Journal of Industrial and Systems Engineering
Vol. 11, No. 2, pp. 21-30
Spring (April) 2018

JISE

A new formulation for cost-sensitive two group support vector machine with multiple error rates

Ali Nedaie[1], Amir Abbas Najafi[1*]

[1]*Faculty of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran*

alinedaie@kntu.ac.ir, aanajafi@kntu.ac.ir

Abstract

Support vector machine (SVM) is a popular classification technique which classifies data using a max-margin separator hyperplane. The normal vector and bias of the mentioned hyperplane is determined by solving a quadratic model implies that SVM training confronts by an optimization problem. Among of the extensions of SVM, cost-sensitive scheme refers to a model with multiple costs which considers different error rates for misclassification. The cost-sensitive scheme is useful when misclassifications cannot be considered equal. For example, it is true for medical diagnosis. In such cases, misclassifying a patient as healthy implies more loss in comparison to the opposite loss. Therefore, cost-sensitive scheme poses as a modified model and hereby aims at minimizing loss function instead of generalization error. This paper, concentrates on a new formulation cost-sensitive classification considering both misclassification cost and accuracy measures. Also, in the training phase a new heuristic algorithm will be used to solve the proposed model. The superiority of the novel method is affirmed after comparing to the traditional ones.

Keywords: Cost-sensitive learning, classification, Support Vector Machine (SVM), supervised learning.

1- Introduction

Since support vector machine (SVM) was introduced by Vapnik (1998), it has been used in various fields by researchers such as pattern recognition (Pontil and Veri, 1998), classification (Chen and Wang, 2003) and detection (Waring and Liu, 2005). To improve the performance of the SVM, many works have been done in current decade, for instance, fuzzy SVM(Yang et al., 2015), least square SVM (Suykens et al., 2002), total margin SVM (Fung and Mangasarian, 2001), Lagrangian SVM (Mangasarian and Musicant, 2001), proximal SVM (Fung and Mangasarian, 2001), polar SVM (Nedaie and Najafi, 2016), Twin SVM (Tian et al, 2013; Qi et al., 2013; Qi et al., 2012) and etc. The base model of support vector machine was introduced for two-class classification problems, although the mentioned model has been extended to multi-class case, such as works done by Platt et al. (2000) and Crammer and Singer (2001). In all aforementioned

*Corresponding author

models, the misclassification rate is considered to be equal in all classes, i.e. traditional models aims at minimizing the generalization error. However, there is some real-case problems that equal misclassification loss may not be efficient for. As an example, in cancer detection if a healthy human considered as a patient one, the inflicted loss is less than the opposite loss. Also, it is true for fraud detection, non-technical loses in electricity etc. Therefore, cost-sensitive scheme was developed to improve the capability and reality aspects of the classification techniques (Turney, 1995). In this regard, Zheng et al. (2006) proposed a cost-sensitive scheme for support vector machine and demonstrated the efficiency of their scheme. Also, other related researches are conducted by Wan et al. (2012) for Laplacian machine in imbalanced classes where most data are unlabeled. Such a case is known as semi-supervised problem.

Determining the misclassification parameters or error rates is a challenging matter which can influence the performance of model. According to our knowledge, three evidences for this statement are (Tran et al., 2005; Chen et al., 2012; Cao et al., 2013). However, we believe that modifying the cost-sensitive SVM model improves its performance and this is the main motivation of our study. More exactly, it means that by a constant misclassification rates, our proposed model outperforms the traditional cost-sensitive SVM in terms of the loss function. Therefore, in this paper a new model of two groups cost-sensitive SVM will be introduced regarding both misclassification cost and accuracy measures. In the proposed model, the error rate of misclassification is a 1×2 vector. Since the proposed method is binary, it is shown that this model can be relaxed and solved by continues variables without less of generality. In numerical experiment, the explained method will be investigated and the superiority of the new model will be affirmed comparing to the traditional ones in terms of the measures. Besides, a new heuristic algorithm is proposed in this study to avoid local optima. To provide a comprehensive description of this work, the rest of these papers is organized as follows:

Section 2 explains the traditional model and cost-sensitive SVM. Our new model is described and formulated in section 3. Numerical results and comparisons are dedicated to be discussing in section 4 and finally section 5 is assigned to conclusion remarks.

