UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 14-11
Statistics and Econometrics Series (07)
May 2014

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A PROJECTION METHOD FOR MULTIOBJECTIVE MULTICLASS SVM

Belén Martín-Barragán[1] , Franciso Javier Prieto[2] and Ling Liu[3]

Abstract

Support Vector Machines (SVMs) have become a very popular technique in the machine-learning field for classification problems. It was originally proposed for classification of two classes. Various multiclass models with a single objective have been proposed mostly based on two families of methods: an all-together approach and a one-against-all approach. However, most of these single-objective models consider neither the different costs of misclassification nor the user's preferences. To overcome these drawbacks, multiobjective models have been proposed.

In this paper we rewrite the different approaches that deal with the multiclass SVM using multiobjective techniques. These multiobjective techniques can give us weakly Pareto-optimal solutions. We propose a multiobjective technique called Projected Multiobjective All-Together (**PMAT**), which works in a higher-dimension space than the object space. With this technique, we can theoretically characterize the Pareto-optimal solution set. For these multiobjective techniques we get approximate sets of the Pareto-optimal solutions. For these sets, we use hypervolume and epsilon indicators to evaluate different multiobjective techniques. From the experimental results, we can see that (**PMAT**) outperfoms the other multiobjective techniques. When facing classification problems with very large numbers of classes, we suggest combining a tree method and multiobjective techniques.

Keywords: Multiclass multiobjective SVM; Weakly Pareto-optimal solution; Pareto-optimal solution.

(1) B. Martín-Barragán, University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, United Kingdom, Email: Belen.Martin@ed.ac.uk
(2) F. Javier Prieto, Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, Email: fjp@est-econ.uc3m.es
(3) L.Liu, Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, Email: lliu@est-econ.uc3m.es

A Projection Method For Multiobjective multiclass SVM

Belén Martín-Barragán* Franciso Javier Prieto† Ling Liu‡

May 9, 2014

Abstract

Support Vector Machines (SVMs) have become a very popular technique in the machine learning field for classification problems. It is originally proposed for classification of two classes. Various multiclass models with a single objective have been proposed mostly based on two families of methods: an all-together approach and a one-against-all approach. However, most of these single-objective models consider neither the different costs of misclassification nor the user's preferences. To overcome these drawbacks, multiobjective models have been proposed.

In this paper we rewrite the different approaches that deal with the multiclass SVM using multiobjective techniques. These multiobjective techniques can give us weakly Pareto-optimal solutions. We propose a multiobjective technique called Projected Multiobjective All-Together (**PMAT**) which works in a higher-dimension space than the object space. With this technique, we can theoretically characterize the Pareto-optimal solution set. For these multiobjective techniques, we try to get approximate sets of the Pareto-optimal solutions. For these sets, we use hypervolume and epsilon indicators to evaluate different multiobjective techniques. From the experimental results, we can see that **PMAT** outperfoms the other multiobjective techniques. When facing classification problems with very large numbers of classes, we suggest to combine a tree method and multiobjective techniques.

KEYWORDS: Multiclass multiobjective SVM; Weakly Pareto-optimal solution; Pareto-optimal solution

1 Introduction

Data mining has become a crucial application area in modern science, industry and society due to the growing size of available databases. One of the main applications in this area is supervised classification: to obtain a model that predicts the value of one variable (class) based on the information from other variables. SVM is a popular approach to solve this problem. In [9], Cortes and Vapnik proposed the classical SVM for classification of two classes. The main idea is to generate a discriminant hyperplane which separates the input objects. During the last couple of decades, hundreds of applications and experiments have shown the high classification accuracy of SVM, e.g.[2, 14, 30]. SVM methods have been proved to be effective not only in applications but also in theory, see [18, 31, 32].

In real life, we have classification problems with more than two classes. It becomes interesting to extend this efficient method to multiclass classification. Researchers have proposed several methods to use SVM for solving multiclass classification problems. These methods

*B. Martín-Barragán, University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, United Kingdom, Email: Belen.Martin@ed.ac.uk

† F. Javier Prieto, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, Email: fjp@est-econ.uc3m.es

‡L. Liu, Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, Email: lliu@est-econ.uc3m.es

can be roughly grouped into two families. The first family constructs and combines several binary (two classes) classification problems, such as one-against-one, one-against-all and directed acyclic graph (DAG) SVMs, see [15, 16, 23, 31]. Alternatively, all-together methods directly find a discriminant function by solving a single optimization problem, which attempts to classify all patterns into the corresponding classes, e.g.[1, 10, 15, 34].

The aforementioned methods are based on solving single-objective optimization problems. They have one main drawback: they do not consider different costs for different misclassification errors, nor a priori information. This difference is important in many applications. For example, in medical diagnosis it is known that the cost of misclassifying a healthy patient as ill is different from misclassifying an ill patient as healthy. To overcome this drawback, a simple way is using a weighted single objective function. These weights are rough indexes for the importance of misclassification costs. But it is hard to associate real numbers with these importances. An alternative way is using a multiobjective approach.

In 2006, Carrizosa and Martin-Barragan proposed biobjective SVM for classification of two classes, see [3]. In that paper, they charaterized all the Pareto-optimal solutions of the biobjective SVM. In 2007, K-Tatsumi et al. used multiobjective multiclass SVM for pattern recognition, see [25]. Based on one-against-all and all-together methods, they proposed a series of multiobjective SVMs to solve multiclass classification problems, e.g.[26, 27, 28, 29]. However, the solutions given by them are weakly Pareto-optimal. Besides, they ignored that the cost of misclassifying class A objects as class B objects may be different from the cost of misclassifying class B objects as class A objects. Not only in medical diagnosis, but also in many other applications, this difference needs to be considered. For example, an investor may need a SVM which can identify high volatility shares as different from low volatility shares, while it may be acceptable to misclassify some of the low volatility shares as high volatility shares. Still, there is another problem that needs to be addressed. When facing classification problems for many classes, the multiobjective SVMs based on the all-together and one-against-all methods proposed in their papers require the values of many parameters to be selected. This may be a big challenge in practice.

In this paper, we first rewrite the multiobjective SVM based on one-against-all and all-together methods in order to consider asymmetric misclassification costs. However, one-against-one is also a widely used method for multiclass classification. It shows comparable results with respect to one-against-all and all-together methods, see [15, 21]. So, we also extend the multiobjective SVM based on a one-against-one method. By using an $\varepsilon-$constraint method, these multiobjective approaches will give us weakly Pareto-optimal solutions.

The second contribution of this work is to provide another model called Projected Multiobjective All-Together (PMAT) for which Pareto-optimal solutions can be characterized. When facing classification problems with a large number of classes, we suggest to use a tree method combined with multiobjective SVMs. Nowadays, the most commonly used methods for this kind of large-class classification problems are based on binary trees and single-objective models, [4, 5, 17]. From the experimental results in this paper,we can see that the proposed projected multiobjective SVM outperforms the other multiobjective approaches mentioned in this paper. Secondly, with a proper division of the classes, combining tree methods with multiobjective approaches performs efficiently. However, how to divide the classes optimally is still an open question that deserves further study.

This paper is organized as follows: Section 2 is devoted to the multiobjective SVMs based on all-together, one-against-all and one-agianst-one methods. Specifically, we introduce the hard-margin versions in Section 2.1 and the soft-margin versions in Section 2.2. To solve these multiobjective SVMs, we suggest to use $\varepsilon-$constraint method in Section 2.3. In Section 3, we

propose PMAT with which we can characterize the corresponding Pareto-optimal solutions. For this approach, we have hard-margin version showed in Section 3.1 and soft-margin version presented in Section 3.2. For large-class classification problems, a multidecision tree method combined with the multiobjective approaches is suggested in Section 4. Computational results are shown in Section 5. Finally, conclusions are presented in Section 6.

2 Multiclass multiobjective SVMs

In what follows we assume that we have a training set $I = \{x_i\}_{i=1}^{k} \subseteq \mathbb{R}^l$ corresponding to m different classes, and let $y_i \in G = \{1, ..., m\}$ denote the class membership of vector x_i. The number of observations in the training set belonging to class p is denoted by k^p. Our aim is to generate a decision function which can help us to predict with high accuracy the class memberships of new objects. To achieve this aim, we generate the discriminant hyperplanes as follows:

- The discriminant hyperplane to separate class p data against class q data:

$$L^{pq}: \quad (\omega^{pq})^T x + b^{pq} = 0, \quad p \neq q, p, q \in G.$$

Ideally, we would like to have all class p objects above hyperplane L^{pq}, and all class q objects below $L^{pq}, p \neq q, p, q \in G$. If we can find hyperplanes such that the training objects satisfy this ideal situation, we say that the training objects are linearly separable. The following figure is an example of linearly separable training objects.

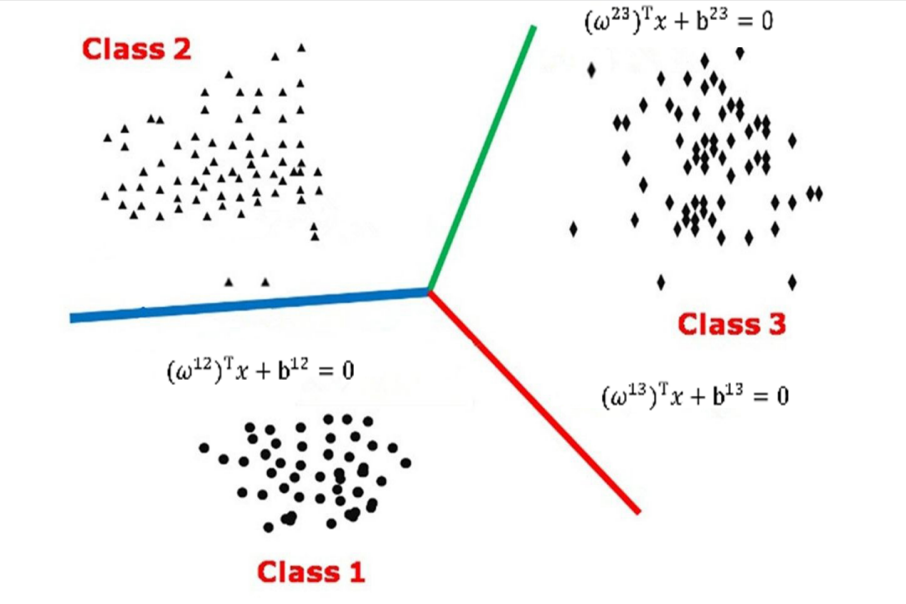

Figure 1. Linearly separable training objects

Before introducing the multiobjective approaches based on all-together, one-against-all and one-against-one models, we briefly review the single-objective approaches for multiclass classification. As in the binary classical SVM [9], the margin between class p objects and class q objects is defined as the distance between two support hyperplanes, which is equal to $2/\|\omega^{pq}\|$. To consider the nonlinearly separable case and the over-fitting phenomenon, we introduce auxiliary variables $\xi^{pq}, p \neq q,\ p, q \in G$ to allow some class p objects to be

misclassified as class q objects. To get a high generalization ability and high classification accuracy, we should maximize all the margins and minimize the classification errors. Following [15, 22, 35] and based on the squares of these auxiliary variables, we can construct the following single-objective problem:

$$\begin{aligned} \min_{\omega,b} \quad & \|\omega\|^2 + \sum_{p=1}^{m} \sum_{q \neq p} c^{pq} \sum_{x \in I_p} (\xi_x^{pq})^2, \\ \text{s.t.} \quad & (\omega^{pq})^T x + b^{pq} + \xi_x^{pq} \geq 1, x \in I_p, p \neq q, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} + \xi_x^{qp} \geq 1, x \in I_q, p \neq q, p, q \in G. \end{aligned} \tag{1}$$

where, $I_p = \{x \in I | x \in \text{ class } p\}$, $\omega = (\omega^{12}, \omega^{13}, \cdots, \omega^{(m-1)m})$ and $c^{pq} \geq 0$. The above program (1) is a strictly convex quadratic problem, so it has an unique optimizer.

After computing these ω^{pq}, $b^{pq}, q > p, p, q \in G$, we need to construct the decision rule to determine the class membership of new objects. To achieve this aim, we take $\frac{(\omega^{pq})^T x + b^{pq}}{\|\omega^{pq}\|}$ as the score obtained from $L^{pq} : (\omega^{pq})^T x + b^{pq} = 0$ for class p when given object x. We then collect all the scores generated for class p as:

$$d^p = \sum_{q \neq p, q \in G} \frac{(\omega^{pq})^T x + b^{pq}}{\|\omega^{pq}\|}, p \in G.$$

Then we can determine the class membership of object x by :

$$x \text{ belongs to class } p, \text{ if and only if } p = arg \left\{ \max_{q \in G} d^q \right\}. \tag{2}$$

Problem (1) is the basic singleobjective problem for classifying all the classes at the same time. Based on this model, we can introduce the multiobjective SVM based on all-together method. A one-against-all method solves m binary SVMs. From these m binary SVMs, we get a series of start points. These starting points are then used in a multiobjective SVM based on a one-against-all method defined from the combination of a one-against-all method and the multiobjective extension of problem (1). In a similar manner, we adapt the multiobjective SVM based on the one-against-one method.

