

# A hierarchical Clustering Method Based on PCA-Clusters Merging for Quasi-linear SVM

Cheng Yang<sup>1</sup>, Keshi Yang<sup>2</sup> & Bo Zhou<sup>3</sup>

<sup>1,2</sup> No. 1 Eastern rd, Feixian Eco Dev Zone, Linyi city, Shandong province, P.R.China.

<sup>3</sup>Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka, Japan

**Keywords:** hierarchical clustering; principal component analysis; quasi-linear kernel; clusters merging.

**Abstract.** This paper proposes an improved hierarchical clustering method based on PCA-clusters merging for quasi-linear SVM. The quasi-linear SVM is an SVM with quasi-linear kernel. It considers a nonlinear separating boundary between class labels as an approximation of multiple local linear boundaries with interpolation and the quasi-linear kernel is composited based the information of local clusters along the boundary. In order to obtain the local clusters, the proposed clustering method, first detects the nonlinear boundary based on the changes of class labels; then obtains small partitions along the nonlinear separating boundary using a hierarchical clustering; and further merges the nearest neighboring clusters distributed in one local linear boundary into one cluster according clusters distributed in one local linear boundary according to PCA-based criterion. The quasi-linear kernel is composited based on the information of local clusters. Experimental results on benchmark datasets demonstrate that the proposed method improves the classification performance efficiently.

## 1.Introduction

Support Vector Machines (SVMs) have been widely used in a variety of application areas. SVMs can solve linearly inseparable classification tasks by mapping the input data into a higher-dimensional feature space through the kernel trick an seeking an optimal separating hyper-plane in the feature space [1]. SVMs with nonlinear kernel functions(RBF, polynomial functions) also face the general over-fitting issue when the number of training samples is small due to their large VC dimensions.

Conventional nonlinear SVM models can not obtain good classification consequences in some genomics datasets(such as FunCat Yeast datasets [2] and Genbase motif-based datasets [3]) because of the disadvantage of having less number of samples in training subspace. In this paper, we propose a quasi-linear SVM model based on composite kernel for solving these classification tasks. The main problem of classification is identifying the separation boundary between different class labels. The proposed quasi-linear SVM model approximately consider a nonlinear separating boundary as an approximation of multi-local linear boundaries with interpolation, in which local linear property of nonlinear boundary is used to construct a quasi-linear kernel. It is very important to preserve the local characteristics of the separating boundary. We propose a hierarchical clustering method follows “divide and conquer” policy and composed of separating boundary detection and clustering technique. The prior knowledge of local subsets is used to construct a composite kernel, and the quasi-linear SVM is implemented by using the composite kernel. This method can find the cluster which can cover a local linear segment of separating boundary and obtain better performance on classification.

In some previous works, ref. [4] presented localized support vector machine which builds multiple linear SVM models from training data and each model is designed to classify a particular test example. Ref. [5] introduced mixtures of linear SVMs by packaging linear SVMs into a probabilistic formulation and embedding them in the mixture of expert model.

The rest parts of the paper are organized as follows: Section 2 describes the overview of the quasi linear SVM model based on composite kernel. The estimation of quasi-linear SVM model has

been discussed in Section 3, Section 4 implements the experiment and results and finally the conclusions in Section 5.

## 2. Quasi-linear SVM with a Composite Kernel

There are labeled training data points of  $N$  samples  $(x_1, y_1), \dots, (x_i, y_i), \dots, (x_N, y_N)$ ,  $x_i \in R^d$  is the input vector corresponding to the  $i$ -th sample labeled by  $y_i \in \{-1, +1\}$  depending on its class.

A nonlinear separating boundary  $f_p(x)$  can which known as a priori knowledge be approximated by  $M$  local linear boundaries with interpolation as showed in Function.1.

$$f_p(x) = \sum_{j=1}^M (\Omega_j^T x + b_j) R_j(x) + b \quad (1)$$

where  $R_j(x)$ 's are interpolation function,  $\Omega_j$  are the parameter vectors of local linear boundaries.

