One-Against-All Multi-Class SVM Classification Using Reliability Measures

Yi Liu and Yuan F. Zheng
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio 43210
Email:{liuyi, zheng}@ece.osu.edu

***Abstract*— Support Vector Machines (SVM) is originally designed for binary classification. The conventional way to extend it to multi-class scenario is to decompose an M-class problem into a series of two-class problems, for which one-against-all is the earliest and one of the most widely used implementations. One drawback of this method, however, is that when the results from the multiple classifiers are combined for the final decision, the outputs of the decision functions are directly compared without considering the competence of the classifiers. To overcome this limitation, this paper introduces reliability measures into the multi-class framework. Two measures are designed: static reliability measure (SRM) and dynamic reliability measure (DRM). SRM works on a collective basis and yields a constant value regardless of the location of the test sample. DRM, on the other hand, accounts for the spatial variation of the classifier's performance. Based on these two reliability measures, a new decision strategy for the one-against-all method is proposed, which is tested on three benchmark data sets and demonstrates its effectiveness.**

***Index Terms*— Support Vector Machines (SVM), multi-class classification, one-against-all classification, static reliability measure (SRM), dynamic reliability measure (DRM).**

I. Introduction

Support Vector Machines (SVM) is a state-of-the-art learning machine based on the *structural risk minimization* induction principle [1], and has achieved superior performance in a wide range of applications [2]–[4]. However, SVM is originally designed for binary classification and the extension of SVM to the multi-class scenario is still an ongoing research topic [5]. The conventional way is to decompose the M-class problem into a series of two-class problems and construct several binary classifiers. The earliest and one of the most widely used implementations is the one-against-all method, which constructs M SVM classifiers with the ith one separating class i from all the remaining classes. One problem with this method, however, is that when the M classifiers are combined to make the final decision, the classifier which generates the highest value from its decision function is selected as the winner and the corresponding class label is assigned without considering the competence of the classifiers. In other words, the outputs of the decision function are employed as the only index to indicate how strong a sample belongs to the class. The underlying assumption for doing so is that the classifiers are totally trustable and equally reliable, which does not always hold in multi-class cases.

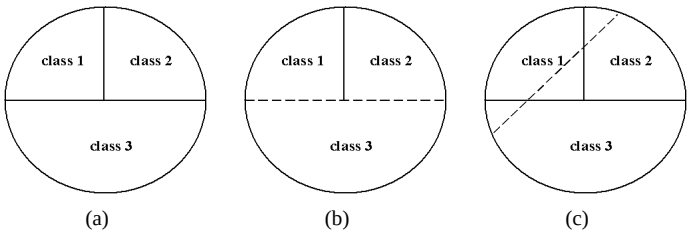

Fig. 1. (a) Three classes and the true boundaries (solid lines). (b) The linear boundary (the dashed line) that separates class 3 and non-class 3. (c) The linear boundary (the dashed line) that separates class 1 and non-class 1.

Fig. 1 shows a 3-class example. The solid lines shown in Fig. 1(a) are the true boundaries. Two linear boundaries obtained using the one-against-all approach are shown as the dashed lines in Fig. 1(b) and Fig. 1(c). Evidentally, the obtained boundary in Fig. 1(b) fits exactly the true boundary and therefore the corresponding classifier is more accurate and reliable than that in Fig. 1(c). However, they are equally trusted at the classification stage by the one-against-all method, which may hurt the overall classification accuracy.

This paper introduces the reliability measure for the binary SVM classifier based on its classification accuracy. Two measures are designed: static reliability measure (SRM) and dynamic reliability measure (DRM). As the name suggests, SRM works in an off-line manner and the result is a constant value regardless of the location of the test sample. DRM, on the other hand, measures the classifier's reliability in a local region surrounding the test sample. As a result, DRM accounts for the spatial variation of the classifier's performance but is not as computationally simple as SRM. Based on these two reliability measures, we suggests to introduce the discrimination among the SVM classifiers and further propose a new decision strategy for the one-against-all approach. The proposed method has been tested on three UCI data sets and better classification performance has been obtained.

