A New Multi-Criteria Quadratic-Programming Linear Classification Model for VIP E-Mail Analysis

Peng Zhang^{1,2}, Juliang Zhang¹, and Yong Shi¹

¹ CAS Research Center on Data Technology and Knowledge Economy,
Beijing 100080, China

zhangpeng04@mails.gucas.ac.cn, yshi@gucas.ac.cn

² School of Information Science and Engineering, Graduate University of
Chinese Academy of Sciences, Beijing 100080, China

Abstract. In the recent years, classification models based on mathematical programming have been widely used in business intelligence. In this paper, a new Multi-Criteria Quadratic-Programming Linear Classification (MQLC) model is proposed and tested with VIP E-Mail dataset. This experimental study uses a variance of K-fold Cross-Validation to demonstrate the accuracy and stability of the model. Finally, we compare our model with the decision tree by using commercial software C5.0. The result indicates that the proposed MQLC model performs better than decision tree on small samples.

Keywords: VIP E-Mail Analysis, Data Mining, Multi-criteria Quadratic-programming Linear Classification, Cross-Validation, Decision Tree.

1 Introduction

Data Mining is a discipline combining a wide range of subjects such as Statistics, Machine Learning, Artificial Intelligence, Neural Network, Database Technology and Pattern recognition [1]. Recently, classification models based on mathematical programming approaches have been introduced in data mining. In 2001, Shi et al [2, 3] proposed a Multiple Criteria Linear Programming (MCLP) model which has been successfully applied to a major US bank credit card portfolio management. In this paper, a new Multi-criteria Quadratic-programming Linear Classification (MQLC) model is proposed and used in the VIP E-Mail dataset provided by a major web hosting company in China. As a new model, the performance and stability of MQLC are focal issues. In order to respond to these challenges, this paper conducts a variance of k-folders cross-validation experiment and compares the prediction accuracy of MQLC with decision tree in software C5.0. Our findings indicate that MQLC is highly stable and performs well in small samples.

This paper is organized as the following. Next section is an overview of MQLC model formulation; the third section describes the characteristics of the VIP E-Mail dataset; the fourth section talks about the process and results of cross validation; the fifth section illustrates the comparison study with decision tree in commercial software C5.0; and the last section concludes the paper with summary of the findings.

2 Formulation of Multi-Criteria Quadratic-Programming Linear Classification Model

Given a set of r attributes $a = (a_1, \dots, a_r)$, let $A_i = (A_{i1}, \dots, A_{ir}) \in R^r$ be one of the sample observations of these properties. Suppose we predefine two groups G_1 and G_2 , we can select a boundary b to separate these two groups. A vector $X = (x_1, x_2, \dots, x_n) \in R^r$ can be identified to establish the following linear inequality $A_i x < b, \forall A_i \in G_1$; $A_i x \geq b, \forall A_i \in G_2$; [2, 3, 4, 5]. We define external measurement α_i to be the overlapping distance of the two-group boundary for record A_i . If $A_i \in G_1$ but we misclassified it into G_2 or vice versa, α_i will equal to $|A_i x - b|$. Then we define internal measurement β_i to be the distance of record A_i from its adjusted boundary b^* . If A_i is correctly classified, distance β_i will equal to $|A_i x - b^*|$, where $b^* = b + \alpha_i$ or $b^* = b - \alpha_i$. Suppose $f(\alpha)$ denotes the relationship of all overlapping α_i while $g(\beta)$ denotes the aggregation of all distances β_i . The final absolute catch rates depend on simultaneously minimizing $f(\alpha)$ and maximizing $g(\beta)$, the most common way of representation is by normal value. Given weights $w_\alpha, w_\beta > 0$ and let $f(\alpha) = \|\alpha + c_1\|_p^p$, $g(\beta) = \|\beta + c_2\|_q^q$, the generalized model can be converted into single criterion mathematical programming model as:

$$(\text{Normal Model}) \text{ Minimize } w_\alpha \|\alpha + c_1\|_p^p - w_\beta \|\beta + c_2\|_q^q \quad (1)$$

Subject to:

$$A_i x - \alpha_i + \beta_i - b = 0, \forall A_i \in G_1$$

$$A_i x + \alpha_i - \beta_i - b = 0, \forall A_i \in G_2$$

$$\alpha_i, \beta_i \geq 0, i = 1, \dots, n$$

Based on the Normal Model, α , β , c_1 , and c_2 can be randomly chosen. When we set $p=2$, $q=1$ and $c_1=1$, $c_2=0$, we can formulate the objective function as

$w_\alpha \sum_{i=1}^n (\alpha_i + 1)^2 - w_\beta \sum_{i=1}^n \beta_i$. After expanding it, we can write the model as follows:

$$(\text{MQLC Model}) \text{ Minimize } w_\alpha \sum_{i=1}^n \alpha_i^2 + 2w_\alpha \sum_{i=1}^n \alpha_i - w_\beta \sum_{i=1}^n \beta_i \quad (2)$$

Subject to:

$$A_i x - \alpha_i + \beta_i - b = 0, \forall A_i \in G_1$$

$$A_i x + \alpha_i - \beta_i - b = 0, \forall A_i \in G_2$$

$$\alpha_i, \beta_i \geq 0, i = 1, \dots, n$$

In MQLC model, the normal value of α is larger than β implies that the penalty for misclassifying A_i is more severe than not enlarging the distance from the adjusted

boundary. Since it is possible that α_i is less than 1, in such scenario the objective value will decrease when we square α_i . To make sure the penalty is aggravated when misclassification occurs, MQLC add 1 to all α_i before squaring them.

