

A new multi-criteria convex quadratic programming model for credit analysis*

Gang Kou¹, Yi Peng^{1,3}, Yong Shi^{1,2}, and Zhengxin Chen¹

¹College of Information Science & Technology, University of Nebraska at Omaha,
Omaha, NE 68182, USA

²Chinese Academy of Sciences Research Center on Data Technology & Knowledge
Economy, Graduate University of the Chinese Academy of Sciences, Beijing 100080, China

³Corresponding author

{gkou, ypeng, yshi, zchen}@mail.unomaha.edu

Abstract. Mathematical programming based methods have been applied to credit risk analysis and have proven to be powerful tools. One challenging issue in mathematical programming is the computation complexity in finding optimal solutions. To overcome this difficulty, this paper proposes a Multi-criteria Convex Quadratic Programming model (MCCQP). Instead of looking for the global optimal solution, the proposed model only needs to solve a set of linear equations. We test the model using three credit risk analysis datasets and compare MCCQP results with four well-known classification methods: LDA, Decision Tree, SVMLight, and LibSVM. The experimental results indicate that the proposed MCCQP model achieves as good as or even better classification accuracies than other methods.

Keywords: Credit Analysis, Classification, Mathematical Programming, Multi-criteria Decision Making.

INTRODUCTION

This paper explores solving classification problem, one of the major sub-fields of data mining, through the use of mathematical programming based methods (Bradley et al 1999, Vapnik 1964 and 2000). Such methods have proven to be powerful in solving a variety of machine learning problems (Chang and Lin 2001, Fung 2003, Joachims 1999, Kou et al 2005, Mangasarian 1999, 2000 and 2005, Mitchell 1997, Zheng et al 2004). However, it is difficult to find the optimal solution of a mathematical programming problem. To overcome this difficulty, a new Multi-criteria Convex Quadratic Programming model (MCCQP) is proposed. In the proposed model, we only need to solve a set of linear equations in order to find the global optimal solution.

This paper is organized as follows: section 2 presents the MCCQP model, section 3 illustrates the numerical implementation and comparison study with several well-

* 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.

established data mining software and reports the results, and section 4 summarizes the paper and discusses future research directions.

MULTI-CRITERIA Convex Quadratic PROGRAMMING MODEL

This section introduces a new MCQP model. This model classifies observations into distinct groups via a hyperplane and based on multiple criteria. The following models represent this concept mathematically (Kou et al 2006):

Each row of a $n \times r$ matrix $A = (A_1, \dots, A_n)^T$ is an r -dimensional attribute vector $A_i = (A_{i1}, \dots, A_{ir}) \in \mathfrak{R}^r$ which corresponds to one of the records in the training dataset of a binary classification problem, $i = 1, \dots, n$; n is the total number of records in the dataset. Two groups, G_1 and G_2 , are predefined while $G_1 \cap G_2 = \Phi$ and $A_i \in \{G_1 \cup G_2\}$. A boundary scalar b can be selected to separate G_1 and G_2 . Let $X = (x_1, \dots, x_r)^T \in \mathfrak{R}^r$ be a vector of real number to be determined. Thus, we can establish the following linear inequations (Fisher 1936, Shi et al. 2001):

$$A_i X < b, \quad \forall A_i \in G_1; \quad (1)$$

$$A_i X \geq b, \quad \forall A_i \in G_2; \quad (2)$$

In the classification problem, $A_i X$ is the score for the i^{th} data record. If all records are linear separable and an element A_i is correctly classified, then let β_i be the distance from A_i to b , and obviously in linear system, $A_i X = b - \beta_i$, $\forall A_i \in G_1$ and $A_i X = b + \beta_i$, $\forall A_i \in G_2$. However, if we consider the case where the two groups are not completely linear separable, there exist some misclassified records. When an element A_i is misclassified, let α_i be the distance from A_i to b , $A_i X = b + \alpha_i$, $\forall A_i \in G_1$ and $A_i X = b - \alpha_i$, $\forall A_i \in G_2$. To complete the definitions of β_i and α_i , let $\beta_i = 0$ for all misclassified elements and α_i equals to zero for all correctly classified elements. Incorporating the definitions of β_i and α_i , (1) and (2) can be reformulated as the following model:

$$A_i X = b - \delta + \alpha_i - \beta_i, \quad \forall A_i \in G_1$$

$$A_i X = b + \delta - \alpha_i + \beta_i, \quad \forall A_i \in G_2$$

δ is a given scalar. $b - \delta$ and $b + \delta$ are two adjusted hyper planes for the model.