2- Support Vector Machine (SVM)

2-1- Traditional Support Vector Machine

The first model of SVM was introduced by Vapnik (1998) for classifying data into the two classes in the form of both linear and non-linear machines. The non-linear strategy in this technique is to map the input vectors into a high dimension feature space corresponding to a kernel $\phi(x)$, and construct a linear decision function in this space to separate the datasets by a hyperplane with maximum margin. Given the training data $X = \{x_1, x_2, \dots, x_n\}$ and classes$y_i \in \{-1,1\}$, the SVM model is as follows:

$$\begin{aligned} Min\ & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i \\ & y_i(\omega^T\,\phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{1}$$

Where, ω and b are the normal vector and bias of the separator hyperplane, respectively. C is the penalty of misclassification and its value can be effected on model training in which the suitable value of Ccan lead to a more accurate and faster classifier (Tran et al., 2005; Hwang et al., 2011) and finally,ξ_i`s are slack variables which can be interpreted as follows:

1- $0 \leq \xi_i \leq 1$: i^{th} point is classified correctly.

2- $\xi_i \geq 1$: i^{th} point is misclassified.

If all of the training datasets are linearly separable (Figure 1), one can be concluded that $\xi_i = 0$ for all $i = 1,2,\dots,n$.

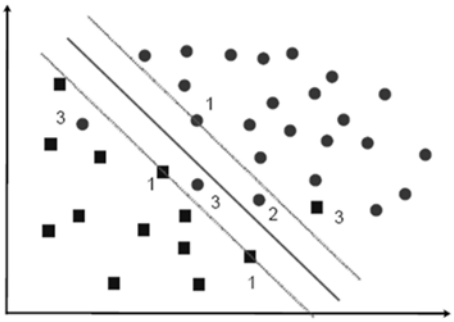

Fig. 1 The SVM separator hyperplane in nonlinear separable case

SVM tries to maximize the margin between the two classes so it can minimize the structural risk instead of empirical risk (Cao et al., 2013).

2-2- Cost-Sensitive Support Vector Machine

In a standard classification problem, the goal is to minimize the probability of misclassification. In many applications, however, some kinds of errors are more important than the others, so it seems that an extended scheme is required for minimizing the errors of interest. In tumor classification, for example, the impact of mistakenly classifying a benign tumor as malignant is much less than of the opposite mistake and so the objective function should be modified to minimize the cost of misclassification and not the number of misclassifications. Support vector machine with cost-sensitive scheme (CSSVM), is one of the extended models of traditional SVM which tries to consider these different error rates in the SVM model. The adjusted model of SVM which leads us to CSSVM is as follows (Zheng et al., 2006):

$$Min_f \frac{1}{2}\|f\|_H^2 + C_1 \sum_{i \in class1} \ell\big(y_i, f(x_i)\big) + C_2 \sum_{i \in class2} \ell\big(y_i, f(x_i)\big) \tag{2}$$

Where,C_1 and C_2 are misclassification rates associated to class 1 and class 2. Also, $\ell\big(y_i, f(x_i)\big)$ is error function and is equivalent toξ_i, usually. The simplified form derived from model (2) will be obtained by considering ξ_i`s as the error rates:

$$\begin{aligned} Min\ & \frac{1}{2}\|\omega\|^2 + C_1 \sum_{i \in class1} \xi_i + C_2 \sum_{i \in class2} \xi_i \\ & y_i(\omega^T\, \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{3}$$

In many applications,C_1 and C_2 can be considered different for finding a suitable separator hyperplane which minimizes the objective function. If $C_1 = C_2$ in the above model, the traditional SVM will be obtained.

3- The proposed cost-sensitive SVM

As described earlier, the cost-sensitive scheme is one of the useful extensions of SVM which can consider the multiple error rates where ξ_i`s have different values. But it is obvious that $\xi_i$`s should not contribute any term in the objective function when they are less than 1. Because all of the data with $0 \leq \xi_i \leq 1$ are correctly classified which do not imply any cost. But it is obvious that model (3) does not consider this fact and should be modified. The performance of the mentioned model may be improved after the modification.

