THE MIXED-NORM PROXIMAL SUPPORT VECTOR CLASSIFIER (M-PSVC)

Wei-Cheng Pao, Leu-Shing Lan, and *Dian-Rong Yang*

Department of Electronics Engineering
National Yunlin University of Science and Technology, Taiwan

ABSTRACT

Support vector machines (SVMs) have been recognized as one of the most powerful tools for machine learning, pattern classification, and function estimation. A number of different variations for the SVMs have been proposed, such as the ν-SVM, least-squares SVM, proximal SVM (PSVM), reduced SVM (RSVM), Lagrangian SVM (LSVM), etc. This paper addresses the issue of generalization of the proximal SVM. We propose a mixed-norm proximal support vector classifier (referred to as the m-PSVC) that combines the characteristics of 1-norm and 2-norm classification errors jointly. Using the method of Lagrange multipliers, we derived a form suitable for efficient implementation. It is found that the decision boundary of the m-PSVC concides with that of the PSVC exactly, while the classification margin of the former is proportional to the $(1 + \frac{C_1}{C_2})^{-1}$ factor. Some demonstrative examples are given to show the relations among the newly developed m-PSVC, standard PSVC, and conventional LS-SVC.

Keywords: support vector machines, support vector classifiers, SVM, PSVC, m-PSVC

1. INTRODUCTION

Support vector machines (SVMs) were introduced by Vapnik and his colleagues [1, 2] and have become one of the most popular tools in machine learning, data minging, pattern classification, and function estimation. Two key ideas are employed in the SVMs, i.e., an implicit kernel mapping trick, and a maximal margin classifier. Since the emergence of SVMs, a wide variety of successful applications have been reported, such as image processing [3], wireless communications[4], computer vision [5], optical character recognition (OCR) [6], text categorization [7], time-series prediction [8], gene expression profile analysis [9], DNA and protein analysis [10], etc. Vapnik *et al.*'s statistical learning theory [2] provides a solid mathematical foundation for the SVMs.

In spite of the merits possessed by the SVMs, they also face some challenges that may limit their usefulness in practical scenarios. The classical QP-based approach is not efficient when dealing with problems with exceedingly large amounts of data because it is computationally expensive and sometimes leads to machines with many support vectors, especially when the classes are significantly overlapped. Lately, Suykens and Vandewalle [11] proposed a modified version of the 2-norm support vector classifier (SVC) which they called least-squares SVC (LS-SVC). In LS-SVC, the nonequality constraints related to soft margins are replaced by equalities. The LS-SVC has a very attractive advantage regarding the computational efficiency of training. For the LS-SVC, training requires only solving a set of linear equations, instead of solving the complex QP. The price paid is a little degradation in the generalization performance. Mangasarian and Fung [12] later proposed another form of the LS-SVC, which they called the proximal support vector classifier (PSVC). The main difference between the PSVC and LS-SVC is the inclusion of a $b^2/2$ term in the design objective function. The net effect is a slight degradation in classification performance; however, the equality constraint $\sum_{i=1}^{l} \alpha_i y_i = 0$ in the dual formulation then disappears.

This paper addresses the issue of generalization of the PSVC. Specifically, we developed a generalized PSVC which we refer to as the mixed-norm proximal support vector classifier (m-PSVC). The m-PSVC is derived by incorporating both 1-norm and 2-norm classification errors into the design objective function. Since both error norms are included, the conventional PSVC can be viewed as one of its special cases. Using the method of Lagrange multipliers, we observe that the solution to the m-PSVC problem is given by a set of linear equations, which are similar to those of the conventional PSVC. However, in the new set of linear equations, a multiplication factor $(1 + \frac{C_1}{C_2})$ controls the overall performance. Through mathematical derivation and experimental justification, we have found that the deci-

sion boundary of the m-PSVC is the same as that of the conventional PSVC, whereas its classification margin is proportional to $(1+\frac{C_1}{C_2})^{-1}$. Some demonstrative examples are given to show the relations among the newly developed m-PSVC, the PSVC, and conventional LS-SVC. To simplify the presentaion, in this paper we only focus on the linear support vector classifier. The extension to the nonlinear case is straightforward.

