Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 03-042

Training Support Vector Machine using Adaptive Clustering

Dongwei Cao and Daniel Boley

October 07, 2003

Training Support Vector Machine using Adaptive Clustering

Dongwei Cao
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
Email: dcao@cs.umn.edu

Daniel Boley
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
Email: boley@cs.umn.edu

Abstract

Training support vector machines involves a huge optimization problem and many specially designed algorithms have been proposed. In this paper, we proposed an algorithm called *ClusterSVM* that accelerates the training process by exploiting the distributional properties of the training data, that is, the natural clustering of the training data and the overall layout of these clusters relative to the decision boundary of support vector machines. The proposed algorithm first partitions the training data into several pair-wise disjoint clusters. Then, the representatives of these clusters are used to train an initial support vector machine, based on which we can approximately identify the support vectors and non-support vectors. After replacing the cluster containing only non-support vectors with its representative, the number of training data can be significantly reduced, thereby speeding up the training process. The proposed *ClusterSVM* has been tested against the popular training algorithm SMO on both the artificial data and the real data, and a significant speedup was observed. The complexity of *ClusterSVM* scales with the square of the number of support vectors and, after a further improvement, it is expected that it will scale with square of the number of non-boundary support vectors.

Keywords: support vector machine, PDDP, clustering, optimization.

I. Introduction

Support vector machines (SVM) (Vapnik [1]) have been successfully applied in a variety of domains, including handwritten digit recognition [2], text document classification [3] and microarray data analysis [4]. In training a support vector machine, one needs to maximize a convex objective function subjecting to box constraints. This kind of optimization problem has been extensively studied and many software packages have been developed. However, the off-the-shelf packages typically require the entire Gram matrix be stored in the main memory and, knowing the fact that the size of the Gram matrix scales with the square of the number of training data, the memory requirement of these packages quickly makes them impractical even for a moderate problem [5]. Thus, many specially tailored optimization algorithms have been proposed. The first class of such algorithms tries to solve the entire optimization problem by solving a series of small problems. The basic techniques include chunking and decomposition, which

were discussed by Boser et al. [2], Osuna et al. [6], Kaufman et al. [7] and Joachims [8]. Especially noteworthy is the SMO (Sequential Minimal Optimization) algorithm by Platt [9] that sequentially optimizes over a subset of size two, for which we can perform the optimization analytically. The success of these algorithms depends on an appropriate criterion for the active set selection and an efficient strategy to cache the Gram matrix. A second class of algorithms tries to approximate the Gram matrix by a smaller matrix either using the low-rank representation (Fine et al. [10]) or by sampling (Williams et al. [11], Achlioptas et al. [12]), thereby reducing the size of the optimization problem and speeding up the training process. Keerthi et al. [13] proposed an algorithm based on observations about the geometrical properties of support vector machines. Based on a hierarchical micro-clustering algorithm, Yu et al. [14] proposed a scalable algorithm to train support vector machines with linear kernels. However, their algorithm currently works for linear kernels only and uses the fact that the original space and the feature space are the same under a linear kernel. Thus, due to the geometric differences between the original space and the feature space under a nonlinear kernel (e.g. [15]), it is quite hard to generalize their algorithm to nonlinear kernels, which is more popular than linear kernels.

In this paper, we proposed a fast training algorithm called *ClusterSVM* whose idea is to speed up the training process by reducing the number of training data. This is accomplished by partitioning the training data into pair-wise disjoint clusters, each of which consists of either only support vectors or only non-support vectors, and replacing the cluster containing only non-support vectors by a representative. In order to identify the cluster that contains only non-support vectors, the training data is first partitioned into several pair-wise disjoint clusters and an initial support vector machine is trained using the representatives of these clusters. Based on this initial SVM, we can judge whether a cluster contains only non-support vectors or not. For the cluster that contains both support vectors and non-support vectors, based on the decision boundary of the initial SVM, we can split it into two subclusters that approximately contain either only non-support vectors or only support vectors. This process is then repeated if one of the subclusters contains both support vectors and non-support vectors. The proposed algorithm works for any type of kernels. The training time of this strategy scales with the square of the number of support vectors and, as shown by experiments, an approximate solution can be found even faster. Further, based on the theory underlying *ClusterSVM*, it is expected that the training time will scale with the number of boundary support vectors after some straightforward extensions to the current work.

The rest of the paper is organized as follows. Section II briefly introduces the optimization problem involved in training SVM, followed by the theoretical results underlying *ClusterSVM*. In section III, the experimental results were reported on both the artificial data and the real data. Finally, section IV concludes the paper with further research topics.

II. CLUSTERSVM

A. Support vector machines

In a two-class classification problem, given a training data set $\mathcal{D}$ of size n

$$\mathcal{D} = \left\{ (\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{R}^N,\ y_i \in \{1, -1\} \right\} \tag{1}$$

where $i = 1, 2, \cdots, n$ and y_i indicates the class membership of the object i represented by vector $\mathbf{x}_i$, the support vector classifier $f(\mathbf{x})$ is defined as [1]

$$f(\mathbf{x}) = sign\left(d(\mathbf{x})\right) = \begin{cases} 1 & : \quad d(\mathbf{x}) \geq 0 \\ -1 & : \quad d(\mathbf{x}) < 0 \end{cases} \tag{2}$$

where $d(\cdot)$ is call the *functional margin* and it is defined as

$$d(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \tag{3}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the dot product of the reproducing kernel Hilbert space $\mathcal{H}$ generated by a symmetric positive definite kernel $K(\cdot, \cdot)$ satisfying the Mercel condition, and $\Phi(\cdot)$ is the mapping associated with $K(\cdot, \cdot)$ [1]. The optimal parameter $\mathbf{w}^*$ and b^* corresponding to the optimal classifier $f^*(\mathbf{x})$ can be obtained by solving the following optimization problem [1]

