Amirkabir University of Tehran
(Polytechnic of Tehran)
School of Computer Engineering and Information Technology

Technical Report of Machine Learning Final Project

Paper Title:

Large-scale support vector machine classification with redundant data reduction

Student Name:

Raghda Al-Taei

Student ID:

94231103

Teacher Name:

Dr. Nazerfard

2017 Feb.

1. Introduction

As it is known officially, SVM classifier is one of the most popular classifiers nowadays. But as the size of the training data increases, the learning procedure takes much more time and it seems that it's raising exponentially, and the thing is that we do not need the whole training data for finding support vectors. Support vectors are the only materials needed for the test procedure, and even in the training step those data samples that are candidates to be support vectors are adequate for training. So in this paper we're trying to find those SV candidates, and remove the redundant training data to reduce the huge time needed for training step.

2. Explanation of the method and parameters

In our proposed method for removing redundant training data, firstly we divide the whole dataset into K clusters by some unsupervised clustering method, like k-means clustering algorithm. The parameter K should be chosen by the user and it is much better to be selected as a large value. Then, we will be trying to divide the clusters into two types. One type is a clusters with the whole data samples belonging to a specific class type (UC set), and the second type is a cluster with data samples belonging to more than one class type (MC set). We will further divide the set MC to unique sub-clusters and then we will have a larger set named UMC.

After this, we should maintain those clusters in UMC set and also find those unique clusters in UC set that are far enough from the hyper-planes gained by SVM classifier in a very smart way. The thing is if we want to use the whole training data for finding those hyper-planes, not a smart job is done! So we better choose a representative point for each cluster and then apply the SVM classifier to find the decision boundaries. Those representative points are the clusters' centroids. By doing this, we will have the approximate decision boundaries and they will be used to remove the redundant clusters in UC set. For this matter, an Algorithm is introduced in the paper and is named "Max-Min cluster distance algorithm".

After removing redundant clusters from US set, we will have the set $S = UC \cup UMC$. Then we have to divide those unique clusters in set S to two parts: clustered data points and scattered data points. Cluster data points are those points nearby the centroid of the cluster and scattered data point are those points far away from the cluster core. There is absolutely a boundary between these two parts and it is one of the cluster points. For doing this, we shall calculate the distance of

any point to the centroid of its own cluster and then we will have a distance set for each unique cluster. After that, we have to find the point that specifies the boundary between the clustered and scattered data points. For this purpose, we use the Fisher Discriminant Ratio (FDR), which is derived from Fisher Discriminant Analysis (FDA) concept and only the last conceptual part of FDA is used in a very brilliant way, which is trying to maximize the $FDR = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$, where the μ_i is the mean of distances of each part of the distance set, and the σ_i^2 is the variance of those distances. Algorithm 2 which is named "Fast Iteration of FDR (FIFDR)" is introduced in the paper for doing this.

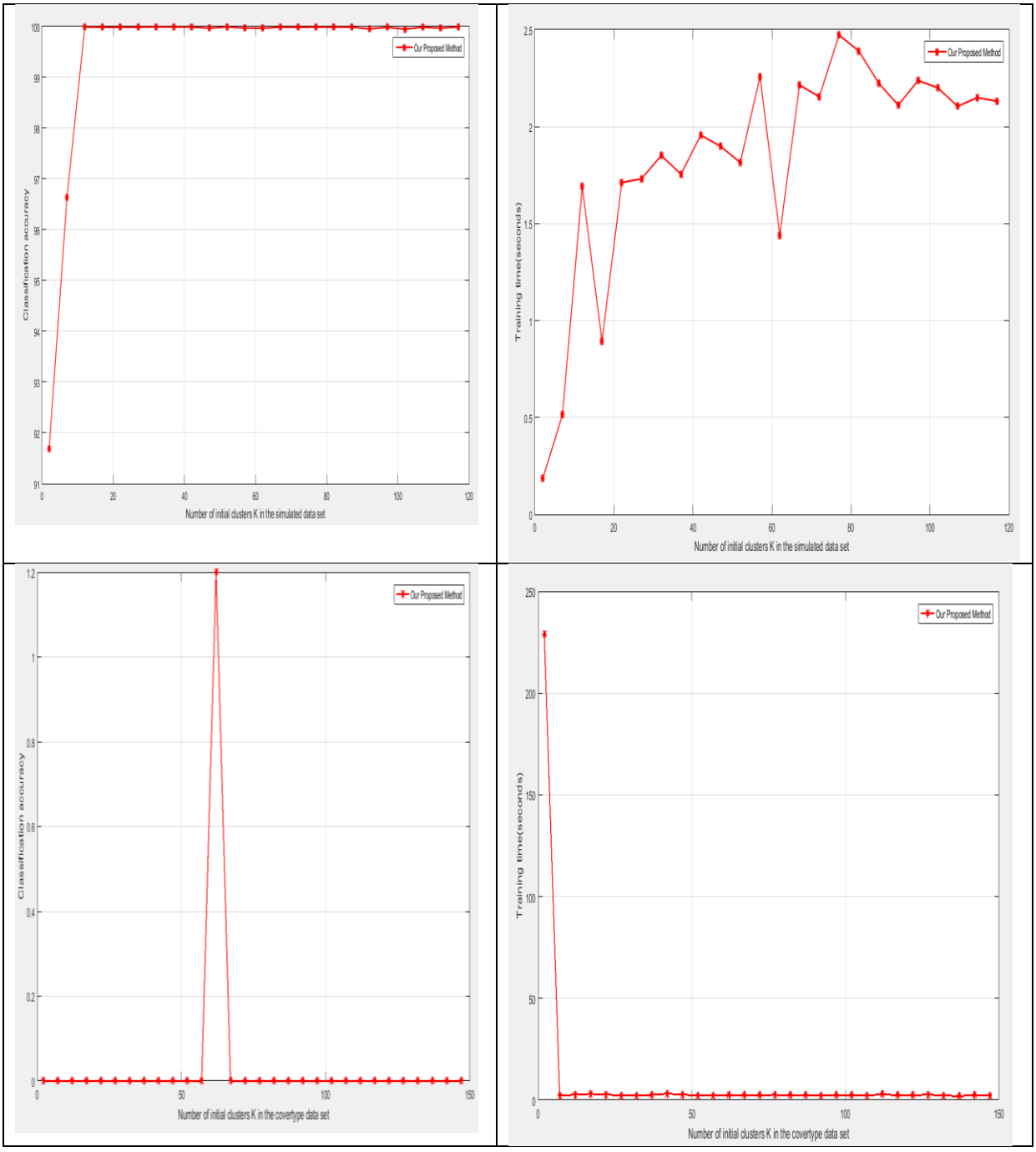
Finally, after removing the whole redundant data from the training data set by using the proposed algorithms, the only thing is to run SVM algorithm on the reduced data set to find the final decision boundaries. The training time of the SVM classifier will be definitely diminished by this method and the experimental results show that the classification accuracy is very close to the result of running the SVM classifier on the whole dataset that is not shrunk.

