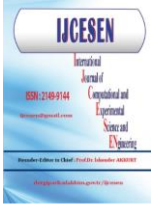
International Journal of Computational and Experimental Science and Engineering
(IJCESEN)
Vol. 4-No.1 (2018) pp. 1-5
http://dergipark.gov.tr/ijcesen



ISSN: 2149-9144

Research Article

Comparison of Different Training Data Reduction Approaches for Fast Support Vector Machines Based on Principal Component Analysis and Distance Based Measurements

Gür Emre GÜRAKSIN[1]*, Harun UĞUZ[2]

[1]Afyon Kocatepe University, Engineering Faculty, Biomedical Engineering Department, Afyonkarahisar-Turkey
[2]Selcuk University, Engineering Faculty, Computer Engineering Department, Konya-Turkey

* Corresponding Author: emreguraksin@aku.edu.tr
ORCID: 0000-0002-1935-2781

(First received 03 January 2018 and in final form 07 March 2018)

Keywords
Support Vector Machines
Principal Component Analysis
Machine Learning

Abstract: Support vector machine is a supervised learning algorithm, which is recommended for classification and nonlinear function approaches. Support vector machines require remarkable amount of memory with time consuming process for large data sets in the training process. The main reason for this is the solving a constrained quadratic programming problem within the algorithm. In this paper, we proposed three approaches for identifying the non-critical points in training set and remove these non-critical points from the original training set for speeding up the training process of support vector machine. For this purpose, we used principal component analysis, Mahalanobis distance and Euclidean distance based measurements for the elimination of non-critical training instances in the training set. We compared the proposed methods in terms of computational time and classification accuracy between each other and conventional support vector machine. Our experimental results show that principal component analysis and Mahalanobis distance based proposed methods have positive effects on computational time without degrading the classification results than the Euclidean distance based proposed method and conventional support vector machine.

1. Introduction

Support vector machines (SVMs), developed by Vapnik [1], have been widely used because of their high generalization to a wide range of applications like text recognition, object recognition, sound recognition and face recognition. The main idea behind SVMs is the optimal separating hyperplane which divides the classes [2]. The optimal separating hyperplane is the constructed decision boundary, which should be the maximum distance from the training data points of the both classes. It can be understood from SVM theory that the equation of this separating hyperplane is determined by the data points close to the hyperplane which are called support vectors [3]. Identifying support vectors from the labeled training set of samples requires the use of an iterative process such as quadratic programming. So finding the classification hyperplane is computationally very expensive. The most important problem of training a SVM is its slow training process for classification problems with large data sets. Therefore, when the number of the training sample set is huge, sometimes it is impossible to use all of the samples for training [2,4]. As such, there is a need to speed up the training process.

In the literature, there are studies in which principal component analyses (PCA) and SVM are used together. In these studies, PCA is widely used for the feature extraction process [5, 6, 7]. In this study, we present three different methods based on PCA, Mahalanobis Distance (MD) and Euclidean Distance (ED) measurements for the reduction of training data of the SVM. By the help of the proposed methods, the non-critical training instances in the training set are removed. Our experimental results show that PCA and MD based proposed methods have positive effects on computational time without degrading the classification results than the Euclidean distance based proposed method and conventional support vector machine.

2. Materials and Methods

In this study, three algorithms based on PCA, MD and ED measurements were proposed for identifying the non-critical points in training set and remove these non-critical points from the original training set for speeding up the training process of SVM. The proposed algorithms were applied four different data sets and the results were compared in terms of computational time and classification accuracy by conventional SVM.

2.1 Data Sets

In this study we used four different data sets for evaluating the performance of the system in terms of computational time and classification accuracy. The first data set used in this study was a bone age data set. The bone age data set consists of x-ray images of children. The X-ray images used in this study were taken from a database related to the study performed by Gertych et al. in 2007 [8] which were collected from a pediatric hospital in Los Angeles, and the images are available via a web site for academic research and educational purposes. For the classification process six features in total were used. The image processing stages and the features extracted by the image processing are described in details in previous study performed by Güraksın et. al. [9].

The other three data sets are well known data sets and publicly available test problems of wine, iris and pima Indian diabetes data sets in repository of machine learning databases [10,11]. The wine data set is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. There are totally 178 instances in three different class types. The iris data set is the best-known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. There are four attributes in total and 150 instances. The last data set used in this study was pima Indian diabetes data set. In particular, all patients in this data set are females at least 21 years old of Pima Indian heritage. There are 768 instances and 8 attributes incident to two class.