$$\text{Minimize} : g(\mathbf{w}, \xi) = \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i \tag{4a}$$

$$\text{Subject to} : y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}} + b) \geq 1 - \xi_i \tag{4b}$$

$$\xi_i \geq 0$$

With the help of Lagrange multipliers, the Wolfe dual form of the above minimization problem is [1]

$$\text{Maximize} : W(\alpha) = \alpha^{\mathbf{T}} \mathbf{1} - \frac{1}{2} \alpha^{\mathbf{T}} \mathbf{H} \alpha \tag{5a}$$

$$\text{Subject to} : 0 \leq \alpha \leq C \tag{5b}$$

$$\alpha^{\mathbf{T}} \mathbf{y} = 0.$$

where $\alpha_i \geq 0$ ($i = 1, 2, ..., n$) are the Lagrange multipliers, $\mathbf{1}$ is a vector of ones and $\mathbf{H}$ is the Gram matrix with component $H_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. The necessary and sufficient condition for a weight vector $\mathbf{w}$ and Lagrange multiplier α to be optimal is the KKT condition [1], which are the primal and dual feasibility constraints plus the following complementarity's conditions

$$\alpha_i \left(y_i \left(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathbf{H}} + b \right) - 1 + \xi_i \right) = 0 \tag{6a}$$

$$\xi_i \left(\alpha_i - C \right) = 0 \tag{6b}$$

Based on the optimal solution α, the functional margin $d(\cdot)$ can also be written as

$$d(\mathbf{x}) = \sum_{x_i \in \mathcal{D}_{SV}} \alpha_i y_i K(\mathbf{x}_i, x) + b \tag{7}$$

where D_{SV} is the set of support vectors, which are the subset of training data that have nonzero α's, that is, $0 < \alpha \leq C$. It is the set of support vectors that determines the decision boundary and all the other training data, that is, non-support vectors, can be removed without influencing the decision boundary.

B. ClusterSVM

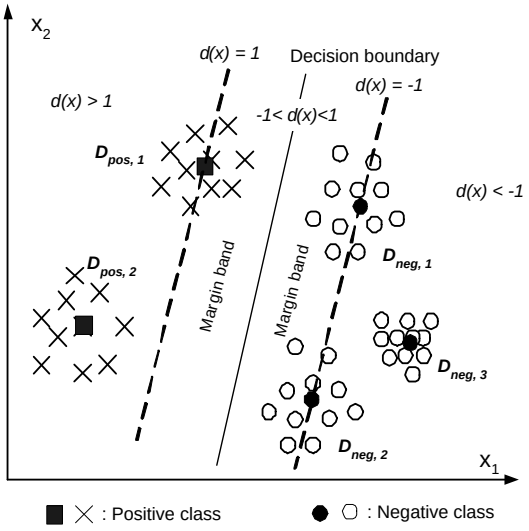

Fig. 1. A toy example. The representative of a cluster is labeled with a solid square/circle. The decision boundary of the initial SVM trained using the representatives of 5 initial clusters is shown.

Figure 1 shows the training data of a two-dimensional two-class classification problem. The training data in the positive class are partitioned into two disjoint clusters and those in the negative class are partitioned into three clusters. The motivation of *ClusterSVM* is to reduce the number of training data by replacing a cluster with an appropriately defined representative. However, not all clusters can be replaced with a representative while yielding the same the SVM as the SVM that would be obtained using the original training data set $\mathcal{D}$. It follows from the following Proposition 1 that there are two kinds of clusters that can be replaced without influencing the solution, including the cluster that contains only non-support vectors ($\alpha = 0$) and the cluster that contains only boundary support vectors ($\alpha = C$).

Since each training datum corresponds to one row and column in the Gram matrix $\mathbf{H}$, replacing data in a cluster with a representative corresponds to replacing the rows and columns associated with these data with a single row and column. Let $\mathcal{D}$ denote the training data set consisting of two disjoint sets $\mathcal{D}_1$ and $\mathcal{D}_2$ and, without losing

generality, assume $\mathcal{D}_1$ is a subset of the training data in class 1. Let α_1 and α_2 be the Lagrange multipliers of the data in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, then the optimization problem (5) is equivalent to

$$\text{Maximize} : W(\alpha) = \left(\alpha_1 \mathbf{1}_1 - \frac{1}{2}\alpha_1^T \mathbf{H}_{11}\alpha_1\right) + \left(\alpha_2 \mathbf{1}_2 - \frac{1}{2}\alpha_2^T \mathbf{H}_{22}\alpha_2\right) - \alpha_1^T \mathbf{H}_{12}\alpha_2 \tag{8a}$$

$$\begin{aligned}\text{Subject to} : & 0 \le \alpha_{1,i} \le C,\ \forall i = 1, 2, ..., n_1 \\ & 0 \le \alpha_{2,j} \le C,\ \forall j = 1, 2, ..., n_2 \\ & \alpha_1^T \mathbf{y}_1 + \alpha_2^T \mathbf{y}_2 = 0.\end{aligned} \tag{8b}$$

Let the index to the row and column that will replace the rows and columns associated with $\mathcal{D}_1$ be 0 and the label $y_0 = 1$, we have the following optimization problem

$$\text{Maximize} : W(\alpha) = \left(\alpha_0 - \frac{1}{2}\alpha_0 H_{00}\alpha_0\right) + \left(\alpha_2 \mathbf{1}_2 - \frac{1}{2}\alpha_2^T \mathbf{H}_{22}\alpha_2\right) - \alpha_0 \mathbf{H}_{02}\alpha_2 \tag{9a}$$

$$\begin{aligned}\text{Subject to} : & 0 \le \alpha_0 \le n_1 C \\ & 0 \le \alpha_{2,j} \le C,\ \forall j = 1, 2, ..., n_2 \\ & \alpha_0 + \alpha_2^T \mathbf{y}_2 = 0.\end{aligned} \tag{9b}$$

where α_0 is the Lagrange multiplier corresponding to the representing row/column, H_{00} is defined as

$$H_{00} = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} H_{ij} \tag{10}$$

and $\mathbf{H}_{02}$ is a row vector of length n_2 with entries defined as

$$H_{0j} = \frac{1}{n_1}\sum_{i=1}^{n_1} H_{ij} \tag{11}$$

where $j = 1, 2, \cdots, n_2$. Then, we have the following proposition.