2.1 Hard-margin multiobjective SVMs

In this section we assume that the training objects are linearly separable. As in [28], we can take $(W^p)^T x + B^p$ to measure the degree of confidence of assigining object x to class p instead of other classes. Then we can get the hyperplane for class p against class q data as:

$$L^{pq} : \quad (W^p - W^q)^T x + B^p - B^q = 0, \quad p \neq q, p, q \in G. \tag{3}$$

That's to say $\omega^{pq} = W^p - W^q, b^{pq} = B^p - B^q, p \neq q, p, q \in G$. As mentioned before, we will consider the asymmetry of misclassification costs. So here, instead of using margins defined as $\frac{2}{\|\omega^{pq}\|} = \frac{2}{\|W^p - W^q\|}$, we take the geometric margins as:

- The geometric margin from class p to class q is:

$$\rho^{pq} = \min_{x \in I_p} \frac{|(W^p - W^q)^T x + B^p - B^q|}{\|W^p - W^q\|}, \quad q \neq p, p, q \in G.$$

2.1.1 Hard-margin multiobjective SVM based on all-together method

In [27], the authors proposed a multiobjective SVM based on an all-together method. They maximize all the pair-wise interclass margins defined as $\min\{\rho^{pq}, \rho^{qp}\}, p \neq q, p, q \in G$. However, this approach ignores any asymmetric misclassification costs. Instead, in this paper, we maximize all the geometric margins $\rho^{pq}, p \neq q, p, q \in G$. With this idea in mind, we can formulate the hard-margin multiobjective SVM based on an all-together method as:

$$\begin{aligned} \max_{W,B} \quad & \left(\rho^{12}, \rho^{21}, \cdots, \rho^{(m-1)m}, \rho^{m(m-1)}\right), \\ \text{s.t.} \quad & (W^p - W^q)^T x + B^p - B^q \geq 1, x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{4}$$

To simplify the above formulation (4), we let

$$\sigma^{pq} = \min_{x \in I_p} (W^p - W^q)^T x + B^p - B^q, p \neq q, p, q \in G.$$

Then, the geometric margins are given by $\rho^{pq} = \frac{\sigma^{pq}}{\|W^p - W^q\|}, p \neq q, p, q \in G$, and we can rewrite (4) as follows:

$$\begin{aligned} \max_{W,B,\sigma} \quad & \left(\frac{\sigma^{12}}{\|W^1 - W^2\|}, \frac{\sigma^{21}}{\|W^1 - W^2\|}, \cdots, \frac{\sigma^{(m-1)m}}{\|W^{m-1} - W^m\|}, \frac{\sigma^{m(m-1)}}{\|W^{m-1} - W^m\|}\right), \\ \text{s.t.} \quad & (W^p - W^q)^T x + B^p - B^q \geq \sigma^{pq}, x \in I_p, q \neq p, p, q \in G, \\ & \sigma^{pq} \geq 1, q \neq p, p, q \in G. \end{aligned} \tag{5}$$

We denote this problem as **HMAT** (Hard-margin Multiobjective All-Together). This is a multiobjective optimization problem with $m(m-1)$ objectives, $m(m+l)$ decision variables and $(m-1)k$ constraints.

2.1.2 Hard-margin multiobjective SVM based on one-against-all method

One-against-all methods solve m binary SVMs, where each of these binary SVMs has low computational cost. The p-th binary SVM classify class p objects against all other objects. The unbalance in the numbers of class objects in these binary SVMs may affect the accuracy of classification and their generalization ability, [28, 29]. There are also some experimental results showing that a one-against-all method may have a worse accuracy for some problems compared with all-together, one-against-one and DAG methods, see [15].

We want to combine the advantages of a one-against-all method and the merits of an all-together method. To achieve this aim, we process the multiclass classification in two phases as in [28]. In the first phase, we use single-objective binary SVMs based on a one-against-all method to get a set of values $\overline{W}^p, \overline{B}^p,\ p \in G$. In the second phase, we define W^p to have the form $\alpha^p \overline{W}^p$. Instead of maximizing all the pair-wise interclass margins [28], we maximize all the geometric margins to get the values for α^p and B^p. With this in mind, we can construct

our hard-margin multiobjective SVM based on one-against-all method as follows:

$$\begin{aligned}\max_{\alpha,B,\sigma} \quad & \left(\frac{\sigma^{12}}{\|\alpha^1\overline{W}^1-\alpha^2\overline{W}^2\|},\frac{\sigma^{21}}{\|\alpha^1\overline{W}^1-\alpha^2\overline{W}^2\|},\cdots,\frac{\sigma^{(m-1)m}}{\|\alpha^{m-1}\overline{W}^{m-1}-\alpha^m\overline{W}^m\|},\right.\\ & \left.\frac{\sigma^{m(m-1)}}{\|\alpha^{m-1}\overline{W}^{m-1}-\alpha^m\overline{W}^m\|}\right),\\ \text{s.t.} \quad & \left(\alpha^p\overline{W}^p-\alpha^q\overline{W}^q\right)^T x+B^p-B^q\geq\sigma^{pq}, x\in I_p, q\neq p, p,q\in G,\\ & \sigma^{pq}\geq 1, q\neq p, p,q\in G.\end{aligned} \tag{6}$$

For convenience, we call the above problem (6) as **HMOAA** (Hard-margin Multiobjective One-Against-All). It has $m(m+1)$ decision variables, $(m-1)k$ constraints and $m(m-1)$ objective functions.

2.1.3 Hard-margin multiobjective SVM based on one-against-one method

A one-against-one method solves $\frac{m(m-1)}{2}$ binary SVMs, where each SVM considers just two of these classes. In some experiments, one-against-one methods show a comparable performance with respect to one-against-all and all-together methods, see [7, 15]. So we believe it is also interesting to extend this method to multiobjective SVMs.

As in Section 2.1.2, we define the multiobjective SVM based on a one-against-one method in two phases. First, from $\frac{m(m-1)}{2}$ binary SVMs, we get a series of values $\bar{\omega}^{pq}$. $\bar{\omega}^{pq}$ is the vector of coefficients of the separating hyperplane for class p against class q. We introduce the combination $\sum_{q\neq p,q\in G}\alpha^{pq}\bar{\omega}^{pq}$ as the coefficients of the hyperplane separating class p against the rest of the classes. Now we can reconstruct the discriminant hyperplane for class p against class q as follows:

$$L^{pq}:\left(\sum_{r\neq p,r\in G}\alpha^{pr}\bar{\omega}^{pr}-\sum_{t\neq q,t\in G}\alpha^{qt}\bar{\omega}^{qt}\right)^T x+B^p-B^q=0,\ p\neq q,p,q\in G.$$

As before, by maximizing all the the pairwise geometric margins, we get our hard-margin multiobjective SVM based on a one-against-one method, as follows:

$$\begin{aligned}\max_{\alpha,B,\sigma} \quad & \left(\frac{\sigma^{12}}{\|\sum_{r\neq 1}\alpha^{1r}\bar{\omega}^{1r}-\sum_{t\neq 2}\alpha^{2t}\bar{\omega}^{2t}\|},\frac{\sigma^{21}}{\|\sum_{r\neq 1}\alpha^{1r}\bar{\omega}^{1r}-\sum_{t\neq 2}\alpha^{2t}\bar{\omega}^{2t}\|},\cdots,\right.\\ & \left.\frac{\sigma^{m(m-1)}}{\|\sum_{r\neq m-1}\alpha^{(m-1)r}\bar{\omega}^{(m-1)r}-\sum_{t\neq m}\alpha^{mt}\bar{\omega}^{mt}\|}\right),\\ \text{s.t.} \quad & \left(\sum_{r\neq p}\alpha^{pr}\bar{\omega}^{pr}-\sum_{t\neq q}\alpha^{qt}\bar{\omega}^{qt}\right)^T x+B^p-B^q\geq\sigma^{pq}, x\in I_p,\ q\neq p, p,q\in G,\\ & \sigma^{pq}\geq 1, q\neq p,\ p,q\in G.\end{aligned} \tag{7}$$

We denote the above optimization problem (7) as **HMOAO** (Hard-margin Multiobjective One-Against-One). It has $m(m-1)$ objectives, $m(2m-1)$ decision variables and $(m-1)k$ constraints.

2.2 Soft-margin multiobjective SVMs

The constraints in the hard-margin methods may be too strict for general problems, as they assume that the training objects are linearly separable and they may lead to the overfitting phenomenon. As in (1), we can add $\xi^{pq}, p \neq q,\ p, q \in G$, to allow some objects from class p to be incorrectly classified as class q. In multiobjective SVMs, we should consider not only maximizing all the geometric margins but also minimizing the misclassification errors.

In order to simplify the problem, unlike [27, 28], we define the geometric margins with ξ^{pq} embedded within the vectors ω. So instead of both maximizing all the geometric margins and minimizing the misclassification errors, we only need to maximize all the geometric margins where these margins are computed for the modified data that incorporates information on the slack variables ξ^{pq}. To achieve this aim, we project the objects onto a higher-dimension space, as in [3]. In that higher dimension space, we can redefine the separating hyperplanes as:

$$L^{pq} : (W^p - W^q, c^{pq}\xi^{pq}, c^{qp}\xi^{qp})^T (x, \delta^{pq}_{\xi x}, \delta^{qp}_{\xi x}) + B^p - B^q = 0,$$

where $\delta^{pq}_{\xi x} = \frac{1}{c^{pq}} e_i$, and e_i is the i-th unit vector, if x is the i-th object in class p; else $\delta^{pq}_{\xi x} = 0$.

We define the distance from object $x \in I_p$ to hyperplane L^{pq} as:

$$\bar{\rho}^{pq}_x = \frac{|(W^p - W^q)^T x + B^p - B^q + \xi^{pq}_x|}{\|\left(W^p - W^q, c^{pq}\xi^{pq}, c^{qp}\xi^{qp}\right)\|},\ p \neq q,\ p, q \in G.$$

We redefine the geometric margin $\bar{\rho}^{pq}$ as the distance from hyperplane L^{pq} to the closest object in class p.

- The geometric margin for class p objects against class q objects is defined as:

$$\bar{\rho}^{pq} = \min_{x \in I_p} \frac{|(W^p - W^q)^T x + B^p - B^q + \xi^{pq}_x|}{\|\left(W^p - W^q, c^{pq}\xi^{pq}, c^{qp}\xi^{qp}\right)^T\|},\ q \neq p, p, q \in G.$$

2.2.1 Soft-margin multiobjective SVM based on all-together method

By maximizing all the geometric margins with slack variables embedded, we can formulate the soft-margin multiobjective SVM based on all-together method as:

$$\begin{aligned} \max_{W,B,\xi} \quad & \left(\bar{\rho}^{12}, \bar{\rho}^{21}, \cdots, \bar{\rho}^{(m-1)m}, \bar{\rho}^{m(m-1)}\right), \\ \text{s.t.} \quad & (W^p - W^q)^T x + B^p - B^q \geq 1, x \in I_p, p \neq q, p, q \in G, \\ & \xi^{pq}_x \geq 0, x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{8}$$

To simplify the formulation of problem (8), we define $\bar{\sigma}^{pq} = \min_{x \in I_p} |(W^p - W^q)^T x + B^p - B^q + \xi^{pq}_x|$. Then a simplified formulation for the soft-margin multiobjective SVM based on an

all-together method is:

$$\begin{aligned}\max_{W,B,\bar{\sigma},\xi} \quad & \left(\frac{\bar{\sigma}^{12}}{\|(W^1-W^2,\xi^{12},\xi^{21})\|_c},\frac{\bar{\sigma}^{21}}{\|(W^1-W^2,\xi^{12},\xi^{21})\|_c},\cdots,\frac{\bar{\sigma}^{m(m-1)}}{\|(W^{m-1}-W^m,\xi^{(m-1)m},\xi^{m(m-1))}\|_c}\right),\\ \text{s.t.} \quad & (W^p-W^q)^T x+B^p-B^q+\xi_x^{pq}\geq\bar{\sigma}^{pq}, x\in I_p, p\neq q, p,q\in G,\\ & \bar{\sigma}^{pq}\geq 1, \xi_x^{pq}\geq 0,\ x\in I_p, p\neq q, p,q\in G,\end{aligned} \tag{9}$$

where, $\|(W^p-W^q,\xi^{pq},\xi^{qp})\|_c=\|(W^p-W^q,c^{pq}\xi^{pq},c^{qp}\xi^{qp})\|,\ q\neq p,\ p,q\in G.$

We refer to the above formulation (9) as **SMAT** (Soft-margin Multiobjective All-Together). This multiobjective optimization problem has $m(m-1)$ objectives, $(m+l)m+(m-1)k$ variables and $(m-1)k$ constraints.

2.2.2 Soft-margin multiobjective SVM based on one-against-all method

As in Section 2.1.2, we construct a soft-margin multiobjective SVM based on a one-against-all method in two phases. In the first phase, we compute values $\overline{W}^p$ by solving a series of binary SVMs. The p-th binary SVM classifies class p objects against all the other objects. Then, we introduce a second phase where all the geometric margins are maximized as in Section 2.2.1. We can formulate a **SMOAA** (Soft-margin Multiobjective One-Against-All) problem as:

$$\begin{aligned}\max_{\alpha,B,\bar{\sigma},\xi} \quad & \left(\frac{\bar{\sigma}^{12}}{\|(\alpha^1\overline{W}^1-\alpha^2\overline{W}^2,\xi^{12},\xi^{21})\|_c},\frac{\bar{\sigma}^{21}}{\|(\alpha^1\overline{W}^1-\alpha^2\overline{W}^2,\xi^{12},\xi^{21})\|_c},\cdots,\right.\\ & \left.\frac{\bar{\sigma}^{m(m-1)}}{\|(\alpha^{m-1}\overline{W}^{m-1}-\alpha^m\overline{W}^m,\xi^{(m-1)m},\xi^{m(m-1)})\|_c}\right),\\ \text{s.t.} \quad & (\alpha^p\overline{W}^p-\alpha^q\overline{W}^q)^T x+B^p-B^q+\xi_x^{pq}\geq\bar{\sigma}^{pq}, x\in I_p, p\neq q, p,q\in G,\\ & \bar{\sigma}^{pq}\geq 1, \xi_x^{pq}\geq 0,\ x\in I_p, p\neq q, p,q\in G.\end{aligned} \tag{10}$$

This is a multiobjective optimization problem with $m(m-1)$ objectives, $(m+1)m+(m-1)k$ variables and $(m-1)k$ constraints.