2.2 Support Vector Machines

SVM was found first by Vapnik in 1979. It was again recommended by Vapnik in 1995 for regression and classification [12]. SVM is a supervised learning algorithm, which is recommended for classification and nonlinear function approaches. Starting to be widely used over the last few years, SVMs have been used in pattern recognition applications like text recognition, object recognition, sound recognition and face recognition.

The main idea in the background of SVM is based on the formation of the equation of Lagrange multipliers. The goal of SVM is to find the optimum separating hyperplane, which is able to classify data points as well as possible and to again separate them into two classification points as much as possible. In other words, it aims to find the state in which the distance between the two classes is the maximum. The hallmarks of this classification reasoning are the support vectors chosen from the training set, and they are located on the closest points of both classes [13]. SVM has a good generalization performance and they work well in practice [14,15]. More information about SVM can be obtained from [12,13].

2.3 Principal Component Analyses

PCA is a statistical technique which is used for extracting information from multivariate data sets. This process is performed with obtaining the principal components of the original variables with linear combinations identified. The first principal component is representing the original data set with maximum variability and the second principal component is representing the remaining data set with maximum variability. This procedure continues consecutively until the last principal component is represented with the remaining data set with maximum variability. It can be understood that the number of the significant principal components of the multi variable data set with the highest variance values among all principal components should be smaller than the number of all principal components. So, PCA can be considered as a data reduction method. In other words, PCA is a technique used to produce a lower-dimensional version of the original data set [16]. More information about PCA can be obtained from [17].

3. Application of Methods Used

In the training phase of the SVM, the SVM is to find the optimum separating hyperplane by using the training set. As we mentioned before, when finding the optimum separating hyperplane, the SVM uses a portion of the training set called support vectors.

Using the entire training set to identify the support vectors from the labeled training set of samples requires the use of an iterative process, such as quadratic programming, that is computationally very expensive. There is therefore no need for the other portion of the training set while training the SVM. In this study for finding the significant portion of the training set for SVM, we proposed some algorithms mentioned in the following sections based on PCA, MD and ED.

3.1 Reduced Training Data for Support Vector Machines Based on Principal Component Analysis

In this section for the reduction of the training samples of SVM, first we proposed a method based on PCA, which is being used as a dimension reduction technique.

In the first stage of our proposed method, we used the PCA method on the samples of the training set for all the classes. There are certain criteria in determining the optimal number of principal components. In our study, the cumulative percentage of variance criteria was used to determine the number of principal components, for its simplicity and high performance [18]. Using this criterion, the principal component with the highest variance is chosen. By the help of this process, our multi-dimensional space is converted to a one-dimensional space.

After the PCA process, for each class the mean values and the standard deviations of the one-dimension data gained from PCA is calculated. SVM is a binary classifier. So for the multiclass classification process, it classifies each class separately. Assume that the mean values of the one-dimensional data of class 1 is M1 and that of the other class is M2 and the standard deviations of each class is Std1 and Std2. In this step, the mean values of each class are compared. If M1<M2, the original training data which has a smaller PCA value than equation 10 is selected for the training of class1 and the original training data which has a higher PCA value than equation 10 again is selected for the training of class 2. If M1>M2, the inverse process carried out by equation 11.

For a training data set with input vector $x_i \in R_n$, i=1,...,l and output labels y_i, y∈{1,-1}, the PCA of the training data corresponding to x_i is x_i'. There are two classes for each classification process. After obtaining the new data xi' from the PCA process, the mean (M) and the standard deviation (s) of the training data of each class is calculated as:

$$M_j = \frac{\sum_{i=1}^{l} x_i'}{l} \qquad j=1,2 \ , i=1,...,l \qquad (8)$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{l} (x_i' - M_j)^2}{(n-1)}} \qquad j=1,2 \ , i=1,...,l \qquad (9)$$

Now the reduced training set can be selected with the equation as:

$$\forall x_i \Rightarrow if\ M_1 < M_2 \begin{cases} x_i' > (M_1 - s_1), & x_i' \in Class1 \\ x_i' < (M_2 + s_2), & x_i' \in Class2 \end{cases} \qquad (10)$$

$$\forall x_i \Rightarrow if\ M_1 > M_2 \begin{cases} x_i' < (M_1 + s_1), & x_i' \in Class1 \\ x_i' > (M_2 - s_2), & x_i' \in Class2 \end{cases} \qquad (11)$$

All x_i 's which provide the appropriate conditions, as we mentioned above, are selected for the training set of the two classes. This procedure is applied to all classes. Thanks to the proposed method, the non-critical training examples in the training set are discharged. We applied the proposed method to four different data sets and the results are discussed in the following section.