Proposition 1: The optimization problem defined by equations (9), (10) and (11) is equivalent to the one obtained by adding a constraint to (5) that requires all Lagrange multipliers corresponding to the data in $\mathcal{D}_1$ be equal.

Proof: Using equation (10), we have

$$\begin{aligned}\alpha_0 H_{00}\alpha_0 &= \alpha_0\alpha_0 \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} H_{ij} \\ &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \frac{\alpha_0}{n_1}\frac{\alpha_0}{n_1} H_{ij}\end{aligned} \tag{12}$$

Let α_1^* be a vector of length n_1 with all components being equal to α_0/n_1, equation (12) can be written as

$$\begin{aligned}\alpha_0 H_{00}\alpha_0 &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \alpha_{1,i}^*\alpha_{1,j}^* H_{ij} \\ &= \alpha_1^{*T}\mathbf{H}_{11}\alpha_1^*\end{aligned} \tag{13}$$

Using the similar arguments, $\alpha_0 \mathbf{H}_{02} \alpha_2$ can be written as

$$\alpha_0 H_{02} \alpha_2 = \alpha_1^{*T} \mathbf{H}_{12} \alpha_2 \tag{14}$$

After substituting equations (13) and (14) into equation (9) and using the fact that $\sum_{i=1}^{n_1} \alpha_{1,i}^* = \alpha_0$, we arrive the following optimization problem

$$\text{Maximize} : W(\alpha_1^*, \alpha_2) = \left(\alpha_1^* \mathbf{1}_1^* - \frac{1}{2} \alpha_1^{*T} \mathbf{H}_{11} \alpha_1^* \right) + \left(\alpha_2 \mathbf{1}_2 - \frac{1}{2} \alpha_2^T \mathbf{H}_{22} \alpha_2 \right) - \alpha_1^{*T} \mathbf{H}_{12} \alpha_2 \tag{15a}$$

$$\begin{aligned} \text{Subject to} : & 0 \leq \alpha_{2,j} \leq C, \ \forall j = 1, 2, ..., n_2 \\ & 0 \leq \alpha_{1,1}^* \leq C \\ & \boldsymbol{\alpha_1^{*T} \mathbf{y_1} + \alpha_2^T \mathbf{y_2} = 0.} \\ & \alpha_{1,1}^* = \alpha_{1,i}^*, \ \forall i = 2, ..., n_1 \end{aligned} \tag{15b}$$

where $0 \leq \alpha_{1,i}^* \leq C$ $(i = 1, 2, ..., n_1)$ follows from the fact that $0 \leq \alpha_0 \leq n_1 C$. The proposition follows by comparing equation (8) and equation (15). ■

The idea of Proposition 1 is illustrated in Figure 2 for a toy problem that has two points in class 1 with Lagrange multipliers α_1 and α_2, and one point in class -1 with Lagrange multiplier α_3. The cube $pqst - ovwu$ is the feasible region of the original optimization problem (8). After replacing two data in class 1 by a representative, the feasible region of the resulting optimization problem (c.f. (15)) is the rectangle $opsw$. Thus, the feasible region of the problem (15) is a subset of that of the problem (8) and, by replacing clusters of training data with their representatives, the solution of the resulting optimization problem is an approximation to the solution of the original optimization problem. Further, it is not hard to show that the optimal solution of (15) is exactly the optimal solution of (8) if $\mathcal{D}_1$ satisfies either of the following conditions. As in Proposition 1, $\mathcal{D}_1$ is a subset of the data in class 1 to be replaced by a representative.

- **Condition 1** All data in $\mathcal{D}_1$ are non-support vectors, which means the corresponding Lagrange multiplier $\alpha_1 = \alpha_2 = 0$. With reference to Figure 2, this means the feasible region of the problem (8) and that of the problem (15) coincides at line op.
- **Condition 2** All data in $\mathcal{D}_1$ are boundary support vectors, which means the corresponding Lagrange multipliers $\alpha_1 = \alpha_2 = C$. With reference to Figure 2, this means that the feasible region of the problem (8) and that of the problem (15) coincides at line ws.

Thus, we can safely replace clusters of above two types with a representative without influencing the solution and such replacement will reduce the size of the optimization problem. In addition, for the cluster $\mathcal{D}_i$ satisfying the above **Condition 1**, we can replace the corresponding rows and columns with the row and column associated with the pseudocenter of this cluster because the Lagrange multiplier of a non-support vector is zero, where the

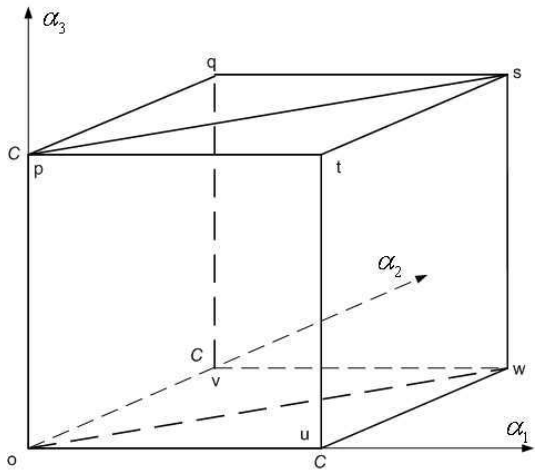

Fig. 2. Illustration of Proposition 1.

pseudocenter $\mathbf{x}^p(\mathcal{D}_i)$ of the cluster $\mathcal{D}_i$ is defined as

$$\mathbf{x}^p(\mathcal{D}_i) = \underset{\mathbf{x}\in\mathcal{D}_i}{\operatorname{argmin}} \left\| \mathbf{x} - \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{x}_k \right\|_2 \tag{16}$$

where $\|\cdot\|_2$ means 2-norm and n_i is the number of data in $\mathcal{D}_i$. As a pilot study, only the cluster satisfying the **Condition 1** is replaced by its representative in the current algorithm.

