Kernel-based multiple criteria linear programming classifier

Zhan Zhang\textsuperscript{a}, Dongling Zhang\textsuperscript{b}, Yingjie Tian\textsuperscript{c},\textsuperscript{*}

\textsuperscript{a}China citic bank, Beijing 100027, China
\textsuperscript{b}Beijing municipal science and technology commission, Beijing 100035, China
\textsuperscript{c}Research center on fictitious economy and data sciences, Chinese academy of sciences, Beijing 100190, China

Abstract

This paper proposed a novel classification model which introduced the kernel function into the original Multiple Criteria Linear Programming (MCLP) model. MCLP model is used as a classification method which can only solve linear separable problems in data mining. However, the proposed kernel-based MCLP model can deal with non-linear cases. Meanwhile, unlike some other complicated models, this model is effective and easy to understand. A couple of experimental tests were conducted to evaluate the performance of the proposed kernel-based MCLP model compared with the existing methods: original MCLP and SVM. The results show that kernel-based MCLP model is a competitive method in dealing with nonlinear separable classification problems.

1. Introduction

Classification is one of the most important data mining tasks, which is to find a set of models that describe and distinguish data classes. Classification algorithms use the sample data from a given classified data set to learn functions or models that map each item of the selected data into one of the predefined classes. The decision function or model is then used to predict the class label of other unclassified data. This classification procedure usually includes three steps: the first step is to build a classification model with one certain classification method based on training data with predetermined data classes, the second step is to test the model on a given test set which can validate the effectiveness of the model, and the third is to use the model on new data for future classification tasks. This process is clearly described in Fig. 1.

Classification can be applied in many areas, such as customer classification for targeted marketing, credit evaluation for safety loan, investment risk analysis and medical clinic decision, etc. As an important problem in research and practice, classification has attracted interests from many fields. As a result, various of methods have been presented to solve this problem. Traditional statistical methods, including linear discriminate function, the quadratic discriminate function and the logistic discriminate function, have been generally accepted by industries.
Recently in data mining and machine learning field, decision trees, fuzzy logic, neural networks, Bayesian and genetic algorithm are the widely used methods for classification.

In recent years, optimization based techniques have been widely used in many classification problems. In [1] and [2] multi-objective linear programming model and quadratic programming model are respectively used for credit card holders’ behavior analysis. [3] utilizes multiple criteria and multiple constraint levels linear programming (MC\(^2\)LP) model with changeable parameters in order to find dynamic best solutions for capital investment to make "taking loss at the ordering time and making profit at the time of delivery" feasible. Support vector machine, which is also based on optimization method, is often used in face detection [4, 5] and text mining application [6] in recent researches. Of all these optimization-based data mining methods, support vector machine (SVM) and multiple criteria linear programming (MCLP), have attracted data mining researchers’ interests recently. The SVM algorithm, which is based on mathematical programming, is originally established by Vapnik [7]. He used a technique known as the "kernel trick" to apply linear classification techniques to non-linear classification problems. By now, it has become a popular classification tool for high-dimensional data. MCLP is also an optimization based classification method [8]. The main feature of this method is that it has multiple objectives in one model. Two commonly used criteria in linear programming for classification are the minimizing of overlapping degree with respect to the discriminated boundary and the maximizing of the sum of the distance of all points to the discriminate boundary. But because these two criteria are contradictory to each other, they can not be optimized simultaneously. In order to combine the two criteria in one model, Shi et al. [9] initiated the compromise solution approach, which is later widely used in many practical problems [1, 10 and 11]. Due to its linear feature, MCLP has the extraordinary advantages in the computational speed. But, when facing the non-linear classification problems, it will not do as well as the linear ones. In this paper, based on the kernel theory and SVM method, we introduced the kernel function into the original MCLP classification model and proposed a kernel based MCLP model to solve non-linear problems.