3.2 Reduced Training Data for Support Vector Machines Based on Mahalanobis Distance and Euclidean Distance

Secondly, we used MD and ED measurements for the reduction of training data of the SVM. Basically, the proposed MD and ED based training data reduction algorithms are same. Only the difference of the algorithms are the distance measurements.

The MD can be seen as a generalization of the Euclidean distance. It differs from ED in its definition which includes the covariance matrix so that the relationships between the variables are taken into account. That is, the correlation between variables or in other words the different dimensions present in the problem is taken into account when computing the distance between the two points [19].

In the proposed methods, we used MD and ED for finding the closest points for the two classes. Then we eliminate the other non critical samples. In figure 1, a sample for this procedure was given.

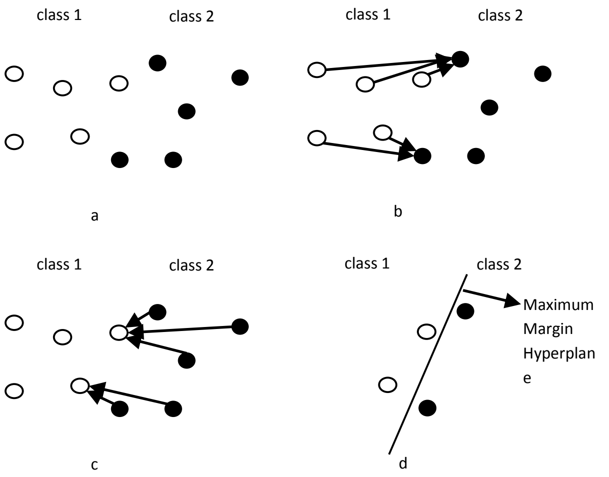


Figure 1. *Distance Based Training Data Reduction*

In the distance based reduction methods, we found the closest point of a training sample in the opposite class by using MD and ED. Then we add this closest point to the new reduced training set. This procedure is applied to all training samples of the SVM. We applied the proposed distance based methods to four different data sets and the results are discussed in the following section.

4. Experimental Results

In this study, we present three different methods based on PCA, MD and ED measurements for the reduction of training data of the SVM. By the help of the proposed methods, the non-critical training instances in the training set are removed. The results of the proposed methods are compared with SVM in terms of accuracy and computational time. All experiments were run on a machine with 2.20 GHz CPU, 8 GB of RAM, 700 GB HDD space and the Windows 7 operating system. All the methods were applied using the Matlab Software Package.

In the classification process of the system linear kernel function is used for all data sets. For multiclass classification of the SVM, one by one approach is used. In bone age data set, the data set was normalized and a 10-fold cross-validation was used during the classification process. In wine data set, in the classification process 30 instances from class 1, 36 instances from class 2 and 24 instances from class 3 were used for the training of the SVM and 29 instances from class 1, 35 instances from class 2 and 24 instances from class 3 were used for the testing process. In the iris data set, 10-fold cross-validation was used during the classification process. In the Pima Indians diabetes data set, 250 instances from class 1 and 134 instances from class 2 were used. 250 instances from class 1 and 134 instances from class 2 were used for the testing process.

Table 1. *Classification Results*

Data Sets		SVM	PCA Based	MD Based	ED Based
Bone Age	CA	%72,82	%72,82	%72,82	%59,49
	CT	8,5099	6,1604	1,9266	0,9744
	TSR	%100	%84,55	%38,86	%20,34
Wine	CA	%100	%100	%100	%95,45
	CT	0,2096	0,1325	0,0705	0,0358
	TSR	%100	%68,89	%48,33	%22,78
Iris	CA	%98	%98	%97,33	%94,67
	CT	4,2410	3,1226	0,7322	0,1898
	TSR	%100	%78,93	%34,96	%8,78
Pima Indians Diabetes	CA	%77,86	%77,60	%77,08	%65,10
	CT	1,0907	0,2507	0,0434	0,0238
	TSR	%100	%87,76	%43,49	%40,10

**CA: Classification Accuracy (%), CT: Computational Time (in Seconds), TSR: The percentage of the training set used for the training process for all the classes with one by one classification method*

We applied the proposed PCA and distance based methods to four different data sets and the results are given in table 1 in terms of computational time and classification accuracy. As seen in table 1, all the proposed methods are faster than the conventional SVM. The classification accuracies of PCA based system and MD based system are very close to the traditional method. The classification accuracies of the ED based system is very low than the other methods.

