Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn

Indranil Bose and Xi Chen

***Abstract* –In this paper, we use two-stage hybrid models consisting of unsupervised clustering techniques and decision trees with boosting on two different data sets and evaluate the models in terms of top decile lift. We examine two different approaches for hybridization of the models for utilizing the results of clustering based on various attributes related to service usage and revenue contribution of customers. The results indicate that the use of clustering led to improved top decile lift for the hybrid models compared to the benchmark case when no clustering is used.**

***Index Terms* – Churn, Clustering, Data mining, Decision trees, Lift, Prediction**

I. Introduction

Preventing customer churn is critical for the survival of mobile service providers because it is estimated that the cost of acquiring a new customer is about $300 or more if the advertising, marketing, and technical support etc are all taken into consideration. On the other hand, the cost of retaining a current customer is usually as low as the cost of a single customer retention call or a single mail solicitation [1]. The high acquisition cost makes it imperative for mobile service providers to devise ways to predict the churn behavior and execute appropriate proactive actions before customers leave the company.

Mobile telecommunication companies have used data mining techniques to identify customers that are likely to churn. Since the main purpose of applying data mining techniques in this area is prediction, supervised learning techniques are popularly used. However, the use of unsupervised learning techniques for churn prediction is rather limited. In this paper, we investigate the issue of how to combine unsupervised learning techniques with supervised learning techniques in the form of hybrid models for the prediction of customer churn. Our goal is to seek answers to the following research questions. Firstly, can clustering algorithms detect patterns that help decision trees in identifying churners better? Secondly, which clustering algorithm(s) are more useful in prediction of churn? Thirdly, what is the best way to combine the results obtained from clustering algorithms with that of decision trees? Finally, what type of behavioral patterns of customers obtained from clustering of customer data is useful for detection of churners?

I. Bose is with the School of Business, University of Hong Kong (corresponding author to provider: phone: (852)22415845; fax: (852)28585614; email: bose@business.hku.hk)

X. Chen is with the School of Business, University of Hong Kong (email: chenxi@Business.hku.hk)

II. Literature Review

There are mainly two types of data mining techniques that are used in practice: supervised learning and unsupervised learning. Supervised learning requires that the data set should contain target variables that represent the classes of data items or the behaviors that are going to be predicted. The most important decision in customer churn management is the separation of churners from non-churners. This is a task that is quite capably handled by supervised learning techniques.

Supervised Learning

Decision tree models are very popular in prediction of churn. Wei and Chiu used different subsets of the whole data set to generate different decision tree models and combined the results of those single decision tree models using a weighted voting approach and generated a final classification decision for churn [2]. They included customer characteristics as well as their contract information in their churn model. Hung et al. clustered customers according to their tenure related data and built

decision trees for each cluster to predict customer churn [3]. Chu et al. used C5.0 decision tree to separate churners from non-churners and to identify key attributes for the prediction of churners [4]. In the second phase of their research, they clustered the detected churners according to the identified key attributes so that retention policies could be designed for each cluster. Decision tree models have also been used to construct hybrid models in combination with other supervised learning techniques. Qi et al. combined decision trees and logistic regression models [5]. They determined different subsets of attributes from customer data based on correlation analysis and then built decision trees using each subset of attributes. Then a logistic regression model was used to predict churn based on the churn likelihood predicted by the decision trees.

Other techniques have also been used for prediction of churn. These included the use of neural networks by Mozer et al. [6], the use of support vector machines (SVM) by Cousement and Van den Poel [7], and the use of evolutionary algorithms by Au et al. [8].

Unsupervised Learning

Unsupervised learning techniques do not require the data set to contain the target variable. Clustering is a type of unsupervised learning technique that can be used to explore data sets in order to discover the natural structure and unknown but valuable behavioral patterns of customers' hidden in it [9]. Various approaches have been used for clustering. Jain et al. presented an overview of unsupervised clustering methods. Clustering techniques group data items based on their similarities. Euclidean distance is a common choice for measuring the similarities. A data item is assigned to a cluster whose center is the most similar to the data item. K-means and K-medoid, self-organizing map (SOM), fuzzy c-means (FCM), and hierarchical clustering represent four different type of clustering techniques. K-means and K-medoid are partitional clustering algorithms that are similar to each other. K-means uses the mean of the data items in a cluster as the center of that cluster whereas K-medoid uses a data item that is at the center of a cluster as the cluster-center. It is reported that K-medoid is less sensitive to the presence of outliers in data sets [10]. SOM is a neural network-based clustering technique that clusters data into a two-dimensional map so that the distribution of clusters can be visualized [11]. FCM is a type of fuzzy clustering algorithm that assigns data items to clusters using membership functions. Hierarchical clustering follows a bottom up approach and forms clusters starting from a single data item.