The rest of this paper is organized as follows. In Section 2, we give a brief summary of the conventional 1-norm, 2-norm, least-squares, and proximal support vector classifiers. Then in Section 3, we present the mixed-norm proximal support vector classifier. Some examples are given in Section 4. Finally Section 5 concludes this paper.

2. 1-NORM, 2-NORM, LEAST-SQUARES, AND PROXIMAL SUPPORT VECTOR CLASSIFIERS

Consider a two-class classification problem with a collection of training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^l$ given, where $\boldsymbol{x}_i \in \mathbf{R}^n$ is the ith input vector, $y_i \in \{+1, -1\}$, l is the total number of training data, and n is the dimension of the input space. A linear classifier is defined by means of a separating hyperplane in the form of $f(\boldsymbol{x}) = (\boldsymbol{w} \cdot \boldsymbol{x}) + b = 0$, where $\boldsymbol{w}$ and b should be determined according to certain optimization criteria. Here $(\boldsymbol{w} \cdot \boldsymbol{x})$ denotes the inner product between $\boldsymbol{w}$ and $\boldsymbol{x}$. To obtain better generalization performance, the separating hyperplane should be placed such that a maximum distance between the two classes of data is achieved. The distance between two classes of data is usually referred to as *margin*. The support vector classifier attempts to maximize this margin while minimizes the classification error. For a linearly separable problem, the SVC design is expressible as a minimization problem

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}(\boldsymbol{w} \cdot \boldsymbol{w}) \tag{1}$$

$$\text{s.t.} \quad y_i[(\boldsymbol{w} \cdot \boldsymbol{x}_i) + b] \geq 1, \ i = 1, \cdots, l \tag{2}$$

Sometimes the classification problems encountered are not linearly separable. Under such circumstances, modifications to the separable formulation are needed. We can introduce a soft margin as follows:

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \quad \frac{1}{2}(\boldsymbol{w} \cdot \boldsymbol{w}) + C\sum_{i=1}^{l} \zeta_i \tag{3}$$

$$\text{s.t.} \quad y_i[(\boldsymbol{w} \cdot \boldsymbol{x}_i) + b] \geq 1 - \zeta_i, \ i = 1, \cdots, l \tag{4}$$

$$\zeta_i \geq 0, \quad i = 1, \cdots, l \tag{5}$$

where C is a regularization parameter to control the balance between the size of margin and the misclassification error. This formulation will lead to a solution that is commonly known as the 1-norm support vector classifier since the l_1 norm error is considered. Another common way to introduce a soft margin into the classification problem is instead to use the l_2 norm error which will yield the 2-norm support vector classifier and gives the following formulation

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \quad \frac{1}{2}(\boldsymbol{w} \cdot \boldsymbol{w}) + \frac{C}{2}\sum_{i=1}^{l} \zeta_i^2 \tag{6}$$

$$\text{s.t.} \quad y_i[(\boldsymbol{w} \cdot \boldsymbol{x}_i) + b] \geq 1 - \zeta_i, \ i = 1, \cdots, l \tag{7}$$

In the 2-norm formulation, the $\zeta_i \geq 0, \ i = 1, \cdots, l$ constraint is discarded since it can be shown to be redundant.

