

# Application of a cost-sensitive method for churn prediction in telecommunication industry

Zhao Wei He Jianmin Wang Chunlin Chen Jinbo

(School of Economics and Management, Southeast University, Nanjing 210096, China)

**Abstract:** To deal with the data mining problem of asymmetry misclassification cost, an innovative churn prediction method is proposed based on existing churn prediction research. This method adjusts the misclassification cost based on the C4.5 decision tree as a baseline classifier, which can obtain the prediction model with a minimum error rate based on the assumption that all misclassifications have the same cost, to realize cost-sensitive learning. Results from customer data of a certain Chinese telecommunication company and the fact that the churners and the non-churners have different misclassification costs demonstrate that by altering the sampling ratio of churners and non-churners, this cost-sensitive learning method can considerably reduce the total misclassification cost produced by traditional classification methods. This method can also play an important role in promoting core competence of Chinese telecommunication industry.

**Key words:** cost-sensitive learning; C4.5; telecommunication industry; customer churn

Due to the deregulation and the great advances in new technologies, the competition in the Chinese telecommunication industry is getting severe. Accordingly, churn prediction and management have become of great concern to the operators. Past research on churn prediction mainly employed classification techniques for the construction of churn prediction models. Hung et al. studied churn management of the wireless industry based on the C5.0 decision tree algorithm<sup>[1]</sup>. Ma et al. used the C4.5 decision tree algorithm to identify churn customers<sup>[2]</sup>. Mozer applied artificial neural networks to predict the customer churn rate in telecommunication companies<sup>[3]</sup>. Yao et al. applied multi-criteria neural networks to recognize unusual telecom customers<sup>[4]</sup>. But the neural network techniques often lack interpretability, and it also requires a long training time to its iterative, which makes it impractical for churn prediction in the telecommunication industry. Consequently, the decision tree algorithm appears to be more appropriate for targeted learning and prediction, because it is capable of efficiently generating interpretable knowledge in an understandable form.

From the above, we can see that churn prediction is always regarded as a common classification problem, which assumes that each instance has the same misclassification cost. However, it is not necessarily the case in telecommunication industry. The cost of predicting a non-churner as a churner is only the retention fee paid by the company, while the cost of predicting a churner as a non-churner is almost equal to the cost of

acquiring a new customer. Therefore, we should consider the misclassification cost of different instances into the classifiers. This classification method is called cost-sensitive learning (CSL). CSL is a challenge to traditional classification methods, which is at its infancy level and only used in such domains as credit card fraud detection<sup>[5]</sup>, text classification<sup>[6]</sup> and bankrupt prediction<sup>[7]</sup>. Zheng et al. proposed a cost-sensitive support vector machine algorithm to modify the misclassification costs of each sample<sup>[8]</sup>.

This paper presents a CSL method to study customers churn in the Chinese telecommunication industry. Choosing under-sampling to modify the misclassification cost of different instances and combining with the C4.5 decision tree as a baseline classifier, we obtain a CSL method (for short CS-C4.5). Based on customer data of a telecommunication company in China, we use the CS-C4.5 to obtain a prediction model with minimum misclassification costs.

## 1 Supervised Learning and Its Evaluations

Given a set of labeled training data  $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ , where  $\mathbf{x}_i$  is a vector of continuous or discrete values called attributes and  $y_i$  is the label of  $\mathbf{x}_i$ . Since churn prediction is a two-classification problem, the labels can be the elements of a discrete set of classes  $y_i \in \{0, 1\}$ . Assuming that all the examples that are presented in the learning algorithm are drawn from the same distribution  $p(\mathbf{x}, y)$ , supervised learning means finding a model or a hypothesis  $h$  which can correctly label a high proportion of unlabeled examples  $\mathbf{x}'$  drawn from the same distribution, that is

$$(\mathbf{x}', h(\mathbf{x}')) \sim p(\mathbf{x}, y) \quad (1)$$

Received 2006-10-26.