The rest of the paper is organized as follows. First, a brief introduction of SVM and the one-against-all approach is given in Section II. In Section III two reliability measures, SRM and DRM, are explained respectively. Section IV presents a new fusion strategy for the one-against-all multi-class SVM classification. The experimental results are given in Section V which is followed by conclusions in Section V.

II. Two-Class Support Vector Machines and the One-Against-All Approach

A. Two-Class SVM

In this section, we give a very brief review of SVM and refer the details to [6] [7]. Consider N training samples: $\{x_1, y_1\}, \ldots, \{x_N, y_N\}$, where $x_i \in R^m$ is a m-dimensional feature vector representing the ith training sample, and $y_i \in \{-1, 1\}$ is the class label of x_i. A hyperplane in the feature space can be described as the equation $w^T x + b = 0$, where $w \in R^m$ and b is a scalar. When the training samples are linearly separable, SVM yields the optimal hyperplane that separates two classes with no training error, and maximizes the minimum distance from the training samples to the hyperplane. It is easy to find that the parameter pair (w, b) corresponding to the optimal hyperplane is the solution to the following optimization problem:

$$\begin{aligned} &\text{minimize :} && L(w) \;=\; \tfrac{1}{2}\,\|w\|^2 \\ &\text{subject to :} && y_i\left(\,w^T x_i \;+\; b\,\right) \geq\; 1,\; i = 1, ..., N. \end{aligned} \tag{1}$$

For linearly nonseparable cases, there is no such a hyperplane that is able to classify every training sample correctly. However the optimization idea can be generalized by introducing the concept of *soft margin*. The new optimization problem thus becomes:

$$\begin{aligned} &\text{minimize :} && L(w, \xi_i) \;=\; \tfrac{1}{2}\,\|w\|^2 \;+\; C\textstyle\sum_{i=1}^{N} \xi_i \\ &\text{subject to :} && y_i\big(w^T x_i) + b\big) \geq 1 - \xi_i, i = 1, \ldots\, N, \end{aligned} \tag{2}$$

where ξ_i are called slack variables that are related to the soft margin, and C is the tuning parameter used to balance the margin and the training error. Both optimization problems (1) and (2) can be solved by introducing the Lagrange multipliers α_i that transform them to quadratic programming problems.

For the applications where linear SVM dose not produce satisfactory performance, nonlinear SVM is suggested. The basic idea is to map x by nonlinearly mapping $\phi(x)$ to a much higher dimensional space in which the optimal hyperplane is found. The nonlinear mapping can be implicitly defined by introducing the so-called kernel function $K(x_i, x_j)$ which computes the inner product of vectors $\phi(x_i)$ and $\phi(x_j)$. The typical kernel functions include the radial basis function $K(x_i, x_j) = \exp\big(-\frac{\|x_i - x_j\|^2}{\sigma^2}\big)$ and the polynomial function $K(x_i, x_j) = \big(x_i^T x_j + 1\big)^d$. If we choose $K(x_i, x_j) = x_i^T x_j$, the non-linear SVM is reduced to its linear version.

At the classification stage, the class label y_{SVM} of a sample x is determined by the sign of the following decision function

$$f(x) = w^T \phi(x) + b = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b. \tag{3}$$

B. One-Against-All Approach

Consider an M-class problem, where we have N training samples: $\{x_1, y_1\}, \ldots, \{x_N, y_N\}$. Here $x_i \in R^m$ is a m-dimensional feature vector and $y_i \in \{1, 2, \ldots, M\}$ is the corresponding class label.