Introducing two vectors  $\Phi(x)$  and  $\Theta$  defined by

$$\Phi(x) = [R_1(x), xR_1(x), \dots, R_M(x), xR_M(x)]^T \quad (2)$$

$$\Theta = [b_1, \Omega_1^T, \dots, b_M, \Omega_M^T]^T$$

$$\text{We further express Eq.(1) as } f_p(x) = \Theta^T \Phi(x) + b \quad (3)$$

Considering the structural risk minimization principle into Eq. (3), the classification problem can be described as QP optimization problem as following:

$$\min_{\Theta, b, \xi} J_p = \frac{1}{2} \Theta^T \Theta + c \sum_{k=1}^N \xi_k \quad s.t. \quad \begin{cases} y_k [\Theta^T \Phi(x_k) + b] \geq 1 - \xi_k, k = 1, \dots, N \\ \xi_k \geq 0, k = 1, \dots, N \end{cases} \quad (4)$$

The Lagrange function has been constructed, via introducing Lagrange multipliers  $(\alpha_k, v_k)$ :

$$L(\Theta, b, \xi; \alpha, v) = J_p(\Theta, \xi) - \sum_{k=1}^N (\alpha_k y_k [\Theta^T \Phi(x_k) + b] - 1 + \xi_k) - \sum_{k=1}^N v_k \xi_k \quad (5)$$

with Lagrange multipliers:  $\alpha_k \geq 0, v_k \geq 0, k = 1, \dots, N$ .

The solution is given by the saddle point of the Lagrange function:  $\max_{\alpha, v} \min_{\Theta, b, \xi} L(\Theta, b, \xi; \alpha, v)$ .

The dual problem becomes as follow:

$$\max_{\alpha} J_D(\alpha) = -\frac{1}{2} \sum_{k,l=1}^N y_k y_l K(x_k, x_l) \alpha_k \alpha_l + \sum_{k=1}^N \alpha_k \quad s.t. \quad \begin{cases} \sum_{k=1}^N \alpha_k y_k = 0 \\ 0 \leq \alpha_k \leq c, k = 1, \dots, N \end{cases} \quad (6)$$

In the quadratic form, the kernel trick is applied.

$$K(x_k, x_l) = \Phi(x_k)^T \Phi(x_l) = (1 + x_k^T x_l) \sum_{j=1}^M R_j(x_k) R_j(x_l). \quad (7)$$

For  $k = 1, \dots, N$ . The nonlinear separating boundary model  $f_p(x)$  is reduced to a standard SVM based on a composite kernel (Eq. 8). Finally the nonlinear SVM classifier takes the form:

$$y = \text{sign}[\sum_{k=1}^N \alpha_k y_k K(x, x_k) + b]. \quad (8)$$

with  $\alpha_k$  positive real constants which are the solution to a QP problem.

The quasi-linear kernel  $K(x_k, x_l)$  (Eq. 7) is the inner product of explicit nonlinear mapping.

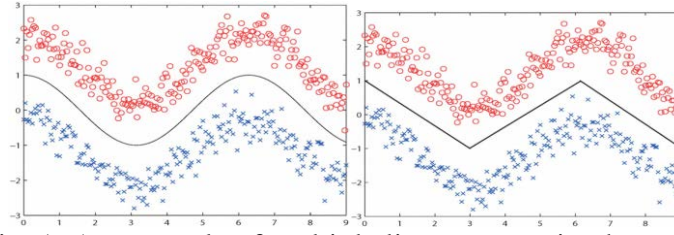


Fig. 1: An example of multiple linear separating boundaries

### 3. Estimation of the Quasi-linear SVM model

#### 3.1 Hierarchical Clustering and PCA

The quasi-linear SVM formulate the nonlinear separating boundary between classes labels is approximately as a combination of local linear boundaries with interpolation. We propose a hierarchical clustering method to obtain clusters with local linear characteristics.

In this paper we use a simple and stable clustering method which is called hierarchical clustering. The results of hierarchical clustering are usually presented in a dendrogram, the result of the method is uniform and symmetrical. The local properties of points depend on how similar they are to each other. This algorithm will traverse all the objects for finding the least dissimilar pair of objects and merging them into a single cluster. Principal Component Analysis (PCA) is introduced for analyzing the local property of each neighboring cluster and obtain multi-dimensional principal component.