Recently, Suykens and Vandewalle [11] proposed a modified version of the 2-norm SVC which they called least-squares SVC (LS-SVC). In LS-SVC, the nonequality constraints related to soft margins are replaced by equalities. The advantage of this reformulation is that the quadratic programming solution procedure is no more needed; instead, the solution comes in a form of linear system of equations. The price paid is a little degradation in the generalization performance. Here we briefly review the problem formulation for the LS-SVC. The optimization problem defined by the LS-SVC is given by

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \quad \frac{1}{2}(\boldsymbol{w} \cdot \boldsymbol{w}) + \frac{C}{2}\sum_{i=1}^{l} \zeta_i^2 \tag{8}$$

$$\text{s.t.} \quad y_i[(\boldsymbol{w} \cdot \boldsymbol{x}_i) + b] = 1 - \zeta_i, \ i = 1, \cdots, l \tag{9}$$

This equality-constrained minimization problem can be solved directly in closed form using the method of Lagrange multipliers. The solution can be expressed as an $(l+1) \times (l+1)$ linear system given by

$$\begin{bmatrix} 0 & \boldsymbol{y}^T \\ \boldsymbol{y} & RR^T + \frac{1}{C} I_{l\times l} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{1}_l \end{bmatrix} \tag{10}$$

where $\boldsymbol{y} = [y_1, y_2, \cdots, y_l]^T$, $R = [y_1\boldsymbol{x}_1, y_2\boldsymbol{x}_2, \cdots, y_l\boldsymbol{x}_l]^T$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_l]^T$ is the corresponding Lagrange multiplier vector, and $\mathbf{1}_l$ is a dimension-l vector with all ones.

The proximal support vector classifier (PSVC) was proposed by Mangasarian and Fung [12] in 2001 as an alternative form of the standard LS-SVC. The key difference between the PSVC and LS-SVC is the inclusion of a $b^2/2$ term in the design objective function. The net effect is a slight degradation in classification performance; however, the equality constraint

$$\textstyle\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{11}$$

in the dual formulation then disappears. The PSVC problem is formally stated as

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \quad \frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}+b^2)+\frac{C}{2}\sum_{i=1}^{l}\zeta_i^2 \tag{12}$$

$$\text{s.t.} \quad y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]=1-\zeta_i,\ i=1,\cdots,l \tag{13}$$

3. THE MIXED-NORM PROXIMAL SUPPORT VECTOR CLASSIFIER (M-PSVC)

Let us now consider a generalized form of the proximal support vector classifier (PSVC) that combines the 1-norm and 2-norm characteristics jointly. This new type of support vector classifier problem can be formally defined as

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \quad \frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}+b^2)+C_1\sum_{i=1}^{l}\zeta_i+\frac{C_2}{2}\sum_{i=1}^{l}\zeta_i^2 \tag{14}$$

$$\text{s.t.} \quad y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]=1-\zeta_i,\ i=1,\cdots,l \tag{15}$$

where C_1 and C_2 are regularization parameters that determine the balance among the decision margin, the l_1-type error, and the l_2-type error. In this optimization problem we use a shorthand notation "s.t." to represent *subject to*. This formulation is a direct extension of the PSVC by adding another 1-norm term to the objective function so that both 1-norm and 2-norm classification errors are considered. As such, the new formulation is referred to as the mixed-norm proximal support vector classifier. To solve this optimization problem, we resort to the method of Lagrange multipliers by first defining the Lagrangian function $\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\zeta},\boldsymbol{\alpha})$ as

$$\begin{aligned}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\zeta},\boldsymbol{\alpha}) \quad &\triangleq \quad \frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}+b^2)+C_1\sum_{i=1}^{l}\zeta_i\\ &+ \quad \frac{C_2}{2}\sum_{i=1}^{l}\zeta_i^2\\ &- \quad \sum_{i=1}^{l}\alpha_i\{y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-1+\zeta_i\}\end{aligned} \tag{16}$$

Letting the partial derivatives of $\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\zeta},\boldsymbol{\alpha})$ be 0 with respect to $\boldsymbol{w}$,b,$\boldsymbol{\zeta}$, and $\boldsymbol{\alpha}$, we obtain

$$\boldsymbol{w}=\textstyle\sum_{i=1}^{l}\alpha_i y_i\boldsymbol{x}_i \tag{17}$$

$$\textstyle\sum_{i=1}^{l}\alpha_i y_i=b \tag{18}$$

$$\zeta_i=\frac{\alpha_i-C_1}{C_2},\ \ i=1,\cdots,l \tag{19}$$

$$y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]=1-\zeta_i,\ i=1,\cdots,l \tag{20}$$

