2011 International Conference of Information Technology, Computer Engineering and Management Sciences

Transfer ensemble model for customer churn prediction with imbalanced class distribution

Yuan Wang
School of Business Administration
Sichuan University
Chengdu 610064, P.R. China
wangyuan9793@gmail.com

Jin Xiao
School of Business Administration
Sichuan University
Chengdu 610064, P.R. China
xjxiaojin@126.com

***Abstract*—Customer churn prediction is an important issue in customer relationship management. The class distribution of customer data is often imbalanced, which may affect the performance of churn prediction model greatly. This paper combines transfer learning and multiple classifiers ensemble, and proposes a transfer ensemble model for imbalanced data (TEMID). This method focuses on using transfer learning and sampling to enlarge the available training set and balance it respectively. What's more, it also uses multiple classifiers ensemble method to implement the classification. The performance of TEMID and some existing transfer learning algorithms are compared in two class imbalanced datasets. The results show that the TEMID methods can actually improve the performance of the customer churn prediction.**

Keywords-customer churn prediction; imbalanced class distribution; transfer ensemble model; transfer learning; multiple classifiers ensemble

I. INTRODUCTION

Customer churn is a frequently discussed issue in the commercial field. Shown by some statistics, the average annual churn rate of customers in the telecommunication industry should be 25% - 30%. And another survey on nine industries of U.S. illustrates that every 5% reduced in customer churn rate, every 25% - 85%[1] the industrial average profit can increase by. Therefore, to predict customer churn exactly and implement customer retention timely is actually significant for enterprises to improve their core competitiveness.

At present, numbers of researches have been made on the customer churn prediction and several important achievements are well verified, in which the commonly used methods consist of Naive Bayesian classifiers[2], support vector machine[3, 4], etc. However, as for a real problem, the churn customers are usually much less than that of non-churn ones, which is called imbalanced class distribution. In order to reduce the imbalance, sampling methods including under-sampling[5] and over-sampling[6] are widely used and their success are mainly due to fully exploiting the available information within the system to increasing the prediction accuracy.

However, the real situation is that the data for customer churn prediction is limited and that for churn class is much less. That situation makes it a challenge to build a suitable prediction model or use the sampling method[7]. On the other hand, we notice that there exist plenty of special data outside the system. They often come from different period, area, and business. Thus, how to make the most of the data outside the system so that we can improve the prediction model and ascend the accuracy is a thought-provoking question.

The transfer learning techniques from machine learning area provide a new way of thinking to solve the problem. Its main idea is using the knowledge acquired from the related field to assist new learning tasks to improve learning outcomes. During decades' development, transfer learning has achieved a lot in theory, algorithm and application but is seldom seen in CRM field.

Taking all the factors above into consideration, this paper proposed a transfer ensemble model for imbalanced data (TEMID). Its core idea can be roughly described in three aspects (1) use sampling method to overcome class-imbalance; (2) through multiple classifiers ensemble, making each classifier play its best in its specialized space; (3) use transfer learning method to make the best of the information outside the system.

II. RELATED WORKS

A. Sampling and Mutiple classifiers ensemble

Sampling techniques including Over-sampling and Under-sampling method[8] could be the most popular method to deal with class-imbalance. For instance, Xia and Jin[4] have applied Over-sampling method to deal with imbalanced class problem, and use SVM to predict customer churn. Verbek[5] has proposed advanced rule induction techniques to build comprehensible churn prediction model, which use Over-sampling method to overcome the class-imbalance among the training models.

Recently, ensemble learning techniques are introduced to the customer churn prediction. For example Lariviere [9] and Buckinx[10] have taken the lead in applying Random Forests algorithm to churn prediction. Lemmens and Croux[11] applied Bagging and Boosting to customer churn

978-0-7695-4522-6/11 $26.00 © 2011 IEEE
DOI 10.1109/ICM.2011.397

prediction combining with Sampling techniques to deal with class-imbalance.

Although all the achievements have made a big contribution to customer churn prediction area, these studies have only used the data within the system. And if the information is limited, it is difficult to obtain satisfactory results. Thus it forces us to utilize transfer learning strategies to take full advantages of the data outside the system.