2.2.3 Soft-margin multiobjective SVM based on one-agianst-one method

In a manner similar to the one presented in Section 2.1.3, after computing a series of values $\bar{\omega}^{pq}$ we can construct a soft-margin multiobjective SVM based on a one-against-one method. Here, $\bar{\omega}^{pq},\ p\neq q, p,q\in G$, is computed from a binary SVM which classifies class p objects

against class q objects:

$$\max_{\alpha,B,\bar{\sigma},\xi} \quad \left(\frac{\bar{\sigma}^{12}}{\|(\sum_{r\neq 1}\alpha^{1r}\bar{\omega}^{1r} - \sum_{t\neq 2}\alpha^{2t}\bar{\omega}^{2t}, \xi^{12}, \xi^{21})\|_c}, \frac{\bar{\sigma}^{21}}{\|(\sum_{r\neq 1}\alpha^{1r}\bar{\omega}^{1r} - \sum_{t\neq 2}\alpha^{2t}\bar{\omega}^{2t}, \xi^{12}, \xi^{21})\|_c}, \cdots, \right.$$
$$\left. \frac{\bar{\sigma}^{m(m-1)}}{\|(\sum_{r\neq m-1}\alpha^{(m-1)r}\bar{\omega}^{(m-1)r} - \sum_{t\neq m}\alpha^{mt}\bar{\omega}^{mt}, \xi^{(m-1)m}, \xi^{m(m-1)})\|_c} \right),$$
$$\text{s.t.} \quad (\sum_{r\neq p}\alpha^{pr}\bar{\omega}^{pr} - \sum_{t\neq q}\alpha^{qt}\bar{\omega}^{qt})^T x + B^p - B^q + \xi_x^{pq} \geq \bar{\sigma}^{pq}, x \in I_p, p \neq q, p, q \in G,$$
$$\bar{\sigma}^{pq} \geq 1, \xi_x^{pq} \geq 0, \ x \in I_p, p \neq q, p, q \in G. \tag{11}$$

We refer to the above problem (11) as **SMOAO** (Soft-margin Multiobjective One-Against-One). It has $m(m-1)$ objectives, $(m-1)k + m(2m-1)$ variables and $(m-1)k$ constraints.

2.3 $\varepsilon-$constraint method to solve multiobjective SVMs

To solve the multiobjective SVMs introduced in Sections 2.1 and 2.2, we first review some basic concepts. For multiobjective optimization problems, as the objectives may be conflicting, it may be impossible to find an optimal solution. Instead, we try to get Pareto-optimal solutions. Following [8, 11, 12], we can define Pareto-optimal solutions and weakly Pareto-optimal solutions as follows:

Given a general multiobjective problem:

$$\max_{\mu\in C}(f_1(\mu), f_2(\mu), \cdots, f_h(\mu)).$$

- A feasible solution μ^* is Pareto-optimal iff there does not exist another feasible solution $\mu \in C$ such that $f_i(\mu) \geq f_i(\mu^*)$ for all $i \in \{1, 2, \cdots, h\}$, and $f_j(\mu) > f_j(\mu^*)$ for at least one $j \in \{1, 2, \cdots, h\}$.
- A feasible solution μ^* is weakly Pareto-optimal iff there does not exist another feasible solution $\mu \in C$ such that $f_i(\mu) > f_i(\mu^*)$ for all $i \in \{1, 2, \cdots, h\}$.

Let's denote $\mathbb{P}$ as the set of all the Pareto-optimal solutions of a multiobjective problem. However, it is hard to compute all these Pareto-optimal solutions. In the past decades, researchers have proposed different methods to obtain an approximating set of Pareto-optimal solutions, such as the weighted-sum method, the $\varepsilon-$constraint method, the hybrid method, Benson's method and so on, see [8, 11, 12]. The weighted-sum method gives solutions that are guaranteed to be Pareto-optimal. However, in the nonconvex case, there is no guarantee that any Pareto-optimal solution is achievable by this method, see [12, 20]. Both the hybrid method and Benson's method need some initial solutions. These initial solutions may be hard to find for some problems. The hybrid method is a combination of weighted sum and $\varepsilon-$constraint methods. Benson's method can be seen as a method to check if the initial solution is Pareto-optimal or not. If the initial solution is not Pareto-optimal, then it will guide us to find a Pareto-optimal solution. As the $\varepsilon-$constraint method may produce more solutions and these

solutions are at least weakly Pareto-optimal [12, 20], we choose the $\varepsilon-$constraint method to solve our multiobjective problems.

The $\varepsilon-$constaint method works in this way: It takes one of the objectives of the multiobjective problem as the objective function of the single-objective problem. The other objectives will be used as constraints.

For example, we try to solve (9) with $\varepsilon-$constraint method as follows:

$$\begin{aligned}
\max_{W,B,\sigma,\xi} \quad & \frac{\sigma^{rs}}{\|(W^r - W^s, \xi^{rs}, \xi^{sr})\|_c}, \\
\text{s.t.} \quad & \frac{\sigma^{pq}}{\|(W^p - W^q, \xi^{pq}, \xi^{qp}\|_c} \geq \varepsilon^{pq},\ (p,q) \neq (r,s), q \neq p, p, q \in G, \\
& (W^p - W^q)^T x + B^p - B^q + \xi_x^{pq} \geq \sigma^{pq},\ x \in I_P, q \neq p, p, q \in G, \\
& \sigma^{pq} \geq 1, \xi_x^{pq} \geq 0,\ x \in I_p, (p,q) \neq (r,s), p \neq q, p, q \in G.
\end{aligned} \tag{12}$$

From [12], we know that each optimal solution of problem (12) is weakly Pareto-optimal for problem (9). Moreover, each Pareto-optimal solution of (9) will also be optimal for problem (12) and some proper choice of ε. However, (12) is still not easy to solve. In [27], they proposed to approximate the Pareto-optimal solutions by fixing the value of $\bar{\sigma}^{rs}$ to a certain value c. Thanks to the homogeneity of the solutions, we approximate the set of Pareto-optimal solutions for (9) by solving:

$$\begin{aligned}
\max_{W,B,\bar{\sigma}^{-rs},\xi} \quad & \frac{c}{\|(W^r - W^s, \xi^{rs}, \xi^{sr})\|_c}, \\
\text{s.t.} \quad & \frac{\bar{\sigma}^{pq}}{\|(W^p - W^q, \xi^{pq}, \xi^{qp})\|_c} \geq \varepsilon^{pq},\ (p,q) \neq (r,s), q \neq p, p, q \in G, \\
& (W^r - W^s)^T x + B^r - B^s + \xi_x^{rs} \geq c,\ x \in I_r, \\
& (W^p - W^q)^T x + B^p - B^q + \xi_x^{pq} \geq \bar{\sigma}^{pq},\ (p,q) \neq (r,s),\ x \in I_P, q \neq p, p, q \in G, \\
& \bar{\sigma}^{pq} \geq 1, (p,q) \neq (r,s), p \neq q, p, q \in G, \\
& \xi_x^{pq} \geq 0,\ x \in I_p, (p,q) \neq (r,s), p \neq q, p, q \in G.
\end{aligned} \tag{13}$$

Problem (13) can be seen as a SOCP (second-order cone program). Using different values of the parameters ε^{pq} we can obtain different solutions. In [27, 28], they suggest to fix the value of ε^{pq} based on the solution of a single objective SVM problem (1). To approximate the set of all Pareto-optimal solutions, we suggest to use different values of ε^{pq} selected to ensure that problem (13) remains feasible.

The other multiobjective SVMs mentioned in Sections 2.1 and 2.2 can also be solved using similar methods. In practice, it is more flexible to use soft-margin multiobjective approaches than to use hard-margin multiobjective approaches. From Sections 2.2.1, 2.2.2 and 2.2.3, we can see that **SMOAA** has the fewest variables. And when the dimension l is larger than the number of classes, **SMOAO** has fewer variables than **SMAT**. What's more, the optimal values $\alpha^p \overline{W}^p$ obtained from **SMOAA** and the optimal $\sum_{r \neq p} \alpha^{pr} \bar{\omega}^{pr}$ obtained from **SMOAO** are also feasible for (9). So we can see that an optimal solution of (9) can't be strictly dominated by optimal solutions of either (10) or (11).

3 Projected multiobjective all-together

In Section 2.3 we have commented how the $\varepsilon-$constraint method provides weakly Pareto-optimal solutions for the multiobjective models introduced in Sections 2.1 and 2.2. In this section, we propose a new and simpler multiobjective approach for these problems, having the property that we can characterize all its Pareto-optimal solutions. It is based on projecting the objective space onto a higher-dimension space, in which we can define the geometric margins in a tractable way. We will refer to the simplified multiobjective SVM based on the use of that projected space as **PMAT** (Projected Multiobjective SVM based on All-Together). As before, the next two sections will introduce the hard-margin **PMAT** and the soft-margin **PMAT** versions of the model.

3.1 Hard-margin projected multiobjective all-together

For linearly separable training objects, we introduce the following (projection) transformation. Let

$$\Delta_x^{pq} = (\delta_x^{12}, \delta_x^{13}, \cdots, \delta_x^{(m-1)m}),\ p < q, p, q \in G,$$

with

$$\delta_x^{ij} = \begin{cases} x, \text{if } (i,j) = (p,q); \\ 0, \text{else.} \end{cases} \tag{14}$$

Then we can express hyperplane L^{pq} in the projected space as: $L^{pq} : \omega^T \Delta_x^{pq} + b^{pq} = 0$, where $\omega = (\omega^{12}, \omega^{13}, \ldots, \omega^{(m-1)m})$.

We redefine the geometric margin from object $x \in I_p$ to hyperplane L^{pq} as the Euclidean distance in the projected space:

$$\varrho_x^{pq}(\omega, b) = \frac{|(\omega)^T \Delta_x^{pq} + b^{pq}|}{\|\omega\|} = \frac{(\omega^{pq})^T x + b^{pq}}{\|\omega\|},\ x \in I_p, p \neq q, p, q \in G.$$

Notice that, in the separable case, we have all class p objects over hyperplane L^{pq}. So we have $(\omega^{pq})^T x + b^{pq} > 0$, for all $x \in I_p, p \neq q, p, q \in G$.

As before, we can define the geometric margin from class p to hyperplane L^{pq} as :

$$\varrho^{pq}(\omega, b) = \min_{x \in I_p} \varrho_x^{pq}(\omega, b), p \neq q, p, q \in G.$$

In order to maximize all the pair-wise geometric margins $\varrho^{pq}(\omega, b)$, we can construct the hard-margin projected multiobjective SVM based on all-together method as:

$$\begin{aligned} \max_{\omega, b} \quad & \left(\varrho^{12}(\omega, b), \varrho^{21}(\omega, b), \cdots, \varrho^{(m-1)m}(\omega, b), \varrho^{m(m-1)}(\omega, b)\right) \\ \text{s.t.} \quad & (\omega^{pq})^T x + b^{pq} > 0, x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} > 0, x \in I_q, q > p, p, q \in G. \end{aligned} \tag{15}$$

We refer to the above multiobjective optimization problem as **HPMAT** (Hard-margin Projected Multiobjective All-Together). For this multiobjective problem, we define the following minimax weighted problem that provides Pareto-optimal solutions of problem (15):

$$\begin{aligned} \max_{\omega, b} \min \ & \left(\varrho^{12}(\omega, b), \theta^{21}\varrho^{21}(\omega, b), \cdots, \theta^{(m-1)m}\varrho^{(m-1)m}(\omega, b), \theta^{m(m-1)}\varrho^{m(m-1)}(\omega, b)\right) \\ \text{s.t.} \ & (\omega^{pq})^T x + b^{pq} > 0, x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} > 0, x \in I_q, q > p, p, q \in G. \end{aligned} \tag{16}$$

The above problem (16) will be a bridge for us to get the characterization of the Pareto-optimal solutions of **HPMAT**. The following lemma establishes the relationship between (16) and **HPMAT**. The values θ^{pq} can be seen as the proportions of the geometric margin ϱ^{12} over the geometric margins ϱ^{pq}.

Lemma 3.1. *(1) The optimal solution of (16) is weakly Pareto-optimal for **HPMAT**;*
*(2) The weakly Pareto-optimal solutions of **HPMAT** are optimal for (16) given some specific values* $\boldsymbol{\theta} = (\theta^{21}, \cdots, \theta^{(m-1)m}, \theta^{m(m-1)}) > 0$.

The proof can be seen in Appendix 1.

Before attempting to characterize the optimal solutions of (16), we introduce the following problem that provides useful information on the optimal solution of (16):

$$\begin{aligned} \min_{\omega,b} \quad & \|\omega\|^2, \\ \text{s.t.} \quad & (\omega^{pq})^T x + b^{pq} \geq 1, x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} \geq 1, x \in I_q, p > q, p, q \in G. \end{aligned} \tag{P1}$$

Problem (16) can be easily replaced with a quadratic problem. By solving that quadratic problem, we can characterize the weakly Pareto-optimal solutions of **HPMAT**, as the following theorem shows.