Figure 2 shows a comparative example of the traditional SVM, cost-sensitive SVM and proposed SVM. In the above graphical example, the error rate associated with class 1 is greater than the opposite misclassification rate. Since traditional SVM con not classify data by considering different error rate, cost-sensitive SVM should be used. But it is obvious that the data with $0 \leq \xi_i \leq 1$ (which are correctly classified) contribute error rate to objective function and lead to a new hyperplane as figure 2(c). But, note that all data with $0 \leq \xi_i \leq 1$ are truly classified and should not move the separator plane because this unduly movement causes to misclassification in opposite class. Then, cost-sensitive scheme may be adjusted for leading to a new model with better characteristics in terms of accuracy (Figure 2(b)). In the other hand the data with $0 \leq \xi_i \leq 1$ should not contribute any term in objective function.
For modeling such problem assume that the error rate matrix is as follows:

$$\gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \tag{4}$$

Where, γ_{ij} is the penalty of misclassifying a data from class i into class j, obviously, $\gamma_{ij} = 0$ for $i = j$.
The new described model will be derived from model (3) after implementing some modifications which eliminate all ξ_i `s associated with the correctly classified data from objective function. It can be implemented by introducing a binary variable, sayα_i which leads us to the model at below:

$$\begin{aligned} Min \quad & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i + \gamma_{12}\sum_{x_i \in class1}\alpha_i + \gamma_{21}\sum_{x_i \in class2}\alpha_i \\ & y_i(\omega^T\,\phi(x_i) + b) \geq 1 - \xi_i \\ & 0 \leq \xi_i \leq 1 + M\alpha_i \qquad i = 1,2,\dots,n \\ & \alpha_i \in \{0,1\} \end{aligned} \tag{5}$$

Note that, *M* is a parameter with near infinite value.

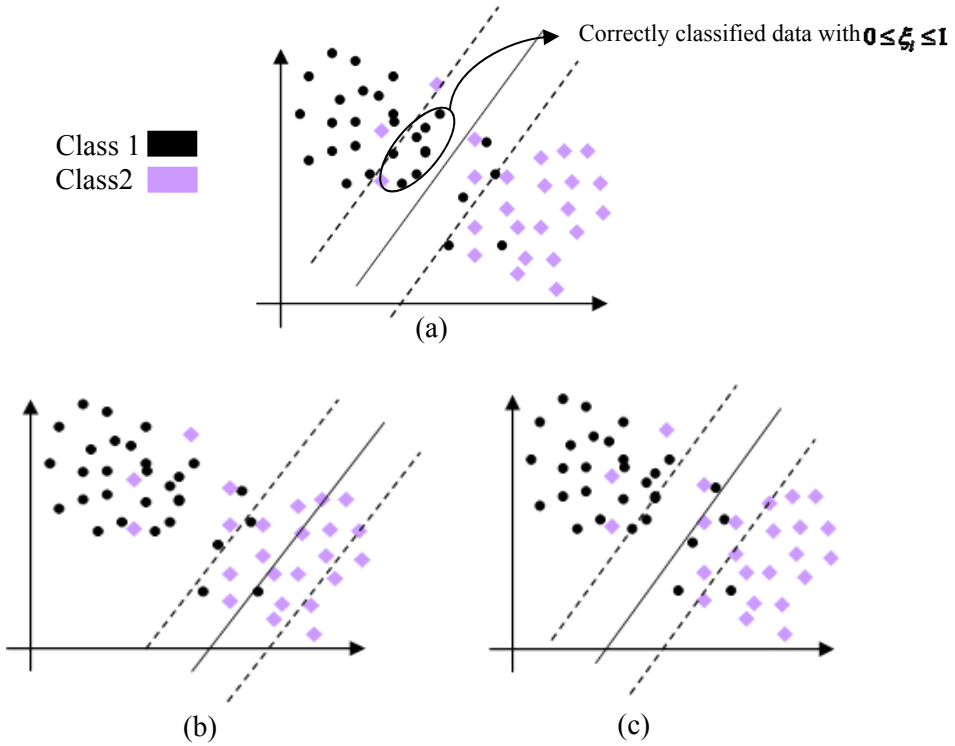

Fig. 2 Support vector machine hyperplanes:
(a) Traditional machine (b) Traditional cost-sensitive machine (c) Proposed cost-sensitive

In model (5), each misclassified data implies that the corresponding α_i should be greater than 0 and then a new term will be contributed to the objective function by error rate γ_{ij}. In such situation if one of the misclassification rates is greater than the other, a more suitable separator hyperplane will be found which minimizes the error rate behind of maximizing the accuracy. But it should be noted that in the modified model the data with $0 \leq \xi_i \leq 1$ do not contribute any term in objective function as error rate.

Lemma 1- *The Model (5) can be relaxed without less of generality. It would be possible by eliminating parameter M and considering* $\alpha_i \geq 0$.