One-against-all approach constructs M binary SVM classifiers, each of which separates one class from all the rest. The ith SVM is trained with all the training examples of the ith class with positive labels, and all the others with negative labels. Mathematically the ith SVM solves the following problem that yields the ith decision function $f_i(x) = w_i^T \phi(x) + b_i$:

$$\begin{aligned} &\text{minimize:} && L(w, \xi_j^i) \;=\; \frac{1}{2}\|w_i\|^2 + C\sum_{l=1}^{N} \xi_j^i \\ &\text{subject to:} && \tilde{y}_j(w_i^T\, \phi(x_j) + b_i) \geq 1 - \xi_j^i,\; \xi_j^i \geq 0, \end{aligned} \tag{4}$$

where $\tilde{y}_j = 1$ if $y_j = i$ and $\tilde{y}_j = -1$ otherwise.

At the classification phase, a sample x is classified as in class i^* whose f_{i^*} produces the largest value

$$i^* = \underset{i=1,\ldots,M}{\arg\max}\; f_i(x) = \underset{i=1,\ldots,M}{\arg\max}\; (w_i^T \phi(x) + b_i). \tag{5}$$

III. Reliability Measures for Two-Class SVM

The performance of a classifier is evaluated by the generalization error R, which is defined as $R = E[Y = \text{sign}(f(X))]$, where $Y \in \{-1, 1\}$ is the true class label of X and f is the decision function. Obviously, a classifier should be considered more reliable if it yields smaller R than the other. Unfortunately R is always not known.

A. Static Reliability Measure

Using the training error R_{emp} to estimate R is a straightforward method which has been adopted in many applications. However, as pointed out in [6] [7], when the number of the training samples is relative small with respect to the dimensionality of the feature vector X, a small R_{emp} does not guarantee a small generalization error R. An upper bound of R is given in [6] [7] and one advantage of SVM is that minimizing the objective function will also minimize this upper bound [6] [7]. In other words, smaller objective function means smaller generalization error, or a more reliable classifier. Following this idea, we rewrite the objective function of SVM as

$$\text{OBJ} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\big(1 - y_i f(x_i)\big)_+, \tag{6}$$

where $(u)_+ = u$ if $u \geq 0$ and 0 if $u \leq 1$, and propose a reliability measure as

$$\lambda_{\text{SRM}} = \exp\Big(-\frac{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\big(1 - y_i f(x_i)\big)_+}{\sigma}\Big), \tag{7}$$

where $f(x_i) = w^T x_i + b$. The parameter $\sigma = CN$ is introduced as a normalization factor to offset the effect of the different regularization parameter C and training size N.

For the linear separable case where $\big(1 - y_i f(x_i)\big)_+ = 0$ for all training samples, the measure λ_{SRM} is reduced to

$$\lambda_{\text{SRM}} = \exp\Big(-\frac{\|w\|^2}{2CN}\Big). \tag{8}$$

Recall that $\frac{2}{\|w\|^2}$ is the classification margin. The classifier with smaller $\|w\|$, which corresponds to larger margin, is considered to be more accurate in generalization and therefore its reliability measure λ_{SRM} is larger.

Note the test sample x does not appear in Eq. (7) and thus λ_{SRM} is the same for all samples. For this reason, it is named *static reliability measure*. The computational load of SRM is not high. When the number of support vectors is relatively smaller than the training size N, which happens most of the time, the complexity of SRM is O(N).

B. Dynamic Reliability Measure

SRM assumes the classifier to be equally effective throughout the entire feature space. In reality, the classifier's performance exhibits spatial variation [8], to accommodate which we extend SRM to DRM, a dynamic reliability measure. The basic idea is to estimate the classifier's reliability in a *local region* of feature space surrounding the test sample x. Here the local region, denoted as $N_k(x)$, is defined as the training samples that compose the k-nearest neighbors of x. Moreover, we are especially interested in the reliability of the classifier with respect to certain output (1 or -1 in this case).

