Multiclass Credit Cardholders' Behaviors Classification Methods*

Gang Kou[1], Yi Peng[1,**], Yong Shi[1, 2], and Zhengxin Chen[1]

[1] College of Information Science & Technology,
University of Nebraska at Omaha, Omaha, NE 68182, USA
[2] Chinese Academy of Sciences Research Center on Data Technology
& Knowledge Economy, Graduate University of the Chinese Academy of Sciences,
Beijing 100080, China
{gkou, ypeng, yshi, zchen}@mail.unomaha.edu

Abstract. In credit card portfolio management a major challenge is to classify and predict credit cardholders' behaviors in a reliable precision because cardholders' behaviors are rather dynamic in nature. Multiclass classification refers to classify data objects into more than two classes. Many real-life applications require multiclass classification. The purpose of this paper is to compare three multiclass classification approaches: decision tree, Multiple Criteria Mathematical Programming (MCMP), and Hierarchical Method for Support Vector Machines (SVM). While MCMP considers all classes at once, SVM was initially designed for binary classification. It is still an ongoing research issue to extend SVM from two-class classification to multiclass classification and many proposed approaches use hierarchical method. In this paper, we focus on one common hierarchical method – one-against-all classification. We compare the performance of See5, MCMP and SVM one-against-all approach using a real-life credit card dataset. Results show that MCMP achieves better overall accuracies than See5 and one-against-all SVM.

Keywords: multi-group classification, decision tree, See5, Multiple criteria mathematical programming (MCMP), one-against-all SVM.

1 Introduction

One of the major tasks in credit card portfolio management is to reliably predict credit cardholders' behaviors. This task has two impacts in credit management: (1) identify potential bankrupt accounts and (2) develop appropriate policies for different categories of credit card accounts. To appreciate the importance of bankrupt accounts prediction, some statistics are helpful: There are about 1.2 billion credit cards in circulation in US. The total credit card holders declared bankruptcy in 2003 are 1,625,208 which are almost twice as many as the number of 812,898 in 1993 (New Generation Research 2004). The total credit card debt at the end of the first quarter

* This work was supported in part by Key Project #70531040, #70472074, National Natural Science Foundation of China; 973 Project #2004CB720103, Ministry of Science and Technology, China and BHP Billion Co., Australia.

** Corresponding author.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part IV, LNCS 3994, pp. 485 – 492, 2006.

2002 is about $660 billion (Cardweb 2004). Bankrupt accounts caused creditors millions of dollars lost each year. In response, credit card lenders have made great effort to improve traditional statistical methods and recognized that more sophisticated analytical tools are needed in this area. Development of appropriate policies for various groups of credit card accounts also has a great impact on credit card issuers' profits. From the creditor's standpoint, the desirable policies should help to keep the profitable customers and minimize the defaults. It is meaningful to conduct multiclass credit cardholders' behaviors classification because it enables card issuers to better manage credit card portfolio.

As one of the major data mining functionalities, classification has broad applications such as credit card portfolio management, medical diagnosis, and fraud detection. Based on historical information, classification builds classifiers to predict categorical class labels for unknown data. Multiclass classification refers to classify data objects into more than two classes.

Researchers have suggested various multiclass classification methods. Multiple Criteria Mathematical Programming (MCMP), decision tree, and Hierarchical Method for Support Vector Machines (SVM) are three of them. Decision tree induction is a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications (Menzies and Hu, 2003). The decision tree software we used in this paper is See5, a Windows95/NT decision tree and rule induction product (RuleQuest 2004). Because See5 is well-known for its high classification accuracy, it is included in this study as a benchmark. MCMP and SVM are both based on mathematical programming and there is no comparison study has been conducted to date. The purpose of this paper is to compare these multiclass classification approaches. While MCMP considers all classes at once, SVM was initially designed for binary classification. It is still an ongoing research issue to extend SVM from two-class classification to multiclass classification and many proposed approaches use hierarchical approach. In this paper, we focus on one common hierarchical method – one-against-all classification. Decision tree induction is a popular classification, so we won't describe it here. For more information about decision tree, please refer to Quinlan (1993).

