accuracy of the model predictions.
The area under curve (AUC) can be used as a measure of the prediction effect. The value of AUC generally ranges between 1.000 and 0.500 and it represents the better prediction if the area value approaches closer to 1.000.

4. Experiment Results

4.1 Data Set

With the speedup of the market competition, maintaining the existing customers is then becoming the core marketing strategy to survive in the telecommunication industry. For the better performance of customer maintaining, it is necessary to predict those who are about to churn in the future. Studying the churn prediction is an important concern in the telecommunication industry. For instants, in the following experiments, the data is collected from China Mobile.

Note that, due to the great uncertainty of the consumer behavior and little data recorded in companies’ operation databases, the records generated by temporary customers and the customers who buy a SIM card and discard the card soon after short-term consumption are cleared. At last, all together 47735 customers are randomly selected from three main sub-branches which located in 3 different cities separately. The observation period is from January 1st, 2008 to May 31st, 2008 and the extracted information is about the activities of the users in using the telecommunication services, such as contract data, consumer behaviors, and billing data.

After data preprocessing such as data clean, integration, transformation, and discretization, the valid customer data is 47365 (99.2% of the total number of samples and noted as dataset $X$), in which, 3421 users are churn customers (the churn rate is 7.2%). In the experiments, the data set $X$ has been separated into two parts: The training data which were generated from January 1st to March 31st, 2008, denoted by $X_i$, and the test data which are generated from April 1st to May 31st, 2008, denoted by $X_o$.

The experiment platform is SPSS Clementine12.0, which provides well-programmed software tools for the classification algorithms of C5.0, logistic, and neural net.

4.2 Attribute Selection and Dataset Splitting

Let total, there are 116 (n=116) variables included in the customer relationship management (CRM) system are extracted as the initial data set $X$.

Implement the cosine similarity based $k$-means clustering method on vectors $A_i (i = 1, 2, \ldots, 116)$ in $X$. Inspired by the customer segmentation in marketing (in conjunction with necessarily experts’ annotations), we cluster the common variables according to their relations in marketing practice. At last, the attributes are clustered into 4 ($l=4$) groups and 4 attributes of brand, area, and bill (having strong correlation with customers’ churn in the telecommunication industry) are chosen as the key attributes, respectively.

Moreover, the values of these four attributes are split into 3 ($k=3$) sub-datasets, respectively, according to the SVD clustering results. The results are summarized in Table 3.

<table>
<thead>
<tr>
<th>Category</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand</td>
<td>GoTone</td>
</tr>
<tr>
<td>Area</td>
<td>Area A</td>
</tr>
<tr>
<td>Age</td>
<td>≤6 months</td>
</tr>
<tr>
<td>Bill</td>
<td>≤50 Yuan</td>
</tr>
</tbody>
</table>

4.3 Ensemble Model Construction

In the following, four ensemble classifiers will be built according to the sub-datasets separated by four attributes of brand, area, age and bill.

The classification algorithms of C5.0, logistic, and neural net algorithms are implemented on each sub-dataset for a series of repeated prediction experiments. The logic view of the ensemble classifier model construction is shown in Fig. 1.

![Fig. 1. Logic view of ensemble classifier models.](image)

For the attribute of brand, the training set $X_i$ is firstly divided into three sub-datasets, namely GoTone, EasyOwn, and M-Zone. Each of them accounts for 7.2%, 80.7%, and 12.1% customers, respectively. In the learning process, each subset is separated firstly into training and test sets according to the ratio of 60.0% and 40.0%.

Among all the classification results reported by each algorithm on the test dataset, the result with the largest AUC area under the ROC curve is selected as the basic model for such a sub-dataset. The AUC results reported by three models on each brand are shown in Table 4.

The comparative results are shown in Fig. 2. The results in
Table 4 show that the neural net algorithm works the best in the prediction of GoTone and EasyOwn sub-datasets, whereas the C5.0 works the best on the M-Zone sub-dataset.