This paper is organized as follows. We will start from giving a brief review of two-class MCLP classification model in section 2. Then section 3 introduces kernel method and the proposed kernel-based MCLP method which is the main part of the paper. Section 4 tests the new model by a small sample with 90 cases. Section 5 shows a series of experimental results of the comparison with the existing methods: MCLP and SVM. The last section gives the conclusions.
2. Two-Class Multiple Criteria Linear Program Classifier

The formulation of classification problem can be described as follows:

Suppose the training set of the classification problem is \(X_{nxr}\), which has \(n\) observations in it. Of each observation, there are \(r\) attributes (or variables) which can be any real value and a two-value class label \(G\) (Good) or \(B\) (Bad). Of the training set, the \(ith\) observation can be described by \(X_i = (X_{i1}, \ldots, X_{ir})\), where \(i\) can be any number from 1 to \(n\). The objective of the classification problem is to learn knowledge from the training set that can classify these two classes, so that when given an unclassified sample \(z = (z_1, \ldots, z_r)\), we can predict its class label with the knowledge.

Multiple criteria linear programming (MCLP) can be used as a classification method. The framework of MCLP is based on the linear discriminate analysis models. In linear discriminate analysis, the purpose is to determine the optimal coefficients (or weights) for the attributes, denoted by \(W = (w_1, \ldots, w_r)\) and a boundary value (scalar) \(b\) to separate two predetermined classes: \(G\) (Good) and \(B\) (Bad); that is

\[
X_{i1}w_1 + \cdots + X_{ir}w_r \leq b, \quad X_i \in B \quad (Bad)
\]

and

\[
X_{i1}w_1 + \cdots + X_{ir}w_r \geq b, \quad X_i \in G \quad (Good)
\]

To formulate the criteria and constraints for data separation, some variables need to be introduced. Fig. 2 demonstrates the basic theory of MCLP. In the classification problem, \(X_iW = X_{i1}w_1 + \cdots + X_{ir}w_r\) is the score for the \(ith\) observation. If all records are linear separable and a sample \(X_i\) is correctly classified, then let \(\beta_i\) be the distance from \(X_i\) to \(b\), and consider the linear system, \(X_iW = b + \beta_i, \forall X_i \in G\) and \(X_iW = b - \beta_i, \forall X_i \in B\). However, if we consider the case where the two groups are not linear separable because of mislabeled records, a “soft margin” and slack distance variable \(\alpha_i\) need to be introduced. \(\alpha_i\) is defined to be the overlapping of the two-class boundary for mislabeled case \(X_i\). Previous equations now can be transformed to \(X_iW = b - \alpha_i + \beta_i, \forall X_i \in G\) and \(X_iW = b + \alpha_i - \beta_i, \forall X_i \in B\). To complete the definitions of \(\beta_i\) and \(\alpha_i\), let \(\beta_i = 0\) for all misclassified samples and \(\alpha_i = 0\) for all correctly classified samples.

Fig. 2. Overlapping of two-class Linear Discriment Analysis

A key idea in linear discriminate classification is that the misclassification of data can be reduced by using two objectives in a linear system. One is to maximize the sum of the distance of all points to the discriminate boundary \((\text{Maximize } b_1 + \cdots + b_r)\) and another is to minimize the overlapping degree \((\text{Minimize } a_1 + \cdots + a_r)\). The two criteria are contradictory to each other, so normally there are two different linear discriminate models MMD and MSD for classification [12].
MCLP combines these two criteria into one single model. It aims at minimizing the sum of $\alpha_i$ and maximizing the sum of $\beta_j$ simultaneously, which forms the two-class MCLP model.

Two-Class MCLP model: [12]

Minimize $\alpha_1 + \cdots + \alpha_n + \beta_1 + \cdots + \beta_n$

Subject to:

\[
\begin{align*}
X_1 w_1 + \cdots + v_r &= b + \alpha_1 - \beta_1, \quad \text{for } X_1 \in B \\
\quad & \text{......} \\
X_n w_1 + \cdots + v_r &= b - \alpha_n + \beta_n, \quad \text{for } X_n \in G \\
\alpha_1, \ldots, \beta_n & \geq 0
\end{align*}
\]

To facilitate the computation, the compromised solution approach has been employed to modify the above model and is finally applied to many practical problems. The compromised solution approach of two-class MCLP model[12] is:

Minimize $d_1^+ + d_1^- + d_\beta^+ + d_\beta^-$

Subject to:

\[
\begin{align*}
\alpha^* + \sum_{i=1}^n \alpha_i &= d_1^+ - d_1^- \\
\beta^* - \sum_{i=1}^n \beta_i &= d_\beta^+ - d_\beta^- \\
X_1 w_1 + \cdots + v_r &= b + \alpha_1 - \beta_1, \quad \text{for } X_1 \in B \\
\quad & \text{......} \\
X_n w_1 + \cdots + v_r &= b - \alpha_n + \beta_n, \quad \text{for } X_n \in G \\
\alpha_1, \ldots, \beta_n, d_1^+, d_\beta^+, d_1^-, d_\beta^- & \geq 0
\end{align*}
\]

Here $\alpha^*$, $\beta^*$ and $X_{n \times r}$ are given, $d_1^+, d_1^-, d_\beta^+, d_\beta^-$, $\alpha_i$, $\beta_j$, $w$ and $b$ are unrestricted. Solving this linear programming problem, we can get the best solution of $w$ and $b$, denoted by $(w^*, b^*)$. And the boundary plane of the two classes is:

\[
xw^* = b^*
\]

where $x$ has $r$ attributes, $w^* = (w_1, \ldots, w_r)^T$ represents the weight for the attributes and $b^*$ is a real number.

3. Kernel-Based MCLP Classifier

MCLP is a linear programming model and is only applicable for linear separable problem. When faced with nonlinear separable data set, it may have bad performance. However, as is commonly known, kernel function is a powerful tool to deal with nonlinear separable data set. By projecting the data into a high dimensional feature space, the data set will become more likely linear separable. So, if we can employ kernel method in MCLP classifier, it might become a nonlinear classifier.

Here, we will briefly introduce the kernel method at first. Any detail about it can be found in [7].

Suppose the training set of the nonlinear classification problem is $X_{n \times n}$, which has $n$ observations and $r$ attributes in it. Each observation belongs to one of the class, labeled $G$ (Good) or $B$ (Bad). Because this data set is nonlinear separable, we need to map it into a higher dimensional space with function $\Phi$ so that it becomes linear separable in
the new space. Then with some linear classification model, i.e. MCLP, we can get the separation function to classify the data in the new space.

But, as is commonly known, the mapping function $\Phi$ is always implicit. If the input data have many attributes, it is hard to perform such mapping operations. Kernel function offers an alternative way to this problem. It is defined to be such a direct computation method.

Definition of kernel function: A kernel is a function $K$, such that for all $x_1, x_2 \in X$, $K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2)$, where $\Phi$ is a mapping from $X$ to a new feature space. $x_1, x_2$ are two samples from $X$.

Although $\Phi$ is always implicit, $K$ is explicit. We can calculate the kernel function directly in some algorithm instead of compute function $\Phi$ of each sample separately. Some commonly used kernel functions are polynomial kernels, Gaussian RBF kernels and Mercer kernels. Here, RBF kernel is used in the experiments: $k(x_1, x_2) = \exp(- g \| x_1 - x_2 \|^2)$.

In data mining field, SVM is the best known and the first to use kernel to construct a nonlinear estimation algorithm. In addition, kernel algorithm is very useful in a number of approaches and fields.

In MCLP model (3), there is no way to use kernel function directly because the training samples are not included in the inner product formulation. Some changes need to be done on this model. Actually, in linear programming problem, given input data set $X$, the solution of $w$ must be driven from these $X$, which makes $w$ to be the linear combination of $x_i$. Assume that the solution of MCLP model can be described in the following form

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$  \hspace{1cm} (5)

and put the $w$ into model (3), we can get the following variation of MCLP model:

Minimize $d_a^+ + d_a^- + d_p^+ + d_p^-$

Subject to:

$$a_i^+ + \hat{\alpha}^a_i a_i = d_a^+ + d_a^-$$

$$b_i^+ + \hat{\alpha}^b_i b_i = d_p^+ + d_p^-$$

$$l_1 y_1(x_1 \times x_1) + l_2 y_2(x_2 \times x_1) + \ldots + l (x_n \times x_1) = b + a_1 - b_1, \text{ for } x_1 \in B,$$

$$\ldots \ldots \ldots$$

$$l_1 y_1(x_1 \times x_n) + l_2 y_2(x_2 \times x_n) + \ldots + l (x_n \times x_n) = b - a_n + b_n, \text{ for } x_n \in G,$$

$$a_i^+ \geq 0, \quad i = 1, \ldots, n;$$

$$b_i^+ \geq 0, \quad i = 1, \ldots, n;$$

$$d_a^+ , d_a^-, d_p^+, d_p^- \geq 0$$

(6)
Consider the same nonlinear classification problem with training set $X_{\text{new}}$, when each sample $x_i$ is mapped into a higher dimensional space with function $\Phi$, the new training set will be $X_{\text{new}} = (\Phi(x_1), \ldots, \Phi(x_n))^T$, where $\Phi(x_i)$ represents each new training sample in the new space. This data set will probably be linear separable. Input this $X_{\text{new}}$ into model (6), we will get the separation function of the two classes in the new space. The model (6) for $X_{\text{new}}$ is:

Minimize $d_\alpha^+ + d_\alpha^- + d_\beta^+ + d_\beta^-$

Subject to:

\[a^* + \bar{\alpha} \alpha_i = d_\alpha^+ - d_\alpha^-\]

\[b^* - \bar{\beta} \beta_i = d_\beta^+ - d_\beta^-\]

\[l_1 y_1 F(x_1) \Phi(x_1) + l_2 y_2 F(x_2) \Phi(x_2) + \ldots + l \Phi(x_n) \Phi(x_n) = b + a_i - b_i, \quad \text{for } x_i \bar{y} B,\]

\[l_1 y_1 F(x_1) \Phi(x_1) + l_2 y_2 F(x_2) \Phi(x_2) + \ldots + l \Phi(x_n) \Phi(x_n) = b - a_i + b_i, \quad \text{for } x_i \bar{y} G,\]

\[a_i \geq 0, \quad i = 1, \ldots, n;\]

\[b_i \geq 0, \quad i = 1, \ldots, n;\]

\[d_\alpha^+, d_\alpha^-, d_\beta^+, d_\beta^- \geq 0\]

Replace the $(\Phi(x_i), \Phi(x_j))$ with $K(x_i, x_j)$ in the above model, we have this kernel-based MCLP model:

Minimize $d_\alpha^+ + d_\alpha^- + d_\beta^+ + d_\beta^-$

Subject to:

\[a^* + \bar{\alpha} \alpha_i = d_\alpha^+ - d_\alpha^-\]

\[b^* - \bar{\beta} \beta_i = d_\beta^+ - d_\beta^-\]

\[l_1 y_1 K(x_1, x_1) + l_2 y_2 K(x_2, x_1) + \ldots + l \quad K(x_n, x_1) = b + a_i - b_i, \quad \text{for } x_i \bar{y} B,\]

\[l_1 y_1 K(x_1, x_n) + l_2 y_2 K(x_2, x_n) + \ldots + l \quad K(x_n, x_n) = b - a_i + b_i, \quad \text{for } x_i \bar{y} G,\]

\[a_i \geq 0, \quad i = 1, \ldots, n;\]

\[b_i \geq 0, \quad i = 1, \ldots, n;\]

\[d_\alpha^+, d_\alpha^-, d_\beta^+, d_\beta^- \geq 0\]

The kernel function in the above model can be any acceptable formulation. Actually, model (6) is the formulation of linear kernel which only applicable for linear separable problem. Here in this paper, we use RBF kernel to solve nonlinear problem. By this way, we incorporate kernel function into a linear programming model, which make it applicable in nonlinear separable problem. In this kernel-based MCLP model, $\alpha^*$, $\beta^*$ and $x_i, y_i$ are given, $d_\alpha^+$, $d_\alpha^-$, $d_\beta^+$, $d_\beta^-$, $\alpha_i$, $\beta_i$, $\lambda$ and $b$ are unrestricted. Solving this linear programming problem, we can get the best solution of $\lambda$ and $b$, denoted by $(\lambda^*, b^*)$. And the boundary plane of the two classes is:

\[\sum_{i=1}^{n} \lambda_i^* y_i K(x_i, x) = b^*\]
where \( x_i \) represents each input sample. And the corresponding decision function is:

\[
y = f(x) = \text{sgn}(\sum_{i=1}^{n} \lambda_i^* y_i K(x_i, x) - b^*)
\]  

(10)

4. Sample Testing

In order to test the validity of the proposed kernel-based MCLP model (8), we conducted an experiment on a small data set with 90 records which is a nonlinear separable set. In the following figures, the cross points and diamond points belong to two different classes respectively.