Suppose $C(x) \in \{1, -1\}$ is the class label assigned to x by the SVM classifier. Let $N_k^{C(x)}(x)$ denote the set of the training samples that locate among the k nearest neighbors of x and are classified to the same class as x

$$N_k^{C(x)}(x) = \{\hat{x}_j | \hat{x}_j \in N_k(x) \text{ and } C(\hat{x}_j) = C(x)\}. \quad (9)$$

By rewriting Eq. (6) as

$$\text{OBJ} = \sum_{i=1}^{N} \Big(\frac{\frac{1}{2}\|w\|^2}{N} + C\big(1 - y_i f(x_i)\big)_+ \Big) = \sum_{i=1}^{N} \text{OBJ}(x_i), \quad (10)$$

we make the training sample x_i contributes to the overall OBJ by OBJ(x_i). In analogy to Eq. (6) which takes the summation of OBJ(x_i) on all samples, we formulate the local version of OBJ as

$$\begin{aligned} \text{OBJ}_{\text{local}} &= \sum \text{OBJ}(\hat{x}_j) \\ &= \sum_{j=1}^{k_x} \Big(\frac{\|w\|^2}{2N} + C\big(1 - \hat{y}_j f(\hat{x}_j)\big)_+ \Big) \\ &= \frac{||w||^2 \cdot k_X}{2N} + C \sum_{i=1}^{k_X} \big(1 - \hat{y}_i f(\hat{x}_i)\big)_+, \quad (11) \end{aligned}$$

where $\hat{x_j} \in N_k^{C(x)}(x)$, $(\hat{x_j}, \hat{y_j})$ is the training pair, and k_x is the number of training samples in the set $N_k^{C(x)}(x)$. Now with $\text{OBJ}_{\text{local}}$ at hand, we can compute the reliability of the decision "x belongs to $C(x)$" by

$$\begin{aligned} \lambda_{\text{DRM}}(x) &= \exp\Big(-\frac{\text{OBJ}_{\text{local}}}{C \cdot k_X}\Big) \\ &= \exp\Big(-\big(\frac{||w||^2}{2CN} + \frac{\sum_{i=1}^{k_X}(1 - \hat{y}_i f(\hat{x}_i))_+}{k_X}\big)\Big) \quad (12) \end{aligned}$$

From the derivations above we can see that, unlike λ_{SRM}, $\lambda_{\text{DRM}}(x)$ is a dynamic function of x, which varies depending on the location of the test sample and the classified label $C(x)$. For this reason, DRM has to be recomputed as new samples come in and is more expensive than SRM. The extra cost is consumed by finding the k nearest neighbors, which is O(N).

IV. New Decision Rule for One-Against-All

Based on the SRM and DRM discussed above, we propose a new decision rule for the one-against-all method.

Suppose that we have trained M support vector machines SVM_1, SVM_1,..., SVM_M, each of which has the decision function f_l. First, we evaluate f_l at the given sample x using Eq. (3), and generate a soft decision $y_{f_l} \in [-1\ 1]$ assuming the classifier is completely trustable

$$y_{f_l} = \text{sign}(f_l(x))(1 - \exp^{|f_l(x)|}). \quad (13)$$

Note that y_{f_l} carries two kinds of information: the sign part encodes the hard decision on "x belongs to class l or not" and its absolute value represents how strong the decision is. The farther x locates away from the decision boundary which yields larger $f_l(x)$, the stronger the decision. Then, instead of comparing y_{f_l} directly, we weight y_{f_l} by the liability measures as $\bar{y}_{f_l} = y_{f_l}\lambda_l$, where λ_l denotes the SRM or DRM of SVM_l. Finally, the sample x is classified as in class l^* that yields the largest $\bar{y}_{f_l}$

$$l^* = \arg\max_{l=1,\ldots,M} \ \bar{y}_{f_l}. \quad (14)$$

It is can be shown that when λ_l are equal, Eq. (14) becomes

$$l^* = \arg\max_{l=1,\ldots,M} \ y_{f_l} = \arg\max_{l=1,\ldots,M} \ f_l(x), \quad (15)$$

which reduces to the decision rule of the conventional one-against-all method.

V. Experimental Results

The proposed approach has been applied to the multi-class data sets obtained from the UCI repository of machine learning [9], and shows its advantage over the conventional one-against-all method. This section reports the experimental results on three data sets which exhibit certain varieties, i.e., the number of classes to be differentiated and the kernel function to be used, which are listed in Table I.