The next issue is to identify clusters that contain only non-support vectors. However, there is a cycle here because the set of support vectors is unknown before training is finished. The solution is to first partition the training data into pair-wise disjoint clusters, then train an initial SVM using the representatives of these clusters. Based on this initial SVM, we can approximately tell the position of each cluster relative to the decision boundary, thereby approximately identifying the clusters containing only non-support vectors. For the cluster that is believed to contain both support vectors and non-support vectors, it is split into two subclusters, one of which is expected to contain only support vectors and the other is expected to contain only non-support vectors. This idea is illustrated in Figure 1 and 3. Figure 1 shows the training data of a two-class classification problem and they are partitioned into 5 pair-wise disjoint clusters ($\mathcal{D}_{pos,1}$, $\mathcal{D}_{pos,2}$, $\mathcal{D}_{neg,1}$, $\mathcal{D}_{neg,2}$ and $\mathcal{D}_{neg,3}$). The representatives (that is, pseudocenters defined in equation 16) of these clusters are labeled with solid squares and solid cycles. An initial SVM was trained using these representatives, and its decision boundary ($d(\mathbf{x}) = 0$) and supporting hyper-planes ($d(\mathbf{x}) = \pm 1$) were shown in Figure 1. For the cluster belonging to the positive class, it is believed to contain only non-support vectors if the functional margin ($d(\mathbf{x})$) of its data are all larger than 1 (e.g. $\mathcal{D}_{pos,2}$) and, hence, it can be replaced by its pseudocenter without being split. However, a cluster belonging to the positive class is believed to contain both support vectors and non-support vectors if it contains some data with functional margin $d(\mathbf{x}) \leq 1$, which are likely to be support vectors, and some data with functional margin $d(\mathbf{x}) > 1$, which are believed to be non-support

vectors. This kind of cluster is partitioned into two subclusters and it is believed that the subcluster having data with $d(\mathbf{x}) > 1$ contains only non-support vectors and can be replaced with its pseudocenter, while the other subcluster ($d\mathbf{x} \leq 1$) is believed to contain only support vectors. An example of such cluster is $\mathcal{D}_{pos,1}$ in Figure 1, which is partitioned into two subclusters along $d(\mathbf{x}) = 1$. Similar arguments apply to the cluster belonging to the negative class. A cluster is believed to contain only non-support vectors if its data all satisfy $d(\mathbf{x}) < -1$, and a cluster is believed to contain both non-support vectors and support vectors if some of its data satisfy $d(\mathbf{x}) < -1$ and the other data satisfy $d(\mathbf{x}) \geq -1$. Using this criterion, cluster $\mathcal{D}_{neg,3}$ needs not to be split, while clusters $\mathcal{D}_{neg,1}$ and $\mathcal{D}_{neg,2}$ need to be split into two subclusters. After splitting some clusters and replacing the clusters and subclusters containing only non-support vectors with a representative, the resulting training data set $\mathcal{D}_{reduced}$ is shown in Figure 3, from which we can see a significant reduction on the number of training data.

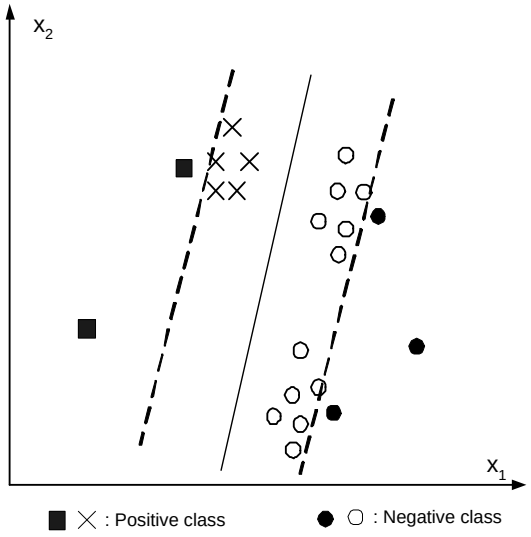

Fig. 3. The reduced training data set $\mathcal{D}_{reduced}$ after splitting some clusters and replacing clusters and subclusters containing only non-support vectors with their representatives. The representatives are labled with solid squares and solid cycles. For clarity purpose, the decision boundary shown in Figure 1 is kept here.

The proposed training algorithm *ClusterSVM* is detailed in Algorithm 1 and its properties are summarized in Proposition 2.

Proposition 2: With reference to Algorithm 1 (*ClusterSVM*) and setting $NP_{max} = \infty$, we have

1) Algorithm 1 will converge after a finite number of passes through the *WHILE* loop (line 5 through 18).
2) The SVM_{new} obtained using only $\mathcal{D}_{reduced}$ is the same as the SVM that would be obtained using $\mathcal{D}$ when Algorithm 1 terminates, that is, when the following condition is satisfied

$$yd(\mathbf{x}) > 1, \quad \forall \mathbf{x} \in \mathcal{D}_{unused} \tag{17}$$

where $\mathcal{D}_{unused}$ contains data that are in $\mathcal{D}$ but not in $\mathcal{D}_{reduced}$.

