Select Representative Samples for Regularized Multiple-Criteria Linear Programming Classification

Peng Zhang[1], Yingjie Tian[1], Xingsen Li[1], Zhiwang Zhang[1], and Yong Shi[1,2,*]

[1] Research Center on Fictitious Economy & Data Science,
Chinese Academy of Sciences, Beijing, China, 100080
[2] College of Information Science & Technology, University of Nebraska at Omaha,
Omaha, NE 68182, USA
zhangpeng04@gmail.com, {tyj,xsli,zzw,yshi}@gucas.ac.cn

Abstract. Regularized multiple-criteria linear programming (RMCLP) model is a new powerful method for classification in data mining. Taking account of every training instance, RMCLP is sensitive to the outliers. In this paper, we propose a sample selection method to seek the representative points for RMCLP model, just as finding the support vectors to support vector machine (SVM). This sample selection method also can exclude the outliers in training set and reduce the quantity of training samples, which can significantly save costs in business world because labeling training samples is usually expensive and sometimes impossible. Experimental results show our method not only reduces the quality of training instances, but also improves the performance of RMCLP.

Keywords: RMCLP, clustering, outlier detection, sample reduction.

1 Introduction

Recently, multiple-criteria mathematical programming models have been widely used for classification in data mining and business intelligence [1]. The original idea of multiple-criteria linear classification is simultaneously maximizing the distances between classes and minimizing the overlapping of each group [2]. Assume there are two groups G_1 and G_2. Multiple-criteria mathematical model try to find a projection direction x and a boundary b, making all the training records a_i satisfy the constraints $a_i x<b$ for $a_i \in G_1$ and $a_i x>b$ for $a_i \in G_2$. This formulation of calculating all the training samples performs pretty well when there are not many outliers, but the disadvantage is also obvious, if there are outliers, it needs much more samples to keep stability. The biggest difference in RMCLP and the well-known support vector machine (SVM) is that RMCLP takes account of all the training instances when draw the classification boundary, and SVM draws the classification boundary only according to a few support vectors. If we can find the representative instances of multiple-criteria models, just as the support vectors of SVM, RMCLP can robustly avoid the effect of noise and reduce the training samples dramatically. This will benefit business application significantly, because in business

[*] Corresponding author.

M. Bubak et al. (Eds.): ICCS 2008, Part II, LNCS 5102, pp. 436–440, 2008.

world, acquiring the training samples is always costly and sometimes impossible, reducing training samples can save much money to do data mining, that is to say, if we find the representative instances, multi-criteria models will more powerful in business data mining.

Clustering is an effective method to remove outliers. With the clustering centers, it also can reduce the training samples to classification. In this paper, we first introduce the RMCLP model, and then we give the algorithm of how to select the representative instances for RMCLP. The following experiment on synthetic dataset demonstrates that finding the representative instances is capable to improve the performance of RMCLP. In the fourth section, we show the empirical study on a real-life credit card dataset. Finally, we conclude this paper.

2 Two Groups RMCLP Model

In 2007, Shi et.al proposed a new regularized multiple-criteria linear programming (RMCLP) model for classification. Now we will give a brief introduction here without demonstration of its mathematical solution, more theoretical explanation of this model is in [3]. Given a set of r-dimensional variables (attribute) vector $a = (a_1,...,a_r)$, let $A_i = (A_{i1},...,A_{ir}) \in R^r$ be one of the sample records of these attributes, where $i = 1,..., n$; n represents the total number of records in the dataset. Suppose two groups, G_1 and G_2, are predefined. A boundary scalar b can be set to separate these two groups. We then give the RMCLP model as follows:

$$
\begin{aligned}
&\min_{z} \ \frac{1}{2}x^T Hx + \frac{1}{2}\alpha^T Q\alpha + d^T\alpha - c^T\beta \\
&\text{s.t.} \\
&\qquad A_i x - \alpha_i + \beta_i = b, \forall A_i \in G_1; \\
&\qquad A_i x + \alpha_i - \beta_i = b, \forall A_i \in G_2; \\
&\qquad \alpha_i, \beta_i \geq 0.
\end{aligned} \tag{1}
$$