Proof- *Define* $\alpha_i' = M\alpha_i$. *Since* $\alpha_i \in \{0,1\}$, *it can be concluded that* $\alpha_i' \geq 0$ *for each feasible solution in the model. By substituting* α_i' *into the Model (5) we have:*

$$
\begin{aligned}
Min \quad & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i + \gamma_{12}\sum_{x_i \in class1}\alpha_i + \gamma_{21}\sum_{x_i \in class2}\alpha_i \\
& y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i \\
& 0 \leq \xi_i \leq 1 + \alpha_i' \qquad i = 1,2,\dots,n \\
& \alpha_i \geq 0
\end{aligned}
\tag{6}
$$

Model (6) is a quadratic programming problem with linear constraints and can be solved by many kinds of algorithms (Shawe-Taylor and Sun, 2011). In the next section the mentioned model will be experimented by using some datasets.

4- Numerical experiments

In this section, for demonstrating the superiority of the proposed model, numerical experiments will be conducted. Two measures, say accuracy and cost of misclassification are considered for comparing the new method with traditional SVM and CCSVM. All of the models are solved in two phases. The first phase is to using interior point algorithm started at $X_0 = 0$ as the first solution and at the next phase the obtained solution is improved by using iterative algorithm sequential quadratic programming (SQP). Stop criterion in this algorithm is considered as 1e-10 for objective function tolerance.

All computations are implemented on MATLAB 2015b. Due to a reasonable CPU time, primal models have been solved instead of dual form; however, it is obvious that dual models are preferable in terms of the complexity and hence consume less time. Train/test sets are selected using 10-fold cross validation technique. In this regard, each dataset is randomly divided into 10 folds and 9 folds are used to train the model and the remaining one is considered as a test set. Accuracy and cost of misclassification measures are reported after calculating the maximum performance across the hold-out predictions (Equations 7 and 8).

$$Accuracy = \frac{N_{11} + N_{22}}{N} \tag{7}$$

$$Cost = \gamma_{12} N_{12} + \gamma_{21} N_{21} \tag{8}$$

Where, N_{ij} is the number of class i data which are classified in class j ($i, j = 1,2$). Obviously, a misclassification is occurred when $i \neq j$.

Finally, a new heuristic method attempts to find a better solution. This current method is used for avoiding the local optima. The pseudo-code of the heuristic is as figure 3.In the other hand, the heuristic algorithm tries to find a better solution by shaking the normal vector and bias of the separator hyperplane. In the iterations of the algorithm, neighbors of the current obtained hyperplane are checked and will be accepted if are more optimal than the previous solutions in terms of the misclassification cost. Note that, the parameter ε is considered equal to $0.01 \times m$ in all datasets.

\\"ω"and "b" are the normal vector and bias of the hyperplane. Also "k" is a constant say neighbor search radius
Begin
Calculate the misclassification cost of the current classifier
While misclassification cost can be decreased or is not less than ε
Do$\omega_i = \omega_i \pm k$ and $b_i = b_i \pm k$
End while
Return"ω"and "b"
End

Fig. 3 Heuristic algorithm pseudo-code for avoiding local optima

Lemma 2- *The heuristic algorithm proposed in Figure 3 has time complexity $O(m\varepsilon)$, where, m is the number of datapoints.*

Proof- *as a worst case, a classifier has 0% accuracy and can be improved as much as ε percent. This means that the heuristic algorithm has maximum 100/ε iterations. Also, in each iteration, the accuracy of*

the hyperplane should be computed over all data. As a result, one can conclude that the algorithm has $O(m\varepsilon)$complexity.

The datasets which are used for comparing the models are shown in table 1. All employed datasets are available at UCI repository[2].In this table, $\gamma_{12} = 2$ and$\gamma_{21} = 1$. It is not difficult to conclude that the proposed CCSVM outperforms traditional machines in terms of the misclassification cost. However, for more reliable comparison, Wilcoxon hypothesis test is used to affirm the advantage of the new machine. Steps for conducting this test are as follows:

- Calculate the paired differencesd_i.
- Rank alld_i`s, ignoring the signs (assign rank 1 to the smallest$|d_i|$, rank 2 to the next, etc.).
- CalculateW^+, the sum of the ranks corresponding to positived_i`s and $W^-$ corresponding to negative$d_i$`s.
- Choose$W = min\{W^+, W^-\}$.
- Use table of critical values for Wilcoxon signed rank test.