Substituting Eqs. (17), (18), and (19) into Eq. (20) yields

$$\textstyle\sum_{i=1}^{l}\alpha_j y_i y_j[(\boldsymbol{x}_j\cdot\boldsymbol{x}_i)+1]+\frac{1}{C_2}\alpha_i=1+\frac{C_1}{C_2},\quad i=1,\cdots,l \tag{21}$$

Putting in matrix form, these equations can be written more compactly as

$$(Q+P+\frac{1}{C_2}I_{l\times l})\,\boldsymbol{\alpha}=(1+\frac{C_1}{C_2})\mathbf{1}_l \tag{22}$$

where $Q_{ij}\stackrel{\triangle}{=}y_iy_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j), 1\le i,j\le l$, $P_{ij}\stackrel{\triangle}{=}y_iy_j, 1\le i,j\le l$, $I_{l\times l}$ is the identity matrix of rank l, and $\mathbf{1}_l$ is a length-l vector of all ones. If the matrix $Q+P+\frac{1}{C_2}I_{l\times l}$ is nonsingular, then we have

$$\boldsymbol{\alpha}=(Q+P+\frac{1}{C_2}I_{l\times l})^{-1}(1+\frac{C_1}{C_2})\mathbf{1}_l \tag{23}$$

Obviously, when C_1 equals 0, Eq. (23) will reduce to the solution of a proximal support vector classifier. Thus by varying the values of C_1 and C_2, the performance of the m-PSVC can be adjusted. In this regard, the proposed m-PSVC offers an extra degree of freedom to the fine tuning of classification performance. Once $\boldsymbol{\alpha}$ is obtained, b can be found using Eq. (18).

4. EXAMPLES

This section presents several examples to demonstrate the behavior of the m-PSVC. Since the m-PSVC can be viewed as a generalized PSVC, we mainly compare the performance between these two SVCs. In addition, the LS-SVC is also included in the comparison to exhibit its close resemblance to the PSVC. The training data for the examples consist of 20 points, 10 of them belong to Class1, while the remaining 10 points belong to Class2. These data are linearly nonseparable. The support vector classifier will try to seek for the best performance by trading-off between the largest separating margin and generalization capability. C_1 and C_2 are tuning parameters (also called hyperparameters) which have to be set before the algorithms start. Here we consider three sets of these parameters: (1) $C_1=1, C_2=1$, (2) $C_1=1, C_2=10$, and (3) $C_1=1, C_2=0.1$. Three different $\frac{C_1}{C_2}$ ratios are chosen as this ratio is the dominating factor that control the overall performance. These parameters are empirically determined. A good theoretical procedure to choose proper hyperparameters is currently an open question and is problem-dependent. The experiments demonstrate where the decision boundary (or classification boundary) is located among the scattered data and how the values of C_1 and C_2 affect the moving of

decision boundary and the width of classification margin. Figs. 1 through 3 show the results for the three sets of tuning parameters. Surprisingly, we see that for these cases the m-PSVC and the PSVC possess the same decision boundary, although their classification margins are different. From Figs. 1, 2, and 3, it is clear that the classification margin increases when the $\frac{C_2}{C_1}$ ratio is increased. We now give a theoretical justification for this kind of behavior. From Eqs. (23) and (18), it is seen that both b and $\boldsymbol{\alpha}$ are proportional to $(1+\frac{C_1}{C_2})$. Suppose that the solution for the PSVC problem is defined by b_P and $\boldsymbol{\alpha}_P$, which correspond to the special case of the m-PSVC solution with $C_1 = 0$. Let the solution to the m-PSVC problem be denoted by b_m and $\boldsymbol{\alpha}_m$, then the m-PSVC solution can be written as