**Biographies:** Zhao Wei (1980—), male, graduate; He Jianmin (corresponding author), male, professor, nj.jian@public1.ptt.js.cn.

For misclassifying an instance being a loss, the goal of the learning algorithm is to minimize the total expected cost about  $h$ :

$$E(h) = \sum_{(\mathbf{x}_i, y_i) \in S} p(\mathbf{x}_i, y_i) c(h(\mathbf{x}_i), y_i) \quad (2)$$

where  $c(h(\mathbf{x}_i), y_i)$  is the cost function representing the loss incurred by  $h$  on an instance  $(\mathbf{x}_i, y_i)$ , and  $E(h)$  is the total expected cost of all the training samples.

In supervised learning,

$$c(h(\mathbf{x}_i), y_i) = \begin{cases} 0 & h(\mathbf{x}_i) = y_i \\ 1 & h(\mathbf{x}_i) \neq y_i \end{cases} \quad (3)$$

It shows that the cost function is 0 when an instance is predicted correctly and 1 otherwise. Obviously, higher accuracy means less cost. Therefore, supervised learning methods are based on 0/1 loss.

Supervised learning algorithms include a decision tree, a logistic regression and an artificial neural network, of which the C4.5 decision tree is the most basic method. The core idea of C4.5 is to select attributes based on a gain ratio, which is described in Ref. [9] in detail. The evaluation indices are accuracy and recall based on a confusion matrix, which is shown in Tab. 1.

**Tab. 1** Confusion matrix for churn prediction

Prediction	Actual churn	Actual non-churn
Churn	True positive (TP)	False positive (FP)
Non-churn	False negative (FN)	True negative (TN)

Note that “actual” represents the true class for an instance, while prediction represents the label assigned by the classifier. Let  $N_{TP}$ ,  $N_{FP}$ ,  $N_{FN}$  and  $N_{TN}$  represent the number of type TP, FP, FN and TN, so the accuracy  $A$  and recall  $R$  are defined as

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}} \quad (4)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (5)$$

From Tab. 1, the cost of FN type error is much higher than that of FP type error, which means we should pay more attention to  $R$  than  $A$ .

## 2 Cost-Sensitive Learning and Its Evaluation

As mentioned above, in the churn prediction task, the misclassification error cost is not equal. Then we should expand the cost matrix to the general case for getting a classifier based on the minimum expected cost. By using Bayes' optimal representation and known probabilities  $p(j | \mathbf{x})$  for class  $j$ , an example  $\mathbf{x}$  should be labeled with  $y_{\text{opt}}^{[10]}$

$$y_{\text{opt}} = \arg \min_{y \in Y} \sum_{j=0}^1 p(j | \mathbf{x}) c(y, j) \quad (6)$$

where  $Y = \{0, 1\}$ ,  $j \in \{0, 1\}$ . Let  $h$  be a classifier, and  $p_h(i, j)$  be the probability that an example is classified

by  $h$  to class  $i$ , which is generated at random and belongs to class  $j$ . Then, according to Eq. (2), the expected cost of  $h$  on  $\mathbf{C}$  is

$$L(h) = \sum_{i=0}^1 \sum_{j=0}^1 p_h(i, j) c(i, j) \quad (7)$$

where  $p_h(i, j) = p_h(i | j) p(j)$ ,  $p(j)$  is the probability that an example belongs to class  $j$ , and  $p_h(i | j)$  is the probability that  $h$  will classify an example from class  $j$  as class  $i$ . CSL is to find the classifier  $h$  to minimize  $L(h)$ .

An element of cost matrix  $c(i, j)$  is positive when it represents an actual cost, whereas, if it is negative, it represents a benefit. In general, we assume that the cost matrix is ① given beforehand; ② independent of  $\mathbf{x}$ ; ③ stationary, i. e. it does not change during learning<sup>[11]</sup>. Writing churner labeled as 0, and non-churner labeled as 1, the cost matrix corresponding to the confusion matrix is shown in Tab. 2.