Redefine X as X/δ , b as b/δ , α_i as α_i/δ , β_i as β_i/δ , and Define a $n \times n$ diagonal matrix Y which only contains “+1” or “-1” indicates the class membership. A “-1” in row i of matrix Y indicates the corresponding record A_i

$\in G_1$ and a “+1” in row i of matrix Y indicates the corresponding record $A_i \in G_2$. The model can be rewritten as:

$$Y(\langle A \cdot X \rangle - eb) = 1 + \alpha - \beta, \quad (4)$$

where $e=(1,1,\dots,1)^T$, $\alpha = (\alpha_1, \dots, \alpha_n)^T$ and $\beta = (\beta_1, \dots, \beta_n)^T$.

The proposed multi-criteria optimization problem contains three objective functions. The first mathematical function $f(\alpha) = \|\alpha\|_p^p = \sum_{i=1}^n (\alpha_i)^p$ (ℓ_p — norm of β_i , $1 \leq q \leq \infty$) describes the summation of total overlapping distance of misclassified records to b . The second function $g(\beta) = \|\beta\|_q^q = \sum_{i=1}^n (\beta_i)^q$ (ℓ_q — norm of β_i , $1 \leq q \leq \infty$) represents the aggregation of total distance of correctly separated records to b . In order to maximize the distance $(\frac{2}{\|X\|_s^s})$ between the two

adjusted bounding hyper planes, the third function $h(X) = \frac{\|X\|_s^s}{2}$ should be minimized. Furthermore, to transform the generalized Multi-Criteria classification model into a single-criterion problem, weights $W_\alpha > 0$ and $W_\beta > 0$ are introduced for $f(\alpha)$ and $g(\beta)$, respectively. A single-criterion mathematical programming model can be set up:

$$\text{(Model 1) Minimize } \frac{1}{2} \|X\|_s^s + W_\alpha \|\alpha\|_p^p - W_\beta \|\beta\|_q^q$$

$$\text{Subject to: } Y(\langle A \cdot X \rangle - eb) = e - \alpha + \beta$$

$$\alpha_i, \beta_i \geq 0, 1 \leq i \leq n$$

where Y is a given $n \times n$ diagonal matrix, $e=(1,1,\dots,1)^T$, $\alpha = (\alpha_1, \dots, \alpha_n)^T$, $\beta = (\beta_1, \dots, \beta_n)^T$, X and b are unrestricted.

Please note that the introduction of β_i is one of the major differences between the proposed model and other existing Support Vectors approaches (Vapnik 1964 and 2000). It is much easier to find optimal solutions for convex quadratic programming form than other forms of nonlinear programming. To make Model 1 a convex quadratic programming form, let $s = 2$, $q = 1$ and $p = 2$. The constraints remain the same and the objective function becomes:

$$\text{(Model 2) Minimize } \frac{1}{2} \|X\|_2^2 + W_\alpha \sum_{i=1}^n \alpha_i^2 - W_\beta \sum_{i=1}^n \beta_i$$

$$\text{Subject to: } Y(\langle A \cdot X \rangle - eb) = e - \alpha + \beta$$

Let $\eta_i = \alpha_i - \beta_i$. According to the definition, $\eta_i = \alpha_i$ for all misclassified records and $\eta_i = -\beta_i$ for all correctly separated records. The definition of η_i is one of the major differences between the proposed model and other existing approaches (Fung 2003, Gonzalez-Castano and Meyer 2000).

Add $\frac{W_b}{2}b^2$ to Model 2's objective function and the weight W_b is an arbitrary positive number.

$$\begin{aligned} \text{(Model 3)} \quad & \text{Minimize } \frac{1}{2} \|X\|_2^2 + \frac{W_\alpha}{2} \sum_{i=1}^n \eta_i^2 + W_\beta \sum_{i=1}^n \eta_i + \frac{W_b}{2} b^2 \\ & \text{Subject to: } Y \langle A \cdot X \rangle - eb = e - \eta \end{aligned}$$

where Y is a given $n \times n$ diagonal matrix, $e=(1,1,\dots,1)^T$, $\eta = (\eta_1, \dots, \eta_n)^T$, η , X and b are unrestricted, $1 \leq i \leq n$.