This paper is structured as follows. The next section discusses the formulation of multiple-group multiple criteria mathematical programming classification model. The third section describes one-against-all SVM multiclass classification method. The fourth section compares the performance of See5, MCMP, and one-against-all SVM using a real-life credit card dataset. The last section concludes the paper.

2 Multi-group Multi-criteria Mathematical Programming Model

This section introduces a MCMP model for multiclass classification. The following models represent this concept mathematically: Given an r-dimensional attribute vector $a = (a_1,...,a_r)$, let $A_i = (A_{i1},..., A_{ir}) \in \Re^r$ be one of the sample records, where $i = 1,..., n$; n represents the total number of records in the dataset. Suppose k groups, G_1, G_2, …, G_k, are predefined. $G_i \cap G_j = \Phi, i \neq j, 1 \leq i, j \leq k$ and

$A_i \in \{G_1 \cup G_2 \cup ... \cup G_k\}$, $i = 1,..., n$. A series of boundary scalars $b_1<b_2<...<b_{k-1}$, can be set to separate these k groups. The boundary b_j is used to separate G_j and G_{j+1}. Let $\boldsymbol{X} = (x_1,..., x_r)^T \in R^r$ be a vector of real number to be determined. Thus, we can establish the following linear inequations (Fisher 1936):

$\boldsymbol{A}_i\boldsymbol{X} < b_1$, $\forall \boldsymbol{A}_i \in G_1$; (1)$b_{j-1} \leq \boldsymbol{A}_i\boldsymbol{X} < b_j$, $\forall \boldsymbol{A}_i \in G_j$; (2)$\boldsymbol{A}_i\boldsymbol{X} \geq b_{k-1}$, $\forall \boldsymbol{A}_i \in G_k$; (1)
$2 \leq j \leq$ k-1, $1 \leq i \leq$n.

In the classification problem, $\boldsymbol{A}_i\boldsymbol{X}$ is the score for the i^{th} data record. If an element $\boldsymbol{A}_i \in G_j$ is misclassified into a group other than G_j, then let $\alpha_{i,j}$ be the distance from $\boldsymbol{A}_i$ to b_j, and $\boldsymbol{A}_i\boldsymbol{X} = b_j + \alpha_{i,j}$, $1 \leq j \leq k-1$ and let $\alpha_{i,j-1}$ be the distance from $\boldsymbol{A}_i \in G_j$ to b_{j-1}, and $\boldsymbol{A}_i\boldsymbol{X} = b_{j-1} - \alpha_{i,j-1}$, $2 \leq j \leq k$. Otherwise, $\alpha_{i,j}, 1 \leq \mathrm{j} \leq \mathrm{k}, 1 \leq \mathrm{i} \leq \mathrm{n}$, equals to zero. Therefore, the total overlapping of data can be represented as $\sum_{j=1}^{k}\sum_{i=1}^{n}(\alpha_{i,j})^p$. If an element $\boldsymbol{A}_i \in G_j$ is correctly classified into G_j , let $\zeta_{i,j}$ be the distance from $\boldsymbol{A}_i$ to b_j, and $\boldsymbol{A}_i\boldsymbol{X} = b_j - \zeta_{i,j}$, $1 \leq j \leq k-1$ and let $\zeta_{i,j-1}$ be the distance from $\boldsymbol{A}_i \in G_j$ to b_{j-1}, and $\boldsymbol{A}_i\boldsymbol{X} = b_{j-1} + \zeta_{i,j-1}$, $2 \leq j \leq k$. Otherwise, $\zeta_{i,j}, 1 \leq \mathrm{j} \leq \mathrm{k}, 1 \leq \mathrm{i} \leq \mathrm{n}$, equals to zero. Thus, the objective is to maximize the distance $\left|\zeta_{i,j}\right|_p$ from $\boldsymbol{A}_i$ to boundary if $\boldsymbol{A}_i \in G_1$ or G_k and is to minimize the distance $\left| \frac{b_j - b_{j-1}}{2} - \zeta_{i,j} \right|_p$ from $\boldsymbol{A}_i$ to the middle of two adjunct boundaries b_{j-1} and b_j if $\boldsymbol{A}_i \in G_j, 2 \leq j \leq k-1$. So the distances of every data to its class boundary or boundaries can be represented as $\sum_{j=1 or k}\sum_{i=1}^{n}\left| \zeta_{i,j} \right|_p$ - $\sum_{j=2}^{k-1}\sum_{i=1}^{n}\left| \frac{b_j - b_{j-1}}{2} - \zeta_{i,j} \right|_p$. As a result, the single-criterion mathematical programming model can be set up as:

(**Model 1**) *Minimize* $w_\alpha \sum_{j=1}^{k}\sum_{i=1}^{n}\left| \alpha_{i,j} \right|_p$ - w_ζ ($\sum_{j=1 or j=k}\sum_{i=1}^{n}\left| \zeta_{i,j} \right|_p$ - $\sum_{j=2}^{k-1}\sum_{i=1}^{n}\left| \frac{b_j - b_{j-1}}{2} - \zeta_{i,j} \right|_p$)

S. T.: $\boldsymbol{A}_i\boldsymbol{X} = b_j + \alpha_{i,j} - \zeta_{i,j}$, $1 \leq j \leq k-1$ (2)

$$A_i X = b_{j-1} - \alpha_{i,j-1} + \zeta_{i,j-1}\ , 2 \le j \le k \tag{3}$$

$$\zeta_{i,j} \le b_j - b_{j-1}\ , 2 \le j \le k \text{ (a)} \quad \zeta_{i,j} \le b_{j+1} - b_j\ , 1 \le j \le k-1 \text{ (b)}$$

where A_i, $i = 1, \ldots, n$ are given, X and b_j are unrestricted, and $\alpha_{i,j}$, $\zeta_{i,j} \ge 0, 1 \le i \le n.$.(a) and (b) are defined as such because the distances from any correctly classified data ($A_i \in G_j, 2 \le j \le k-1$) to two adjunct boundaries b_{j-1} and b_j must be less than $b_j - b_{j-1}$. Let $p = 2$, then objective function in Model 1 can now be a quadratic objective and we have:

(Model 2) *Minimize* $w_\alpha \sum_{j=1}^{k}\sum_{i=1}^{n}(\alpha_{i,j})^2$ - w_ζ ($\sum_{j=1 or j=k}\sum_{i=1}^{n}(\zeta_{i,j})^2$ - $\sum_{j=2}^{k-1}\sum_{i=1}^{n}[(\zeta_{i,j})^2 - (b_j - b_{j-1})\zeta_{i,j}]$) (4)

Subject to: (4), (5), (c) and (d)

3 SVM One-Against-All Multiclass Classification

Statistical Learning Theory was proposed by Vapnik and Chervonenkis in the 1960s. Support Vector Machine (SVM) is one of the Kernel Machine based Statistical Learning Methods that can be applied on various types of data and can detect the internal relations among the data objectives. Given a set of data, one can define the kernel matrix to construct SVM and compute an optimal hyperplane in the feature space which is induced by a kernel (Vapnik, 1995). There exist different multi-class training strategies for SVM such as one-against-all classification, one-against-one (pairwise) classification, and Error correcting output codes (ECOC).