In this model, α and β are n-dimension vectors. α denotes the misclassification distance while β denotes the correct classification distance. As far as the constraints be considered, it means when we misclassify $A_i \in G_1$ to G_2 or vice versa, there is a distance α_1 and the value equals $| A_i x - b |$; when we correctly classify A_i, there is a distance β_i and the value equals $| A_i x - b^* |$, where $b^* = b + \alpha_i$ or $b^* = b - \alpha_i$. If we let the objective function be the linear combination of α and β, that is to say, minimize $\sum \alpha_i$ and $-\sum \beta_i$ simultaneously, we can get the original MCLP model. To assure the MCLP model always has a solution, we add the $\frac{1}{2}x^T Hx$ and $\frac{1}{2}\alpha^T Q\alpha$ to the objective function to formulate the two groups RMCLP model. $H \in R^{r*r}$, $Q \in R^{n*n}$ are symmetric positive definite matrices. $d^T, c^T \in R^n$.The RMCLP model is a convex quadratic program.

3 Algorithm for Sample Selection

Data preprocess is an important step for data mining. Without pure and right data source, data mining models will lose its power in discovering useful knowledge. Traditionally, bagging and boosting methods are popular in sample selection. In this section, a clustering based sample selection algorithm (Algorithm 1) is introduced as follows:

```
Input: training samples Tr, testing samples Ts, parameter $\varepsilon$ , exclusion percentage s
Output: selected samples Tr'
Begin
 1. Set Tr'=Tr
  2. While ( |PrevClusteringCenter-CurrClusteringCenter| $<\varepsilon$ ) {
    2.1. Calculate current clustering center; $cent = \frac{1}{|Tr'|}\sum_i x_i$
    2.2. For each instances $i \in Tr$ do {
        2.2.1 Calculate the Euclidean distance of the clustering center,
              $dis_i = \sqrt{\sum_{r\in R} | cent_r - Tr_r^i |^2}$
        2.2.2 get the s% of the instances which are farthest to the center, denoted
              as the set {P}
       2.2.2 exclude the noisy instances, Tr'=Tr\{P}.
    }
  }
  3. Return the selected samples Tr'
 End
```

Algorithm 1. Clustering method to get the representative samples

We now create a synthetic dataset and show how the effectiveness of our method on RMCLP. Assume there is tow groups classification problem, all of the samples follows the distribution $x \sim N(\mu, \Sigma)$, more specifically, $G_1 \sim N(1,\Sigma)$, $G_2 \sim N(5,\Sigma)$. There are 6 instances in G_1 and 6 instances in G_2, additionally, a noisy instance -100 is added in G_1, and a noisy point -200 is add in G_2. We will see how these two noisy instances affect the boundary of RMCLP:

G_1= {1.11, 1.20, 1.19, 0.82, 0.90, -50}
G_2 = {5.20, 5.22, 4.76, 4.99, -100}

Case 1: We directly build RMCLP model on this dataset, the optimal parameters of RMCLP are H=10^5, Q=1, d=10^5, c=100. The projection director x=-0.798034, and boundary b=7.33516. The objective value of each element in G_1 is {-0.886, -0.958, -0.950, -0.654, -0.718, 39.902}, in G_2 is {-4.150, -4.166, -3.799, -3.982, -4.086, 79.803}. We can figure out that the accuracy is 83% on G_1 and 16% on G_2.

Case 2: Now we call algorithm 1 first to take out the noisy -50 from G_1 and -100 from G_2, and then we build RMCLP model on the purified dataset and get the optimal parameters are H=10^5, Q=1, d=10^5 , c=10^4. The x=1.98336 and b=5.50112. The objective value of G_1 is {2.202, 2.380, 2.360, 1.626, 1.785} and of G_1 is {10.313, 10.353, 9.441, 9.897, 10.155}. So the accuracies of two groups are both 100%. This means we've gotten a fine representative training sample.