Theorem 3.2. *The set of weakly Pareto-optimal solutions for **HPMAT** is :*

$$\left\{(\boldsymbol{\omega}, \boldsymbol{b}) = \left(\mu\omega_\theta^{12}, \cdots, \mu\omega_\theta^{(m-1)m}, \mu b_\theta^{12}, \cdots, , \mu b_\theta^{(m-1)m}\right) | \mu > 0, \theta^{pq} > 0, p < q, p, q \in G\right\},$$

where $\theta^{12} = 1$, $\omega_\theta^{pq} = \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}}\omega_1^{pq}$ *and* $b_\theta^{pq} = \frac{\theta^{qp}-\theta^{pq}}{2\theta^{pq}\theta^{qp}} + \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}} b_1^{pq}$ *for all* $p \neq q, p, q \in G$, *with* (w_1, b_1) *being optimal to* $(P1)$.

Proof. First, using the definition of the geometric margins, we can rewrite (16) as:

$$\begin{aligned} \min_{\omega,b} \quad & \frac{\|\omega\|}{\min\left\{\min_{x\in I_1} \omega^{12}x + b^{12}, \theta^{21} \min_{x\in I_2} -\omega^{12}x - b^{12}, \cdots, \theta^{m(m-1)} \min_{x\in I_m} -\omega^{(m-1)m}x - b^{(m-1)m}\right\}} \\ \text{s.t.} \quad & \omega^{pq}x + b^{pq} > 0, x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} > 0, x \in I_q, q > p, p, q \in G. \end{aligned} \tag{17}$$

We can see that (ω, b) is optimal for (17) iff $(\mu\omega, \mu b)$ is optimal for (17) for any $\mu > 0$. So we can standardize the denominator of the objective. Then we can solve the following problem to get the optimal solution of (17):

$$\begin{aligned} \min_{\omega,b} \quad & \|\omega\| \\ \text{s.t.} \quad & \min\left\{\min_{x\in I_1} (\omega^{12})^T x + b^{12}, \theta^{21} \min_{x\in I_2} (-\omega^{12})^T x - b^{12}, \cdots, \theta^{m(m-1)} \min_{x\in I_m} (-\omega^{(m-1)m})^T x - b^{(m-1)m}\right\} = 1. \end{aligned} \tag{18}$$

Easily we can see the above problem (18) is equivalent to

$$\begin{aligned} \min_{\omega,b} \quad & \|\omega\| \\ \text{s.t.} \quad & \theta^{pq}[(\omega^{pq})^T x + b^{pq}] \geq 1, x \in I_p, q > p, p, q \in G, \\ & \theta^{qp}[-(\omega^{pq})^T x - b^{pq}] \geq 1, x \in I_q, q > p, p, q \in G. \end{aligned} \tag{19}$$

This problem is equivalent to:

$$\begin{aligned} \min_{\omega,b} \quad & \|\omega\|^2 \\ \text{s.t.} \quad & \theta^{pq}[(\omega^{pq})^T x + b^{pq}] \geq 1, x \in I_p, q > p, p, q \in G, \\ & \theta^{qp}[-(\omega^{pq})^T x - b^{pq}] \geq 1, x \in I_q, q > p, p, q \in G. \end{aligned} \tag{20}$$

As the objective function of (20) is strictly convex, we can see that the optimal solution ω_θ is unique. Besides, considering that the objective of (20) is quadratic (positive definite) and the constraints are affine functions, KKT conditions are necessary and sufficient for optimality. The KKT conditions for (20) are:

$$\begin{aligned} & 2\omega^{pq} = \theta^{pq} \sum_{x \in I_p} \lambda_x^{pq} x - \theta^{qp} \sum_{x \in I_q} \lambda_x^{qp} x, \ q > p, \ p, q \in G, \\ & \theta^{pq} \sum_{x \in I_p} \lambda_x^{pq} - \theta^{qp} \sum_{x \in I_q} \lambda_x^{qp} = 0, q > p, \ p, q \in G \\ & \lambda_x^{pq}[\theta^{pq}(\omega^{pq})^T x + \theta^{pq} b^{pq} - 1] = 0, x \in I_p, \ q > q, \ p, q \in G, \\ & \lambda_x^{qp}[\theta^{qp}(-\omega^{pq})^T x - \theta^{qp} b^{pq} - 1] = 0, x \in I_q, \ q > p, \ p, q \in G, \\ & \lambda_x^{pq} \geq 0, p \neq q, p, q \in G, \forall x \in I, \\ & \theta^{pq}[(\omega^{pq})^T x + b^{pq}] \geq 1, x \in I_p, q > p, p, q \in G, \\ & \theta^{qp}[-(\omega^{pq})^T x - b^{pq}] \geq 1, x \in I_q, q > p, p, q \in G \end{aligned} \tag{21}$$

From these KKT conditions, we can see that $(\lambda^{pq}, \lambda^{qp}) \neq 0, q > p, p, q \in G$. Without loss of generality, we can say that, for each $p, q \in G$ with $q > p$, there exist some $x_{pq} \in I_p$ such that $\lambda_{x_{pq}}^{pq} \neq 0$. Then we get

$$b^{pq} = \frac{1}{\theta^{pq}} - (\omega^{pq})^T x_{pq}, q > p, p, q \in G.$$

So we can see that the set of optimal solutions for (20) is nonempty. Considering the convexity of the objective function, we have that (20) has a unique optimal solution.

(ω_1, b_1) is optimal for (P1). Let λ_1 be the corresponding KKT multiplier vector. Then take:

$$\begin{aligned} & \omega_\theta^{pq} = \frac{\theta^{pq} + \theta^{qp}}{2\theta^{pq}\theta^{qp}} \omega_1^{pq}, \ q > p, \ p, q \in G, \\ & b_\theta^{pq} = \frac{\theta^{qp} - \theta^{pq}}{2\theta^{pq}\theta^{qp}} + \frac{\theta^{qp} + \theta^{pq}}{2\theta^{pq}\theta^{qp}} \times b_1^{pq}, q > p, q, p \in G, \\ & \lambda_{\theta x}^{pq} = \frac{\theta^{pq} + \theta^{qp}}{2\theta^{pq}\theta^{qp}} \times \frac{1}{\theta^{pq}} \lambda_{1x}^{pq}, x \in I_p, \ p \neq q, p, q \in G. \end{aligned} \tag{22}$$

Then $(\omega_\theta, b_\theta)$ will be the unique optimal solution of (20), since it satisfies the KKT conditions. Then, for any $\mu > 0$ we have that $(\mu\omega_\theta, \mu b_\theta)$ is optimal for (17). Using Lemma 3.1, we conclude that $(\mu\omega_\theta, \mu b_\theta)$ is weakly Pareto-optimal for **HPMAT**. □

After characterizing these weakly Pareto-optimal solutions of **HPMAT**, we try to identify its Pareto-optimal solutions. We now show that these weakly Pareto-optimal solutions will also be Pareto-optimal for **HPMAT**.

Corollary 3.3. *The Pareto-optimal solution set of **HPMAT** will be:*

$$\left\{ (\boldsymbol{\omega}, \boldsymbol{b}) = \left(\mu\omega_\theta^{12}, \cdots, \mu\omega_\theta^{(m-1)m}, \mu b_\theta^{12}, \cdots, \mu b_\theta^{(m-1)m} \right) | \mu > 0, \theta^{pq} > 0, p < q, p, q \in G \right\},$$

where $\theta^{12} = 1$, $\omega_\theta^{pq} = \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}} \omega_1^{pq}$ *and* $b_\theta^{pq} = \frac{\theta^{qp}-\theta^{pq}}{2\theta^{pq}\theta^{qp}} + \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}} b_1^{pq}$ *for all* $q > p, p, q \in G$, *with* (w_1, b_1) *being optimal to* $(P1)$.

Proof. From the definitions of Pareto-optimal and weakly Pareto-optimal solutions, we know that the Pareto-optimal solutions will also be weakly Pareto-optimal. So we only need to prove that the weakly Pareto-optimal solutions of **HPMAT** will also be Pareto-optimal.

Let (ω_*, b_*) be a weakly Pareto-optimal solution of **HPMAT**. Then, there exist some $\theta > 0$ and $\mu > 0$ such that $(\mu\omega_*, \mu b_*)$ will be optimal for (20). Suppose (ω_*, b_*) is not Pareto-optimal for **HPMAT**. For any $\mu > 0$ we have $\varrho^{pq}(\omega, b) = \varrho^{pq}(\mu\omega, \mu b)$. So $(\mu\omega_*, \mu b_*), \forall \mu > 0$ will not be Pareto-optimal for **HPMAT**. Then there exist (ω_0, b_0) such that:

$$\varrho^{pq}(\omega_0, b_0) \geq \varrho^{pq}(\mu\omega_*, \mu b_*),\ p \neq q, p, q \in G, \tag{23}$$

and at least one $(i, j), i \neq j, i, j \in G$, such that $\varrho^{ij}(\omega_0, b_0) > \varrho^{ij}(\mu\omega_*, \mu b_*)$.

Without loss of generality, we can take $\|\omega_0\| = \|\mu\omega_*\|$. Then we have:

$$(\omega_0^{pq})^T x + b_0^{pq} \geq (\mu\omega_*^{pq})^T x + \mu b_*^{pq}, x \in I_p, p \neq q, p, q \in G.$$

As $(\mu\omega_*, \mu b_*)$ is optimal for (20), we have that (ω_0, b_0) is also feasible for (20). As $\|\omega_0\| = \|\mu\omega_*\|$, we can say that (ω_0, b_0) is optimal for (20). Since (20) has a unique optimal solution, we must have $\omega_0 = \mu\omega_*,\ b_0 = \mu b_*$. Thus, we have:

$$\varrho^{pq}(\omega_0, b_0) = \varrho^{pq}(\mu\omega_*, \mu b_*),\ \forall p \neq q, p, q \in G.$$

This contradicts our assumption that (23) has at least one strict inequality. We then conclude that (ω_*, b_*) is Pareto-optimal for **HPMAT**. □

3.2 Soft-margin projected multiobjective all-together

In Section 3.1 we have introduced the **HPMAT** problem. As before, in order to consider the overfitting problem and nonlinearly separable training objects, we derive a soft-margin variant for that problem. We need to properly define the geometric margins so that we can characterize the Pareto-optimal solutions for the resulting soft-margin multiobjective problem. We consider the following projection:

$$\Delta_{\xi x}^{pq} = (\delta_{\xi x}^{12}, \delta_{\xi x}^{21}, \cdots, \delta_{\xi x}^{m(m-1)}), q > p, p, q \in G,$$

where

$$\delta_{\xi x}^{ij} = \begin{cases} \frac{1}{c^{pq}} e_i & \text{if } (i, j) = (p, q) \text{ and } x \text{ is the } i\text{-th object in class } p, \\ 0 & \text{if } (i, j) \neq (p, q), i \neq j, i, j \in G, \end{cases}$$

and e_i is the i-th unit vector.

In the projected space we can construct the hyperplane classifying class p objects against class q objects as:

$$L^{pq} : (\omega, c\xi)^T (\Delta_x^{pq}, \Delta_{\xi x}^{pq}) + b^{pq} = 0, q > p, p, q \in G,$$

where, $c\xi = (c^{12}\xi^{12}, c^{21}\xi^{21}, \cdots, c^{m(m-1)}\xi^{m(m-1)})$ and Δ_x^{pq} defined as in Section 3.1.

We define the geometric margin from object x to hyperplane L^{pq} as the Euclidean distance in the projected space.

- The geometric margin from object $x \in I_p$ to hyperplane L^{pq} is:

$$\bar{\varrho}_x^{pq}(\omega, c\xi, b) = \frac{|(\omega, c\xi)^T (\Delta_x^{pq}, \Delta_{\xi x}^{pq}) + b^{pq}|}{\|(\omega, c\xi)\|} = \frac{(\omega^{pq})^T x + \xi^{pq} + b^{pq}}{\|(\omega, c\xi)\|}, x \in I_p, p \neq q, p, q \in G.$$

- The geometric margin for class p objects against class q objects is:

$$\bar{\varrho}^{pq}(\omega, c\xi, b) = \min_{x \in I_p} \bar{\varrho}_x^{pq}(\omega, c\xi, b), p \neq q, p, q \in G.$$

We wish to maximize all the geometric margins defined with the slack variables embedded. We formulate the following multiobjective problem:

$$\begin{aligned} \max_{\omega,b} \quad & \left(\bar{\varrho}^{12}(\omega, b), \bar{\varrho}^{21}(\omega, b), \cdots, \bar{\varrho}^{(m-1)m}(\omega, b), \bar{\varrho}^{m(m-1)}(\omega, b)\right) \\ \text{s.t.} \quad & (\omega^{pq})^T x + b^{pq} + \xi_x^{pq} > 0, x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} + \xi_x^{qp} > 0, x \in I_q, q > p, p, q \in G, \\ & \xi_x^{pq} \geq 0, x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{24}$$

We refer to the above multiobjective optimization problem (24) as **SPMAT** (Soft-margin Projected Multiobjective All-Together). By applying procedures similar to the ones used in Theorem 3.2 and Corollary 3.3, we can characterize the weakly Pareto-optimal and Pareto-optimal solutions for **SPMAT**.

Theorem 3.4. *The set of weakly Pareto-optimal solutions for **SPMAT** is :*

$$\left\{(\boldsymbol{\omega}, \boldsymbol{b}) = \left(\mu\omega_\theta^{12}, \cdots, \mu\omega_\theta^{(m-1)m}, \mu b_\theta^{12}, \cdots, , \mu b_\theta^{(m-1)m}\right) | \mu > 0, \theta^{pq} > 0, p < q, p, q \in G\right\},$$

where $\theta^{12} = 1$, $\omega_\theta^{pq} = \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}}\omega_1^{pq}$ *and* $b_\theta^{pq} = \frac{\theta^{qp}-\theta^{pq}}{2\theta^{pq}\theta^{qp}} + \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}} b_1^{pq}$ *for all* $q > p, p, q \in G$, *with* (ω_1, b_1) *being optimal to* (1).