SVM-light (Joachims 2004) is a well known software package for support vector machine *binary* classification. It is not designed to perform multiclass classification. We apply SVM-light to two-group classifications, then implement a one-against-all procedure for a four-class classification. Suppose the four groups are A, B, C and D. The four-class one-against-all procedure is: ABCD⇒ A|B+C+D ⇒ A|B|C+D ⇒ A|B|C|D. Table 1 shows the classification results and is displayed in the format of confusion matrices, which pinpoint classification accuracies. Table 2 gives an analysis of classification accuracies and false alarm rates (the percentage of misclassified records to all records which are classified to a group). The assumption of one-against-all procedure is described as following:

The classification accuracy is stable. The classification accuracy of the forecasting dataset is equal to the classification accuracy of the testing dataset as well as the classification accuracy of the training dataset. The following symbols are used in this section.

N_x Number of records in group x

N_{xyz} Number of records in group x, y and z

4 Credit Cardholders' Behaviors Classification

The model proposed can be used in many fields, such as general bioinformatics, antibody and antigen, credit fraud detection, network security, text mining, etc. This research will focus on credit card classification. The real-life credit card dataset used in this paper is come from a US bank. It contains 6000 records and 7 variables. The variables are Interest charge, Interest charge as percent of credit line, Number of months since last payment, Credit line, Average payment of revolving accounts, Last balance to payment ratio, and Average OBT revolving accounts. This dataset has been used as a classic working dataset for various data analyses to support the bank's business intelligence. We define four classes for this dataset using a label variable: The Number of Over-limits. The four classes are: Bankrupt charge-off accounts (Number of Over-Limits≥ 12), Non-bankrupt charge-off accounts (7 ≤Number of Over-Limits ≤ 11), Delinquent accounts (2 ≤ Number of Over-Limits ≤ 6), and Current accounts (0 ≤ Number of Over-Limits ≤ 2). Bankrupt charge-off accounts are accounts that have been written off by credit card issuers because of cardholders' bankrupt claims. Non-bankrupt charge-off accounts are accounts that have been written off by credit card issuers due to reasons other than bankrupt claims. The charge-off policy may vary among authorized institutions. Delinquent accounts are accounts that haven't paid the minimum balances for more than 90 days. Current accounts are accounts that have paid the minimum balances or have not balances. For decision tree method, we use See5.0. MCMP is solved by LINGO 8.0, a software tool for solving nonlinear models (LINDO Systems Inc.). SVM one-against-all is implemented using SVM-light version 6.01 (Joachims 2004), a well-known SVM software.

Table 1. A example of one-against-all 4-classes classification results

1st step	A	B+C+D
Classified as Group A	a	$N_{bcd} - bcd$
Classified as Group B,C,D	$N_a - a$	bcd
2nd step	B	C+D
Classified as Group B	b	$\frac{bcd}{N_{bcd}} \times N_{cd} - cd$
Classified as Group C,D	$\frac{bcd}{N_{bcd}} \times N_b - b$	cd
3rd step	C	D
Classified as Group C	c	$\frac{bcd}{N_{bcd}} \times \frac{cd}{\frac{bcd}{N_{bcd}} \times N_{cd}} \times N_d - d$
Classified as Group D	$\frac{bcd}{N_{bcd}} \times \frac{cd}{\frac{bcd}{N_{bcd}} \times N_{cd}} \times N_c - c$	d

The four-group classification results of See5, MCMP, and SVM-light on the credit card data are summarized in Table 3, 4, and 5, respectively. In addition, we compute Type I and II error rates. Type I error is defined as the rate of records that are misclassified as Current to records that are classified as Current. Type II error is defined as the rate of records that are actually Current but are misclassified as the other three classes (Bankrupt charge-off, Non-bankrupt charge-off, and Delinquent) to records that are classified as the other three classes. Since misclassified Current accounts contribute to huge lost in credit card business and thus creditors are more concern about Type I error than Type II error. From the confusion matrices in Table 3, 4, and 5, we observe that (1) MCMP achieves the lowest test Type I error rate: 1.65%. SVM-light has the second lowest test Type I error rate: 1.7%. See5 has the highest test Type I error rate: 2.2%; (2) Among the three classification methods, MCMP has the best test classification accuracies for Delinquent, Charge-off, and Bankrupt classes. See5 has the best test classification accuracy for Current class.