$$b_m = b_P(1+\frac{C_1}{C_2}) \tag{24}$$

$$\boldsymbol{\alpha}_m = \boldsymbol{\alpha}_P(1+\frac{C_1}{C_2}) \tag{25}$$

The decision boundary for the LS m-SVC problem is given by

$$(\textstyle\sum_{i=1}^{l}\alpha_{i,m}y_i\boldsymbol{x}_i)^T\boldsymbol{x}+b_m=0 \tag{26}$$

which reduces to

$$(\textstyle\sum_{i=1}^{l}\alpha_{i,P}y_i\boldsymbol{x}_i)^T\boldsymbol{x}+b_P=0 \tag{27}$$

Therefore, we conclude that the m-PSVC and PSVC possess the same decision boundary, whereas the separation margin of the m-PSVC is directly proportional to $(1+\frac{C_1}{C_2})^{-1}$. The relation between the m-PSVC and the LS-SVC is also interesting. For smaller values of $\frac{C_1}{C_2}$, the 2-norm error dominates, and therefore the decision boundaries of the m-PSVC and the LS-SVC coincide with each other. For larger values of $\frac{C_1}{C_2}$, the 1-norm error dominates, the discrepancy between the m-PSVC and the LS-SVC becomes distinct. These effects can be seen clearly from Figs. 1 through 3.

5. CONCLUSION

This paper presents a new form of proximal support vector classifier (PSVC) that combines both the 1-norm and 2-norm error characteristics. The mixed-norm PSVC (m-PSVC) can be viewed as a generalized proximal support vector classifier that includes the standard PSVC as one of its special cases. It is found that the decision boundary of the m-PSVC exactly concides with that of the PSVC, while the classification margin of the former is proportional to the $(1+\frac{C_1}{C_2})^{-1}$ factor. Some demonstrative examples are given to show the relations among the newly developed m-PSVC, PSVC, and conventional LS-SVC. Some issues regarding the m-PSVC deserve further investigation, including analysis of properties, tuning of hyperparameters, and applications.

6. REFERENCES

[1] C. Cortes and V. N. Vapnik, "Support VectorNetworks," *Machine Learning*, Vol. 20, pp. 273-297, 1995.

[2] V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[3] D. Li, R. M. Mersereau, and S. Simske, "Blind Image Deconvolution Using Support Vector Regression," *Proc. 2005 IEEE Int'l Conf. Acoustics, Speech, and Image Processing*, pp. II-113 116, 2005.

[4] I. Santamaría *et al.*, "Blind Equalization of Constant Modulus Signals Using Support Vector Machines," *IEEE Tr. Signal Processing*, Vol. 52, No. 6, 2004.

[5] B. Castañeda and J. C. Cockburn, "Reduced Support Vector Machines Applied to Real-Time Face Tracking," *Proc. 2005 IEEE Int'l Conf. Acoustics, Speech, and Image Processing*, pp. II-673 676, 2005.

[6] Y. A. LeCun, *et al.*, "Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition," *Neural Networks*, pp. 261-276, 1995.

[7] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector Machines for Spam Categorization," *IEEE Tr. Neural Networks*, Vol. 10, pp. 1048-1054, 1999.

[8] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear Prediction of Chaotic Time Series Using a Support Vector Machine," *Proc. 1997 IEEE Workshop*, pp. 511-520, 1997.

[9] T. Furey, *et al.*, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics*, Vol. 16, pp. 906-914, 2000.

[10] A. Zien, *et al.*, "Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites in DNA," *Bioinformatics*, Vol. 16, pp. 799-807, 2000.

[11] J. A. K. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Process. Lett.*, Vol. 9, pp.293-300, 1999.

[12] O. L. Mangasarian and G. Fung, "Proximal Support Vector Machine Classifiers," *Proc. 2001 Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 64-70, 2001.