The proof is similar to the proof of Theorem 3.2. The details can be found in Appendix 2.

Corollary 3.5. *The Pareto-optimal solution set of **SPMAT** will be:*

$$\left\{(\boldsymbol{\omega}, \boldsymbol{b}) = \left(\mu\omega_\theta^{12}, \cdots, \mu\omega_\theta^{(m-1)m}, \mu b_\theta^{12}, \cdots, , \mu b_\theta^{(m-1)m}\right) | \mu > 0, \theta^{pq} > 0, p < q, p, q \in G\right\},$$

where $\theta^{12} = 1$, $\omega_\theta^{pq} = \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}}\omega_1^{pq}$ *and* $b_\theta^{pq} = \frac{\theta^{qp}-\theta^{pq}}{2\theta^{pq}\theta^{qp}} + \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}} b_1^{pq}$ *for all* $q > p, p, q \in G$, *with* (ω_1, b_1) *being optimal to* (1).

The proof of this result is identical to the proof for Corollary 3.3.

4 Multiobjective approaches for many classes

In Sections 2 and 3 we have introduced the multiobjective approaches for multiclass classification problems. We have also argued that the soft-margin multiobjective versions are more suitable for their application to general data. For **SMAT**, **SMOAA** and **SMOAO**, we use an $\varepsilon-$constraint method to get their weakly Pareto-optimal solutions. In practical cases we have many constraints and a large number of objectives, since m is large, This implies that we need to fix $m(m-1)-1$ values corresponding to $\varepsilon^{pq}, (p, q) \neq (r, s), p \neq q, p, q \in G$. For all the soft-margin multiobjective problems, we need to find $m(m-1)$ values for $c^{pq}, p \neq q, p, q \in G$. Besides, considering too many classes at one time may lead us to low classification accuracies. So we have chosen an approach based on dividing the classes into groups. In this way, at each node we will just need to solve a classification problem for a small number of classes (or groups).

Researchers have already used efficient tree-based methods for multiclassification problems, such as [4, 5, 17]. In these papers, the researchers have focused on binary-tree methods. This will also be problematic when m is very large, because many SVMs will need to be solved, requiring a large computational effort. With the goal of generating an efficient tree structure, our choice is a multidecision tree that uses multiobjective multiclass SVMs in each node. Consider the following tree structure as an illustration:

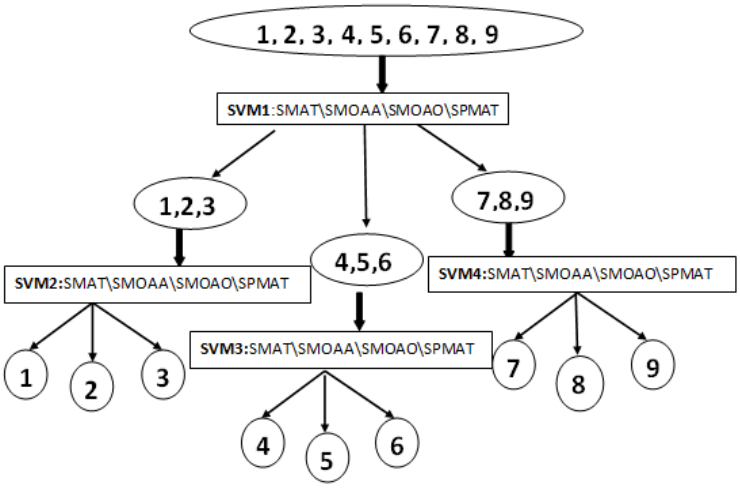

Figure 2. The tree proceeding for classification of many classes

With different choices of class divisions, we will have different numbers of layers and of SVMs to be solved. The dividing method will also affect the classification accuracies. To divide classes into two groups, we can use a simple method: first select a linear combination of the attributes, and then for the projected values get the mean for all the data in all classes. Then collect the classes which have a mean smaller than the total mean in one group and the rest of the classes in the other group. Of course, we can also use the median or quartiles to divide the classes into different groups. How to select this linear combination, and in general how to divide the classes in an optimal manner is still an open problem that we plan to study in the future.

5 Computation experiment

For multiobjective SVMs we aim to approximate the Pareto-optimal solution set. For **SMAT**, **SMOAA** and **SMOAO**, we can obtain the weakly Pareto-optimal solutions by using a $\varepsilon-$constraint method. Still, even though we can characterize the Pareto-optimal solution set for **SPMAT**, it is still computationally hard to get all the Pareto-optimal solutions, because we may have an infinite number of Pareto-optimal solutions.

We will compare different methods with respect to the quality of their approximations for the Pareto-optimal set; we will say that a method outperforms another when it approximates this set better than the other. Test accuracies are also important, since we want to construct a decision function to predict the class membership of new objects. We use the solutions generated from the soft-margin multiobjective SVMs corresponding to **SMAT**, **SMOAA**, **SMOAO** and **SPMAT** to obtain the corresponding test accuracies, and to compare them. In this paper, we use the epsilon and hypervolume indicators, together with the test accuracies as objectives to measure the performance of these multiobjective SVMs. Following [13], the

hypervolume indicator $I_H(A)$ measures the hypervolume of that portion of the objective space that is weakly dominated by an approximating set of Pareto-optimal solutions A. And the epsilon indicator is defined as $I_{\epsilon+} = \inf_{\epsilon\in\mathbb{R}}\{\forall z^2 \in R, \exists z^1 \in A \text{ such that } z^1 \preceq_{\epsilon+} z^2\}$, where R is an reference set. These indicators are Pareto compliant.

We have used the following data sets: IRIS, WINE, SEEDS, CAR (Car Evaluation), SCC (Synthetic Control Chart Time Series) and CTG (Cardiotocography). All of them are available in the UCI Machine Learning Repository. A summary of the information of these data sets is listed in the following table:

Table 1. Data set description

Data set	size of the data set	No. of Dim.	No. of classes
IRIS	150	4	3
WINE	178	13	3
SEEDS	210	7	3
CAR	1728	16	4
SCC	600	60	6
CTG	2126	35	10

For the epsilon indicator, we select the reference set R as $\{(1, 1, \ldots, 1, 1)\} \subset \mathbb{R}^m$, which is the ideal test accuracy. And for the hypervolume indicator, we take the reference point as $(0, 0, \ldots, 0, 0) \in \mathbb{R}^m$. Based on the definition of the hypervolume indicator in [13], the method which has the largest hypervolume indicator (closest to 1 among all the methods) outperforms the others. Similarly, if one method has an epsilon indicator value that is closer to 0 than that for the other methods, we indicate that this method outperforms the others.

To explore the performance of **SMAT**, **SMOAA**, **SMOAO** and **SPMAT** with these experimental data sets, we get 50 approximate test accuracy sets for each of these methods. Then we generate 50 indicator values for each of the methods. For **SMAT**, **SMOAA** and **SMOAO**, to get the first indicator value we complete the following steps:

- Step 1: We arrange the objects in a random order. We then choose the last 20% objects as test objects and leave the rest as training objects.

- Step 2: We use a $\varepsilon-$constraint method to solve **SMAT**, **SMOAA** and **SMOAO** as described in (13) for **SMAT**, and their analogous for **SMOAA** and **SMOAO**. But before solving these SOCPs, we need to fix the values of (r, s), $(c^{12}, c^{21}, \cdots, c^{(m-1)m}, c^{m(m-1)})$ and $\varepsilon^{pq}, (p, q) \neq (r, s), p \neq q, p, q \in G$.

  For $(c^{12}, c^{21}, \cdots, c^{(m-1)m}, c^{m(m-1)})$ we apply a 10-fold cross-validation method to problem (1). The $\varepsilon-$constraint method will give us a solution x that is at least weakly Pareto-optimal, [12]. And x is Pareto-optimal if and only if there exists a ε_* such that x is an optimal solution of the SOCP (13) for all $(r, s), r, s \in G$. So we can fix $(r, s) = (1, 2)$. Then we choose $\varepsilon^{21}, \varepsilon^{13}, \cdots, \varepsilon^{m(m-1)}$ as uniform random values from $(0, r^{pq})$. Here r^{pq} is an upper bound for ε^{pq}. It should be chosen properly to ensure that the corresponding single objective problem (13) is feasible. After solving these SOCPs, we get a solution which is at least weakly Pareto-optimal for each of these three methods.

- Step 3: For each of the three methods, and the weakly Pareto-optimal solutions obtained in Step 2, we use (2) to get the corresponding test classification accuracy vector $(a^1, a^2, \cdots, a^m)$. Here a^p is the test classification accuracy for class p.

- Step 4: We repeat Steps 2 and 3 a large enough number of times (in this paper, we repeat the steps 100 times) to get a set of test classification accuracy vectors for each of these three methods. We use Matlab and Mosek to solve the SOCPs. In many cases we obtain the optimal solutions, but sometimes it will give us near optimal solutions and sometimes it will fail to compute a solution. We only keep the optimal solutions, and we may end up with a set of test classification accuracy vectors having less than 100 vectors.

- Step 5: For each of these three methods, with the set of test accuracy vectors that we get in Step 4, we calculate the corresponding indicator values for the given reference set and reference point.

Note that in order to obtain an indicator for **SMAT**, **SMOAA** and **SMOAO**, we need a set of test accuracy vectors, requiring the solution of a large number of SVMs, and a large computational effort. We repeat the preceding five steps 50 times to get the 50 hypervolume and epsilon indicators.

From the theoretical results of Corollary 3.5, we know that using **SPMAT** we get Pareto-optimal solutions. As before, to use the epsilon and hypervolume indicators to evaluate the performance of **SPMAT**, we apply the following method to get the first epsilon and hypervolume indicators:

- Step 1: Using the same training and test objects and values of $(c^{12}, c^{21}, \cdots, c^{(m-1)m}, c^{m(m-1)})$ that we chose for **SMAT**, **SMOAA** and **SMOAO**, we solve (1) to get (ω_1, b_1).

- Step 2: From Lemma 3.1, we have $\theta^{pq} = \frac{\varrho_*^{12}}{\varrho_*^{pq}}$. We generate uniform random values z^{pq} for all $p \neq q, p, q \in G$, from $(0, 1)$, and we let $\theta^{pq} = \frac{z^{12}}{z^{pq}}$. By using Corollary 3.5 with (ω_1, b_1), we obtain a Pareto-optimal solution of **SPMAT**.

- Step 3: With the solution from Step 2, we use (2) to get a test accuracy vector.

- Step 4: Repeat Step 2 and Step 3 100 times (or 10000 times) to get a set of test accuracy vectors which contains exactly 100 vectors (or 10000 vectors).

- Step 5: With the given reference set and reference point, we calculate the epsilon and hypervolume indicators for **SPMAT**.

Notice that to obtain an indicator for **SPMAT** we only need to solve one single objective SVM (1). This saves a lot of computational cost. Besides, we can get a larger approximating set which has exactly 100 (or 10000) test accuracy vectors. This process is repeated for each test set, as done for the other methods, obtaining 50 hypervolume and epsilon indicators.

For IRIS, WINE and SEEDS, as they have only three classes, following the above steps, we get the results presented in the following figures (Figure 3 to Figure 8) and tables (Table 2 to Table 7). As in Section 4, for classification problems with many classes, we suggest to use multidecision trees that use multiobjective SVMs in each node. For example, in this paper, for CAR,SCC and CTG, we combine a tree method and multiobjective SVMs to get the epsilon and hypervolume indicators. The corresponding results can be seen in Figure 9 to Figure 14 and Table 8 to Table 13. In Appendix 3, we show the partitioning ways that we have used for these data sets.

In the following tables and figures, **SPMAT1** denotes the indicators calculated from a set of 100 test accuracy vectors from **SPMAT**, and **SPMAT2** denotes the indicators

calculated from a set of 10000 test accuracy vectors from **SPMAT**. Each of the following figures contains five boxplots of indicators gotten by **SMAT**, **SMOAA**, **SMOAO**, **SPMAT1** and **SPMAT2** separately. Each of the following tables contains ten columns. The first colum lists the methods that we have used. The following seven columns (column 2 to column 8) show the mean values, variances, minimums, 25 percentiles, medians, 75 percentiles and the maximums of the corresponding indicators. In the ninth column, 'set size' refers to the average approximate set size for each of these soft-margin multiobjective approaches. And in the last colum, 'time' refers to the average time for getting a hypervolume and epsilon indicator with respect to each of these soft-margin multiobjective approaches.