Table 4: AUC of prediction on brand sub-datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>GoTone</th>
<th>EasyOwn</th>
<th>M-Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
<td>Training</td>
</tr>
<tr>
<td>C5.0</td>
<td>0.583</td>
<td>0.597</td>
<td>0.911</td>
</tr>
<tr>
<td>Neural net</td>
<td>0.796</td>
<td>0.803</td>
<td>0.851</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.868</td>
<td>0.713</td>
<td>0.843</td>
</tr>
</tbody>
</table>

4.4 Result Evaluation

The previous experiments on sub-datasets separated by different key attributes provide 4 hybrid models. Next, we will use these four models to make prediction on dataset $X_2$. Also, the measurements of precision and ROC curve are used to evaluate the performance of each model.

1) Comparison of precision

The average accuracy of prediction provided by the four of each model based on $X_2$ is summarized in Table 8. It shows that there is the highest precision (86.1%) reported while using the key attribute of area for data segmentation to build a hybrid model, followed by the result generated with the attribute bill (85.9%). However, the performance of hybrid models constructed by the attributes brand and age for data segmentation is lower (81.2% and 76.2%).

Similarly, the performances of classification (prediction) on sub-datasets split by attributes of area, age, and bill are reported in Tables 5 to 7, respectively. Accordingly, the visualized results are shown in Figs. 3 to 5.

Fig. 2. AUC of prediction on brand sub-datasets: (a) GoTone, (b) EasyOwn, and (c) M-Zone.

Fig. 3. AUC of prediction on area sub-datasets: (a) area $A$, (b) area $B$, and (c) area $C$.

Fig. 4. AUC of prediction on age sub-datasets: (a) net age low, (b) net age middle, and (c) net age high.


2) Comparison of ROC

The ROC curves for the prediction results provided by the four hybrid models on testing set \( X_2 \) is shown in Fig. 6. The area under the ROC curve of each hybrid model is calculated in Table 9.

Comparing the results in Fig. 6 and Table 9, we know that the two hybrid models constructed based on attributes of area and bill would generate a better AUC (0.888 and 0.855) than based on brand and age (0.828 and 0.845).

According to the experiment results, we can conclude that using the attribute of area as the segment variable would get the best prediction results, which are followed by those of the bill attribute. However, the key attributes age and brand would perform relatively poorly. Therefore, in practice of customer churn prediction, it is recommended that telecommunication companies use the consumers’ bill information as the key attribute to build the customer churn prediction hybrid model for each area separately. Moreover, it is necessary to strengthen brand management and to improve the customer segmentation effect of different brands.

3) Limitations

The main idea of the method proposed in this work is to construct an ensemble classifier for higher precision and managerial insights. We should note some limitations of this work. First, there is a lack of criteria for how many base classifiers should be selected in the hybrid classifier. Second, the proposed method has involved some time consumption preprocessing processes in ensemble classifier construction, for example, the PCA and SVD methods, which would cause the higher complexity of computation.

5. Conclusions

Classification analysis has been widely used in the study of
decision problems. However, with the increasing complexity of modern management and the diversity of related data, the results provided by a single classifier are suspected of having poor semantics, thus are hard to understand in the management practice, especially for the prediction tasks with the very complex data and managerial scenarios.

Regarding to the management issues of classification and prediction, an ensemble of single classifiers is an effective way to improve the prediction results. In order to solve the problems of poor precision and management semantics caused by the ordinary ensemble classifiers, in this paper, we proposed the ensemble classifier construction method based on the key attributes in the data set. The experimental results based on the real data collected from China Mobile show that the key-attributes-based ensemble classifier has the advantages on both prediction accuracy and result comprehensibility.

References


Yu Qian is currently an associate professor with University of Electronic Science and Technology of China (UESTC), Chengdu. Her research interests include operation management and information economics. Her papers have been published and presented in journals and conferences such as the Flexible Services and Manufacturing Journal, Decision Support Systems, and POMS annual conference.
Liang-Qiang Li received the B.S. degree from Sichuan Normal University, Chengdu in 2004. He is currently pursuing the Ph.D. degree with the Department of Management Science and e-Commerce, UESTC. His research interests include business intelligence, e-commerce, information system management, and urban computing.

Jian-Rong Ran received the M.S. degree from UESTC, in 2009. His research interests include customer relationship management and e-commerce.

Pei-Ji Shao is currently a professor with the School of Economics and Management, UESTC. He authored the textbook Management Information System and has published articles in the areas of information management and e-business. In addition to higher education, he is an expert specifically for the Government of Sichuan Province of China in the Internet-related sector.