B. Transfer Learning

The concept of transfer learning originated in the "Psychology". To implement it, there must be two domains; one is target domain denoted by *T* and the other is source domain denoted by *S*. The model based on transfer learning can be roughly described as using the information from the *S* to enrich the *T* to complete the classifying task[12].

1) TrBagg algorithm: TrBagg[13] is a modification of Bagging[14] proposed by Bremian in 1996. Via *bootstrap(T)*, the Bagging comes up with mutiple training subsets which train numerous weak classifiers. Finally, the results of all the classifiers are combined through majority voting algorithm.

Two phases are generated on the Bagging by TrBagg, learning phase and filter phase. In the learning phase, it generated a merged dataset of *T* and *S* as the training set *T'*. In the filer phase, a set of classifiers, $C_T^* \subseteq C_T$, is selected so that labels of the target concept can be accurately predicted through them. Once C_T^* is given, the classification procedure of TrBagg is the same as that of standard Bagging. But only the result given by classifiers in C_T^* can take part in the voting process.

*2) TrAdaboosting algorithm.*TrAdaboosting[15] is a transfer learning strategies based on Adaboost algorithm[16]. The traditional Adaboost defaults that the distributions of test set and training set are the same. It assigns each training sample some certain initial weight. When someone is misclassified, the Adaboosting would increase its weight to emphasize it. Though emphasis, the probability of misclassification when next training is coming will descend. Therefore, after several times iteration, the training quality of model will reach a high level. When the concepts of source domain and transfer learning are introduced, the principles of Adaboosting keep available as well. However, because of the difference between *T* and *S*, *TrAdaboosting* would descend its weight to control the training process when the sample from *S* is misclassified, besides that operation is designed based on Hedge (β).

Although either Trbagging or TrAdaboosting has its own advantages, neither of them takes imbalanced class distribution of data into consideration. So they can hardly predict churn exactly when they are involves in CRM field. Thus the transfer ensemble model for imbalanced data (TEMID) is proposed in this paper.

III. TRANSFER ENSEMBLE MODEL FOR IMBALANCED DATA-TEMID

A. The basic idea of TEMID

Primarily, the overall process of TEMID can be stated as follows: in TEMID, we firstly balance the class distributions of *T* and *S* by over-sampling them respectively. Then, we under-sample the merge set of balanced *T* and balanced *S* and the new training sets are born. Thirdly, a number of classifiers are trained by those new born training sets. There is a local area in training sets corresponded with every sample to be classified and the quality of each trained classifier should be checked in those local areas, to be exact, TEMID will give each classifier a certain weight due to its classification accuracy in local areas. Lastly, considering its weight, ensemble classifiers will come up with final results.

In fact, the merge set of balanced *T* and balanced *S* could be recognized as the result of *T'*s learning from *S* after over-sampling. Since in the training process, the importance of samples from *S* and *T* is very different, we must be aware of the fact that there is plenty of redundant information in the training set after the learning. So, the under-sampling of the merge set can remove some useless information to some extend.

B. Algorithm description

Denote target domain of the churn prediction problem by *T*, which consists of $m1$ samples, the related source domain by *S*, $m2$ samples contained and there $m2>m1$, the training set in target domain by *Tr*, the test set by *Te*, $n1$ and $n2$ samples respectively and there $m1= n1+n2$, the sample to be classified in *Te* by xi^*, $xi^* \in Te$ and there $i=1, 2, \ldots, n2$, the available base classifier set $C = \{C_1, C_2, \ldots, C_N\}$, and the class label 0 and 1, where 1 refers to churn customer and 0 refers to non-churn one. In addition, the labels of *Tr* and *S* are known while the label of Te is unknown.

Input: *Tr*, *Te* and *S*.

Output: the prediction of the class label of *Te* in this model.