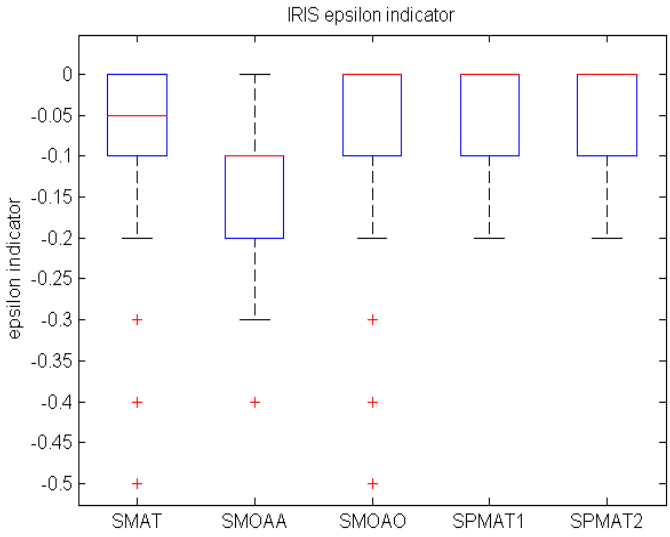

Figure 3. The epsilon indicators for IRIS data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	-0.092	0.0175	-0.5	-0.1	-0.05	0	0	79.3	38.65
SMOAA	-0.124	0.0108	-0.4	-0.2	-0.1	-0.1	0	64.3	40.74
SMOAO	-0.09	0.0177	-0.5	-0.1	0	0	0	79.16	40.75
SPMAT1	-0.036	0.0028	-0.2	-0.1	0	0	0	100	0.47
SPMAT2	-0.03	0.0026	-0.2	-0.1	0	0	0	10000	3.49

Table 2. Epsilon indicator statistic information for IRIS data

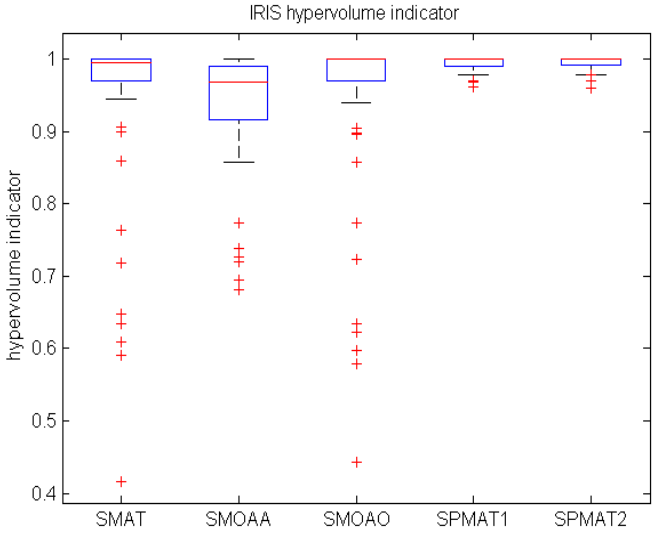

Figure 4. The hypervolume indicators for IRIS data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	0.9338	0.0182	0.4161	0.9705	0.9955	1	1	79.3	38.65
SMOAA	0.9350	0.0077	0.6806	0.9161	0.9688	0.9899	1	64.3	40.74
SMOAO	0.9337	0.0182	0.4424	0.9700	1	1	1	79.16	40.75
SPMAT1	0.9943	0.0001	0.9609	0.9903	1	1	1	100	0.47
SPMAT2	0.9954	0.0001	0.9603	0.9913	1	1	1	10000	3.49

Table 3. Hypervolume indicator statistic information for IRIS data

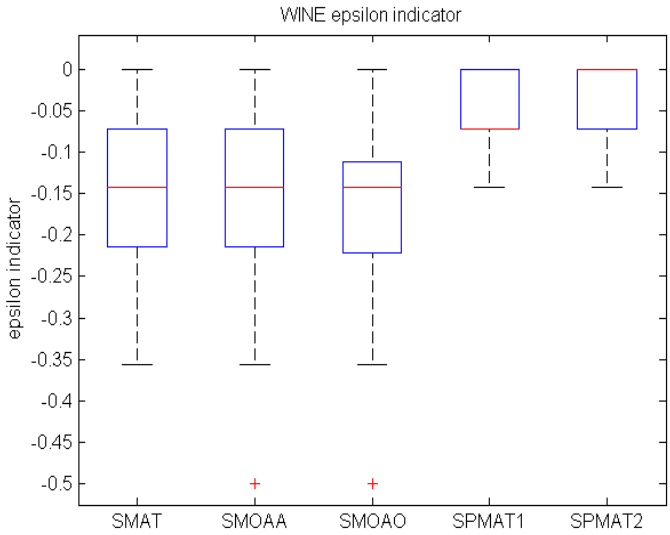

Figure 5. The epsilon indicators for WINE data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	-0.1397	0.0072	-0.3571	-0.2143	-0.1429	-0.0714	0	53.9	28.69
SMOAA	-0.1566	0.0095	-0.5	-0.2143	-0.1429	-0.0714	0	60.74	31.05
SMOAO	-0.1747	0.0104	-0.5	-0.222	-0.1429	-0.1111	0	68.12	30.21
SPMAT1	-0.048	0.0025	-0.1429	-0.0714	-0.0714	0	0	100	0.33
SPMAT2	-0.039	0.0024	-0.1429	-0.0714	0	0	0	10000	3.69

Table 4. Epsilon indicator statistic information for WINE data

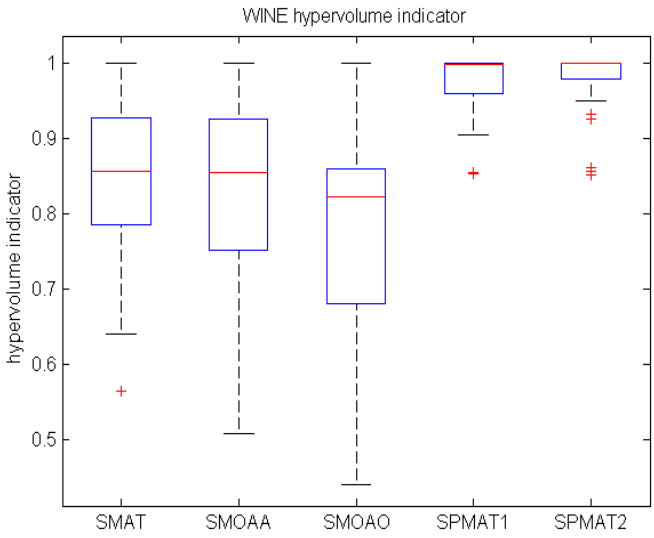

Figure 6. The hypervolume indicators for WINE data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	0.8502	0.0094	0.5648	0.7848	0.8571	0.9284	1	53.9	28.69
SMOAA	0.8163	0.0145	0.5073	0.7523	0.8549	0.9252	1	60.74	31.05
SMOAO	0.7788	0.0165	0.4404	0.6814	0.8233	0.8596	1	68.12	30.21
SPMAT1	0.9742	0.0017	0.8539	0.9604	0.9993	1	1	100	0.33
SPMAT2	0.9802	0.0014	0.8515	0.9797	1	1	1	10000	3.69

Table 5. Hypervolume indicator statistic information for WINE data

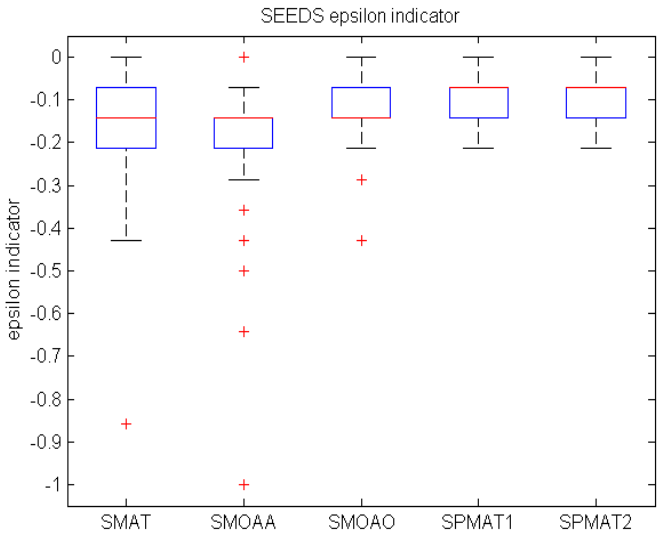

Figure 7. The epsilon indicators for SEEDS data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	-0.14	0.0187	-0.8571	-0.2143	-0.1429	-0.0714	0	93.32	45.79
SMOAA	-0.1914	0.0272	-1	-0.2143	-0.1429	-0.1429	0	83.54	48.61
SMOAO	-0.1314	0.0088	-0.4286	-0.1429	-0.1429	-0.0714	0	84.92	48.71
SPMAT1	-0.0929	0.0017	-0.2143	-0.0714	-0.0714	-0.0714	0	100	0.52
SPMAT2	-0.0829	0.0022	-0.2143	-0.0714	-0.0714	-0.0714	0	10000	3.79

Table 6. Epsilon indicator statistic information for SEEDS data

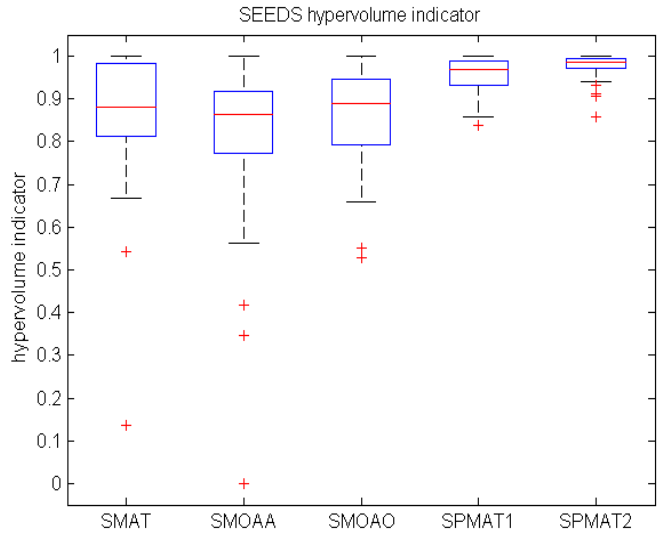

Figure 8. The hypervolume indicators for SEEDS data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	0.8625	0.022	0.1352	0.8114	0.8813	0.9844	1	93.32	45.79
SMOAA	0.8135	0.033	0	0.7741	0.8650	0.9177	1	83.54	48.61
SMOAO	0.8699	0.0122	0.5291	0.7930	0.8898	0.9455	1	84.92	48.71
SPMAT1	0.9577	0.0015	0.8390	0.9325	0.9684	0.9888	1	100	0.52
SPMAT2	0.9763	0.0007	0.8595	0.9709	0.9847	0.9933	1	10000	3.79

Table 7. Hypervolume indicator statistic information for SEEDS data

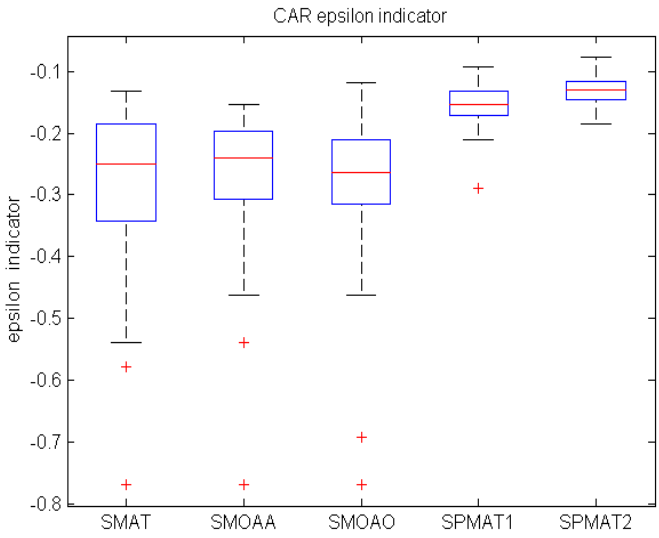

Figure 9. The epsilon indicators for CAR data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	-0.285	0.0181	-0.7692	-0.3421	-0.25	-0.1842	-0.1316	35.34	65.7
SMOAA	-0.2717	0.0141	-0.7692	-0.3077	-0.2404	-0.1974	-0.1538	30.4	60.22
SMOAO	-0.2894	0.0181	-0.7692	-0.3158	-0.2632	-0.2105	-0.1184	31.56	60.98
SPMAT1	-0.154	0.0009	-0.2895	-0.1711	-0.1538	-0.1322	-0.0921	100	1
SPMAT2	-0.1283	0.0005	-0.1842	-0.1447	-0.1298	-0.1157	-0.0769	10000	3.6

Table 8. Epsilon indicator statistic information for CAR data

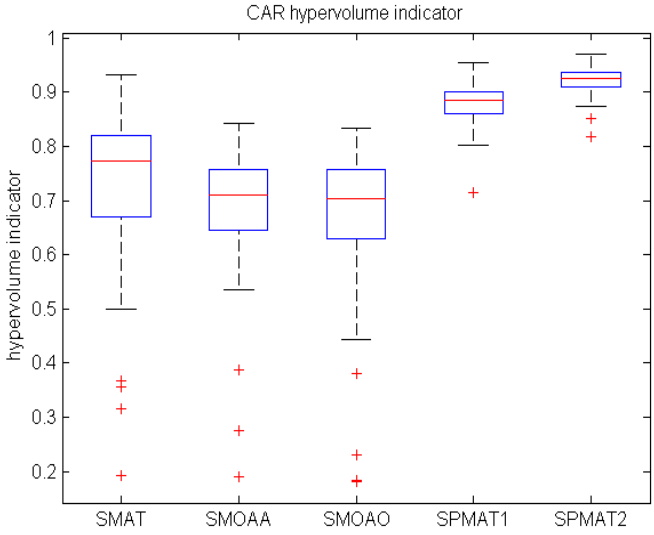

Figure 10. The hypervolume indicators for CAR data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	0.7259	0.0243	0.1932	0.6708	0.7719	0.8206	0.9322	35.34	65.7
SMOAA	0.6801	0.0166	0.1903	0.6443	0.7109	0.7572	0.8426	30.4	60.22
SMOAO	0.6632	0.0237	0.1817	0.6302	0.7042	0.7574	0.8339	31.56	60.98
SPMAT1	0.8770	0.0014	0.7156	0.8604	0.8845	0.8997	0.9543	100	1
SPMAT2	0.9212	0.0007	0.8168	0.9105	0.9256	0.9363	0.9698	10000	3.6