#### 3.2 Detect the Nonlinear Separating Boundary

The proposed partitioning approach consists of separating boundary detection and cluster merging technique is proposed to partition training samples in to clusters with local linear characteristics. A separating boundary detection method considering sample label changes in neighbor area is used to select samples which are near to the separating boundary firstly. The cluster merging technique is used to partition the selected separating boundary into local subsets for estimating parameters of SVM model.

We take an example on synthetic data for demonstrating the proposed guided partitioning approach. As shown in Fig. 2-1, a two-class synthetic training samples have 339 samples in each class. By implementing unsupervised clustering method on training data may emerge mussy and unintended subsets. The partitioning demonstration of implementing hierarchical clustering (the number of cluster is set as 4 empirically) on synthetic data (678 samples) is shown in Fig. 2-2, where different clusters are plotted in various colors. The clusters can't capture distribution information of local subsets properly.

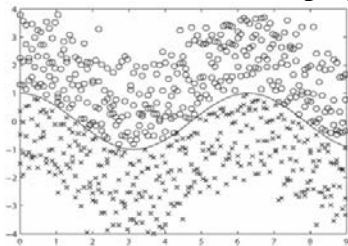


Fig. 2-1

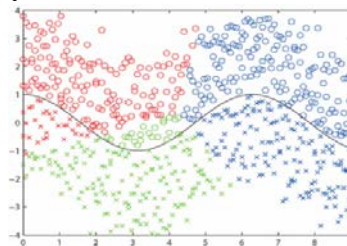


Fig. 2-2

Fig. 2: The example of unsupervised clustering method. (1) the synthetic data (678 samples); (2) the clustering results of applying hierarchical clustering

In order to solve above problem, in the proposed partitioning method, a separating boundary detection technique is introduced to filter samples before implementing cluster merging. Samples around the boundary are selected by detecting changes of different class labels. The detecting procedure is described as follows. For a certain training sample P, seek the n nearest neighbor

samples, if one of these neighbor samples has a different class label, this sample P is selected as a sample around the separating boundary to be reserved. The result of detecting process on training samples can produce a subset of samples that distribute around the separating boundary between class labels. This detection can filter out information which may be considered as less relevant, while conserving the important distribution properties near to the separating boundary.

### 3.3 Cluster Merging according to the PCA Analysis

Then, the main line trend of each cluster in high dimension can be approximately described by principal component of each cluster with adopting PCA method. The included angle between principal components is transformed into calculating multidimensional angles between the vectors. If the include angle between these two neighboring clusters is less than a threshold value, the merging operation will be implemented through the hierarchical clustering method among neighboring clusters. Then, repeat the merging process after all the neighboring clusters are identified. In this case, cluster can exactly cover the separating boundary on each segment. In addition, the merge of minority clusters also avoid the over-fitting problem on edge detection.

### 3.4 Training of Local SVMs

In simulation, the negative Euclidean distance is used to measure similarity as common cases. The shared value is set as the median of the input similarities. Then the centers have already obtained. K is the number of clusters. Obviously, the set of centroids  $D_c = \{c_1, \dots, c_k\}$  is also a sub set of the training data set.

The clusters  $D_{Ei}$  and the set of centroids  $D_c$ , training data sets  $D_i$  for local SVMs are obtained:  $D_i = D_{Ei} \cup D_c, i=1, \dots, M$ ,  $M$  is the number of datasets.

Assume input-target training pairs of the  $m$ -th training dataset  $D_m$  is denoted as  $\{x_{mi}, \psi_{mi}\}$ ,  $m=1, \dots, M$ .  $x_{mi}$  is the  $i$ -th input vector and  $\psi_{mi}$  is class label.

In order to find the separating boundary in the training data set  $D_m$ , the input data set is firstly transformed from a low-dimension space to a high-dimensional feature space by a nonlinear mapping function  $\varphi(\bullet)$  [6]. By introducing a vector of Lagrange multipliers  $\alpha = (\alpha_1, \dots, \alpha_N)$ , the optimal separating boundary task is constructed as a QP optimization problem in the dual space[7]:

$$\max_{\alpha} Q(\alpha) = -\frac{1}{2} \sum_{i,j=1}^N \psi_{mi} \psi_{mj} K(x_{mi}, x_{mj}) \alpha_i \alpha_j + \sum_{j=1}^N \alpha_j \text{ s.t. } \begin{cases} \sum_{i=1}^N \alpha_i \psi_{mi} = 0 \\ 0 \leq \alpha_i \leq C, \forall i \end{cases} \quad (9)$$

where  $K(x_{mi}, x_{mj}) = \varphi(x_{mi})^T \varphi(x_{mj})$  is the kernel function[8].