3. Experimental Results

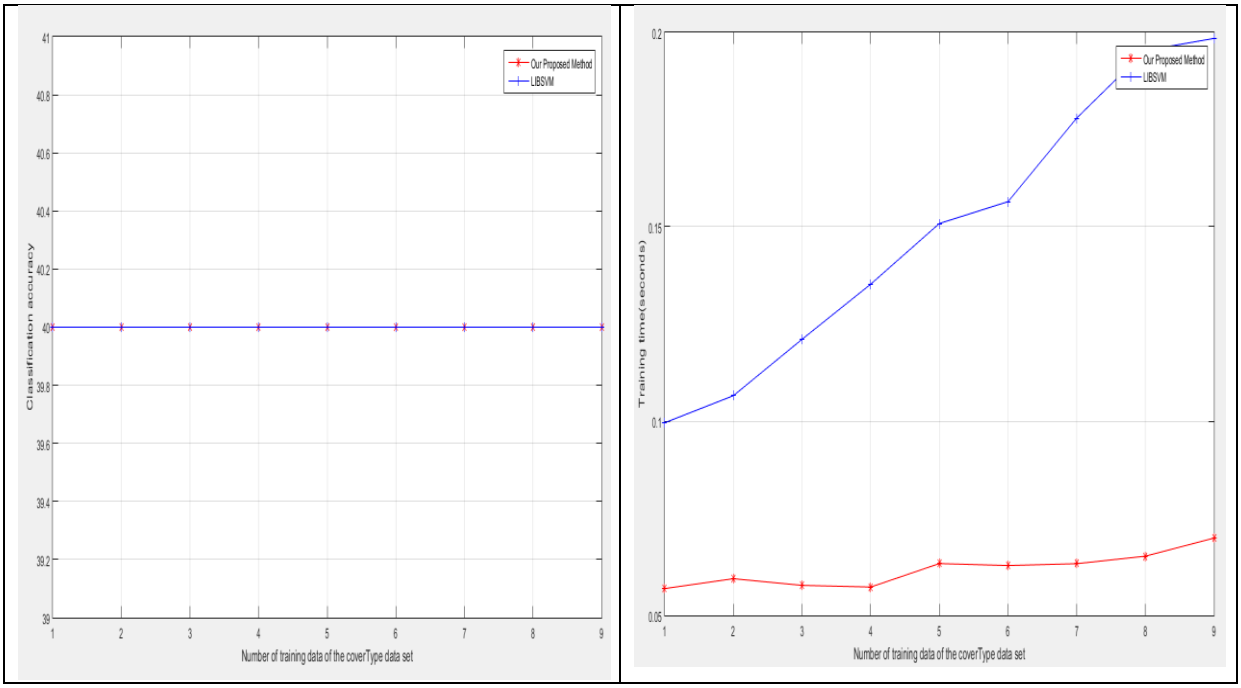
Here, we will use two of the data sets introduced in the paper. One is the simulated data set and the second is the Cover Type dataset. For the simulated data set we will calculate the 'classification accuracy' and 'Training time (seconds)' versus 'number of initial clusters'. For the second data set we will do two kinds of experiments. One is that for the simulated data set, and the second is 'classification accuracy' and 'Training time (seconds)' versus 'number of training data'.

Note: For the Cover Type data set because of its huge size we couldn't gain the final result because of lack of time! so we shrank this data set to a very smaller size and only mention the final results that may not be totally like those in paper. The Final results are as follows:

Experiment 1:



Experiment 2:



As it is obvious from the second experiment, the training time of the proposed method is much less than the mere SVM as the number of training data increases.

4. References

[1] **Shen, Xiang-Jun, et al. "Large-scale support vector machine classification with redundant data reduction."** ***Neurocomputing*** **172 (2016): 189-197.**