Table 2. Accuracy and False Alarm Rate analysis of 4-classes classification results

	Accuracy	False Alarm Rate
A	$\frac{a}{N_a}$	$\frac{N_{bcd}-bcd}{a+N_{bcd}-bcd}$
B	$\frac{b}{N_b}$	$\frac{\frac{bcd}{N_{bcd}}\times N_{cd}-cd+\frac{N_a-a}{bcd+N_a-a}\times b}{\frac{bcd}{N_{bcd}}\times N_{cd}-cd+b}$
C	$\frac{c}{N_c}$	$\frac{\frac{bcd}{N_{bcd}}\times\frac{cd}{\frac{bcd}{N_{bcd}}\times N_{cd}}\times N_d-d+\frac{\frac{bcd}{N_{bcd}}\times N_b-b}{cd+\frac{bcd}{N_{bcd}}\times N_b-b}\times c+\frac{N_a-a}{bcd+N_a-a}\times\frac{cd}{cd+\frac{bcd}{N_{bcd}}\times N}}{\frac{bcd}{N_{bcd}}\times\frac{cd}{\frac{bcd}{N_{bcd}}\times N_{cd}}\times N_d-d+c}$
D	$\frac{d}{N_d}$	$\frac{\frac{bcd}{N_{bcd}}\times\frac{cd}{\frac{bcd}{N_{bcd}}\times N_{cd}}\times N_c-c+\frac{\frac{bcd}{N_{bcd}}\times N_b-b}{cd+\frac{bcd}{N_{bcd}}\times N_b-b}\times d+\frac{N_a-a}{bcd+N_a-a}\times\frac{cd}{cd+\frac{bcd}{N_{bcd}}\times N}}{\frac{bcd}{N_{bcd}}\times\frac{cd}{\frac{bcd}{N_{bcd}}\times N_{cd}}\times N_c-c+d}$

Table 3. See5 Credit Card Classification Results

Evaluation on training data (280 cases):					Accuracy	Error Rate
(1)	(2)	(3)	(4)	<-classified as		
66	3	0	1	(1): Current	94.29%	Type I
6	35	27	2	(2): Delinquent	50.00%	9.59%
1	5	56	8	(3): Charge-off	80.00%	Type II
0	6	37	27	(4): Bankrupt	38.57%	1.93%
Evaluation on test data (5720 cases):						
(1)	(2)	(3)	(4)	<-classified as		
3830	609	455	87	(1): Current	76.89%	Type I
83	182	289	48	(2): Delinquent	30.23%	2.20%
3	16	83	24	(3): Charge-off	65.87%	Type II
0	1	7	3	(4): Bankrupt	27.27%	63.80%

Table 4. MCMP Credit Card Classification Results

Evaluation on training data (280 cases):					Accuracy	Error Rate
(1)	(2)	(3)	(4)	<-classified as		
50	12	8	0	(1): Current	71.43%	Type I
5	55	10	0	(2): Delinquent	78.57%	13.79%
2	5	55	8	(3): Charge-off	78.57%	Type II
1	1	5	63	(4): Bankrupt	90.00%	9.01%
Evaluation on test data (5720 cases):						
(1)	(2)	(3)	(4)	<-classified as		
3406	1012	559	4	(1): Current	68.38%	Type I
53	440	100	9	(2): Delinquent	73.09%	1.65%
4	23	92	7	(3): Charge-off	73.02%	Type II
0	0	2	9	(4): Bankrupt	81.82%	69.78%