3 VIP E-Mail Dataset

Our partner company's VIP E-Mail data are mainly stored in two kinds of repository systems; one is databases manually recorded by employee, which was initially produced to meet the needs of every kind of business service; the other is log files recorded automatically by machine, which contains the information about customer login, E-Mail transaction and so on. After integrating all these data with the keyword SSN, we finally acquire the Large Table which has 185 features from log files and 65 features from database. Considering the customer records integrity, we eventually extracted two groups of customer record, the current customers and the lost customers, 4998 records in each group respectively. Combining these 9996 SSN with the 230 features, we eventually acquired the Large Table for data mining.

4 Empirical Study of Cross-Validation

In this paper, a variance of k-fold cross-validation is used on VIP E-Mail dataset; each calculation is consisting of training with one of the subsets and testing with the other k-1 subsets. We have computed 10 group dataset. The accuracy of 10 group training set is extremely high, with the average accuracy on the lost user of 88.48% and on the Current user of 91.4%. When we concentrate on the 10 group testing set, the worst and best classification catch rates are 78.72% and 82.17% for the lost customers and 84.59% and 89.04 % for the Current customers. That is to say, the absolute catch rates of the lost class are above 78% and the absolute catch rates of the Current class are above 84%. The result indicates that a good separation of the lost class and Current class is observed with this method.

5 Comparison of MQLC and Decision Tree

The following experiment compares MQLC and Decision Tree in the commercial software C5.0. From the results of Table II, we can see that when the percentage of training set is increased from 2% to 25%, the accuracy of Decision Tree increases from 71.9% to 83.2% while the accuracy of MQLC increases from 78.54% to 83.56%. The accuracy of MQLC is slightly better than Decision Tree. Moreover, when training set is 10% of the whole dataset, the accuracy of MQLC peaked at 83.75%. That is to say, MQLC can perform well even on small samples. The reason maybe due to the fact that MQLC solves a convex quadratic problem, which can acquire the global optimal solution easily, on the other hand, Decision Tree merely selects the better tree from a limited built tree, which might not be the best tree. In addition, the pruning procedure of Decision Tree may further eliminate some better branches. In conclusion, MQLC performs better than Decision Tree on small samples.

Table 1. Comparison of MQLC and Decision Tree

Percentage of training	Decision Tree		MQLC	
	Training	Testing	Training	Testing
2%	92.7%	71.9%	96.0%	78.54%
5%	94.9%	77.3%	92.8%	82.95%
10%	95.2%	80.8%	90.7%	83.75%
25%	95.9%	83.2%	88.95%	83.56%

6 Conclusion

In this paper, a new Multi-criteria Quadratic-programming Linear Classification (MQLC) Model has been proposed to the classification problems in data mining. 230 features are extracted from the original data source to depict all the VIP E-Mail users, and 9996 records are chosen to test the performance of MQLC. Through the results of cross-validation, we can see that the model is extremely stable for multiple groups of randomly generated training set and testing set. The comparison of MQLC and Decision Tree in C5.0 tells us that MQLC performs better than Decision Tree on small samples. There have been experiments that show the accuracy of MQLC is not affected by the parameters of the objective function, further research will include mathematically proving this phenomenon.

Acknowledgments. This research has been partially supported by a grant from National Natural Science Foundation of China (#70621001, #70531040, #70501030, #70472074), 973 Project #2004CB720103, Ministry of Science and Technology, China, NSFB(No.9073020) and BHP Billiton Co., Australia.

References

1. Han, J. W., M. Kamber: Data Mining: Concepts and Techniques. San Diego. CA: Academic Press (2000) ch.1.

2. Shi, Y., Wise, M., Luo, M. and Lin, Y.: Data mining in credit card portfolio management: a multiple criteria decision making approach. in Multiple Criteria Decision Making in the New Millennium, M. Koksalan and S.Zionts, eds., Berlin: Springer (2001) 427-436

3. Shi, Y., Peng, Y., Xu, W., Tang, X.: Data Mining via Multiple Criteria Linear Programming: Applications in Credit Card Portfolio Management. International Journal of Information Technology and Decision Making, vol. 1(2002) 131-151.

4. Gang Kou, Yi Peng, Yong Shi, Weixuan Xu: A Set of Data Mining Models to Classify Credit Cardholder Behavior. International Conference on Computational Science (2003) 54-63

5. Yi Peng, Gang kou, Zhengxin Chen, Yong Shi: Cross-Validation and Ensemble Analyses on Multiple-Criteria Linear Programming Classification for Credit Cardholder Behavior. International Conference on Computational Science (2004) 931-939