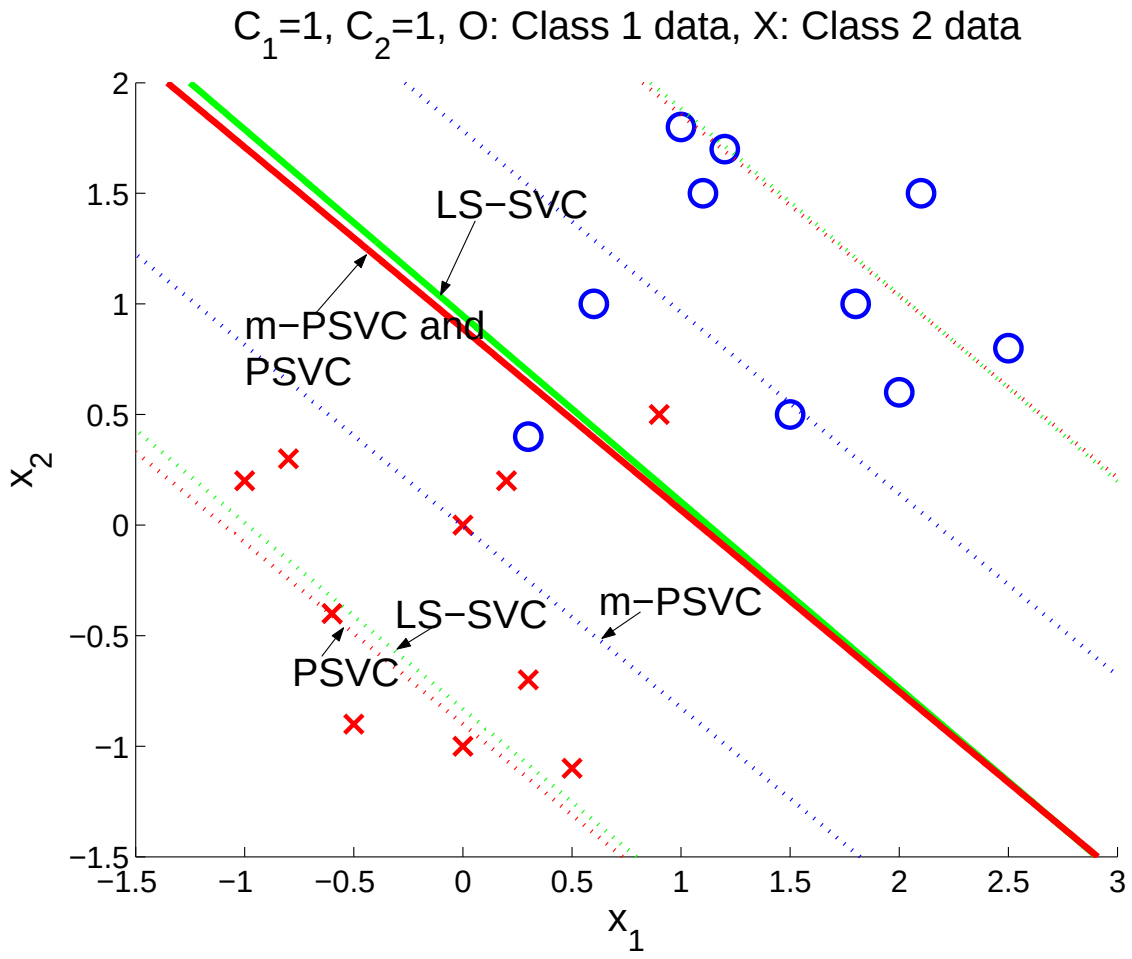

Fig. 1. A demonstrative example to show the relations among the mixed-norm proximal support vector classifier (m-PSVC), the proximal support vector classifier (PSVC), and the least-squares support vector classifier (LS-SVC). In this example, $C_1 = 1$ and $C_2 = 1$.