1. Divide *Tr* into two sets, *Tr*1 for churn class and *Tr*2 for non-churn class. Divide *S* into two sets, *S*1 for churn class and *S*2 for non-churn class.
2. Carry out $|Tr2|-|Tr1|$ times random sampling with replacement in *Tr*1, one sample for each time, and added each sample to *Tr*1 right after sampled, and a new dataset *U*1 occurs. Carry out $|S2|-|S1|$ times random sampling with replacement in *S*1, one sample for each time, and added each sample to *S*1 right after sampled, and a new dataset *U*2 occurs
3. Random sample $|U2|-|U1|$ samples without replacement in *U*2 and combine remain samples with *U*1 to get the new training set *Tr'*. Use *Tr'* to train a classifier, a trained classifier $\hat{C}_j$ is acquired.

4. Repeat Step 3 N times until N classifiers are trained, $\hat{C} = \{\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_N\}$

5. For each sample to be classified, $xi^* \in Te$, i=1, 2, …, $|Te|$

5.1 Find k-neighbors of xi^* in Tr and compose them to be a new training set, which is a local area of xi^*, denoted by $D_k^i = (x_1, x_2, \ldots, x_k)$.

5.2 Classify D_k^i by $\hat{C} = \{\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_N\}$ respectively and in terms of xi^*, each classification accuracy caused by $\hat{C}_j$ classifying D_k^i can be acquired as Acc_j^i, $\mathbf{Acc}^i = (Acc_1^i, Acc_2^i \ldots Acc_N^i)$

5.3 Denote the weight of $\hat{C}_j$ for xi^* by w_j^i, there $w_j^i = Acc_j^i / \sum \mathbf{Acc}^i$, $\mathbf{w}^i = (w_1^i, w_2^i \ldots w_N^i)$. Get the classification result of xi^* under each classifier $\hat{C}_j$, r_j^i is acquired and $\mathbf{r}^i = (r_1^i, r_2^i \ldots r_N^i)$, where r_j^i=0 or 1.

5.4 Denote the final classification result for xi^* by R^i

$$R^i = \begin{cases} 0 & \sum_{k=1}^{N} w_k^i r_k^i < 0.5 \\ 1 & \sum_{k=1}^{N} w_k^i r_k^i \geq 0.5 \end{cases} \quad (1)$$

IV. EXPERIMENTAL ANALYSIS

In order to evaluate the performance of the TEMID used in customer churn prediction, we selected two dataset with highly class imbalanced distribution as experimental data set. What's more we compared it to three usual models

A. Data description

1) "Churn" dataset: "Churn" originated in a famous machine learning database-UCI database in California University. In this dataset, it defines the churn by one mobile customer's giving up all the services of one telecommunication company，where the time window is three continuous months. There are 19 characteristic (8 of which are continuous variables, 10 are discrete variables and 1 is class label), and 3333 samples (2850 of which is for churn and 483 for non-churn, the proportion of them is 5.9006), in the dataset. Unquestionably, the class distribution of churn is highly imbalance.

2) "CBC" dataset: "CBC" dataset (Churn of Bank in Chongqing) comes from a commercial bank credit card customer database in Chongqing and all the data are collected within 2010 May-Oct. There are 25 variables for each sample, 8 of which are continuous variables and 17 of which are discrete variables. We define the churn by that one customer canceled his or her credit card in 2010 May-Oct. or did not spend any during 3 consecutive months, After simple preprocess of data, we found 1244 sample as target domain, in which 1151 samples are for churn while 104 samples are for non-churn. Meanwhile, we found another dataset which came from another credit business, where 1802 of them is for churn while 198 of them is for non-churn and it could be treated as source domain.

B. Experimental setup

1) Dataset preprocess: Because the distribution of "churn" is independent as a whole, we must divide it into T and S purposely. Two datasets have different distribution and there $|S|:|T| \approx 7:3$. Exactly, that operation is as follows: selecting a variable v, a certain region A, we can acquire: $S = \{x \mid x \in churn \text{ and } v(x) \in A\}$, $T = \{x \mid x \in churn \text{ and } v(x) \notin A\}$. In that way the independence and otherness of S and T can be assured, meanwhile the similarity is assured as well. As for the dataset "CBC", S and T have been given in advance.

