Applied Soft Computing 16 (2014) 201–209

Contents lists available at ScienceDirect

Applied Soft Computing

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

Neighbors' distribution property and sample reduction for support vector machines

Fa Zhu [a,b,*], Jian Yang [b], Ning Ye [a], Cong Gao [c], Guobao Li [a], Tongmin Yin [d]

[a] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, PR China
[b] Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, PR China
[c] Department of Computer Science, University of Regina, Canada
[d] College of Forest Resources and Environment, Nanjing Forestry University, PR China

ARTICLE INFO

Article history:
Received 26 April 2013
Received in revised form 1 November 2013
Accepted 16 December 2013
Available online 25 December 2013

Keywords:
Neighbors' distribution property
Sample-neighbor angle
Cosine sum
Sample reduction
SVM

ABSTRACT

For data pre-processing of SVMs, many scholars tried to find those samples, which would become support vectors. Generally, support vectors locate in the overlap regions, which are between different classes. But overlap region does not always exist. In this paper, a new method is proposed to find the boundary regions of each class instead of overlap regions. This method could deal with the dataset without overlap regions. Summing the cosine of the sample-neighbor angle, the sum ranges from 0 to k. When the sample locates in the boundary region of data distribution, the sum would be close to k; when the sample locates in the interior of the data distribution, the sum would be close to 0. Using cosine sum, the samples locating in the interior of each class can be disposed before SVMs training. Experimental results show that the proposed method can solve the problem, which the methods based on finding overlap regions cannot deal with.

1. Introduction

Support vector machine is a classical machine learning technology [1,2]. It has been widely used in many fields such as face recognition [3–5], gene selection [6], text categorization [7], and novelty detection [8]. Since Support vector machine needs to solve a quadratic programming [1,2]. The time complexity and space complexity are both very high. This limits the application of the support vector machine in large-scale datasets. In order to solve this problem, many scholars researched it and proposed many methods. Osuna proposed the decomposition algorithm, which fixed the working set size [9]. Platt proposed SMO algorithm, which only preserved 2 samples in the working set [10]. It can be deemed as a special case of decomposition algorithm. Tsang proposed core vector machine, which is based on minimum core-set [11]. The aim of those algorithms is to reduce the size of working set. Fung G replaced the inequality constraints of quadratic programming with equality constraints in order to avoid solving quadratic programming [12]. Joachims applied cutting plane methods to SVM [13]. Keerthi proved that solving the problem of maximum margin is equivalent to solving the problem of the minimum distance between two convex polytopes [14]. Only the samples on the convex hull influence on solving the problem of the minimum distance between two convex polytopes.

Another solution is to select a subset of training set. It can reduce the complexity of the original problem. A novel method, finding boundary regions of each class is proposed in this paper. Using boundary regions of each class to represent original set could avoid the assumption that there are overlap regions in training set, which proposed by Shin [19].

The remaining part of this paper is organized as follows. Related work will be introduced in Section 2. A novel method for reducing training set is proposed in Section 3. Section 4 provides the experimental results of the proposed method. Discussion and conclusive remarks are provided in Section 5.

2. Related work

The final classifier learnt by support vector machine is only relevant to support vectors. The other non-support vectors do not affect it. Merely preserving the support vectors, the accuracy of support vector machine does not deteriorate. Many scholars tried to select a special subset to represent the training set. Lee chose the subset by random selection and proposed Reduced SVM algorithm [15]. Almeida clustered the training set into several categories and used the center of each category to represent them [16]. The cluster method he used is K-means. Koggalage [17] and Zheng [18] also had

* Corresponding author at: College of Information Science and Technology, Nanjing Forestry University, PR China; Department of Computer Science, Nanjing Univerisity of Science and Technology, PR China. Tel.: +86 13512513764.
E-mail addresses: zhufag@gmail.com, zf_0902@163.com (F. Zhu).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.12.009

done similar research. The disadvantage of those clustering-based methods is that the performance of cluster algorithm is usually unstable. A related performance comparison was given by Liu and Nakagawa [29].

Generally, the samples become support vectors usually locate near the separation plane. Shin and Cho defined overlap region and proved that there is no difference between the original dataset and the selected subset, which consists of the samples in overlap regions. He found that there are more heterogeneous samples among k-nearest neighbors when the sample locates near the separation plane [19]. According to neighborhood properties, he proposed two measures: Neighbors Entropy and Neighbors Match.[1] The Neighbors Entropy and Neighbors Match are both greater than 0 when the sample locates near the separation plane. Neighbors Entropy is defined by formula (1).

$$Neighbors\_Entropy(x_i, k) = \sum_{j=1}^{J} P_j \cdot \log_J \frac{1}{P_j} \tag{1}$$

The parameter 'k' represents the number of nearest neighbors. The 'P_j' is defined by (k_j/k), 'k_j' represents the number of neighbors, which belong to class $\boldsymbol{j}$.

Neighbors Match is defined by formula (2).

$$Neighbors\_Match(x_i, k) = \frac{\left|\{x_i | label(x_i) = label(x_j), x_j \in kNN(x_i)\}\right|}{k} \tag{2}$$

The $kNN(x_i)$ is the list of k-nearest neighbors.

