DOI: 10.21276/sjet.2017.5.3.3

Scholars Journal of Engineering and Technology (SJET)
Sch. J. Eng. Tech., 2017; 5(3):91-95
ISSN 2321-435X (Online)
ISSN 2347-9523 (Print)

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

An Improving Mulit-category Classification Method Based on the Binary Tree Support Vector Machine

BU Qing-chao
Mechanical Science and Engineering Institute, Northeast Petroleum University, Daqing 163318, China

***Corresponding author**
BU Qing-chao
Email: 906736181@qq.com

Abstract: Aiming at the problems of the slow convergence speed of general partial binary tree support vector machine (SVM) classifier and the fault samples easy to accumulate caused by complete binary tree and partial binary tree SVM classifier. The thesis proposes a multi-category classification method based on the unbalanced binary tree support vector machine (SVM),which construct a unbalanced binary tree SVM, making the easy to distinguish category can split out step by step from the root node, and reducing the accumulated errors caused by previous classification by analyzing the distribution of sample space. The results show that, comparing this method with the method of complete- and partial-binary tree ,an unbalanced binary tree SVM built in this paper has a strong ability of autonomous learning, and can easily distinguish separate classes first, thus improving classification accuracy.
Keywords: binary tree; support vector machine; lustering analysis; many kinds of points; pattern recognition.

INTRODUCTION

Support vector machine (Support Vector Machine, SVM Cortes and Vapnik1995 Corinna) is first proposed, it solve the small sample, nonlinear and high dimensional pattern recognition shows many unique advantages, and can be applied to the function fitting other machine learning problems [1]. SVM is based on statistical theory and structural risk minimization principle to ensure a good generalization ability [2]. However, as a new learning machine, support vector machine has some areas to be improved. The traditional support vector machine is to solve two types of sample classification problems, but the actual problem is often a multi class classification problem; therefore, how to effectively promote the two classification of multi class classification is one of the important research contents of the support vector machine [3].

At present, the support vector machine to solve the multi classification method has two kinds: one is the multi class problem solving, reflected the multi class classification problem is in an optimization formula, this method due to the unknown parameters, the sample structure is not clear, the large amount of calculation, high error rate, rarely used in practical problems in. Two is the use of logical thinking two category classifier according to a certain rule combination, form a multi category classifier, this approach is more common, more common are: classification method of [4] (OAO), a method of [5-6] (OAA), directed acyclic graph (DDAG) multi classification algorithm [7-8] the two fork tree (BT) [9-10], etc.. Among them, two binary tree algorithm compared with other algorithms, because of its simple structure, less the number of classifier classification accuracy and repeated training, the advantages of fewer samples are widely used, is a multi classification algorithm is very suitable.

IMPROVED TWO TREE SUPPORT VECTOR MACHINE

Traditional two tree support vector machine

More than two tree classification method (BT) is proposed by Vural and Dy [11], for multi class classification problem, this method will be divided into two sub categories, and then further divided into two sub categories of the second class, and so on, until all nodes contain only one category that is sample completely so far apart. A complex multi class classification problem is transformed into several two classification problems to solve. Due to the different forms of the two classification problems, the structure of the corresponding binary tree is also different, and the different levels of the structure have certain influence on the accuracy of the two fork tree. At present, there are two kinds of SVM classification in the tree: the partial tree and the complete two fork tree.

Each method has its own characteristics, such as partial two tree algorithm is the K class in the k-1 class as a class, the rest of the class as another major category, the establishment of a classifier of the two values. And then in the k-1 class, the -1 class (k-1) to be considered as a large class, the k-1 class of the rest of the class as another class, the establishment of a two value classifier. And so on, until the last two categories to build a classifier of the two. The advantage of this classification is that the classification accuracy is high, the computational complexity is small, and the classification model is easy to create. But it also has its own shortcomings, that is the partial structure of the two fork tree is generated randomly, based on the two fork tree structure has been determined by cannot be most easily the first class segmentation segmentation, affect the generalization ability, while increasing the error accumulation, and the classification time than completely two forks tree. The other is a complete two binary tree classifier, this method is all class divided into two sub categories, and then further divided into two sub categories of the second class, and so on, until all the samples so far apart. Some literatures refer to these two methods as "multi class support vector machine based on decision tree or hierarchical clustering". This method has the advantages of simple structure, small number of classifiers, short classification time, no non separable region. The problem is that all the samples are randomly classified after the model is built, which is easy to cause error accumulation, and the error of the method is higher than that of partial two tree, which affects the classification accuracy and generalization [12].