**Tab. 2** Cost matrix for churn prediction

Type	Cost
TP	$c(0, 0)$
FP	$c(0, 1)$
FN	$c(1, 0)$
TN	$c(1, 1)$

When the cost matrix complies with the three conditions, Margineantu pointed out that changing the measurement unit for the costs would not alter the optimal decision. Moreover, any such cost matrix  $\mathbf{C}$  can be transformed into an equivalent cost matrix  $\mathbf{C}'$  with zero diagonal elements and positive values for the off-diagonal elements by adding  $-c(j, j)$  to every column  $j$  entry of  $\mathbf{C}$ . So we have

$$c'(0, 0) = 0, c'(1, 1) = 0 \quad (8)$$

$$c'(1, 0) = c(1, 0) - c(0, 0), c'(0, 1) = c(0, 1) - c(1, 1) \quad (9)$$

Then Eq. (6) can be transformed into

$$y_{\text{opt}} = \arg \min_{y \in \{0, 1\}} \sum_{j=0}^1 p(j | \mathbf{x}) c'(y, j) \quad (10)$$

Noting that  $p(\mathbf{x})$  does not depend on the target label, and, therefore, does not affect the classification decision, based on the Bayes' theorem we have

$$y_{\text{opt}} = \arg \min_{y \in \{0, 1\}} \sum_{j=0}^1 p(j) p(\mathbf{x} | j) c'(y, j) \quad (11)$$

Let  $p'(j) = p(j) c'(y, j) / \sum_{j=0}^1 p(j) c'(y, j)$ , then

Eq. (11) can be reformulated as

$$y_{\text{opt}} = \arg \min_{y \in \{0, 1\}} \sum_{j=0}^1 p'(j) p(\mathbf{x} | j) \quad (12)$$

Under-sampling is to retain the highest  $p'(j)$  and select a fraction  $p'(i) / p'(j)$  of the training data instances in the resampled training set.

Under-sampling can get a good result based on

traditional classifiers, while avoiding the complex modification on the algorithms themselves. In this paper, we take the under-sampling approach by combining the C4.5 decision tree as the base classifier to realize CSL (CS-C4.5).

The weighed accuracy and the total cost are usual cost-sensitive evaluation indices, which can be calculated by the cost ratio ( $R_C$ )<sup>[12]</sup>.  $R_C$  is defined as

$$R_C = \frac{c(0, 1) - c(1, 1)}{c(1, 0) - c(0, 0)} \quad (13)$$

Then the calculation formulae of the weighed accuracy  $A_w$  and the total cost  $C_T$  are<sup>[12]</sup>

$$A_w = \frac{R_C N_{TN} N_{TP}}{R_C (N_{TN} + N_{FP}) + N_{TP} + N_{FN}} \quad (14)$$

$$C_T = N_{FN} (c(1, 0) - c(0, 0)) + N_{FP} (c(0, 1) - c(1, 1)) \quad (15)$$

### 3 Empirical Studies

Call details collected from a certain Chinese telecommunication company were investigated. Specifically, the call records made between January 2005 and April 2005, time interval from May to June in 2005, and prediction period was July 2005. We excluded those subscribers whose services were terminated by the company during this data collection period because their payments were delinquent. In this paper, we only take the voluntary churners into consideration. These are defined as those customers who have consumed nothing in a consecutive three months period or have cancelled their accounts by themselves. As a result, the customer database consists of  $51 \times 10^3$  customers, including 1 000 churners and  $50 \times 10^3$  non-churners.