The Lagrange function corresponding to Model 5 is

$$L(X, b, \eta, \theta) = \frac{1}{2} \|X\|_2^2 + \frac{W_\alpha}{2} \sum_{i=1}^n \eta_i^2 + W_\beta \sum_{i=1}^n \eta_i + \frac{W_b}{2} b^2 - \theta^T (Y \langle A \cdot X \rangle - eb) - e^T \eta$$

where $\theta = (\theta_1, \dots, \theta_n)^T$, $\eta = (\eta_1, \dots, \eta_n)^T$, $\theta_i, \eta_i \in \mathcal{R}$.

According to Wolfe Dual Theorem, $\nabla_X L(X, b, \eta, \theta) = X - A^T Y \theta = 0$,
 $\nabla_b L(X, b, \eta, \theta) = W_b b + e^T Y \theta = 0$, $\nabla_\eta L(X, b, \eta, \theta) = W_\alpha \eta + W_\beta e - \theta = 0$.

Introduce the above 3 equations to the constraints of Model 5, we can get:

$$\begin{aligned} Y((A \cdot A^T)Y\theta + \frac{1}{W_b} e(e^T Y \theta)) + \frac{1}{W_\alpha} (\theta - W_\beta e) &= e \\ \Rightarrow \theta &= \frac{(1 + \frac{W_\beta}{W_\alpha})e}{\frac{I}{W_\alpha} + Y((A \cdot A^T) + \frac{1}{W_b} ee^T)Y} \quad (5) \end{aligned}$$

Proposition 1: For some $W_\alpha > 0$, $\theta = \frac{(1 + \frac{W_\beta}{W_\alpha})e}{\frac{I}{W_\alpha} + Y((A \cdot A^T) + \frac{1}{W_b} ee^T)Y}$ exists.

Proof: Let $H = Y[A - (\frac{1}{W_b})^{\frac{1}{2}} e]$, we get:

$$\theta = (1 + \frac{W_\beta}{W_\alpha}) (\frac{I}{W_\alpha} + HH^T)^{-1} e \quad (5)$$

$\forall H, \exists W_\alpha > 0$, when W_α is small enough, the inversion of $(\frac{I}{W_\alpha} + HH^T)$ exists.

$$\text{So } \theta = \frac{(1 + \frac{W_\beta}{W_\alpha})e}{\frac{I}{W_\alpha} + Y((A \cdot A^T) + \frac{1}{W_b} ee^T)Y} \text{ exists. } \square$$

Algorithm 1

Input: a $n \times r$ matrix A as the training dataset, a $n \times n$ diagonal matrix Y labels the class of each record.

Output: classification accuracies for each group in the training dataset, score for every record, decision function $\text{sgn}((X^* \cdot A_i) - b^*) \begin{cases} > 0, \Rightarrow A_i \in G_1 \\ \leq 0, \Rightarrow A_i \in G_2 \end{cases}$

Step 1 compute $\theta^* = (\theta_1, \dots, \theta_n)^T$ by one of (5) or (6). W_β, W_α and W_b are chosen by standard 10-fold cross-validation.

Step 2 compute $X^* = A^T Y \theta^*$, $b^* = \frac{-1}{W_b} e^T Y \theta^*$.

Step 3 classify a incoming A_i by using decision function $\text{sgn}((X^* \cdot A_i) - b^*) \begin{cases} > 0, \Rightarrow A_i \in G_1 \\ \leq 0, \Rightarrow A_i \in G_2 \end{cases}$.

END

Numerical experiments in credit risk analysis

The model proposed can be used in many fields, such as general bioinformatics, antibody and antigen, credit fraud detection, network security, text mining, etc. We conducted three numerical experiments to evaluate the proposed MCCQP model. All experiments are concerned about credit risk analysis. Each record in these three sets has a class label to indicate its' financial status: either Normal or Bad. Bad indicates a bankrupt credit or firm account and Normal indicates a current status account. The result of MCCQP is compared with the results of 4 widely accepted classification methods: Linear Discriminant Analysis (LDA) (SPSS 2004), Decision Tree based See5 (Quinlan 2003), SVM light (Joachims 1999) and LibSVM (Chang and Lin 2001).