Comparison of complete two tree and partial binary tree algorithm

The total number of training samples for m, each category contains m/N samples, two partial binary tree and the two fork tree training time were Ta= $c\sum_{i-1}(\frac{N-i+1}{N}m)$, Tb= $c\sum_{i=1}^{\log 2^N} 2^{i-1}\left(\frac{m}{2^{i-1}}\right)^r$, C for R for the proportional constant, the power-law index, usually 2 ≦r ≦3. Ta>Tb can be obtained, that is, complete two fork tree training time is less than the bias of the training time of the fork tree, the other two tree training time between the two between the two. On the average number of classifiers, the average number of classifiers for the complete two fork tree is $\log 2^N$, and the average number of classifiers used by the partial binary tree is (N+1/2+1/N). The other two binary tree classifier number average value between the two, totally two fork tree classification classifier required less fast classification, which can be seen completely two forks tree and approximately two tree completely in the training time and classification speed is better than the partial binary tree two. But because of all the samples were divided into two groups is complete binary tree, no unified allocation rules, so the classification error completely two binary tree is higher than the two partial binary tree, and two tree is more likely to cause the error accumulation effect, generalization ability.

Common multi class classification analysis

1.Class N samples, two kinds of minimum separation measure, and then merge them together, as a category, denoted as X1, calculate the separability measure between X1 and the other N-2 samples, select the minimum of the measure, it will merge with class X1 as a category X2. And so on, until the last class, he will merge and N-1 as positive and negative samples, training support vector machine, is taken as the root node, and then take the last one with the class and all other classes as positive and negative samples from the remaining categories, training support vector machine, the second layer as a node. And so on, until the sample is completely separated from [13]. Although this classification has good classification effect in some samples, it has its own defects. In Figure 1, according to the above classification (1) as the initial interface, can be visually seen（2）The initial classification result is better than the interface (1). So for this distribution, the method will not separability strong preference separation, classification accuracy and generalization ability influence.

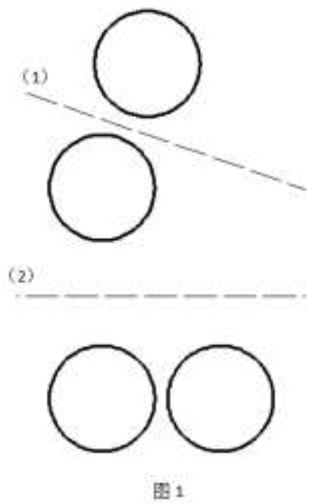

Fig-1:

2.In the N sample, select the smallest two class separability, they will be placed in the set S1, and then the rest of the sample from the N-2 class to select a kind of sample in S2 and S1 maximum separability, separability measure to calculate the remaining N-3 samples with S1, S2, and S1 S2, respectively of small sample in which, until all the sample classification is completed, the establishment of S1, S2 is about the subclass two binary tree classifier, SVM classifier training, structure parameters. And so on, until all samples are completely separated [14]. This kind of classification is commonly used in multi class classification, but it is not ideal for some special samples. In Figure 2, according to the above classification (1) as the initial interface, from the figure can visually see (2) for the initial interface classification effect is stronger than (1). Therefore, for the sample distribution in Figure 2, this method does not have a good classification effect.

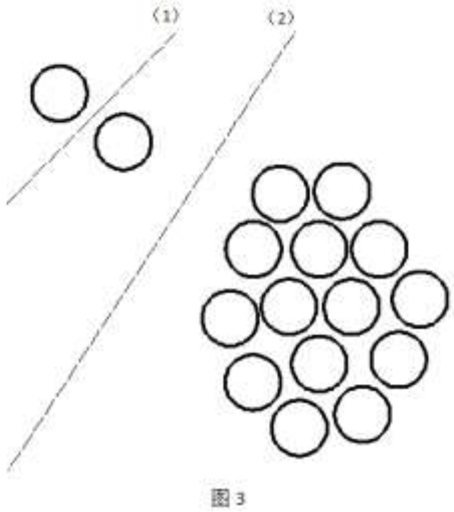

Fig-2:

3.The N sample class separability measure calculation, the separability from big to small order, the sample according to the separability (which gives off the first large separability out), and so on, until all the samples all separate [15]. This classification is suitable for most sample classification, but for some special samples, the classification effect is not good. In Figure 3, according to the above classification (1) as the initial interface, but it can be seen from the figure, (2) is stronger than the effect of the initial classification interface (1). As a result, the method does not have good generalization ability for the distribution shown in figure 3.