Obtain the vector of Lagrange multipliers  $\alpha$ . If  $\alpha_i \neq 0$ , sample-I is a SV[9].

The local SVM decision function by using the  $m$ -th training data is obtained:

$$f_m(x) = \text{sign}[\sum_{i=1}^N \alpha_i \psi_{mi} K(x, x_{mi}) + b]. \quad (10)$$

where  $x_{mi}$  are samples from the training data and  $x$  is the test vector. The local SVMs are used as local separating boundary clusters with interpolation.

## 4. Experiment and Results

First, we present on the proposed method on the synthetic Gaussian data set at begin. Second, we carry out extensive experiments on 5 benchmark datasets.

#### 4.1. Synthetic Data Set

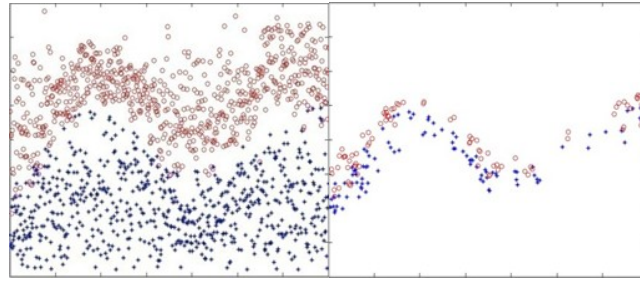


Fig. 3-1

Fig. 3-2

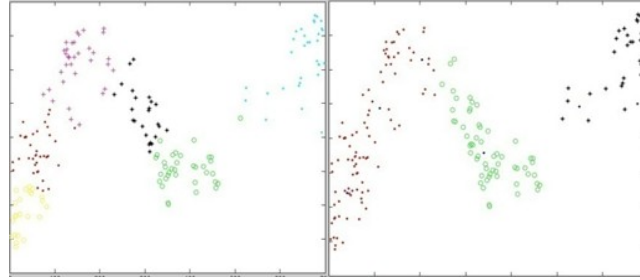


Fig. 3-3

Fig. 3-4

Fig. 3: the merging process of the proposed method, Fig. 3-3 shows selected samples (219 samples) around the separating boundary detection. Fig.3-3and 4 shows clustering results of AP clustering and proposed method.

Firstly, we compare SVM based on conventional clustering methods and SVM based on proposed method on the synthetic Gaussian data set. In datasets, training and testing sets both have 700 samples. We can easily see from Fig. 3-4 that there are three clusters and each of them can cover a local linear boundary.

#### 4.2. Simulation on Benchmark Data Set

In order to evaluate the performance of the method further, we compare different algorithms on 4 well-known benchmark problems. These data sets include 4 benchmark real-world data sets from the University of California at Irvine (UCI) Machine Learning Repository[10]. They are Pima: Pima Indians Diabetes dataset; Breast: Breast Cancer Wisconsin dataset; Heart: Heart Disease dataset and Yeast: Yeast dataset. A summary of datasets is presented in Table 1.

Table 1: the summary of four benchmark data sets

Data	No. Train	No. Test	Positive %	Negative %	dimension
Pima	500	268	268	500	8
Breast	500	199	241	458	10
Heart	170	100	120	150	13
Yeast	625	339	496	468	436

Four classical evaluation metrics of Accuracy, Precision, Recall and F-score are used to evaluate the efficiency of proposed method. The formula is:

$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$ ;  $Precision = \frac{TP}{TP + FP}$ ;  $Recall = \frac{TP}{TP + FN}$ ;  $F-score = \frac{2 * Precision * Recall}{Precision + Recall}$  TP the number of true positives (correctly predicted positive samples), FP the number of false positives (positive predictions that are incorrect), and FN the number of false negatives (positive that are incorrectly predicted negative).