The first benchmark set is a German credit card application dataset from UCI Machine Learning databases (UCI 2005). The German set contains 1000 records (700

Normal and 300 Bad) and 24 variables. The second set is an Australian credit approval dataset from See5 (Quinlan 2003). The Australian set has 383 negative cases (Normal) and 307 positive cases (Bad) with 15 attributes. The last set is a Japanese firm bankruptcy set (Kwak et al 2005). The Japanese set includes Japanese bankrupt (Bad) sample firms (37) and non-bankrupt (Normal) sample firms (111) between 1989 and 1999. Each record has 13 variables.

Credit Classification Process

Input: The Credit Card dataset $A = \{ A_1, A_2, A_3, \dots, A_n \}$, a $n \times n$ diagonal matrix Y

Output: Average classification accuracies for Bad and Normal of the test set in 10-fold cross-validation; scores for all records; decision function.

Step 1 Apply several classification methods: LDA, Decision Tree, SVM, MCCQP, to A using 10-fold cross-validation. The outputs are a set of decision functions, one for each classification method.

Step 2 Compute the classification accuracies using the decision functions.

END

The following tables (Table 1, 2, and 3) summarize the averages of 10-fold cross-validation test-sets accuracies of Linear Discriminant Analysis (LDA) (SPSS 2004), Decision Tree base See5 (Quinlan 2003), SVM light (Joachims 1999), LibSVM (Chang and Lin 2001), and MCCQP for each dataset. ‘‘Type I Error’’ is defined as the percentage of predicted Normal records that are actually Bad records and ‘‘Type II Error’’ is defined as the percentage of predicted Bad records that are actually Normal records. Since Type I error indicates the potential charge-off lost of credit issuers, it is considered more costly than Type II error. In addition, a popular measurement, KS score, in credit risk analysis is calculated. The higher the KS score, the better the classification methods. The KS (Kolmogorov-Smirnov) value is defined as:

KS value = $Max |Cumulative\ distribution\ of\ Bad - Cumulative\ distribution\ of\ Normal|$.

Table 1 reports the results of the five classification methods for German set. Among the five methods, LibSVM achieves the best results for all the measurements and MCCQP achieves the second best results.

Table 2 summarizes the results for the Australian set. Among the five methods, MCCQP achieves the best overall accuracy, Normal accuracy and Type II error while LDA achieves the lowest Type I error rate and highest Bad classification accuracy and KS-score.

Table 1 10-fold cross-validation result of German set

Method	Classification Accuracy			Error Rate		KS-Score
	Overall	Normal	Bad	Type I	Type II	
LDA	72.20%	72.57%	71.33%	28.32%	27.77%	43.90
See5	72.20%	84.00%	44.67%	39.71%	26.37%	28.67
SVMlight	66.50%	77.00%	42.00%	42.96%	35.38%	19.00
LibSVM	94.00%	100.00%	80.00%	16.67%	0.00%	80.00
MCCQP	73.50%	74.38%	72.00%	27.35%	26.24%	46.38

Table 2 10-fold cross-validation result of Australian set

Method	Classification Accuracy			Error Rate		KS-Score
	Overall	Normal	Bad	Type I	Type II	
LDA	85.80%	80.68%	92.18%	8.83%	17.33%	72.86
See5	86.52%	87.99%	84.69%	14.82%	12.42%	72.68
SVMLight	44.83%	18.03%	90.65%	34.14%	47.48%	8.69
LibSVM	44.83%	86.89%	27.10%	45.62%	32.61%	13.99
MCCQP	86.38%	87.00%	85.52%	14.27%	13.20%	72.52

Table 3 summarizes the result for Japanese set. Among the five methods, MCCQP achieves the highest classification accuracies for overall, Normal, and Bad. In addition, MCCQP has the highest KS-score and lowest Type I and II error rates. Although See5 got the highest classification accuracy for Normal class, its classification accuracy for Bad is only 35.14%.