In spite of the popularity of unsupervised learning techniques, there is little literature devoted to the utilization of the natural patterns detected by clustering algorithms in the building of churn classification models. Chu et al. applied the hierarchical SOM clustering technique to cluster churners. However, clustering was performed after prediction was made by the decision tree model and the results of SOM did not improve the performance of the decision tree models in any way [4]. In a different area of application, Thomassey and Fiordaliso used cluster labels obtained by K-means as target variables for decision trees for sales forecasting [12]. In their research, the decision tree model is used to find rules that could explain the formation of the clusters. The research conducted by Hung et al. is most closely related to this paper [3]. They clustered customers according to a single variable (i.e. tenure) and built decision trees for each cluster. They used the decision trees on the same testing data in order to find which cluster could generate decision trees with better prediction accuracy. In this paper, we use multiple variables for clustering and examine different approaches of hybridization for utilizing the results of clustering in order to build better supervised learning models (using decision trees) for prediction of customer churn.

III. Data Description

The three customer churn data sets used in this research are obtained from the Teradata Center at Duke University, USA [13]. The first data set contains 100000 records of customers. The ratio of churning customers to non churning customers is about 50%. The second data set contains 50000 records of customers and the third data set contains 100000 records of customers. The churn ratio of customers in the second and third data set is about 1.8%. In the numerical experiments reported in this paper, we use the first data set as the training data and refer to it as the calibration data. The second and the third data sets are used as testing data and are subsequently referred to

as current data and future data, respectively.

IV. Experiments

Decision Trees

Clustering is used as the first stage in the hybrid method and the second stage is conducted using decision trees. Although several supervised learning techniques could be chosen for the second stage, the C5.0 decision tree model with boosting is adopted in this research. There are a number of reasons for that. In general, decision trees are found to be efficient and fast in prediction of churn and compared to other supervised learning techniques, they can automatically decide the importance of attributes. Also, decision trees can tolerate the presence of outliers and missing data and so minimum effort is required for data preprocessing. C5.0 is an upgraded version of C4.5 developed by Quinlan [14]. Compared to C4.5, C5.0 is faster, more accurate, and less memory intensive. To enhance the performance of C5.0, it is extended with the boosting algorithm. The boosting approach combines different classifiers by assigning weights to them. The weights are then iteratively adjusted over several trials according to the performance of the classifiers. Although each single classifier may not have good performance, the combination of them using the boosting approach can improve the overall performance of classification models significantly. At the same time, the boosting approach can avoid the problem of overfitting so that classification models can have good performance not only for the training data but also for unknown testing data [10].

Clustering Techniques

Five different clustering algorithms are examined in this research as the first stage of the hybrid method. They are K-means, K-medoid, SOM, FCM, and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [15]. BIRCH belongs to the family of hierarchical clustering algorithms and has been found to be efficient for large data sets. It can automatically identify the optimal number of clusters during clustering. For K-means, K-medoid, and FCM, Dunn's index is used for identification of optimal number of clusters [16]. For SOM, data is clustered into a two dimensional map.

Selection of Attributes for Clustering

For mobile telecommunication services, the most important information is minutes of use of mobile services and revenue contribution for those services. The minutes of use of mobile services is decided by the customers themselves whereas the revenue contribution is influenced by the pricing plan adopted by the mobile service providers. Clustering is performed on two types of attributes related to voice calls: service usage and revenue contribution. Both service usage and revenue contribution are characterized by multiple attributes (7 attributes for each group) and clustering is used on them so that customers could be segmented using multivariate information.

Choice of Performance Metric

Due to the highly skewed distribution of the target variable – 'churn' in the current and future data sets, the traditional method of assessing classification accuracy of models could not be applied in this research. In fact, we could achieve accuracy as high as 98.2% by classifying all customers as non churners. However, this result would not be meaningful. We needed models that could identify customers who were most likely to churn so that appropriate actions could be taken to retain them. We used top decile lift as the metric of choice to compare the performance of the different hybrid models because it is popularly used in the literature [7, 8] to compare different models, which were used for churn prediction, in terms of their ability to capture customers with high risk of churn. The higher the top decile lift, the better is the model.

Alternative Methods of Hybridization

The five clustering algorithms were applied on the calibration data. The result included two cluster labels. One indicated the identity of the segment obtained using information on service usage and the other indicated the identity of the segment obtained using information on revenue contribution. We examined two methods of hybridization for utilizing the results of the clustering techniques. The first method used the labels that represented the identity of clusters as input to the decision tree model for prediction of churn. The second method separated the customers into different clusters and then built decision tree models for each cluster. In the first method, the decision tree models with boosting were trained for each clustering technique and used labels of service usage clusters (UseLbl), labels of revenue

contribution clusters (RevLbl), and both type of labels (TwoLbl) as input. In the second method, two types of C5.0 decision tree models with boosting were trained: training models for each service usage cluster (UseClst), and training models for each revenue contribution cluster (RevClst). Finally, a benchmark model was used and that was a C5.0 decision tree model that did not utilize any results from the clustering techniques (NoLbl). As the last stage of the experimental procedure, the trained hybrid models were used on the two testing data sets and the top decile lifts were computed.