Proof: From line 13 and 14 in Algorithm 1, we can see that the size of the reduced training data set $\mathcal{D}_{reduced}$ is strictly increasing after each pass through the *WHILE* loop. Since there is a finite number of training data in $\mathcal{D}$, $\mathcal{D}_{reduced}$ will be the same as $\mathcal{D}$ after a finite number of passes through the *WHILE* loop, which means that Algorithm 1 will converge after a finite number of passes.

For the second part of the proposition, we need only to show that, for the weight vector $\mathbf{w}$ of SVM_{new}, the KKT conditions are satisfied for all training data in $\mathcal{D}_{unused}$, that is, the Lagrange multiplier is zero. For a given $\mathbf{x}_i \in \mathcal{D}_{unused}$, we have

$$y_i d(\mathbf{x}_i) > 1 \Longrightarrow y_i \left(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \right) > 1 \tag{18}$$

Knowing the fact that $\xi_i \geq 0$, this means that the constraint specified by equation (4b) will not be active, thus $\alpha_i = 0$. ■

It should be noted that the second conclusion of Proposition 2 does not depend on how to partition $\mathcal{D}$ into $\mathcal{D}_{reduced}$ and $\mathcal{D}_{unused}$. As long as the condition specified in equation (17) is satisfied for all data in $\mathcal{D}_{unused}$, the SVM obtained using $\mathcal{D}_{reduced}$ is the same as that would be obtained using $\mathcal{D}$.

III. Experiments

A. Implementations

Due to its popularity, the training algorithm $\mathbb{A}$ we use in Algorithm 1 is Platt's SMO [9], and the *ClusterSVM* is compared with SMO. An implementation of SMO by Chang et al. [16] and its Matlab® wrapper by Ma et al. [17] were used in this paper. However, it should be pointed out that, being used as a meta-algorithm, *ClusterSVM* could accelerate *any* training algorithm. The clustering algorithm $\mathbb{C}$ used here is the PDDP (Principal Direction Divisive Partition) by Boley [18] because it is one the most efficient clustering algorithms.

The number of initial clusters k^+ (k^-) can be any number between one and the number of training data n^+ (n^-) in $\mathcal{D}^+$ ($\mathcal{D}^-$). However, knowing the fact that the initial SVM will be trained using the representatives of the initial clusters and all subsequent partitions will depend on the initial SVM, the number of initial clusters should be large enough so that the initial SVM can approximate the true SVM reasonably well. At the same time, it should not be too large since letting $k^+ = n^+$ and $k^- = n^-$ would make $\mathcal{D}_{reduced} = \mathcal{D}$, and there would be no speedup. Another reason for preferring small k^+ (k^-) is that both clustering the training data $\mathcal{D}$ and training the initial SVM needs to be performed very quickly. In this paper, the following square root heuristic is suggested

$$k^+ = round(\sqrt{n^+}) \text{ and } k^- = round(\sqrt{n^-}) \tag{19}$$

There are primarily two motivations for this heuristic. First, knowing the fact that the time for clustering typically scales linearly with the number of data [19], the square root heuristic can make the total time to obtain the initial SVM scale linearly with the number of training data. The second reason is that this heuristic has been suggested in

Algorithm 1 ClusterSVM: Two class SVM

Require: A SVM training algorithm $\mathbb{A}$; A clustering algorithm $\mathbb{C}$; Training data set $\mathcal{D} = \mathcal{D}^+ \cup \mathcal{D}^-$, where $\mathcal{D}^+$ ($\mathcal{D}^-$) is the set of the training data in class 1 (-1); The number of initial clusters k^+ (k^-) into which $\mathcal{D}^+$ ($\mathcal{D}^-$) is partitioned; The maximum number of passes NP_{max} through the $WHILE$ loop.

Call the clustering algorithm $\mathbb{C}$ to partition $\mathcal{D}^+$ ($\mathcal{D}^-$) into k^+ (k^-) clusters, that is

$$\mathcal{D}^+ = \bigcup_{i=1}^{k^+} \mathcal{D}_i^+ \quad \text{and} \quad \mathcal{D}^- = \bigcup_{i=1}^{k^-} \mathcal{D}_i^-$$

Define the set $\mathcal{G}$ of clusters as

$$\mathcal{G} \leftarrow \{\mathcal{D}_1^+, \cdots, \mathcal{D}_{k^+}^+, \mathcal{D}_1^-, \cdots, \mathcal{D}_{k^-}^-\}$$

Define the reduced training data set $\mathcal{D}_{reduced}$ as (c.f. (16))

$$\mathcal{D}_{reduced} \leftarrow \{\mathbf{x}^p(\mathcal{D}'), \mathcal{D}' \in \mathcal{G}\}$$

$Flag \leftarrow 1$, $NP \leftarrow 0$

while $Flag = 1$ and $NP < NP_{max}$ **do**

  $Flag \leftarrow 0$, $NP \leftarrow NP + 1$

  Train SVM_{new} using $\mathcal{D}_{reduced}$ and the training algorithm $\mathbb{A}$

  $\mathcal{G}^{old} \leftarrow \mathcal{G}$ and $\mathcal{G} \leftarrow \emptyset$

  for all $\mathcal{D}' \in \mathcal{G}^{old}$ **do**

    if $\exists \mathbf{x} \in \mathcal{D}'$ such that $yd(\mathbf{x}) \leq 1$ according to SVM_{new}, where y is the label of $\mathbf{x}$ **then**

      $Flag \leftarrow 1$

      Split $\mathcal{D}'$ into $\mathcal{D}'_{sv}$ and $\mathcal{D}'_{nsv}$

$$\mathcal{D}'_{sv} \leftarrow \{\mathbf{x} | \mathbf{x} \in \mathcal{D}' \quad \text{and} \quad yd(\mathbf{x}) \leq 1\}$$

$$\mathcal{D}'_{nsv} \leftarrow \{\mathbf{x} | \mathbf{x} \in \mathcal{D}' \quad \text{and} \quad yd(\mathbf{x}) > 1\}$$