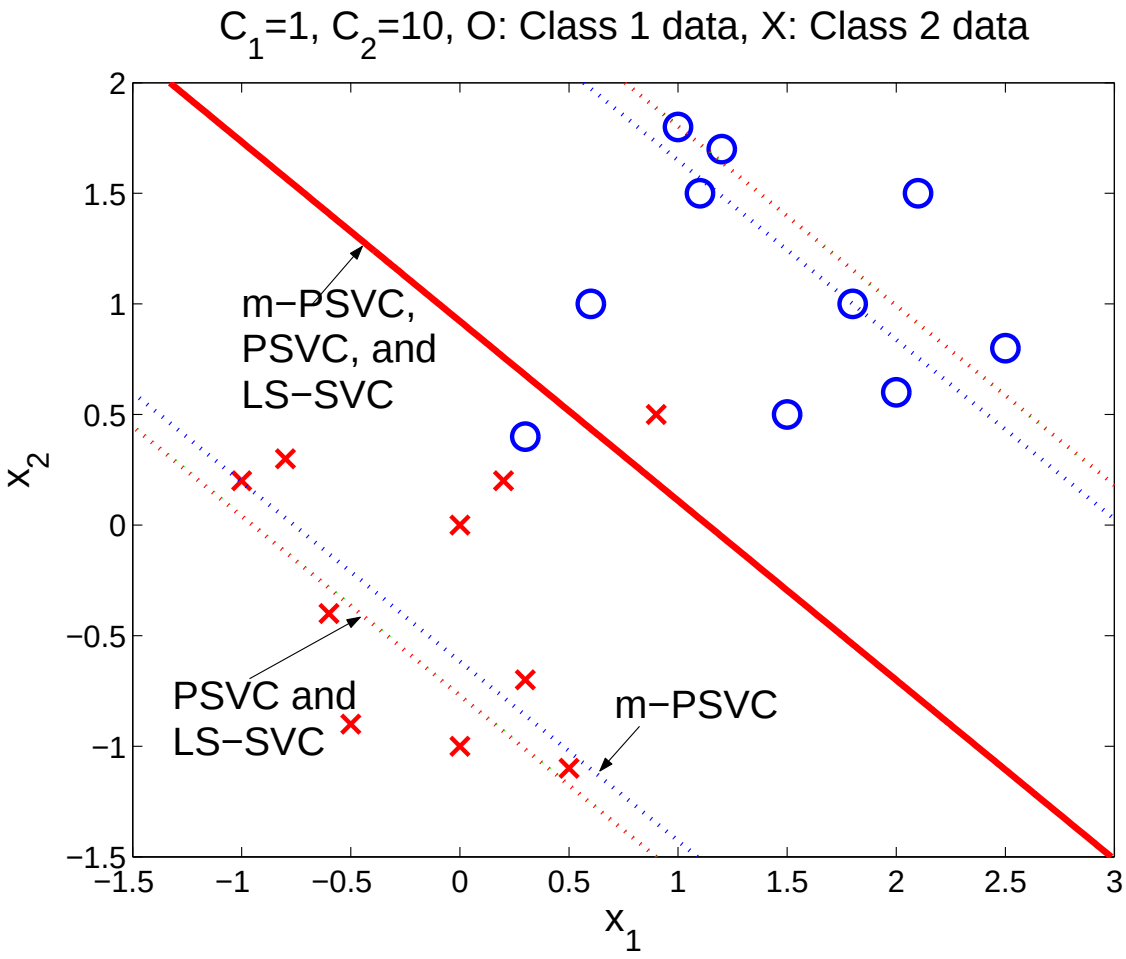

Fig. 2. A demonstrative example to show the relations among the mixed-norm proximal support vector classifier (m-PSVC), the proximal support vector classifier (PSVC), and the least-squares support vector classifier (LS-SVC). In this example, $C_1 = 1$ and $C_2 = 10$.