In our experiments, the LibSVM[11] is taken as a basis. In guided partitioning approach, we set neighbor parameter  $n=5$  for separating boundary detection. About the parameters of AP clustering (defined in Ref. [12]), the maximum number of iterations is set as 1000, and early terminate parameter is set as 100; the damping factor, which may be needed if oscillations occur, is set as 0.9; Euclidean distance is used for the similarity metric of samples.

Table 2: the comparisons between the AP and the proposed method

	AP				Proposed method			
	Acc.	Prec.	Rec.	Fscore	Acc.	Prec.	Rec.	Fscore
Pima	75	0.592	0.701	0.646	78.3	0.637	0.756	0.692
Breast	95.9	0.952	0.952	0.952	96.5	0.952	0.964	0.959
Heart	85	0.837	0.818	0.827	87	0.86	0.84	0.85
Yeast	69.7	0.723	0.742	0.732	71.5	0.755	0.731	0.75

Table 3: the comparisons between the K-mean and the proposed method

	K-mean				Proposed method			
	Acc.	Prec.	Rec.	Fscore	Acc.	Prec.	Rec.	Fscore
Pima	73.5	0.5777	0.651	0.612	78.3	0.637	0.756	0.692
Breast	91.9	0.936	0.869	0.9012	96.5	0.952	0.964	0.959
Heart	86	0.857	0.818	0.837	87	0.86	0.84	0.85
Yeast	67.9	0.781	0.6	0.678	71.5	0.755	0.731	0.75

We compared our method with the k-mean clustering and AP clustering method. The experimental results for four labels are presented in Tab.2.

## Conclusions

In this paper, we present a hierarchical clustering method of PCA-based cluster merging as a preprocessing with the quasi-linear SVMs. The proposed method which partitions the nonlinear separating boundary and merges the clusters in a local linear separating boundary by using improved hierarchical clustering method, is demonstrated effectively for approximately express the nonlinear separating boundary.

The experimental results demonstrate that the proposed method can solve the nonlinear classification tasks in biological functional classification problem, which conventional nonlinear kernel SVMs can not work efficiently. In the future, we will implement quasi-linear SVM in computer image and vision fields.

The research work was supported by instruction research project of Qingdao Technological University under Grant No. 13-2.

## References

- [1] V. Vapnik: The Nature of Statistical Learning Theory. Springer, Verlag Berlin (1999)
- [2] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya in: Hierarchical multi-label prediction of gene function, vol. 22(7), pp. 830-836, (2006).
- [3] S. Diplaris, G. T soumakas, P. Mitkas, and I. V lahavas, in Protein classification with multiple algorithms, in Proc. of the 10th Panhellenic Conference on Informatics (PCI'05)(Volos, Greece), pp. 448-456,(2005).
- [4] H. Cheng, P. Tan, and R. Jin in: Efficient algorithm for localized support vector machine, IEEE Transactions on Knowledge and Data Engineering, Vol. 22(4), pp. 537-549(2010)
- [5] Z. Fu, A. Robles-Kelly: On mixtures of linear SVMs for nonlinear classification, Lecture Notes in Computer Science, Vol. 5342 ,p.489-9 (2010)
- [6] E. Osuna, R. Freund and F. Girosi, Support Vector Machines: Training and Applications. A.I. Memo 1602, MIT A.I.Lab, (1997)
- [7] J. Suykens: Least Squares Support Vector Machines. Tutorial IJCNN, (2003)
- [8] A. J. Smola and B. Schölkopf, On a kernel based method for pattern recognition, regression, approximation and operator inversion. Algorithmica, 1998. Em Technical Report 1064, GMD FIRST, (April 1997)
- [9] S. R. Gunn: Support Vector Machines for Classification and Regression. Technical Report, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, 10 May, (1998)

- [10] P. M. Murphy, and D. W. Aha: UCI Repository of Machine Learning Databases. technical report, Department of Information and Computer Science, University of California, Irvine, California. Available: <http://archive.ics.uci.edu/ml/>
- [11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines, Software at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, (2001)
- [12] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," Science, vol. 315, pp. 972-976, (2007).