      Remove $\mathbf{x}^p(\mathcal{D}')$ from $\mathcal{D}_{reduced}$

      $\mathcal{D}_{reduced} \leftarrow \mathcal{D}_{reduced} \cup \mathcal{D}'_{sv} \cup \{\mathbf{x}^p(\mathcal{D}'_{nsv})\}$

      $\mathcal{G} \leftarrow \mathcal{G} \cup \{\mathcal{D}'_{nsv}\}$

    end if

  end for

end while

Return the SVM_{new}.

the study of clustering algorithms (e.g. [20]). The effectiveness of this heuristic was demonstrated experimentally.

Since the initial SVM can approximate the true SVM quite well and each pass through the outer *WHILE* loop (line 5 to 18 in Algorithm 1) involves training a SVM, the next issue is how many times the *WHILE* loop should be performed. Based on the experiments, it is enough to carry out the *WHILE* loop once.

The last implementation issue is the strategy for the multi-class classification problem. There are many strategies for multi-class classification problem and, in this paper, the "one versus the rest" strategy is used. In this strategy, assuming there are m classes, m classifiers are trained and each of them discriminates one class from all the other classes. A test data is classified to the class that has the maximum functional margin $d(\cdot)$. In order to avoid repeated clusterings, the clustering algorithm is applied to the data of each class before any classifier is trained. Then, to train the classifier that discriminates the class i from the remaining $m-1$ classes, the clusters corresponding to class i are used as the partition of the data in class i, and the clusters corresponding to the remaining $m-1$ classes are put together and used as the partition for the data in those $m-1$ classes. All experiments were run on a PC running Windows 2000 Server with one Pentium 4 2.8GHz processor and 1GB RAM, and the algorithm was implemented using Matlab® [21].

B. Data sets

There are three data sets examined in this paper.

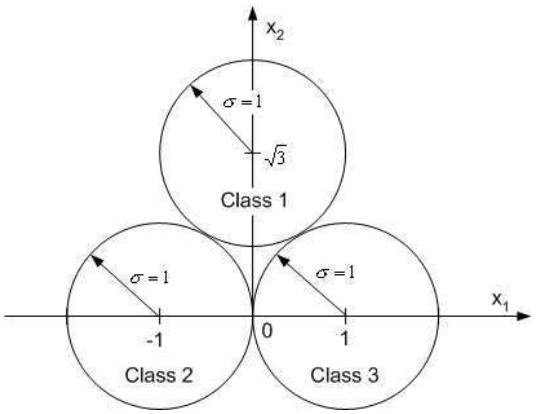

Fig. 4. Artificial data set.

- **Artificial data set** As shown in Figure 4, this is a three-class classification problem and each class consists of data drawn from a 2D normal distribution with covariance matrix being identity matrix. The centers of three classes are $(0, \sqrt{3})$, $(-1, 0)$ and $(1, 0)$. The same number of training data are drawn for each class and the size of the training data $\mathcal{D}$ varies from 300 to 6000. The test data set is of the same size as the training data set and is constructed in the same way. All 3 classifiers are obtained using the regularization coefficient $C = 10000$ (c.f. equation (4a)) and the linear kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{20}$$

- **USPS data set** This is the US Postal Service (USPS) handwritten zip code recognition data set and there are 7291 training data and 2007 test data, all of which were collected from mail envelopes in Buffalo [22]. Each digit is represented as a 16×16 matrix whose entry ranges from -1 to 1. As suggested by [23], a smoothing operation using a Gaussian kernel with width 0.75 was applied to the image as a preprocessing step. The regularization coefficient $C = 10$ and the kernel is a homogeneous polynomial kernel of degree 3 defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{256} \right)^3 \tag{21}$$

- **Isolet data set** This data set was downloaded from UCI machine learning repository [24] and the goal is to recognize 26 spoken letters. There are 6238 training data and 1559 test data. Each datum has 617 attributes and each attribute is a real number between -1 and 1. All 26 classifiers were obtained using the regularization coefficient $C = 0.02$ and the linear kernel defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j. \tag{22}$$

C. Experimental results

The effect of the number of initial clusters k was studied through the artificial data set. There 2000 training data in each class (6000 total) and the number of initial clusters k varies from 1 to 81 with an interval of 2. For each value of k, 10 randomly generated training data set were tried. Figure 5 compares the relative difference between the error rate of the initial SVM with that of the true SVM for different values of k. The relative difference RD is defined as

$$RD = \frac{|ER_{initial} - ER_{true}|}{ER_{true}} \tag{23}$$

where $ER_{initial}$ and ER_{true} are the error rate of the initial SVM and the true SVM on the same test data set. Figure 6 shows the time to obtain the initial SVM $T_{Initial\ SVM}$ as a function of k. $T_{Initial\ SVM}$ consists of time for clustering and the time for training the initial SVM. From Figure 5 and Figure 6, we can see that the square root heuristic, corresponding to $k = 45$ in this experiment, gives a reasonable good trade-off between the accuracy and the complexity, although it is a rather gross heuristic.

With the number of initial clusters being specified by the square root heuristic, the effect of the number of passes NP through the *WHILE* loop (line 5 to 18 in Algorithm 1) is shown in Table I for the artificial data set with 6000 training data. The SVM trained after 3 passes is the true SVM, which would be obtained using the original training data set, thus the corresponding error rate can be taken as the reference. From Table I, it can be seen that one pass through the *WHILE* loop is enough to give a good performance. Thus, the maximum number of passes NP_{max} in Algorithm 1 is set to 1. In addition, it can be seen from Table I that, with $NP = 0$, the initial SVM also gives pretty good result.