The P-value measure is reported at table1. In this regard, 0.0001, for instance, means that misclassification cost for Proposed CCSVM is less than traditional one in 0.9999 of confidence. Note that, in traditional SVM, the penalty parameter C is considered as$(\gamma_{12} + \gamma_{21})/2$.

Table 1. The results for different models by $\gamma_{12} = 2$ and $\gamma_{21} = 1$

Dataset	SVM		CCSVM		Proposed CCSVM	
	Accuracy (%)	Misclassification Cost	Accuracy (%)	Misclassification Cost	Accuracy (%)	Misclassification Cost
Balloons	100	0	100	0	100	0
Credit	86.6	157	86.6	157	86.6	157
Pima	77.4	231	73.7	222	76.6	219
Liver	71.5	151	57.9	145	66.9	120
Ionosphere	92.4	77	90.1	54	92.4	47
Wine	100	0	100	0	100	0
Sonar	89.9	171	85.6	164	88.9	161
Haberman	73.8	212	59.7	204	68.9	193
Wholesale	91.6	121	89.3	116	91.5	100
CMC	70.5	257	67.1	246	69.7	242
Breast	96.3	198	77.9	98	90.0	89
Sphere	64.8	624	63.1	602	64.7	523
Titanic	77.3	319	73.6	306	76.4	301
Spam	89.4	349	72.3	326	83.5	269
Wilt	93.4	199	91.0	228	93.3	198
Banana	58.4	832	55.6	799	57.7	788
Phoneme	72.8	791	58.9	786	68.0	650
P-Value	-	0.0000	-	0.0001	-	-

Another experiment is established using different values for misclassification parameters. In this regard, tables 2 and 3 illustrate the results.

[2]*http://archive.ics.uci.edu/ml*

Table 2- The results for different models by $\gamma_{12} = 8$ and $\gamma_{21} = 1$

Dataset	SVM		CCSVM		Proposed CCSVM	
	Accuracy (%)	Misclassification Cost	Accuracy (%)	Misclassification Cost	Accuracy (%)	Misclassification Cost
Balloons	100	0	100	0	100	0
Credit	86.6	577	83.90	259	85.74	254
Pima	77.4	579	65.10	268	65.10	268
Liver	71.5	329	57.97	145	57.97	145
Ionosphere	92.4	173	85.47	65	87.47	64
Wine	100	0	100	0	100	0
Sonar	89.9	530	75.6	245	75.6	245
Haberman	73.8	657	59.8	289	59.8	289
Wholesale	91.6	313	88.7	140	90.6	137
CMC	70.5	799	59.2	369	59.2	369
Breast	96.3	304	78.07	133	78.0	133
Sphere	64.8	1631	59.9	612	61.3	603
Titanic	77.3	994	74.8	446	76.5	437
Spam	89.4	920	75.1	425	75.1	425
Wilt	93.4	621	75.7	273	75.7	273
Banana	58.4	2598	54.0	976	55.2	961
Phoneme	72.8	2222	70.5	997	72.0	978
P-Value	-	0.0000	-	0.0078	-	-

Comparing the employed machines, it can be conclude that the traditional SVM is superior one in terms of the accuracy. However, this model is not preferable regarding the misclassification cost. In other side, the proposed CCSVM is outstanding machine from misclassification viewpoint. In table 2, for instance, Phoneme dataset affirms the excellence of the novel model considering both accuracy and cost results.

Table 3- The results for different models by $\gamma_{12} = 10$ and $\gamma_{21} = 1$

Dataset	SVM		CCSVM		Proposed CCSVM	
	Accuracy (%)	Misclassification Cost	Accuracy (%)	Misclassification Cost	Accuracy (%)	Misclassification Cost
Balloons	100	0	100	0	100	0
Credit	86.6	717	83.90	303	83.90	303
Pima	77.4	695	65.10	268	65.10	268
Liver	71.5	395	57.97	145	57.97	145
Ionosphere	92.4	215	85.47	69	85.47	69
Wine	100	0	100	0	100	0
Sonar	89.9	658	75.6	253	75.6	253
Haberman	73.8	788	59.8	289	59.8	289
Wholesale	91.6	375	84.7	120	84.7	120
CMC	70.5	992	68.3	419	68.3	419
Breast	96.3	377	80.9	145	80.9	145
Sphere	64.8	1957	52.5	718	52.5	718
Titanic	77.3	1193	71.5	382	71.5	382
Spam	89.4	1143	86.6	483	86.6	483
Wilt	93.4	771	78.5	297	78.5	297
Banana	58.4	3118	47.3	1144	47.3	1144
Phoneme	72.8	2667	67.3	855	67.3	855
P-Value	-	0.0000	-	1	-	-

Table 3, reveals another fact. In this table, the traditional and proposed CCSVM models behave similar, meaning that when the different between errors are great, the performance of the two mentioned approaches

tends to be same. Therefore, one can truly conclude that the proposed machine boasts in slightly different errors.