V. RESULTS

Tables 1 and 2 represent the results obtained for the current and future data sets respectively. In these tables each row represents a clustering techniaue and each column represents a method of utilizing the results of clustering under the two methods of hybridization. There are totally 25 combinations of clustering techniques and methods of utilizing clustering results. 15 combinations belonged to the first method of hybridization and the remaining 10 combinations belonged to the second method of hybridization. The value in each cell of the two tables is the top decile lift of the corresponding model for prediction of churn. It is observed that the best model for current data was the hybrid model comprising SOM and C5.0 tree with boosting, using 'RevLbl'. For future data, the best models were the BIRCH and C5.0 tree with boosting using 'RevLbl' and the FCM and C5.0 tree with boosting using 'TwoLbl'. It can also be observed from Tables 1 and 2 that among the five clustering techniques that were used in these experiments, the hybrid models using K-means performed the best with 3 models beating the benchmark model for the current data and 4 models beating the benchmark model for the future data. The hybrid models using FCM performed the worst.

Among the 15 models that belonged to the first method of hybridization, 10 could beat the performance of the benchmark model in terms of the top decile lift for current data and 13 could beat the performance of benchmark model for future data. On the other hand, among the 10 models that belonged to the second method of hybridization, only 1 could beat the benchmark model for both current and future data. This indicated that the first method of hybridization performed better than the second in terms of top decile lift and thus it was better to include the cluster label as an additional input item rather than forming the clusters first and then using C5.0 decision trees with boosting on each cluster.

Table 1. Top Decile Lift for Current Data

	First method			Second methodn		
	UseLbl	RevLbl	TwoLbl	VceClst	RevClst	NoLbl
BIRCH	2.49	2.61	2.40	1.99	1.78	2.44
FCM	2.47	2.50	2.61	2.44	2.42	2.44
KM	2.51	2.56	2.52	2.18	2.48	2.44
KMD	2.54	2.46	2.47	2.37	2.10	2.44
SOM	2.47	2.42	2.53	2.23	2.37	2.44

Table 2. Top Decile Lift for Future Data

	First method			Second method		
	UseLbl	RevLbl	TwoLbl	VceClst	RevClst	NoLbl
BIRCH	2.61	2.56	2.61	2.10	1.66	2.52
FCM	2.46	2.53	2.51	2.10	2.41	2.52
KM	2.55	2.57	2.46	2.25	2.53	2.52
KMD	2.47	2.54	2.60	2.23	2.22	2.52
SOM	2.62	2.65	2.48	2.23	2.41	2.52

Figures 1 and 2 represent the lift curves for the best models for current and future data respectively. In these figures the y-axis represents the value of the lift and the x-axis represents the deciles. The lift curve for the benchmark models are also shown in the two figures for ease of comparison. From Figure 1, it can be observed that the hybrid SOM and C5.0 tree model with boosting that used 'RevLbl' had higher lift values not only for the top decile but also for the second and the third deciles. From Figure 2, it can be observed that the hybrid FCM and C5.0 tree model with boosting that used 'TwoLbl' had similar top decile lift as the hybrid BIRCH and C5.0 tree model with boosting that used 'RevLbl'. However, the performance of the hybrid FCM and C5.0 tree model with boosting that used 'TwoLbl' deteriorated fast for the subsequent deciles and was even worse than the benchmark model (e.g., second and third deciles). In comparison, the hybrid BIRCH and C5.0 tree model with boosting that used 'RevLbl' continued to perform better than the benchmark model upto the sixth decile. This leads us to the conclusion that the hybrid BIRCH and

C5.0 tree model with boosting that used 'RevLbl' was the best model for future data although its top decile lift was exactly same as the hybrid FCM and C5.0 tree model with boosting that used 'TwoLbl'.

