Neurocomputing 172 (2016) 189–197

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Large-scale support vector machine classification with redundant data reduction

Xiang-Jun Shen [a,*], Lei Mu [a], Zhen Li [a], Hao-Xiang Wu [a], Jian-Ping Gou [a], Xin Chen [b]

[a] *School of Computer Science and Communication Engineering, JiangSu University, JiangSu 212013, China*
[b] *Hermes Microvision Inc., San Jose, CA, USA*

ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form
12 July 2014
Accepted 27 October 2014
Available online 8 May 2015

Keywords:
Support vector machine
Classification
Clustering
Redundant data reduction

ABSTRACT

Large-scale image classification has shown great importance in object recognition and image retrieval as the vast amounts of social multimedia sharing on the networks. While the time and memory requirements for SVM training surge with an increase in the sample size, which makes SVM impractical even for a moderate problem as the number of training data reaches to the extent of hundreds of thousands. To solve this problem, many specially designed algorithms are proposed such as clustering-based SVM training which attempts to remove the clustered data points that lie far away from support vectors. In this paper, we further explore that there exist clustered and scattered data points in a cluster. The clustered data points that lie around the clustering centroid are the dense data points, which are in the inner layer of a cluster. Those data points are viewed as having no SVs and removed. While the scattered data points are the sparse data points in the outside layer of a cluster. Those data points are viewed as having SVs and thus reserved. The Fisher Discriminant Ratio is employed to determine a boundary between the clustered and scattered data points in one cluster, which is computed based on the distance densities of data points to the cluster centroid. The redundant clustered data points in clusters are thus removed to speed up SVM training process. Several experimental results show that our proposed method has good classification accuracy while training time is significantly reduced. The training time in our proposed method only accounts for about 17 percent of the time in LIBSVM on the large data set of Covertype.

1. Introduction

Recent years have witnessed an explosion in the amount of images and videos in social media sharing websites, such as Flickr and Youtube, due to the spread of digital cameras, mobile devices and networking technology. An urgent need is how to effectively search these huge amounts of data efficiently. And this need has been recognized in the computer vision and machine learning research communities and large-scale classification methods have become an active topic of research [1–3] in recent years. Many classifiers, such as Nearest neighbor classifiers, decision tree, Bayesian classifiers, support vector machine (SVM), and the ensemble methods, have been proposed for image classification [4]. Among these technologies, SVM is the most prominent technology proposed by Vapnik [5] for its solid mathematical basis. It improves the generalization ability of a classifier by maximizing the margin between the two classes and thus achieves better classification accuracy on test data than using other classifiers on most popular benchmark data sets [6].

Despite its good theoretical foundations, SVM is not applicable for classification of large data set owing to vast time to train these large data set to obtain support vectors (SVs). And the SVs are the data points that are closest to the separating hyperplane in the training set. The decision for new data to be classified is solely based on the SVs. In order to obtain SVs, it is necessary to solve the quadratic programming (QP) problem, which depends heavily on the cardinality of the training set. And the optimization problem with big data causes an intensive computational complexity due to an increase in the number of data points used for training. This constraint makes SVM impractical even for a moderate problem [7] as the number of training data reaches to the extent of hundreds of thousands.

Much effort is devoted to reducing the time and space complexities when training large data sets. In general, these algorithms are divided into three categories: (I) dividing the original QP problem in SVM training into smaller QP sub-problems, (II) selecting a small number of representative training samples from the large data set to reduce the number of training data points, and (III) developing paralleled approaches which divide a large

[*] Corresponding author.
E-mail address: xjshen@ujs.edu.cn (X.-J. Shen).

http://dx.doi.org/10.1016/j.neucom.2014.10.102
0925-2312/© 2015 Elsevier B.V. All rights reserved.

data set into smaller one, each independently running on separate computer nodes.

For the first class of algorithms, it includes chunking and decomposition methods, which are discussed by Osuna et al. [8], Boser et al. [9] and Kaufman [10]. Platt [11] proposes the Sequential Minimal Optimization (SMO) algorithm that transforms the large QP problem into a series of small QP problems, each of which optimizes only a subset of size two. Platt's SMO algorithm is further accelerated by Pavlov et al. [12] using Boost-SMO algorithm.

The second class of SVM training approaches is to choose representative samples to scale down the entire training data before SVM training. Lee and Mangasarian [13] propose the Reduced SVM (RSVM) that uses the random sampling technique to select a random subset of training data. Systematic Sampling RSVM (SSRSVM) [14,15] is further proposed to select the informative data points to form the reduced set. Active learning is another technique used in SVM training to reduce the number of training data points [16–18]. And Tsang [19] proposes core vector machine algorithm that samples the data points on Minimum enclosing ball. Meanwhile, another big sub-class of such algorithms is clustering. Clustering [20] is an unsupervised learning technique that classifies similar objects into groups (clusters), according to some criteria such as distance metrics. Clustering algorithms aim to removing the training data that form clusters far away from a separating hyperplane. For example, hierarchical clustering [21,22], adaptive clustering [23] and fuzzy clustering [24] techniques are used to decrease the complexity of SVM training. Cervantes et al. [25] further presents a minimum enclosing ball clustering algorithm to classify large data. And the clustering and RSVM techniques [26,27] are combined. As for the third class of distributed and parallel algorithms [28–33], they utilize the computing and storage capacities in distributed or parallel nodes and thus can complete the SVM training task that cannot be completed in one node.