As for the division of Te and Tr in the target domain T, we use the method of random sampling without replacement. Sampling about $0.3|T|$ samples randomly from T, Te is acquired, while remain data is made up of the training set Tr. Another vital point is that the method of multi-interval discretization of continuous-valued attributes for classification learning proposed by Fayyad and Irani [17] should be used to discretize the continuous data so that dataset fit the classification models.

2) Experimental setup: According to the idea of transfer learning, we design 3 algorithms to be compared with TEMID: (1) Tr-SVM(Transfer Support vector machine) is an improved traditional machine learning algorithm and two phases have been added to SVM, in which one is Over-sampling phases to overcome class-imbalance and the other is learning phase that generate a merge set Tr' of T and S, $Tr' = S \cup Tr$, as the new training set. (2) TrBagg and TrAdaboosting have been improved by generating Over-sampling phase respectively and both of them use SVM as basic classifier. In that way, including TEMID, all of them acquire the captivity to overcome class-imbalance.

3) Evaluation criteria： In order to evaluate the performance of the model, a confusion matrix has been introduced (Table 1). And based on it, we generate 6 usual criterion:1) the accuracy of classification, Acc=（$D1$+ $D4$）/ ($D1$+$D2$+$D3$+$D4$)×100%; 2) Specificity, which describes the classification accuracy of majority, Sp= $D1$/ ($D1$+ $D2$) ×100%; 3) Sensitivity, which describe the classification accuracy of minority, Se= $D4$/ ($D3$+ $D4$) ×100%; 4) Hit Ratio = $D4$/ ($D2$+ $D4$) ×100%; 5) Intensification Factor = Hit Ratio /the churn of the test dataset; 6) The area under the receiver operating characteristic curve (AUC).

Table 1 Customer churn prediction confusion matrix

Sample status	Forecast for Negative class	Forecast for positive class	Total
Negative （churn）	*D*1	*D*2	*D*1+ *D*2
Positive （non-churn）	*D*3	*D*4	*D*3+ *D*4
Total	*D*1+ *D*3	*D*2+ *D*4	*D*1+*D*2+*D*3+*D*4

4) Parameter of TEMID setup：In TEMID, *k* and *N* can not be determined optimally. For (*N*, *k*）, we traverse all the optional values, there *N*=5,10,15,20,25,30 and *k*=3,5,7,9,11. Lastly, we select the group with the best result , （*N* *, *k**） , and regard its result as final output of

C. Experimental results and analysis

1) "Churn" dataset：Through a number of experiments, we found out that TEMID can perform the best when N=15 and k=11, The results with other three algorithms are as follows (Table 2).As illustrated, Tr-SVM, Trbagg and TEMID can solve problem of customer churn prediction with imbalanced class distribution because all their *Acc* are over 80% and their *Se* can reach about 70%. Especial for TEMID, it is actually satisfying because the *Se*-the most significant index in fact- reaches the peak at 77% which is the highest one in all the algorithms. However, the serious class-imbalance causes the Hit ratio a little low to some degree.

2) "CBC" dataset: The results are follows(Table 4). Since "CBC" came from the real situation, it can explain the performance better. As shown by Table 4, performance of all four algorithms is very excellent, where *Acc*, *Se*, *Sp* and AUC are over 90%; In comparison, the *Se* of TEMID is extremely excellent which reaches the peak at 92% (even higher than *Sp*). On the other hand, despite that four algorithms have their own advantages, shown in Table 4, TEMID perform the best as a whole.

Table 2：The comparison of performance in "churn" dataset

	Acc	Se	Sp	Hit ratio	Intensification factor	AUC	Remark
Tr-SVM	0.8239	0.7404	0.8325	0.2886	3.6075	0.8918	
TrAdaboosting	0.8386	0.3432	0.8872	0.2317	2.8963	0.5968	
Trbagging	0.8017	0.6717	0.8147	0.2644	3.3050	0.9125	
TEMID	**0.8124**	**0.7710**	**0.8166**	**0.2828**	**3.5355**	**0.9066**	***k*=11， *N*=15**