Table 5. SVM-light Credit Card Classification Results

Evaluation on training data (280 cases):					Accuracy	Error Rate
(1)	(2)	(3)	(4)	<-classified as		
40	22	8	0	(1): Current	57.14%	Type I
1	69	0	0	(2): Delinquent	98.57%	4.76%
0	0	70	0	(3): Charge-off	100.00%	Type II
1	1	2	66	(4): Bankrupt	94.29%	12.61%
Evaluation on test data (5720 cases):						
(1)	(2)	(3)	(4)	<-classified as		
3411	1135	136	299	(1): Current	68.48%	Type I
55	199	66	282	(2): Delinquent	33.06%	1.70%
3	27	23	73	(3): Charge-off	18.25%	Type II
1	0	1	9	(4): Bankrupt	81.82%	69.78%

5 Conclusion

This is the first time that we investigate the differences among decision tree, MCMP, and one-against-all SVM for multiclass classification using a real-life credit card dataset. The results indicate that MCMP achieves better classification accuracy than See5 and one-against-all SVM. In our future research, we will focus on the theoretical differences between MCMP and one-against-all SVM. Another topic of interest is to study the subject of reducing computational cost and improving algorithm efficiency for high dimensional or massive datasets.

References

[1] Bradley, P.S., Fayyad, U.M., Mangasarian, O.L. (1999) Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing,* 11, 217-238.

[2] Cardweb.com, The U.S. Payment Card Information Network, accessed April 23, 2004, [available at: http://www.cardweb.com/cardlearn/stat.html].

[3] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning, Morgan Kauffman Publication.

[4] Hsu, C. W. and Lin, C. J. (2002) A comparison of methods for multi-class support vector machines, *IEEE Transactions on Neural Networks*, 13(2), 415-425.

[5] Joachims, T. (2004) SVM-light: Support Vector Machine, available at: http://svmlight.joachims.org/.

[6] Knerr, S., Personnaz, L., and Dreyfus, G. (1990), “Single-layer learning revisited: A stepwise procedure for building and training a neural network”, in *Neurocomputing: Algorithms, Architectures and Applications*, J. Fogelman, Ed. New York: Springer-Verlag.

[7] Kou, G., Peng, Y., Shi, Y., Chen, Z. and Chen X. (2004b) “A Multiple-Criteria Quadratic Programming Approach to Network Intrusion Detection” in Y. Shi, et al (Eds.): CASDMKM 2004, LNAI 3327, Springer-Verlag Berlin Heidelberg, 145–153.

[8] Li, J.P, Liu, J.L, Xu, W.X., Shi, Y. *Support Vector Machines Approach to Credit Assessment.* In Bubak, M., Albada, et al (Eds.) , ICCS 2004, LNCS 3039, Springer-Verlag, Berlin, 892-899, 2004.

[9] LINDO Systems Inc., *An overview of LINGO 8.0*, http://www.lindo.com/cgi/frameset.cgi?leftlingo.html;lingof.html.

[10] New Generation Research, Inc., April 2004, [available at: http://www.bankruptcydata.com/default.asp].

[11] Menties, T. and Hu, Y. (2003) Data Mining for Very Busy People, IEEE Computer, p. 18-25.

[12] RuleQuest research (2004) [available at: http://www.rulequest.com/see5-info.html]. See 5.0. (2004) [available at:http://www.rulequest.com/see5-info.html].

[13] Stolfo, S.J., Fan, W., Lee, W., Prodromidis, A. and Chan, P.K. (2000) Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection: Results from the JAM Project, *DARPA Information Survivability Conference*.

[14] Vapnik, V. N. and Chervonenkis (1964), On one class of perceptrons, *Autom. And Remote Contr. 25(1).*

[15] Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory,* Springer, New York.

[16] Zhu, D., Premkumar, G., Zhang, X. and Chu, C.H. (2001) *Data Mining for Network Intrusion Detection: A comparison of Alternativest Methods*, Decision Sciences, Volume 32 No. 4, Fall 2001.