Table II through IV compares the performance of *ClusterSVM* with that of SMO, where the number of initial clusters is specified by the square root heuristic and the maximum number of passes through the *WHILE* loop

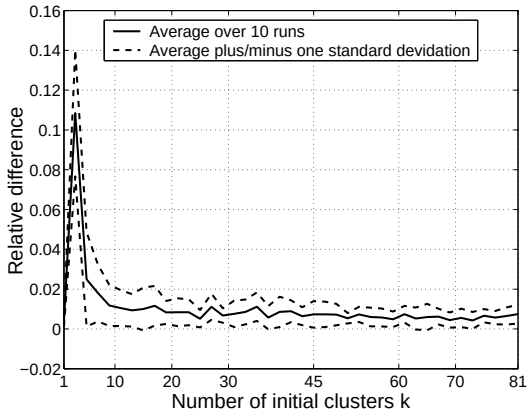

Fig. 5. The performance of the initial SVM compared to that of true SVM as a function of the number of initial clusters k. There are 2000 training data in each class (6000 in total) and the square root heuristic corresponds to $k = 45$. The seemingly good performance of $k = 1$ comes from the symmetry of this problem and it has no general implications.

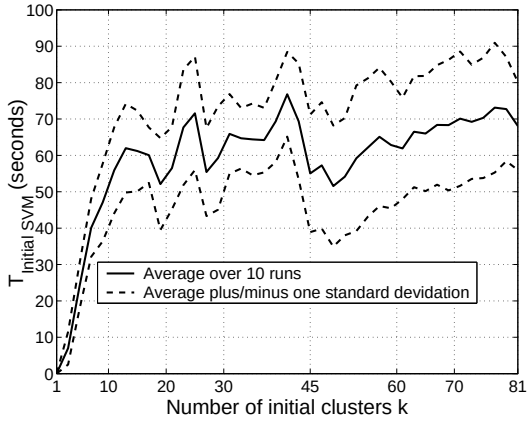

Fig. 6. Time to obtain the initial SVM as a function of the number of initial clusters k. There are 2000 training data in each class (6000 in total) and the square root heuristic corresponds to $k = 45$.

TABLE I

EFFECTS OF NP ON THE ARTIFICIAL DATA SET (6000 TRAINING DATA). NP IS THE NUMBER OF PASSES THROUGH THE *WHILE* LOOP IN ALGORITHM 1.

NP	0	1	2	3
Error rate (%)	25.87	25.43	25.48	25.47

$NP_{max} = 1$. In these tables, the speed up is defined as

$$Speedup = \frac{T_{SMO}}{T_{ClusterSVM}} \tag{24}$$

where T_{SMO} is the training time of SMO and $T_{ClusterSVM}$ is the training time of *ClusterSVM*. The clustering time is the time used for the clustering all training data. Based on these tables, we have the following observations.

- $N_{train,i}$ ($i = 1, 2, \cdots, m$) is the actual number of training data used to train the i-th classifier. For SMO, this number is the number of training data of all classes and it is independent of which classifier is being trained. For *ClusterSVM*, $N_{train,i}$ is the number of training data after replacing every cluster containing only non-support vectors with its representative. For the artificial data set shown in Table II, $N_{train,i}$ is almost the same for all three classifiers when *ClusterSVM* is used. This is within our expectations because of the symmetry of the artificial data set. However, for the USPS data set shown in Table III, $N_{train,i}$ varies from one classifier to another when *ClusterSVM* is used. This is reasonable because all ten classifiers are inherently different. For example, discriminating digit 1 from the other digits is different from discriminating digit 0 from the other digits. Similarly, for the Isolet data set shown in Table III, different classifier has different number of training data when *ClusterSVM* is used. Thus, the *ClusterSVM* reduces the number of training data in a task dependent way.
- N_{train} is the average number of training data over all k classifiers and, comparing *ClusterSVM* with SMO, it can be seen that *ClusterSVM* reduce the number of training data significantly. It is this data reduction that help accelerating the training process.
- The speed up of *ClusterSVM* over SMO is 3.2 for the artificial data set, 1.5 for the USPS data set and 1.9 for the Isolet data set.
- Comparing the error rate of SMO and that of *ClusterSVM*, it can be seen that the speedup of *ClusterSVM* sacrifices little performance. This nice property is attributed to the good initial clustering that makes the initial SVM approximate the true SVM quite well. At the same time, as shown in these tables, the overhead induced by clustering is only a small faction of total training time.

Finally, the scaling performance of *ClusterSVM* was shown in Figure 7 for the artificial data set, which shows the average training time over 10 runs. It can be seen that *ClusterSVM* scales better than SMO.

IV. Conclusions

An efficient SVM training algorithm *ClusterSVM* was proposed in this paper and a significant speedup over SMO was observed on both the artificial data set and the real data set. The possible extensions to *ClusterSVM* are the follows.

- The second sufficient condition mentioned after the Proposition 1 has not been used in *ClusterSVM*. It is not hard to incorporate this condition into *ClusterSVM* and this would make the training time scale with the number of non-boundary support vectors. This would definitely speed up the SVM training further.
- With the help of a clustering algorithm, *ClusterSVM* effectively incorporate the distributional property of the training data into the training process. It is expected that the similar idea can be used to improve other supervised learning algorithm like neural networks.

TABLE II

ARTIFICIAL DATA SET. $N_{train,i}$ IS THE ACTUAL NUMBER OF TRAINING DATA TO TRAIN THE i-TH CLASSIFIER.