Besides this synthetic dataset, we will test the effectiveness of our clustering based sample selection method on a real life dataset in next section.

4 Empirical Study on Real Life Dataset

In this section, we will do experiments on a real life credit card dataset, more detailed introduction of this dataset is in [4]. According to the previous research work on this dataset, we first select a benchmark training size of randomly selected 700 bankruptcy records and 700 normal records, and the remained 4600 records are used to testing the performance. Now what we need to do is to examine three assumptions: First, is the randomly selected 1400 points are suitable for build model? Second, are there any noise points in this randomly selected dataset? Third, can we reduce the 1400 points in a much smaller size and improve the accuracy synchronously? Experimental results in Table 1 tell us the results. The first column is the current training samples' size, from the 1400 instances to 140 instances, column two and three is the performance on training samples and the column four and five is the performance on 4600 testing records. The experiment is conducted as follows: first, we build RMCLP model on all the 1400 samples, the result is 73,78% and we take it as benchmark accuracy. Then

Table 1. Comparison of different percentage of training samples

Percentage (number) of training samples	Training Samples		Testing Samples (4600 instances)	
	Right instances	Accuracy	Right instances	Accuracy
100%(1400)	1096	78.29%	3394	73.78%
90%(1260)	998	79.20%	3295	71.63%
80%(1120)	912	81.43%	3292	71.57%
70%(980)	789	80.51%	3571	77.63%
60% (840)	667	79.40%	3761	81.76%
50% (700)	559	79.86%	3881	84.37%
40% (560)	449	80.18%	3964	86.17%
30% (420)	331	78.81%	4050	88.04%
20% (280)	232	82.86%	4073	88.54%
10% (140)	116	82.86%	1971	42.85%

we call algorithm 1 with s=1, the clustering method will drop one points after one iteration. We remark 9 special sets, 10%, 20%, …, to 90% of the original 1400 samples respectively, and build RMCLP model on these selected samples. We finally list the performance of RMCLP with each different sample. Intuitionally, we though the larger the training samples, the more information we could get, and more accurate when prediction. However, from the result, we can see that, the 1400 randomly selected samples is not the best training set for RMCLP model, there exists noisy and useless points. The clustering method reduces the training samples continuously. When we get 20%, that is 280 samples in the benchmark 1400 samples, we can build RMCLP with 88.54% on testing set, which is the best.

5 Conclusions

As a new promising data mining tool, RMCLP has been extensively used in data mining and business intelligence. To label lots of training samples before building RMCLP model is expensive and sometime impossible. Even if we could label all of them, there would be lots of noisy that impact the performance of RMCLP. So we should find representative instances of RMCLP, just as find the support vectors of SVM. In this paper, we have proposed a clustering based method to wipe off the noisy data and further more, to reduce training samples. From empirical study, we can see that our method is eligible to improve the performance of RMCLP on the credit card dataset. Whether this clustering method gets the optimal representative samples of RMCLP has not been discussed here and will be remained as our further research topic.

Acknowledgments

This research has been partially supported by a grant from National Natural Science Foundation of China (#70621001, #70531040, #70501030, #10601064, #70472074), National Natural Science Foundation of Beijing #9073020, 973 Project #2004CB720103, Ministry of Science and Technology, China and BHP Billiton Co., Australia.

References

1. Olson, D., Shi, Y.: Introduction to Business Data Mining. McGraw-Hill, New York (2007)
2. Shi, Y., Peng, Y., Kou, G., 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 4, 581–600 (2005)
3. Shi, Y., Chen, X., Tian, Y., Zhang, P.: A Regularized Multiple Criteria Linear Program for Classification. In: IEEE ICDM 2007 (2007).
4. Peng, Y., Kou, G., Chen, Z., Shi, Y.: Cross-Validation and Ensemble Analyses on Multiple-Criteria Linear Programming Classification for Credit Cardholder Behavior. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3039, pp. 931–939. Springer, Heidelberg (2004)