5- Conclusions

In this paper, a new formulation of cost-sensitive support vector machine was proposed. Despite of the traditional cost-sensitive model, the objective function of the proposed model considered only misclassified data points instead of in margin ones for minimizing the error rate. This fact was shown that the approach can improve the generalization performance of the support vector machine. In the other hand, the proposed model considered both misclassification cost and accuracy measures simultaneously and so leads to a more accurate model with lower misclassification cost. The performance of this model was compared to traditional cost-sensitive scheme and traditional linear SVM. The results affirmed the superiority of the new machine in terms of the misclassification cost.

References

Cao, P., Zhao, D., & Zaiane, O. (2013, April). An optimized cost-sensitive SVM for imbalanced data learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 280-292). Springer Berlin Heidelberg.

Chen, X. L., Jiang, Y., Chen, M. J., Yu, Y., Nie, H. P., & Li, M. (2012). A Dynamic Cost Sensitive Support Vector Machine. In *Advanced Materials Research* (Vol. 424, pp. 1342-1346). Trans Tech Publications.

Chen, Y., & Wang, J. Z. (2003). Support vector learning for fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, *11*(6), 716-728.

Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, *2*(Dec), 265-292.

Fung, G. M., & Mangasarian, O. L. (2005). Multi category proximal support vector machine classifiers. *Machine learning*, *59*(1-2), 77-97.

Hwang, J. P., Park, S., & Kim, E. (2011). A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. *Expert Systems with Applications*, *38*(7), 8580-8585.

Mangasarian, O. L., & Musicant, D. R. (2001). Lagrangian support vector machines. *Journal of Machine Learning Research*, *1*(Mar), 161-177.

Mangasarian, O. L., & Wild, E. W. (2001). Proximal support vector machine classifiers. In *Proceedings KDD-2001: Knowledge Discovery and Data Mining*.

Nedaie, A., & Najafi, A. A. (2016). Polar support vector machine: Single and multiple outputs. *Neurocomputing*, *171*, 118-126.

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (1999, November). Large Margin DAGs for Multiclass Classification. In *nips* (Vol. 12, pp. 547-553).

Pontil, M., & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE transactions on pattern analysis and machine intelligence*, *20*(6), 637-646.

Qi, Z., Tian, Y., & Shi, Y. (2012). Laplacian twin support vector machine for semi-supervised classification. *Neural Networks*, *35*, 46-53.

Qi, Z., Tian, Y., & Shi, Y. (2013). Robust twin support vector machine for pattern classification. *Pattern Recognition*, *46*(1), 305-316.

Shawe-Taylor, J., & Sun, S. (2011). A review of optimization methodologies in support vector machines. *Neurocomputing*, *74*(17), 3609-3618.

Suykens, J. A., De Brabanter, J., Lukas, L., & Vandewalle, J. (2002). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, *48*(1), 85-105.

Tian, Y., Qi, Z., Ju, X., Shi, Y., & Liu, X. (2014). Nonparallel support vector machines for pattern classification. *IEEE transactions on cybernetics*, *44*(7), 1067-1079.

Tran, Q. A., Li, X., & Duan, H. (2005). Efficient performance estimate for one-class support vector machine. *Pattern Recognition Letters*, *26*(8), 1174-1182.

Turney, P. D. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of artificial intelligence research*, *2*, 369-409.

Vapnik, V. (1998). Statistical Learning Theory, New York, Wiley.

Wan, J. W., Yang, M., & Chen, Y. J. (2012). Cost sensitive semi-supervised Laplacian support vector machine. *Acta Electronica Sinica*, *40*(7), 1410-1415.

Waring, C. A., & Liu, X. (2005). Face detection using spectral histograms and SVMs. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *35*(3), 467-476.

Yang, C. Y., Wang, J. J., Chou, J. J., & Lian, F. L. (2015). Confirming robustness of fuzzy support vector machine via ξ–α bound. *Neurocomputing*, *162*, 256-266.

Zheng, E. H., Li, P., & Song, Z. H. (2006). Cost sensitive support vector machines. *Control and decision*, *21*(4), 473.