The investigated company collects and maintains plenty of information about its customers. For each customer, the attributes consist of contract data, payment type data, contract starting data and contract termination data. Moreover, we also design many derived variables to describe the behavior characteristics of telecommunication customers, and acquire 79 attributes in all. Then we can use the CS-C4.5 method to the simplified data by choosing 80% samples as training data and 20% as test data. First, the cost matrix must be given. Assuming in churn prediction of telecommunication customers, the cost matrix may be

$$C = \begin{bmatrix} -300 & 5 \\ 0 & 0 \end{bmatrix}$$

Predicting a churner correctly has a benefit of 300 yuan that takes a negative in the cost matrix. Predicting a non-churner will loss the retention fee 5 yuan that takes a positive in the cost matrix. Other situations in the matrix have no loss or benefit, and thus take zero value. In order to get  $R_C$ , Margineantu's transformation is applied to get  $C'$  with zero diagonal elements.

$$C' = \begin{bmatrix} 0 & 5 \\ 300 & 0 \end{bmatrix}$$

Obviously,  $R_C = 1/60$ , which means the cost of an FN type error is 60 times larger than an FP type error. For convenience, the two types of misclassification costs are set 60 and 1, respectively, then  $C_T = 60N_{FN} + N_{FP}$ .

Since the training misclassification cost ratio (TCR) is always unequal to the evaluation cost ratio (ECR)<sup>[12]</sup>, this paper uses under-sampling to modulate the TCR with 50:1, 40:1, 30:1, 20:1, 10:1, 5:1, 4:1, 3:1, 2:1, 1:1, 1:2, 1:3, 1:4, 1:5. Tab. 3 shows the evaluation results. When the TCR and ECR are both equal to 50:1, the accuracy is even greater than 98%, while the recall is only 19%, which means more than 80% of churners have not been detected. Therefore, traditional classifiers would not get the ideal results for neglecting different classification costs. Meanwhile, the total cost of the model attains to 9 727. With the decrease of TCR, the accuracy falls continuously after a short rising, while the recall increases monotonously. The maximum accuracy 98.43% occurs in the TCR of value of 20:1, but the total cost is relatively high.

Fig. 1 gives the curve for the total cost changing with the TCR. The TCR value that generates the lowest total cost is 1:1; thus when the TCR value is closer to 1:1, suboptimal results are produced. Returning to Tab. 3, the accuracy of the CS-C4.5 (81.46%) is much lower than that of the C4.5 (98.34%). But taking more practical objectives into consideration,  $C_T$  of the CS-C4.5 (3 779) is the ideal result, compared with  $C_T$  of the C4.5 (9 727). The CSL gets lower total cost with lower accuracy, which is incompatible with traditional classifiers. In addition, the weighted accuracy almost has the same trend as  $C_T$  for adding the cost ratio into the accuracy. All the above have proved the validity of our method.

**Tab. 3** Evaluation results for CS-C4.5

TCR	$A/\%$	$R/\%$	$A_w/\%$	$C_T$
50:1	98.34	19.00	55.79	9 727
40:1	98.36	20.50	56.60	9 548
30:1	98.41	23.50	58.23	9 189
20:1	98.43	24.50	58.78	9 069
10:1	97.92	34.50	63.90	7 941
5:1	96.83	45.50	69.30	6 754
4:1	94.17	60.00	75.84	5 315
3:1	93.13	63.00	76.97	5 067
2:1	89.92	69.00	78.70	4 686
1:1	81.46	84.00	82.82	3 779
1:2	69.68	93.00	82.19	3 919
1:3	63.07	96.50	81.00	4 180
1:4	56.84	98.50	79.19	4 579
1:5	54.23	97.50	77.44	4 964

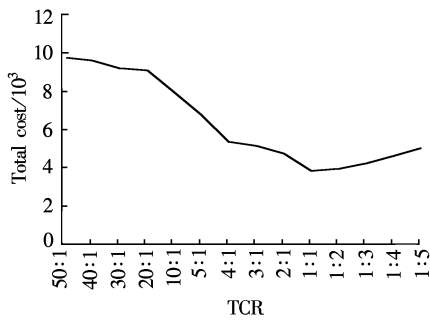


Fig. 1 Curve for the total cost changing with TCR

## 4 Conclusion

Churn prediction and management are critical in the fast changing, fiercely competitive telecommunication industry. To be able to improve customer retention, a telecommunication service provider has to be able to predict customer churn. In response to the limitations of existing churn prediction methods, we propose a cost-sensitive churn prediction technique CS-C4.5, which combines the under-sampling method and the C4.5 decision tree to address the challenge of asymmetry misclassification costs between churners and non-churners. The empirical evaluation results suggest that the CS-C4.5 outperforms traditional classifiers with lower total cost.