Table 3：The comparison of performance in "CBC" dataset

	Acc	Se	Sp	Hit ratio	Intensification factor	AUC	Remark
Tr-SVM	0.9165	0.8792	0.9208	0.5515	5.6805	0.9440	
TrAdaboosting	0.9156	0.9045	0.9171	0.5333	5.4930	0.9592	
Trbagging	0.9257	0.8527	0.9339	0.5903	6.0801	0.9178	
TEMID	**0.9142**	**0.9209**	**0.9136**	**0.5336**	**5.4965**	**0.9639**	***k*=9， *N*=25**

CONCLUSION

In summary, according the experiments above, we can draw the conclusion that, in the TEMID, the transfer learning strategy can factually be used to enrich the information within the system and therefore the sampling can overcome the class-imbalance more reliably. In addition, using the multiple classifiers ensemble method could make the most of classifiers and implement the classification. Thus Transfer ensemble model for imbalanced data is a considerable approach to achieve the purpose for customer churn prediction with imbalanced class distribution.

ACKNOWLEDGMENT

This research is supported by the Natural Science Foundation of China under Grant Nos. 70771067 and 71071101, Soft Science Program of Sichuan Province under No. 2010ZR0132 and Research Start-up Project of Sichuan University under No. 2010SCU11012. Corresponding author Jin Xiao..

REFERENCES

[1] F. F. Reichheld and T. Teal, The loyalty effect: The hidden force behind growth, profits, and lasting value, MA: Harvard Business Press, 2001.

[2] S. V. Nath , "Data warehousing and mining: customer churn analysis in the wireless industry," PhD thesis, Florida: Florida Atlantic University, 2003.

[3] W. H. Au, K. C. C. Chan, and X. Yao, "A novel evolutionary data mining algorithm with applications to churn prediction," *IEEE Transactions on Evolutionary Computation*, vol. 7, pp. 532-545, 2003.

[4] Guoen Xia and Weidong Jin, "Model of customer churn prediction on support vector machine," Systems Engineering-Theory & Practice, vol.28, pp. 71-77, 2008 (In Chinese).

[5] W. Verbeke, D. Martens, et al., "Building comprehensible customer churn prediction models with advanced rule induction techniques," Expert Systems with Applications, vol. 38, pp. 2354-2364, 2011.

[6] C. S. Lin, G. H. Tzeng, and Y.C. Chin, “Combined rough set theory and flow network graph to predict customer churn in credit card accounts,” Expert Systems with Applications, vol. 38, pp. 8-15, 2011.

[7] V. N. Vapnik, Statistical learning theory, Wiley-Blackwell 1998.

[8] G. E. Batista, ,R.C. Prati, and M.C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” ACM SIGKDD Explorations Newsletter, vol. 6, pp. 20-29, 2004.

[9] B. Lariviere and D. Van den Poel, “Predicting customer retention and profitability by using random forests and regression forests techniques,” Expert Systems with Applications, vol. 29, pp. 472-484, 2005.

[10] W. Buckinx and D. Van den Poel, “Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting,” European Journal of Operational Research, vol. 164, pp. 252-268, 2005.

[11] A. Lemmens and C. Croux, “Bagging and boosting classification trees to predict churn,” Journal of Marketing Research, vol. 43, pp. 276-286. 2006.

[12] S.J. Pan, and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1345-1359, 2009.

[13] T. Kamishima, M. Hamasaki, and S. Akaho. “TrBagg: a simple transfer learning method and its application to personalization in collaborative tagging,” *Proc. 9th IEEE International Conference on Data Mining(ICDM)*, Dec. 2009.

[14] L. Breiman, “Bagging predictors,” Machine learning, vol. 24, pp. 123-140. 1996.

[15] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for Transfer Learning," *Proc. 24th Int'l Conf. Machine Learning(ICMA)*, June 2007, pp. 193-200.

[16] Y. Freund, and R. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, pp. 119-139, 1997.

[17] U.M. Fayyad, and K.B. Irani, “Multi-interval discretization of continuous-valued attributes for classification learning,” *Proc. 13th International Joint Conference on Artificial Intelligence (IJCAI)*, 1993: pp. 1022–1027.