5. Conclusion

In this paper, three training data reduction algorithms has been proposed for SVM classifier to eliminate the non-critical training examples in the training set of the SVM. For this purpose, PCA, MD and ED based algorithms were performed. These algorithms have been used with different data sets and compared with the conventional SVM. The experimental results show that by using PCA and MD based algorithms, a significant amount of execution time in the training process can be saved without degrading or with a few amounts of loss in the performance of the resulting SVM classifier. When we examined the algorithms used in this study, ED based data reduction method has a big influence on execution times than the conventional SVM and the other proposed algorithms. But in ED based algorithm there is substantially loss in the performance. By using PCA based proposed algorithm, there is only a small of loss in the performance in Pima Indians diabetes data set. The performances in other data sets are same as conventional SVM. By using MD based proposed algorithm, there are small of losses in the performance in Pima Indians diabetes and iris data sets. The performances in other two data sets are same as conventional SVM. As a result, PCA and MD based data reduction algorithms significantly has not got any loss or in the performance on the SVM classifier. So PCA and MD based algorithms seems more useful than the ED based algorithm and conventional SVM.This study showed that the amount of training data is very important for SVM classifier in computational time. So, the PCA and MD based proposed methods could be very useful for huge number of datasets in different problems for the future studies.

Acknowledgement

This paper presented in ”4rd International Conference on Computational and Experimental Science and Engineering (ICCESEN-2017)”

References

[1] Vapnik V. “The nature of statistical learning theory” (Springer-Verlag, New York, 1995).

[2] Cervantes J., X. Li, W. Yu, “Support vector classification for large data sets by reducing training data with change of classes”, International conference on systems, man. and cybernetics, IEEE, 2008, 2609-2614.

[3] Javed I., M.N. Ayyaz, W. Mehmood “Efficient training data reduction for SVM based handwritten digits’ recognition”, International conference on electrical engineering, 2007, 1-4.

[4] Koggalage R., S. Halgamuge “Reducing the number of training samples for fast support vector machine classification”, Neural information processing - letters and reviews, vol. 2, no. 3, 2004, 57-65.

[5] Fortuna J., D. Capson, “Improved support vector classification using PCA and ICA feature space modification”, Pattern recognition, 37, 2004, 1117-1129.

[6] Cao L. J., K.S. Chua, W.K. Chong, H.P. Lee, Q.M. Gu “A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine”, Neurocomputing, 55, 2003, 321-336.

[7] Subasi A., M.I. Gursoy “EEG signal classification using PCA, ICA, LDA and support vector machines”, Expert systems with applications, vol. 37, no. 12, 2010, 8659-8666.

[8] Gertych A., A. Zhang, J. Sayre, S.P. Kurkowska, H.K. Huang “Bone age assessment of children using a digital hand atlas”, Computerized Medical Imaging and Graphics, 31, 2007, 322-331.

[9] Güraksın G. E., H. Uğuz, Ö.K. Baykan “Bone age determination in young children from newborn to 6 year-old using support vector machines”, Turkish Journal of Electrical Engineering and Computer Sciences, 24, 1693-1708.

[10] Frank A., A. Asuncion “UCI Machine Learning Repository” [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[11] Edla D. R., V. Gondlekar, V. Gauns “HK-Means:A Heuristic Approach to Initialize and estimate the number of clusters in biological data”, ICCESEN2016, ACTA PHYSICA POLONICA A, 130(1), 2016,78-82.

[12] Grimaldi M., P. Cunningham, A. Kokaram “An evaluation of alternative feature selection strategies and ensemble techniques for classifying music”, Workshop in Multimedia Discovery and Mining, ECML/PKDD03, Dubrovnik, Croatia, 2003.

[13] Tamura H., K. Tanno “Midpoint validation method for support vector machines with margin adjustment technique”, International Journal of Innovative Computing, Information and Control, 5(11(A)), 2009, 4025-4032.

[14] Yüksel A. S., Ş.F. Çankaya, İ.S. Üncü “Design of a machine learning based predictive analytic system for spam problem”, ICCESEN2016, ACTA PHYSICA POLONICA A, 132 (2), 2017, 500-504.

[15] Cömert Z., A.F. Kocamaz “Comparison of Machine Learning Technişques for Fetal Heart Rate Classification”, ICCESEN2016, ACTA PHYSICA POLONICA A, 132(3), 2017, 451-454.

[16] Uğuz H. “A Biomedical System Based on Artificial Neural Network and Principal Component Analysis for Diagnosis of the Heart Valve Diseases”, J. Med. Syst., 36, 2012, 61-72.

[17] Lindsay I., A. Smith “A tutorial on principal components analysis”, < http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial>, 2002.

[18] Valle S., W. Li, S.J. Qin “Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods”, Ind. Eng. Chem. Res. 38, 1999, 4389–4401.

[19] Torra V., Y. Narukawa “On a comparison between Mahalanobis distance and Choquet integral: The Choquet–Mahalanobis operatör”, Information Sciences, 190, 2012, 56-63.