The first data set is the image segmentation data, where each sample has 19 continuous attributes collected from a 3×3 region of an outdoor image. There are seven classes to classify: *brickface*, *sky*, *foliage*, *cement*, *window*, *path*, and *grass*. The training set consists of 210 samples with 30 per class while the size of the test set is 2100 with 300 samples per class.

The polynomial kernel with $d = 1$, which has been reported as a good choice for this data set [10], is first adopted. Fig. 2(a) shows the classification errors yielded by using the decision functions $f(x)$ (the conventional one-against-all method), SRM and DRM. The errors are plotted by including the classes one by one (in alphabet order). As one can see, SRM and DRM always lower the error percentage and DRM performs the best. In order to see how the new approach performs when the kernel function is not well chosen, we replace the kernel function with the 2-degree polynomial. As expected, the one-against-all method degrades a lot since in this case the classifier is not so trustable any more. Meanwhile, the proposed method degrades too. However, with the the classifier's reliability taken into account, the magnitude of the degrading is much smaller (Fig. 2(b)).

The iris plant data set is a small set yet one of the best known data sets to be found in the pattern recognition literature. It contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The class *setosa* is linearly separable from *Versicolour* and *Virginica* while the latter two are linearly nonseparable from each other. Linear SVM is sufficient for this set, and the errors obtained by leave-one-out cross validation are plotted in Fig. 2(c).

The last data set is the letter recognition data. This data set contains 20000 samples, each of which corresponds to one of the 26 capital letters in the English alphabet. 16 integer-valued features are provided to represent each letters. Typically