Table 9. Hypervolume indicator statistic information for CAR data

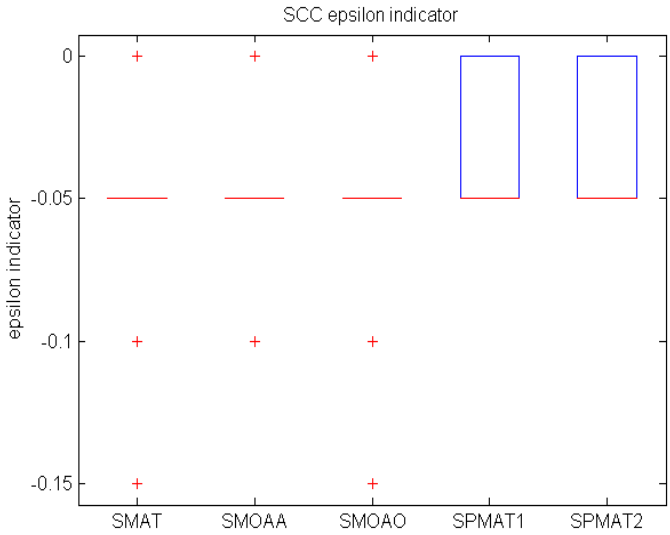

Figure 11. The epsilon indicators for SCC data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	-0.057	0.0009	-0.15	-0.05	-0.05	-0.05	0	91.31	411.2
SMOAA	-0.054	0.0007	-0.1	-0.05	-0.05	-0.05	0	32.27	407
SMOAO	-0.056	0.00088	-0.15	-0.05	-0.05	-0.05	0	36.71	296.6
SPMAT1	-0.037	0.0005	-0.05	-0.05	-0.05	0	0	100	2.83
SPMAT2	-0.029	0.0006	-0.05	-0.05	-0.05	0	0	10000	4.63

Table 10. Epsilon indicator statistic information for SCC data

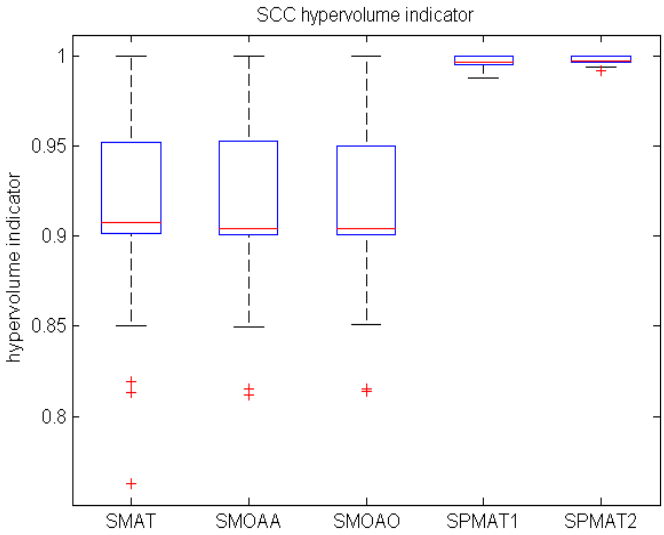

Figure 12. The hypervolume indicators for SCC data set

Method	mean	variance	min	25%	median	75%	max	set size	time(s)
SMAT	0.9151	0.0026	0.7627	0.9015	0.9074	0.9525	1	91.31	411.2
SMOAA	0.9179	0.022	0.8121	0.9012	0.9040	0.9526	1	32.27	407
SMOAO	0.9161	0.0023	0.8141	0.9010	0.9041	0.9503	1	36.71	296.6
SPMAT1	0.9968	0.0000	0.9882	0.9951	0.9965	1	1	100	2.83
SPMAT2	0.9977	0.0000	0.9921	0.9970	0.9975	1	1	10000	4.63

Table 11. Hypervolume indicator statistic information for SCC data

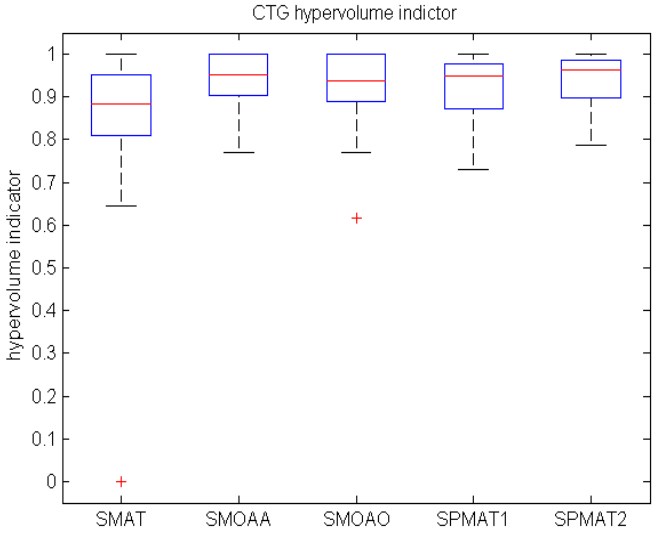

Figure 13. The epsilon indicators for CTG data set

Method	mean	varaince	min	25%	median	75%	max	set size	time(s)
SMAT	-0.1473	0.0623	-1	-0.1	-0.0885	-0.0476	0	38.8	315.6
SMOAA	-0.0535	0.0023	-0.1429	-0.1	-0.0476	0	0	17.3	290.8
SMOAO	-0.0547	0.0022	-0.1429	-0.1	-0.0476	0	0	12.2	130.0
SPMAT1	-0.0783	0.0024	-0.25	-0.1	-0.0714	-0.0513	-0.0087	100	5.06
SPMAT2	-0.0702	0.0021	-0.2143	-0.1	-0.0625	-0.0455	-0.0087	10000	10.66

Table 12. Epsilon indicator statistic information for CTG data

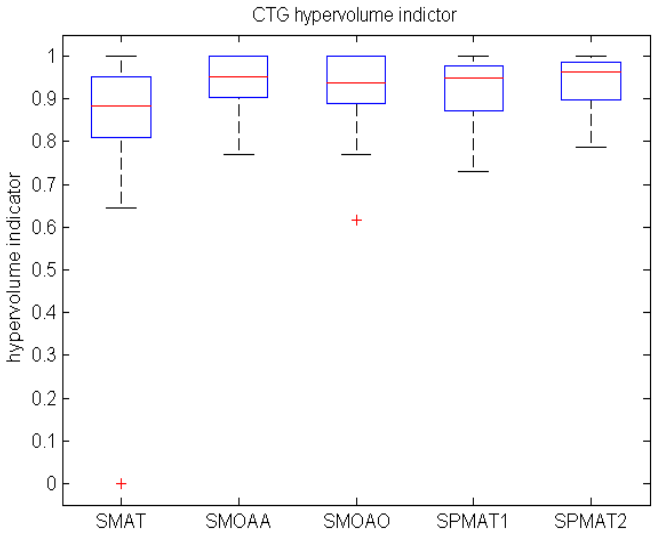

Figure 14. The hypervolume indicators for CTG data set

Method	mean	varaince	min	25%	median	75%	max	set size	time(s)
SMAT	0.8131	0.0654	0	0.8099	0.8834	0.9533	1	38.8	315.6
SMOAA	0.9357	0.0041	0.7700	0.9034	0.9527	1		17.3	290.8
SMOAO	0.9281	0.0061	0.6176	0.8889	0.9380	1	1	12.2	130.0
SPMAT1	0.9230	0.0049	0.7307	0.8727	0.9497	0.9765	0.9989	100	5.06
SPMAT2	0.9392	0.0036	0.7880	0.8974	0.9627	0.9857	0.9998	10000	10.66

Table 13. Hypervolume indicator statistic information for CTG data

6 Conclusion

The results in Section 5 show that **SPMAT** outperforms the other methods in most cases. For IRIS, WINE, SEEDS, CAR and SCC, we can see **SPMAT** has the largest mean indicators and the largest minimal indicators (hypervolume indicators and epsilon indicators), compared with **SMAT**, **SMOAA** and **SMOAO**. For CTG data set, **SPMAT** shows comparable performance with respect to the other three multiobjective approaches considered in this paper.

The experimental results with the IRIS, SEEDS, WINE, CAR, SCC and CTG data sets show that **SPMAT** is efficient. Besides, as we mentioned, **SPMAT** is able to provide exact Pareto-optimal solutions while the other methods only give us weakly Pareto-optimal solutions. Finally, using **SPMAT** the computational costs to get the approximating set of Pareto-optimal solutions will be much lower than those required by **SMAT**, **SMOAA** and **SMOAO**.

The experimental results also suggest that combining a tree method with multiobjective SVMs may be efficient. However, how to divide the classes optimally for general data remains an open problem.

Appendix 1
Proof of Lemma 3.1

First, assume that (ω^*, b^*) is optimal for (16). Notice that the feasible region of (16) and the feasible region of **HPMAT** are the same.

If (ω^*, b^*) is not weakly Pareto-optimal for **HPMAT**, there will exist a feasible (ω_0, b_0) such that

$$\varrho^{12}(\omega_0, b_0) > \varrho^{12}(\omega^*, b^*), \varrho^{21}(\omega_0, b_0) > \varrho^{21}(\omega^*, b^*) \cdots, \varrho^{m(m-1)}(\omega_0, b_0) > \varrho^{m(m-1)}(\omega^*, b^*).$$

As $\theta^{pq} > 0$, we have:

$$\begin{aligned}&\varrho^{12}(\omega_0, b_0) > \varrho^{12}(\omega^*, b^*), \theta^{21}\varrho^{21}(\omega_0, b_0) > \theta^{21}\varrho^{21}(\omega^*, b^*), \cdots, \\ &\theta^{m(m-1)}\varrho^{m(m-1)}(\omega_0, b_0) > \theta^{m(m-1)}\varrho^{m(m-1)}(\omega^*, b^*).\end{aligned}$$

This contradicts our assumption that (ω^*, b^*) is optimal for (16).

As a consequence, (ω^*, b^*) must be a weakly Pareto-optimal solution of **HPMAT**. Then, for any feasible (ω_0, b_0), there exists some $i \neq j, i, j \in G$ such that $\varrho^{ij}(\omega_0, b_0) \leq \varrho^{ij}(\omega^*, b^*)$. Let

$$\varrho_* = \max\left(\varrho_*^{12}, \varrho_*^{21}, \cdots, \varrho_*^{(m-1)m}, \varrho_*^{m(m-1)}\right),$$

where $\varrho_*^{pq} = \varrho^{pq}(\omega^*, b^*), p \neq q, p, q \in G$.

Formulate the following problem:

$$\begin{aligned}\max_{\omega,b} \min &\left(\frac{\varrho_*}{\varrho_*^{12}}\varrho^{12}(\omega, b), \frac{\varrho_*}{\varrho_*^{21}}\varrho^{21}(\omega, b), \cdots, \frac{\varrho_*}{\varrho_*^{m(m-1)}}\varrho^{m(m-1)}(\omega, b),\right), \\ \text{s.t.} \quad &(\omega^{pq})^T x + b^{pq} > 0, \quad x \in I_p, p < q, p, q \in G, \\ &-(\omega^{pq})^T x - b^{pq} > 0, \quad x \in I_q, p < q, p, q \in G.\end{aligned} \tag{25}$$

It is easy to see that (ω^*, b^*) is optimal for (25). By diving all the objectives in (25) by $\frac{\varrho_*}{\varrho_*^{12}}$, we get the equivalent optimization problem:

$$\begin{aligned}\max_{\omega,b} \min &\left(\varrho^{12}(\omega, b), \frac{\varrho_*^{12}}{\varrho_*^{21}}\varrho^{21}(\omega, b), \cdots, \frac{\varrho_*^{12}}{\varrho_*^{m(m-1)}}\varrho^{m(m-1)}(\omega, b),\right), \\ \text{s.t.} \quad &(\omega^{pq})^T x + b^{pq} > 0, \quad x \in I_p, p < q, p, q \in G, \\ &-(\omega^{pq})^T x - b^{pq} > 0, \quad x \in I_q, p < q, p, q \in G.\end{aligned} \tag{26}$$

Thus, (ω^*, b^*) is also optimal for (26).