This study benefits not only churn prediction research and practice but also other data mining applications with identical or similar characteristics. Further research can be applied in those industries (e. g. credit card issuers and internet service providers) where fierce competition provides incentives for customers to switch. Therefore, expanding the developed technique to other industries suggests interesting directions for future research.

## References

[1] Hung S Y, Yen C D, Wang H Y. Applying data mining to

telecom churn management[J]. *Expert Systems with Applications*, 2006, **31**(3): 515 – 524.

- [2] Ma H M, Yin H B, Guo X. The mining model of defecting customers[J]. *Journal of Huazhong University of Science and Technology: Nature Science Edition*, 2003, **31**(9): 28 – 30. (in Chinese)
- [3] Mozer M C. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry [J]. *IEEE Trans on Neural Networks*, 2000, **11**(3): 690 – 696.
- [4] Yao M, Shen B, Li M F. A kind of unusual customers recognition system based on multi-criteria neural network and CART in telecom system[J]. *Systems Engineering — Theory & Practice*, 2004, **24**(5): 78 – 83. (in Chinese)
- [5] Chan P, Stolfo S. Towards scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection[C]//*Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. New York, USA, 1998: 164 – 168.
- [6] Liu Y, Yang Y, Carbonell J. Boosting to correct inductive bias in text classification[C]//*Proceedings of the Eleventh International Conference on Information and Knowledge Management*. Pittsburgh, USA, 2002: 348 – 355.
- [7] Foster D P, Stine R A. Variable selection in data mining: building a predictive model for bankruptcy[J]. *Journal of the American Statistical Association*, 2004, **99**(466): 303 – 313.
- [8] Zheng E H, Li P, Song Z H. SVM-based cost sensitive mining[J]. *Information and Control*, 2006, **35**(3): 294 – 298. (in Chinese)
- [9] Quinlan J R. *C4.5: Programs for machine learning*[M]. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
- [10] Duda R O, Hart P E, Stork D G. *Pattern classification* [M]. 2nd ed. John Wiley and Sons, Inc., 2000.
- [11] Viaene S, Dedene G. Cost-sensitive learning and decision making revisited [J]. *European Journal of Operational Research*, 2005, **166**(11): 212 – 220.
- [12] Ciraco M, Rogalewski M, Weiss G. Improving classifier utility by altering the misclassification cost ratio[C]//*ACM SIGKDD Workshop on Utility-Based Data Mining*. Chicago, USA, 2005: 46 – 52.

# 一种代价敏感学习方法在电信业流失预测中的应用

赵巍 何建敏 王纯麟 陈金波

(东南大学经济管理学院, 南京 210096)

**摘要:**根据已有的流失预测方法,提出新的流失预测方法解决数据挖掘中的非对称错分代价问题。该方法以传统 C4.5 决策树算法为基准分类器,融合代价调整方法实现代价敏感学习。相比之下, C4.5 决策树算法仅是基于样本错分代价相同假定,建立了一种错分率最低而非总错分代价最低的预测模型。基于某电信企业的客户数据,及流失客户和非流失客户代价非对称的实际,实证研究结果表明,CS-C4.5 通过调整流失类和非流失类样本的比例,大大降低了传统分类算法的样本错分总代价。该方法对于提高电信企业的核心竞争力具有重要的现实意义。

**关键词:**代价敏感学习; C4.5; 电信业; 客户流失

**中图分类号:** F626; TP391