If there is no sample with class $\boldsymbol{j}$ ($k_j = 0$) in $kNN(x_i)$ ($P_j = 0$), the definition of Neighbors Entropy would be no sense.

Chang proposed scoring function to detect concept boundary [20].[2] And scoring function was defined by formula (3). The 'τ_j' is defined as the square of the distance from x_i to the nearest neighbor with the opposite class.

$$c(x_i, x_j) = \exp\left(-\frac{||x_i - x_j|| - \tau_j}{\gamma}\right), \quad x_j \in kNN(x_i) \tag{3}$$

It is determined by formula (4) whether a sample is preserved or not. When 'S_{x_i}' is greater than 0, the sample is reserved.

$$S_{x_i} = \frac{1}{\#x_j}_{x_i} \sum_{\text{s.t. } x_j \in kNN(x_i)} c(x_i, x_j) \tag{4}$$

The '$\#x_j$' is defined as the number of neighbors with the opposite class. The 'γ' is defined as the mean of $||x_i - x_j||_2^2 - \tau_j$. Similarly to Neighbors Entropy, the definition of 'τ_j' is also meaningless when there is no neighbor with opposite class among k-nearest neighbors.

The above two methods could find the overlap regions of a dataset. But in some cases, there is no overlap region between different classes, such as in Fig. 1. These methods are invalid.

In this paper, we try to find the boundary regions of each class instead. It could avoid assuming that there are overlap regions between different classes.

3. A novel method to reduce the training set

The information of a sample contains is relevant to its location. And the location can be determined by its neighbors' distribution. We first define mass center and then discuss neighbors' distribution properties.

[1] We use NPPS to represent his method.
[2] We use CBD to represent his method.

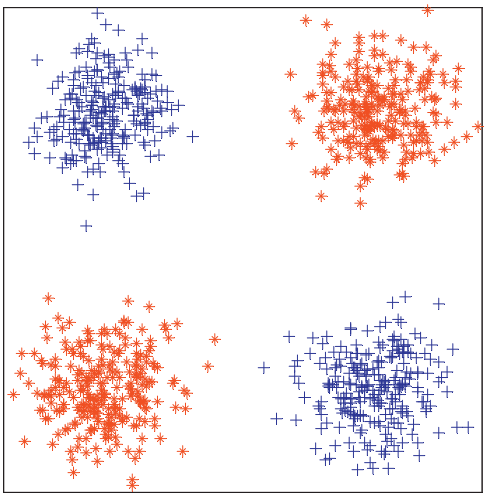

Fig. 1. A two-class dataset without overlap region ('+' is one class; '*' is the other class).

Definition 3-1(mass center). D is a dataset and $x_i (i = 1, 2, \ldots, k) \in D$ are the k-nearest neighbors of $x_0 \in D$. The mass center of the neighborhood sphere, which is formed by $x_i (i = 1, 2, \ldots, k) \in D$, is defined as: $\bar{x}_k = (1/k)\sum_{i=1}^{k} x_i$.

In Fig. 2, the neighborhood sphere of x_0 is divided into 2 parts by a hyperplane (AB), which is perpendicular to $x_0 - \bar{x}_k$. 'x_0' is a sample in dataset D, while $\bar{x}_k$ is mass center. We define the part, which mass center locates in, as $D1$ and the other part as $D2$.

Obviously, the distribution of neighbors in $D1$ ($D2$) can be classified into two cases: (1) the number of neighbors in $D1$ is close to the number of neighbors in $D2$, as shown in Fig. 3; (2) the number of neighbors in $D1$ is far more than the number of neighbors in $D2$, as shown in Fig. 4. Because the mass center locates in $D1$, it is impossible that the number of neighbors in $D1$ is far less than the number of neighbors in $D2$.

In case 1, the sample usually locates in the interior of the data distribution; in case 2, the sample usually locates near the boundary of the data distribution. We proposed the following two definitions.

Definition 3-2(sample-neighbor angle). θ is defined as the angle between $x_0 - \bar{x}_k$ and $x_0 - x_i$. ($\bar{x}_k$ is mass center, $\bar{x}_k = (1/k)\sum_{i=1}^{k} x_i$).

The cosine of θ is defined by formula (5).

$$c_{0,i} = \cos(\theta) = \frac{\langle x_0 - \bar{x}_k, x_0 - x_i \rangle}{||x_0 - \bar{x}_k|| ||x_0 - x_i||} \tag{5}$$

According to the Definition 3-2, it can be easily obtained that when the neighbor locates in $D1$, the angle $\theta < (\pi/2)$ and formula

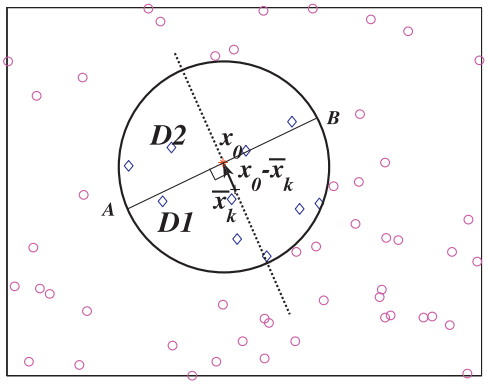

Fig. 2. The definitions of $D1$ and $D2$.