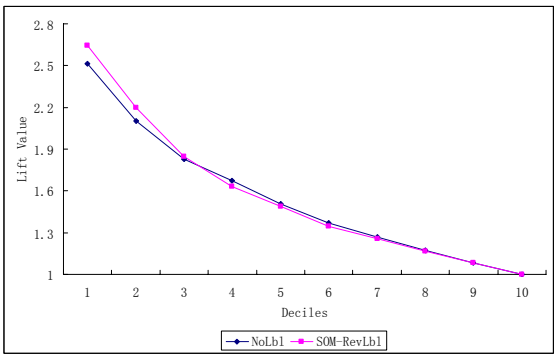

Figure 1. Lift Curve of the Best Hybrid Model for Current Data

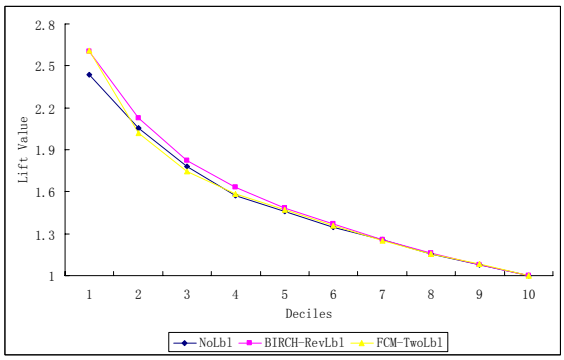

Figure 2. Lift curve of the Best Hybrid Models for Future data

VI. Discussion

In this paper, we answered the four research questions listed in the introduction about hybrid models that combined unsupervised clustering techniques with decision trees. For the first question, our results showed that including cluster labels as inputs to C5.0 decision tree models with boosting improved the performance of those models in terms of top decile lift. Hence, we can say that the patterns detected by the clustering techniques helped C5.0 decision trees to detect the phenomenon of churning better. The clustering results represented multivariate splitting of customers into different groups. In contrast, decision trees can only split the customers based on single attributes at one time. When splitting of customers using those clustering labels, it could be regarded as giving decision trees the ability to split customers based on multivariate information at one time which might explain the reason for the improvement in performance of decision trees. For the second question, SOM helped generate the best hybrid model for current data and BIRCH helped generate the best hybrid model for future data, and KM helped generate the most number of models that could beat the benchmark model for the two data sets. Because of the mixed result, we recommend that if marketing experts want to predict customer churn at a nearby point in time, they should use the hybrid model with SOM and C5.0 decision tree with boosting. If they want to predict customer churn at farther point in time, they should use the hybrid model with BIRCH and C5.0 decision tree with boosting. Finally, if they don't have accurate knowledge about the time frame of decision making, then they should consider the hybrid model with KM and C5.0 decision tree with boosting. For the third question, the results illustrated that including cluster labels as input to the decision trees was always a better method of hybridization than clustering the customers and then using decision trees on each customer cluster. It was difficult to find a consistent answer to the fourth question. However, it is worth noting that the two best models for the current and future data used revenue cluster labels. Also, for current data, models including revenue cluster labels were always better than the benchmark model whereas for the future data, 4 out of 5 models including revenue cluster labels performed better than the benchmark model. Therefore, it is safe to recommend revenue cluster labels as input to the decision trees for these two data sets.

References

[1] A. Berson, S. Smith, and K. Thearling, *Building Data Mining Applications for CRM*, New York: McGrawHill, 2002.

[2] C. –P. Wei and I. –T. Chiu, "Turning telecommunications call details to churn prediction: a data mining approach," *Expert Systems with Applications*, vol. 23, no. 2, pp. 103-112, 2002.

[3] S. –Y. Hung, D. C. Yen, and H. –Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, no. 3, pp. 515-524, 2006.

[4] B. –H. Chu, M. –S. Tsai, and C. –S. Ho, "Toward a hybrid data mining model for customer retention," *Knowledge Based Systems*, forthcoming, 2006.

[5] J. Y. Qi, Y. M. Zhang, Y. Y. Zhang, and S. Shi, "TreeLogit model for customer churn prediction," *Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing*, Dec. 12-15, 2006, Guangzhou, China, pp. 70-75.

[6] M. C. Mozer, R. Wolniewicz, and D. B. Grimes, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 690-696, 2000.

[7] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Application*, vol. 34, no. 1, pp. 313-327, 2008.

[8] W. –H. Au, K. C. C. Chan, and X. Yao, "A novel evolutionary data mining algorithm with application to churn prediction," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 6, pp. 532-545, 2003.

[9] G. Punj and D. W. Stewart. "Cluster analysis in marketing research: review and suggestions for application," *Journal of Marketing Research*, vol. 20, no.2, pp. 134-148, 1983.

[10] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, San Diego: Elsevier, 2006.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[12] S. Thomassey and A. Fiordaliso, "A hybrid sales forecasting system based on clustering and decision trees," *Decision Support Systems*, vol. 42, no. 1, pp. 408-421, 2006.

[13] Duke, "Case studies, presentations and video modules," 2005, (retrieved September 2007 from http://www.fuqua.duke.edu/centers/ccrm/datasets/download.html#data).

[14] J. R. Quinlan, *C4.5 Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993.

[15] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *Proceedings of the ACM SIGMOD Conference on Management of DATA*, 1996, Montreal, Canada, pp. 103-114.

[16] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 28, no.3, pp. 301-315, 1998.