	SMO	*ClusterSVM*
$N_{train,1}$	6000	3260
$N_{train,2}$	6000	3026
$N_{train,3}$	6000	3022
N_{train}	6000	3103
Training time (sec.)	8344	2588
Clustering time (sec.)	NA	3
Speedup	3.2	
Error rate (%)	25.47	25.43

TABLE III

USPS DATA SET. $N_{train,i}$ IS THE ACTUAL NUMBER OF TRAINING DATA TO TRAIN THE i-TH CLASSIFIER.

	SMO	*ClusterSVM*
$N_{train,1}$	7291	788
$N_{train,2}$	7291	2364
$N_{train,3}$	7291	1975
$N_{train,4}$	7291	1544
$N_{train,5}$	7291	2259
$N_{train,6}$	7291	1621
$N_{train,7}$	7291	1206
$N_{train,8}$	7291	2407
$N_{train,9}$	7291	1560
$N_{train,10}$	7291	2638
N_{train}	7291	1836
Training time (sec.)	105	68
Clustering time (sec.)	NA	18
Speedup	1.5	
Error rate (%)	5.43	5.28

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. NY: Wiley, 1998.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," 1992, Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, ACM.

[3] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning*. Springer, 1998.

[4] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, M. Ares, and D. Haussler, "Support vector machine classification of microarray gene expression data," University of California, Santa Cruz, Tech. Rep., 1999.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

TABLE IV

ISOLET DATA SET. $N_{train,i}$ IS THE ACTUAL NUMBER OF TRAINING DATA TO TRAIN THE i-TH CLASSIFIER.

	SMO	*ClusterSVM*
$N_{train,1}$	6238	1102
$N_{train,2}$	6238	1237
$N_{train,3}$	6238	840
$N_{train,4}$	6238	1242
$N_{train,5}$	6238	1123
$N_{train,6}$	6238	1102
$N_{train,7}$	6238	1016
$N_{train,8}$	6238	932
$N_{train,9}$	6238	957
$N_{train,10}$	6238	1039
$N_{train,11}$	6238	1048
$N_{train,12}$	6238	914
$N_{train,13}$	6238	945
$N_{train,14}$	6238	1159
$N_{train,15}$	6238	946
$N_{train,16}$	6238	1435
$N_{train,17}$	6238	1018
$N_{train,18}$	6238	883
$N_{train,19}$	6238	762
$N_{train,20}$	6238	1225
$N_{train,21}$	6238	980
$N_{train,22}$	6238	1346
$N_{train,23}$	6238	1258
$N_{train,24}$	6238	837
$N_{train,25}$	6238	852
$N_{train,26}$	6238	864
N_{train}	6238	1041
Training time (sec.)	278	144
Clustering time (sec.)	NA	24
Speedup	1.9	
Error rate (%)	4.55	4.55

[6] E. Osuna, R. Freund, and F. Girosi, "An improved training algorithm for support vector machines," in *IEEE NNSP*, A. Island, Ed., 1997.

[7] L. Kaufman, "Solving the quadratic programming problem arising in support vector classification," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. MIT Press, 1999.

[8] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. MIT Press, 1999.

[9] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. MIT Press, 1999.

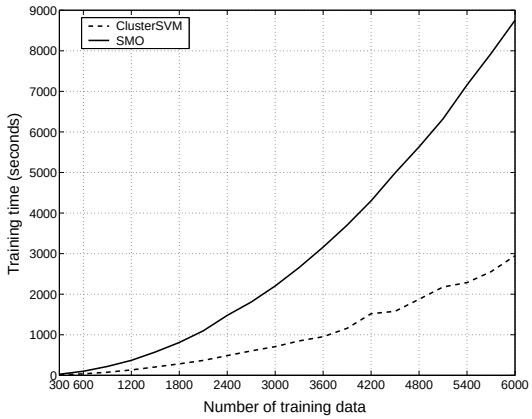

Fig. 7. Comparison of the scaling performance of *ClusterSVM* and that of SMO on the artificial data set. The solid line represents SMO and the dashed line represents *ClusterSVM*. The number of training data varies from 300 to 6000 with an interval of 300. For clarity purpose, labels on the horizontal axis only show every 600.

[10] S. Fine and K. Scheinberg, "Efficient svm training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2001.

[11] C. K. I. Williams and M. Seeger, "Using the nystrom method to speed up kernel machines," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Diettrich, and V. Tresp, Eds. MIT Press, 2001.

[12] D. Achlioptas, F. McSherry, and B. Schölkopf, "Sampling techniques for kernel methods," in *Advances in Neural Information Processing Systems 14*, S. B. Thomas, G. Dietterich, and Z. Ghahramani, Eds., 2002.

[13] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "A fast iterative nearest point algorithm for support vector machine classifier design," *IEEE Transactions on Neural Networks*, vol. 11, no. 1, January 2000.

[14] H. Yu, J. Yang, and J. Han, "Classifying large data sets using svm with hierarchical clusters," in *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[15] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, no. 12, pp. 783–789, 1999.

[16] C. C. Chang and C. J. Lin, "Libsvm: a library for support vector machines," 2001, version 2.33. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[17] J. Ma, Y. Zhao, and S. Ahalt, "OSU SVM classifier matlab toolbox 3.00." [Online]. Available: http://eewww.eng.ohio-state.edu/∼maj/osu_svm/

[18] D. L. Boley, "Principal direction divisive partitioning," *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 325–344, 1998.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

[20] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "Scatter/gather: a cluster-based approach to browsing large document collections." in *15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)*, 1992, pp. 318–329.

[21] The Mathworks Inc., "Matlab 6.1," http://www.mathworks.com. [Online]. Available: http://www.mathworks.com/

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. J. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, pp. 541–551, 1989.

[23] B. Schölkopf, C. Burges, and V. Vapnik, "Extracting support data for a given task," in *First International Conference on Knowledge Discovery and Data Mining*, U. M. Fayyad and R. Uthurusamy, Eds. AAAI Press, 1995.

[24] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html