Consider a typical two-class SVM training. As the SVs lie on the boundary of the convex hulls of binary classes, the data points that are far away from the hyperplane are not useful for classification and thus can be removed from the training data set. So the aim of such clustering-based SVM training algorithms is to remove the clusters that are far away from the hyperplane.

While compared with the work of clustering-based SVM training algorithms mentioned above, the novelty in our work is that we consider not only to remove the redundant clusters, but we further consider data distributions in clusters and the redundant data points in every cluster are removed. Thus we explore that the data points in every cluster can be classified as clustered and scattered data points. The clustered data points that lie around the clustering centroid are the dense data points, which are in the inner layer of a cluster. Those data points are viewed as having no SVs and removed. While the scattered data points are the sparse data points in the outside layer of a cluster. Those data points are viewed as having SVs and thus reserved. So our aim is to find boundaries between the clustered and scattered data points in clusters. In this paper, the Fisher Discriminant Ratio (FDR) criterion [4] is applied to find such boundaries in every cluster based on the distance densities of data points to cluster centroids. The redundant clustered data points in clusters are thus removed and finally speed the SVM training process much more. Several experimental results on simulated and real data sets show that, compared with LIBSVM[34], our proposed method keep good classification accuracy while the training time is significantly reduced. The training time in our proposed method only accounts for about 17 percent of the time in LIBSVM on the large data set of Covertype.

The remainder of this paper is organized as follows: Section 2 briefly introduces the optimization problem involved in SVM and two-class classification in Fisher Discriminant Analysis (FDA). Section 3 describes the proposed method based on data redundancy reduction. In Section 4, our experimental results are reported on both the simulated and real data sets. Section 5 gives a conclusion.

2. Support vector machine and Fisher discriminant analysis

In this section, the fundamental concepts of SVM and FDA are described briefly.

2.1. Support vector machine

Assume that a training data set of binary classes is $\mathbf{D}\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, -1\}\}$, $i = 1, 2, \ldots, n$. $\mathbf{x}_i$ is the vector data point, n is the number of data points in the set $\mathbf{D}$ and the dimension of data point $\mathbf{x}_i$ is $\mathbb{R}^d$. y_i is the class membership of $\mathbf{x}_i$. Training SVM is to find the maximum margin hyperplane that separates the training data set.

The hyperplane is determined by a vector $\mathbf{w}$ with minimal norm and an offset b. A quadratic problem [5] needs to be resolved to find such an optimal hyperplane:

$$\min_{\mathbf{w},\boldsymbol{\xi}} : G(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \tag{1}$$

subject to

$$y_i(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where ξ_i is a slack variable to tolerate mis-classifications. C is a parameter that determines the cost of the constraint violation. The variable b determines the offset of the hyperplane from the origin. $\phi(\mathbf{x}_i) = (\mathbf{x}_i^1, \ldots, \mathbf{x}_i^m)$ is a mapping from vector field $\mathbb{R}^d$ into feature space $\mathbb{R}^m$ which is a higher dimension Hilbert space H. And $\langle \cdot, \cdot \rangle$ denotes the dot product in H. $\phi(\cdot)$ is a nonlinear function by using a kernel $K(\cdot, \cdot)$. And the kernel must satisfy the Mercer condition [5]: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$.

With the help of Lagrangian multipliers [5], the dual form of the above minimization problem in Eq. (1) is equivalent to

$$\max_{\boldsymbol{\alpha}} : W(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{n} y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \alpha_i \alpha_j + \sum_{i=1}^{n} \alpha_i \tag{2}$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_k = 0, \quad 0 \leq \alpha_i \leq C$$

where $\boldsymbol{\alpha}$ is a vector with components α_i that is the Lagrange multiplier.

The necessary and sufficient conditions for a weight vector $\mathbf{w}$ and Lagrange multipliers $\boldsymbol{\alpha}$ to optimal are the KKT conditions [5]. Many solutions of Eq. (2) are zero, which mean most α_i are zero, Based on the non-zero α_i, the optimal vector $\mathbf{w}$ is:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i) \tag{3}$$

where $\mathbf{w}$ is expressed by a linear combination of support vectors (SVs). The elements in the set SVs are the subset of training sample vectors $\{\mathbf{x}_i\}_{i=1}^{l}$, which have the l non-zero Lagrange multipliers $\{\alpha_i\}_{i=1}^{l}$. And the resulting classifier is

$$y(\mathbf{x}) = \sum_{i \in SVs} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b,$$

$$\text{sign}(y(\mathbf{x})) = \begin{cases} +1 : y(\mathbf{x}) > 0 \\ -1 : y(\mathbf{x}) < 0 \end{cases} \tag{4}$$