Fig. 3. (a) Original data set; (b) Result with original MCLP

Fig. 4. (a) Result with linear kernel MCLP; (b) Result with RBF kernel MCLP.

Fig. 3(a) shows the original data set, in which two classes appear to be easily separated from each other. Fig. 3(b) shows the classification result of MCLP model (3). Because MCLP is a linear model, it can only generate a straight line to classify the two classes which leads to the result in Fig. 3(b) that the data at the bottom-left belong to one class and the upper-right to the other. Fig. 4(a) describes the classification result by linear kernel MCLP, which is equivalent to model (6). We can see Fig. 4(a) has similar result with Fig. 3(b), which also proves that MCLP model (3) and linear kernel model (6) is identical. It is obvious from the results that MCLP and linear kernel MCLP model can not properly solve this nonlinear problem. But if we use RBF kernel in kernel-based MCLP model (8), the classification result will be perfect, and the precision will be 100% in this problem, see Fig. 4(b).

This test also shows that with an appropriate kernel function, the proposed kernel-based MCLP model (8) can effectively solve the nonlinear separable problem.
5. Experimental Results

In this section, we will use two public classification data sets Splice and Breast Cancer data sets in UCI repository [13] to compare the performance of kernel-based MCLP with the other two methods: SVM and original MCLP. Here we only use the linear kernel and RBF kernel on SVM and MCLP.

Before doing the classification, we perform some data preprocessing: scale each feature to [0, 1] and split each data set into 10 groups. We will use 10 cross-validation to estimate the accuracy of prediction. Every time we choose 9 data sets together to be the training set, and the one leftover is the test set. For every data set we can get 10 accuracies and at last the average accuracy is the whole accuracy, shown in table 1.

In every training process, parameters in every algorithm are chosen in some discrete set in order to get the best accuracy. We select the parameter $\gamma$ from $\{0.001, 0.01, 0.1, 1, 16, 128\}$. The best $\gamma$ of certain algorithm on a given training set is chosen according to the best accuracy, which is also the same value used for prediction on the test set.

Table 1. Accuracy with different algorithm

<table>
<thead>
<tr>
<th></th>
<th>linear kernel</th>
<th>RBF kernel</th>
<th>linear kernel</th>
<th>RBF kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accur.</td>
<td>SVM MCLP</td>
<td>SVM MCLP</td>
<td>SVM MCLP</td>
<td>SVM MCLP</td>
</tr>
<tr>
<td>splice</td>
<td>72.70% 79.60%</td>
<td>88.50% 80.90%</td>
<td>98%</td>
<td></td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>93% 88.50%</td>
<td>90.50% 96.50%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each data set, we draw a broken line chart to show the results of the five different algorithms as follows:

Fig. 5. Accuracy on data set Splice.

Fig. 6. Accuracy on data set Breast cancer.
From the above figures, we can see that for each data set, RBF kernel MCLP plays much better than the other algorithms especially original MCLP and linear kernel SVM. This also shows that kernel-based MCLP is a competitive method in classification.

6. Conclusions

In this paper, we introduced the kernel function into Multiple Criteria Linear Program (MCLP), which allows the model to deal with not only linear separable problems, but also nonlinear ones. Our model is simple in its formulation and effective in terms of performance in comparison with other algorithm such as SVM. And the experimental results on some benchmark classification data sets show that the kernel-based MCLP model performs even better than SVM classifier.

Acknowledgements

This research has been partially supported by grants from National Natural Science Foundation of China (#70921061, #10601064, #70531040, #70621001).

References