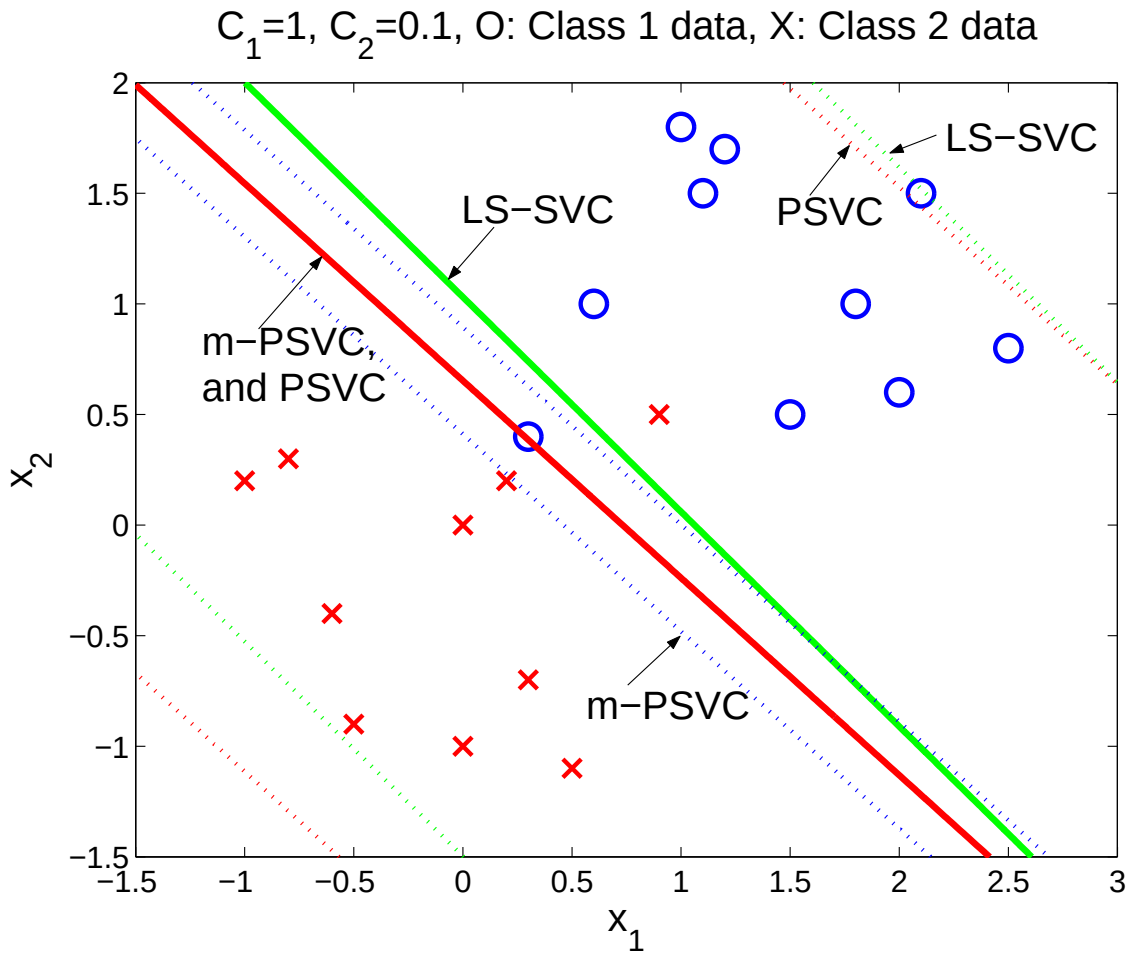

Fig. 3. A demonstrative example to show the relations among the mixed-norm proximal support vector classifier (m-PSVC), the proximal support vector classifier (PSVC), and the least-squares support vector classifier (LS-SVC). In this example, $C_1 = 1$ and $C_2 = 0.1$.