Appendix 2

Proof of Theorem 3.4

As before, the weakly Pareto-optimal solution of **SPMAT** can be found by solving the following problem:

$$\begin{aligned} \max_{\omega,b,\xi} \min \quad & \Big(\bar{\varrho}^{12}(\omega,b), \theta^{21}\bar{\varrho}^{21}(\omega,b), \cdots, \theta^{(m-1)m}\bar{\varrho}^{(m-1)m}(\omega,b), \theta^{m(m-1)}\bar{\varrho}^{m(m-1)}(\omega,b)\Big) \\ \text{s.t.} \quad & (\omega^{pq})^T x + b^{pq} + \xi^{pq}(x) > 0, \quad x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} + \xi^{qp}(x) > 0, \quad x \in I_p, q > p, p, q \in G, \\ & \xi^{pq}(x) \geq 0, \quad x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{27}$$

Problem (27) is equivalent to

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \frac{\|(\omega, c\xi)\|}{\min\{\min_{x\in I_1}(\omega^{12})^T x + b^{12} + \xi^{12}(x), \cdots, \theta^{m(m-1)} \min_{x\in I_m}(\omega^{m(m-1)})^T x + b^{m(m-1)} + \xi^{m(m-1)}(x)\}} \\ \text{s.t.} \quad & (\omega^{pq})^T x + b^{pq} + \xi^{pq}(x) > 0, \quad x \in I_p, q > p, p, q \in G, \\ & -(\omega^{pq})^T x - b^{pq} + \xi^{qp}(x) > 0, \quad x \in I_p, q > p, p, q \in G, \\ & \xi^{pq}(x) \geq 0, \quad x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{28}$$

By introducing a condition to bound away from zero the denominator of the objective function, we obtain the equivalent problem

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \|(\omega, c\xi)\|, \\ \text{s.t.} \quad & \theta^{pq}((\omega^{pq})^T x + b^{pq} + \xi^{pq}(x)) \geq 1, \quad x \in I_p, q > p, p, q \in G, \\ & \theta^{pq}(-(\omega^{pq})^T x - b^{pq} + \xi^{qp}(x)) \geq 1, \quad x \in I_q, q > p, p, q \in G, \\ & \xi^{pq}(x) \geq 0, \quad x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{29}$$

Problem (29) is also equivalent to

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \|(\omega, c\xi)\|^2, \\ \text{s.t.} \quad & \theta^{pq}((\omega^{pq})^T x + b^{pq} + \xi^{pq}(x)) \geq 1, \quad x \in I_p, q > p, p, q \in G, \\ & \theta^{pq}(-(\omega^{pq})^T x - b^{pq} + \xi^{qp}(x)) \geq 1, \quad x \in I_q, q > p, p, q \in G, \\ & \xi^{pq}(x) \geq 0, \quad x \in I_p, p \neq q, p, q \in G. \end{aligned} \tag{30}$$

From the strict convexity of the objective function of (30) its optimal solution (ω_*, ξ_*) is unique. As the constraints are affine functions and the objective is quadratic (and positive definite), the KKT conditions are necessary and sufficient for optimality.

These KKT conditions are:

$$
\begin{aligned}
&2\omega^{pq} = \theta^{pq}\sum_{x\in I_p}\lambda_x^{pq}x - \theta^{qp}\sum_{x\in I_q}\lambda_x^{qp}x,\ q>p,\ p,q\in G,\\
&\sum_{x\in I_p}\theta^{pq}\lambda_x^{pq} - \theta^{qp}\sum_{x\in I_q}\lambda_x^{qp} = 0, \quad q>p,\ p,q\in G\\
&2c^{pq}\xi^{pq}(x) = \theta^{pq}\lambda_x^{pq}, \quad x\in I_p,\ p\neq q,\ p,q\in G,\\
&\lambda_x^{pq}[\theta^{pq}(\omega^{pq})^T x + \theta^{pq}b^{pq} + \theta^{pq}\xi^{pq}(x) - 1] = 0, \quad x\in I_p,\ q>p,\ p,q\in G,\\
&\lambda_x^{qp}[-\theta^{qp}(\omega^{pq})^T x - \theta^{qp}b^{pq} + \theta^{qp}\xi^{qp}(x) - 1] = 0, \quad x\in I_q,\ q>p,\ p,q\in G,\\
&\lambda_x^{pq}\geq 0, \quad p\neq q, p,q\in G, \forall x\in I_p,\\
&\theta^{pq}[(\omega^{pq})^T x + b^{pq} + \xi^{pq}(x)]\geq 1, \quad x\in I_p, q>p, p,q\in G,\\
&\theta^{qp}[-(\omega^{pq})^T x - b^{pq} + \xi^{qp}(x)]\geq 1, \quad x\in I_q, q>p, p,q\in G.
\end{aligned}
\tag{31}
$$

From these conditions we can see that $(\lambda^{pq},\lambda^{qp})\neq 0, q>p,\ p,q\in G$. Then, there exists some $x_{pq}\in I_p($ without loss of generality), such that

$$
b^{pq} = \frac{1}{\theta^{pq}} - (\omega^{pq})^T x_{pq} - \xi^{pq}(x_{pq}), \quad q>p, p,q\in G.
$$

From this characterization, the set of optimal solutions for (30) is nonempty. From the convexity of the objective function, we have that (30) has a unique optimal solution. When $\theta = (1,1,\cdots,1,1)$, we have (30) $\Longleftrightarrow$ (1).

Suppose (ω_1, b_1) is optimal for (1) and λ_1 are the corresponding KKT multipliers. Then let

$$
\begin{aligned}
\omega_\theta^{pq} &= \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}}\omega_1^{pq}, \quad q>p,\ p,q\in G\\
b_\theta^{pq} &= \frac{\theta^{qp}-\theta^{pq}}{2\theta^{pq}\theta^{qp}} + \frac{\theta^{qp}+\theta^{pq}}{2\theta^{pq}\theta^{qp}}b_1^{pq}, \quad q>p,q,p\in G,\\
\xi_{\theta x}^{pq} &= \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}}\xi_1^{pq}(x), \quad p\neq q,\ p,q\in G,\\
\lambda_{\theta x}^{pq} &= \frac{\theta^{pq}+\theta^{qp}}{2\theta^{pq}\theta^{qp}}\frac{1}{\theta^{pq}}\lambda_{1x}^{pq}, \quad x\in I_p,\ p\neq q, p,q\in G.
\end{aligned}
\tag{32}
$$

These values $(\omega_\theta, b_\theta, \xi_\theta)$ are the unique optimal solution of (30), since they satisfy the KKT conditions (31).

Appendix 3

The trees used for CAR, SCC and CTG data sets

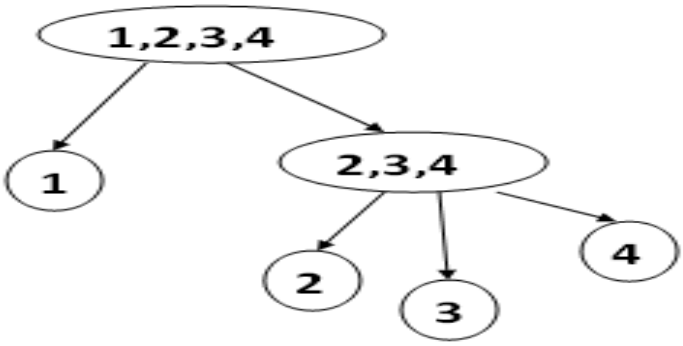

Figure 15. The dividing for CAR data set

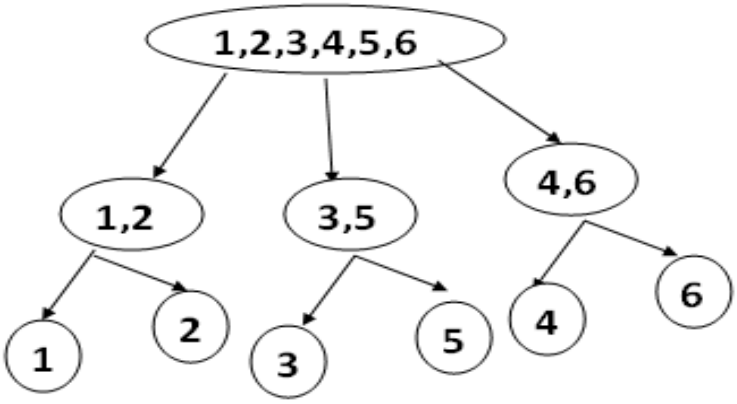

Figure 16. The dividing for SCC data set

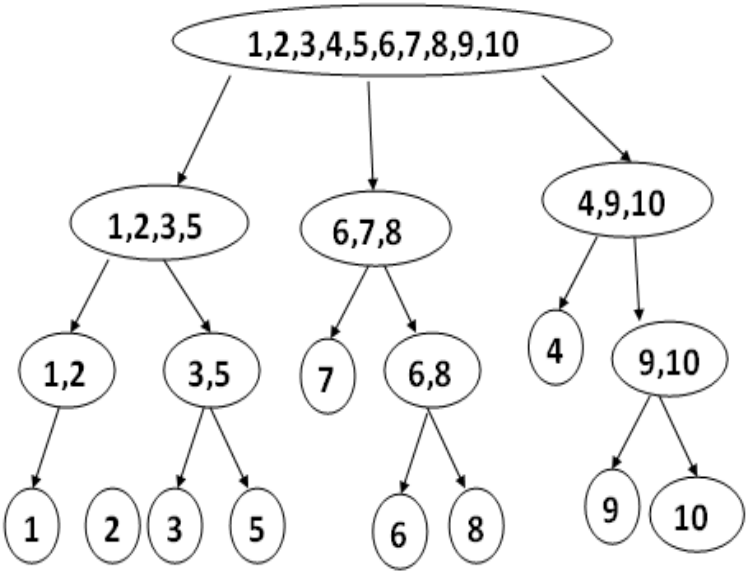

Figure 17. The dividing for CTG data set

References

[1] E.J. Bredensteiner and K.P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1):53–79, 1999.

[2] Michael PS Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S Furey, Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

[3] E. Carrizosa and B. Martin-Barragan. Two-group classification via a biobjective margin maximization model. *European Journal of Operational Research*, 173(3):746–761, 2006.

[4] Jin Chen, Cheng Wang, and Runsheng Wang. Combining support vector machines with a pairwise decision tree. *Geoscience and Remote Sensing Letters, IEEE*, 5(3):409–413, 2008.

[5] Jin Chen, Cheng Wang, and Runsheng Wang. Adaptive binary tree for fast svm multiclass classification. *Neurocomputing*, 72(13):3370–3375, 2009.

[6] Sungmoon Cheong, Sang Hoon Oh, and Soo-Young Lee. Support vector machines with binary tree architecture for multi-class classification. *Neural Information Processing-Letters and Reviews*, 2(3):47–51, 2004.

[7] KK Chin. Support vector machines applied to speech pattern classification. *Mphil. In Computer Speech and Language Processing, Cambridge University Engineering Department*, 1999.

[8] Altannar Chinchuluun and Panos M Pardalos. A survey of recent developments in multiobjective optimization. *Annals of Operations Research*, 154(1):29–50, 2007.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[10] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[11] Kalyanmoy Deb. Multi-objective optimization. *Multi-objective optimization using evolutionary algorithms*, pages 13–46, 2001.

[12] M. Ehrgott. *Multicriteria optimization*, volume 491. Springer Verlag, 2005.

[13] Carlos M Fonseca, Joshua D Knowles, Lothar Thiele, and Eckart Zitzler. A tutorial on the performance assessment of stochastic multiobjective optimizers. In *Third International Conference on Evolutionary Multi-Criterion Optimization (EMO 2005)*, volume 216, page 240, 2005.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[15] C.W. Hsu and C.J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[16] Ulrich H-G Kreßel. Pairwise classification and support vector machines. In *Advances in kernel methods*, pages 255–268. MIT Press, 1999.

[17] Hansheng Lei and Venu Govindaraju. Half-against-half multi-class support vector machines. In *Multiple Classifier Systems*, pages 156–164. Springer, 2005.

[18] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.

[19] R Timothy Marler and Jasbir S Arora. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395, 2004.

[20] George Mavrotas. Effective implementation of the< i> ε</i>-constraint method in multi-objective mathematical programming problems. *Applied Mathematics and Computation*, 213(2):455–465, 2009.

[21] Jonathan Milgram, Mohamed Cheriet, Robert Sabourin, et al. ąřone against oneąś or ąřone against alląś: Which one is better for handwriting recognition with svms? In *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[22] Edgar Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. 1997.

[23] John C Platt, Nello Cristianini, and John Shawe-Taylor. Large margin dags for multiclass classification. In *NIPS*, volume 12, pages 547–553, 1999.

[24] Friedhelm Schwenker. Hierarchical support vector machines for multi-class pattern recognition. In *Knowledge-based intelligent engineering systems and allied technologies, 2000. Proceedings. Fourth international conference on*, volume 2, pages 561–565. IEEE, 2000.

[25] K. Tatsumi, K. Hayashida, H. Higashi, and T. Tanino. Multi-objective multiclass support vector machine for pattern recognition. In *SICE, 2007 Annual Conference*, pages 1095–1098. IEEE, 2007.

[26] K. Tatsumi, K. Hayashida, and T. Tanino. Multi-objective multiclass support vector machine maximizing exact margins. *FRONTIERS SCIENCE SERIES*, 49:381, 2007.

[27] K. Tatsumi, R. Kawachi, K. Hayashida, and T. Tanino. Multiobjective multiclass soft-margin support vector machine maximizing pair-wise interclass margins. *Advances in Neuro-Information Processing*, pages 970–977, 2009.

[28] K. Tatsumi, M. Tai, and T. Tanino. Multiobjective multiclass support vector machine based on the one-against-all method. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–7. IEEE, 2010.

[29] K. Tatsumi, M. Tai, and T. Tanino. Nonlinear extension of multiobjective multiclass support vector machine based on the one-against-all method. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1570–1576. IEEE, 2011.

[30] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[31] V. Vapnik. Statistical learning theory. 1998, 1998.

[32] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York Inc, 2000.

[33] Konstantinos Veropoulos, Colin Campbell, Nello Cristianini, et al. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on artificial intelligence*, volume 1999, pages 55–60. Citeseer, 1999.

[34] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the seventh European symposium on artificial neural networks*, volume 4, pages 219–224, 1999.

[35] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[36] Liu Zhigang, Shi Wenzhong, Qin Qianqing, Li Xiaowen, and Xie Donghui. Hierarchical support vector machines. In *Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International*, volume 1, pages 4–pp. IEEE, 2005.