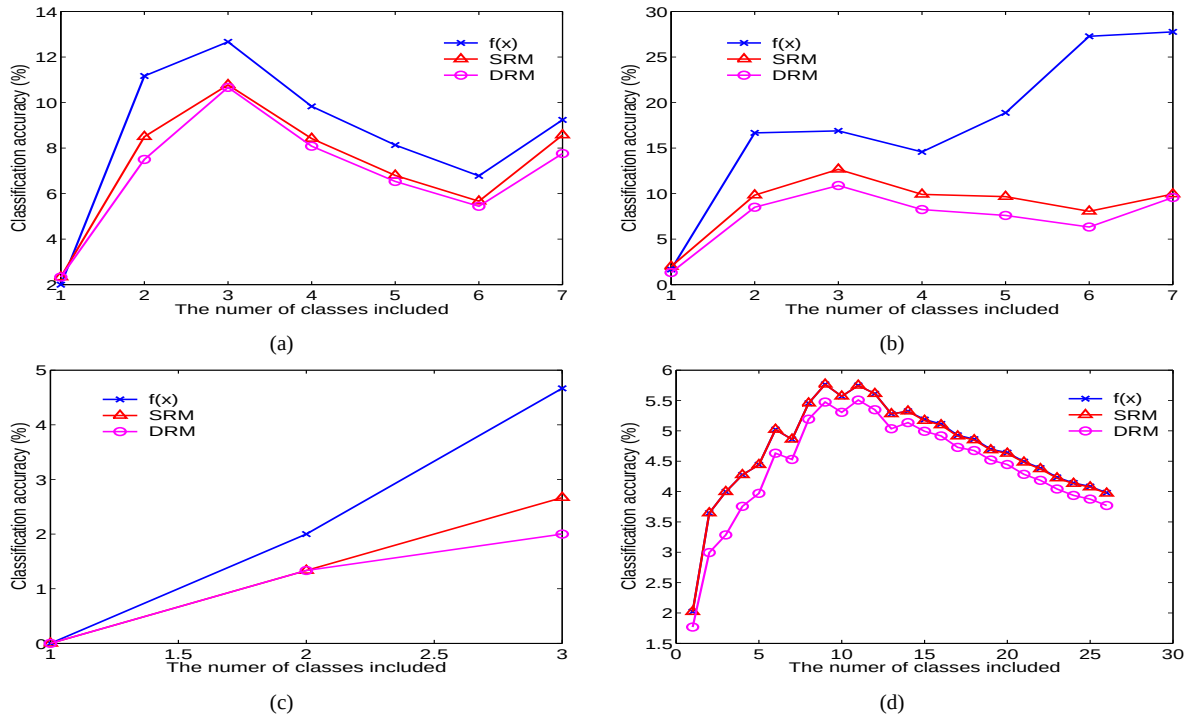

Fig. 2. The comparison of the performance by including the classes one by one. (a) Image segmentation data set using 1-degree polynomial as the kernel function. (b) Image segmentation data set using 2-degree polynomial as the kernel function. (c) Letter recognition data set using RBF as the kernel function. (d) Iris Plant data set using linear kernel function.

TABLE I

COMPARISON OF CLASSIFICATION ERRORS (BOLDFACE INDICATES THE BEST PERFORMANCE).

	number of classes	number of attributes	kernel function	classification errors (%)			relative error reduction	
				$f(x)$	SRM	DRM	SRM	DRM
Image Segmentation[1]	7	19	1-degree polynomial	9.24	8.57	**7.76**	7.25%	16.0%
Image Segmentation[2]	7	19	2-degree polynomial	27.8	9.95	**9.57**	64.2%	65.6%
Letter Recognition	26	16	RBF	3.98	3.97	**3.77**	0.25%	5.3%
Iris Plant	3	4	linear	4.0	2.67	**2**	33.3%	50%

the first 16000 samples are used as the training data and the remaining 4000 as the test data. After experimenting with different kernel functions, the RBF is found to be the best choice for this 26-class problem. Yielding a total of 3.98% misclassifications, all the 26 classifiers are very competent, and SRM and DRM are only able to reduce the errors to 3.97% and 3.77% respectively.

VI. CONCLUSIONS

One-against-all, which constructs M binary classifiers to differentiate each class from the rest, is a conventional method to extend SVM from the binary to M-class classification. At the classification stage, the test sample is assigned to the class whose decision function produces the largest value, implicitly assuming that all the SVM classifiers are equally reliable which fails in many situations. This paper proposes two methods, SRM and DRM, to measure the reliability of classifiers based on their estimated generalization accuracy. This paper also suggests a new decision strategy for one-against-all method that introduces the discrimination among the SVM classifiers based on the reliability measures. The experimental results have demonstrated that the proposed approach is able to improve the classification accuracy without imposing high computational cost.

REFERENCES

[1] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support Vector Machines", *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.

[2] M. Pontil and A. Verri, "Support Vector Machines for 3D Object Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637-646, June 1998.

[3] D.J. Sebald and J.A. Bucklew, "Support Vector Machine Techniques for Nonlinear Equalization", *IEEE Tran. on Signal Processing*, vol. 48, no. 11, pp. 3217-3226, Nov. 2000.

[4] G.D. Guo, A.K. Jain, W.Y. Ma, and H.J. Zhang, "Learning Similarity Measure for Natural Image Retrieval with Relevance Feedback", *IEEE Tran. on Neural Networks*, vol. 13, no. 4, pp. 811-820, July 2002.

[5] C.W. Hsu, and C,J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines", *IEEE Tran. on Neural Networks*, vol. 13, no. 2, pp. 415-425, Mar. 2002.

[6] V.N. Vapnik, "An Overview of Statistical Learning Theory", *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 988-999, Sept. 1999.

[7] C. Cortes and V.N. Vapnik, "Support Vector Networks", *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[8] K. Woods, W.P. Kegelmeyer, and K. Bowyer, "Combination of Multiple Classifier Using Local Accuracy Estimates", *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 405-410, April 1997.

[9] C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases", Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/ mlearn/MLRepository.html, 1998.

[10] J.T. Kwok, "Moderating the Outputs of Support Vecotr Machine Classifiers", *IEEE Tran. on Neural Networks*, vol. 10, no. 5, pp. 1018-1013, Sept. 1999.