Table 3 10-fold cross-validation result of Japanese set

Method	Classification Accuracy			Error Rate		KS-Score
	Overall	Normal	Bad	Type I	Type II	
LDA	68.92%	68.47%	70.27%	30.28%	30.97%	38.74
See5	72.30%	84.68%	35.14%	43.37%	30.36%	19.82
SVMLight	48.15%	47.25%	52.94%	49.90%	49.91%	0.19
LibSVM	50.46%	49.45%	55.88%	47.15%	47.49%	5.33
MCCQP	72.30%	72.30%	72.47%	27.58%	27.65%	44.77

Conclusion

In this paper, a new MCCQP model for classification problem has been presented. In order to validate the model, we apply the model to three credit risk analysis datasets and compare MCCQP results with four well-known classification methods: LDA, Decision Tree, SVMLight, and LibSVM. The experimental results indicate that the proposed MCCQP model achieves as good as or even better classification accuracies than other methods.

There are still many aspects that need further investigation in this research. Theoretically, MCQP is highly efficient method in both computation time and space on large-scale problems. Since all 4 datasets used are relatively small, it will be a nature extension to apply MCQP in massive dataset. $(A_i \cdot A_j)$ in Model 3 and 4 is inner product in the vector space and it can be substituted by a kernel $K(A_i, A_j)$, which will extend the applicability of the proposed model to linear inseparable datasets. Future studies may be done on establishing a theoretical guideline for selection of kernel that is optimal in achieving a satisfactory credit analysis result.

References

- Bradley, P.S., Fayyad, U.M., Mangasarian, O.L., Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11, 217-238, 1999.
- Chang, Chih-Chung and Lin, Chih-Jen, LIBSVM : a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Fung , G. “Machine learning and data mining via mathematical programming-based support vector machines”, Ph.D thesis, The University of Wisconsin - Madison. 2003
- Fung, G. and Mangasarian, O. L. Multicategory Proximal Support Vector Machine Classifiers, *Machine Learning* 59, 2005, 77-97.
- Gonzalez-Castano, F. and Meyer, R. Projection support vector machines. Technical Report 00-05, Computer Sciences Department, The University of Wisconsin – Madison, 2000
- 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.), LNCS 3039, Springer-Verlag, Berlin, 892-899, 2004.
- LINDO Systems Inc., *An overview of LINGO 8.0*, <http://www.lindo.com/cgi/frameset.cgi?leftlingo.html;lingof.html>.
- Joachims, T. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- Joachims, T. (2004) SVM-light: Support Vector Machine, available at: <http://svmlight.joachims.org/>.
- Kou, G., X. Liu, Y. Peng, Y. Shi, M. Wise and W. Xu, “Multiple Criteria Linear Programming to Data Mining: Models, Algorithm Designs and Software Developments” *Optimization Methods and Software* 18 (4): 453-473, Part 2 AUG 2003
- Kou, G., Y. Peng, Y. Shi, M. Wise and W. Xu, “Discovering Credit Cardholders’ Behavior by Multiple Criteria Linear Programming” *Annals of Operations Research* 135 (1): 261-274, JAN 2005
- MATLAB. User’s Guide. The MathWorks, Inc., Natick, MA 01760, 1994-2005.
- Mitchell, T. M. *Machine Learning*. McGraw-Hill, Boston, 1997.
- Murphy, P. M. and Aha, D. W. UCI repository of machine learning databases, 1992. www.ics.uci.edu/~mlearn/MLRepository.html.
- Mangasarian, O. L. and Musicant, D. R. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10:1032-1037, 1999.
- Mangasarian, O. L. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135-146, Cambridge, MA, 2000. MIT Press.
- Mangasarian, O. L. Support Vector Machine Classification via Parameterless Robust Linear Programming, *Optimization Methods and Software* 20, 2005, 115-125.
- Quinlan, J. See5.0. (2004) [available at: <http://www.rulequest.com/see5-info.html>].
- Shi, Y., Peng, Y., Kou, G. and Chen, Z “Classifying Credit Card Accounts for Business Intelligence and Decision Making: A Multiple-Criteria Quadratic Programming Approach” *International Journal of Information Technology and Decision Making*, Vol. 4, No. 4 (2005) 1-19.
- Vapnik, V. N. and Chervonenkis (1964), On one class of perceptrons, *Autom. And Remote Contr.* 25(1).
- Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 2000.
- Zheng, J., Zhuang, W., Yan, N., Kou, G., Peng, H., McNally, C., Erichsen, D., Cheloha, A., Herek, S., Shi, C. and Shi, Y., “Classification of HIV-1 Mediated Neuronal Dendritic and Synaptic Damage Using Multiple Criteria Linear Programming” *Neuroinformatics* 2 (3): 303-326 Fall 2004.