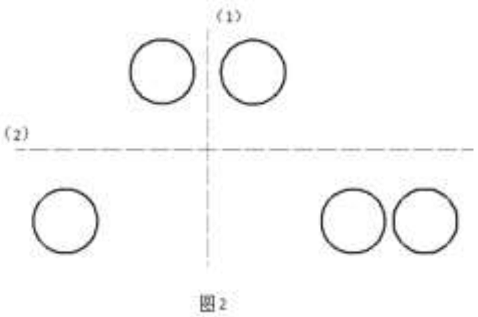

Fig-3:

To sum up, in the classification of samples, not only to consider a sample with other sample separability, but also consider the two different samples or between multiple samples that separability from multiple directions to the real separability of large sample was divided so that the classification error from the root node. To improve the classification accuracy.

Improved two tree support vector machine

The three classification methods introduced in the previous section are from one point, along a certain direction, without considering the integrity of the sample. This paper puts forward a classification method of multi categories, both with the sample separability measure and consider the integrity of the sample, the separability of large sample was divided to establish non balance two binary tree support vector machine, multi category classification is accurate and efficient.

The separability measure between samples of the common with Euclidean distance method [15] and method [16-17] ball structure, because the structure of the ball SVM classifier has the advantages of simple structure, high precision, is currently widely used classification algorithm. Here the ball structure method, the formula is as follows:

$$d_{i,j} = \left\|(a^i - a^j)\right\|_2 - (R^i + R^j) \qquad ①$$

$a^i - a^j$ Sample i and sample j over the distance between the center of sphere; R^i super sphere radius sample i; R^j super sphere radius sample j (hyper sphere center and radius by literature [18] type (10), (15) obtained).

By using formula separability measure between samples, i.e. the distance between the samples in each class, the minimum distance between the sample will be classified as a class (if there is some distance between samples with equal to the minimum, then put them into a category), using the literature obtained after the sample classification [18] hyper sphere center and radius at this time, the measure of separability between samples by formula; the increase of the distance will be a certain value (in order to ensure that most samples are separated, the value is defined as the maximum distance and the minimum distance difference and two times the sample taking), similarly, will be less than the distance from the sample is classified as a class, until last two samples (if not last remaining two, put down to the original 1/2, repeat the above steps). Experiments show that this method has better classification ability, the separability of large sample priority if using this method, separation, figure 1, figure 2, figure 3 priority classification surfaces (2), which shows that the sample of the algorithm for spatial position of unbalanced classification ability. Specific steps are as follows:

1. N samples, according to the [18] that each sample Super Center of sphere AI and radius Ri, the distance between the samples obtained publicity, remember the minimum distance Dmin between the samples and the maximum distance of dmax.

2. Calculate step： $d = \dfrac{d_{\max} - d_{\min}}{2 * \text{Sample category}}$

3.In the interval [dmin, dmax] with Dmin as the starting point, end point Dmax, D steps, calculating the distance between samples using value formula. If the distance between samples is less than or equal to Dmin, the sample is classified as a class (there are a few of the distance between the samples is less than or equal to Dmin, grouped into several categories; if the three types of samples, two samples of the distance is less than Dmin, a more than Dmin, the three samples were classified as a class). Calculate the classification hyper sphere centre and radius of the.

4. When the distance between all of the samples are larger than the Dmin, the dmin=dmin+d, the publicity obtained from the new sample of 3 steps that the set of samples, and compared with the Dmin, if the distance between samples is less than or equal to Dmin, the sample is classified as a class, and find the center of sphere radius and surplus.

5. If there is no final two kinds of samples, only to get a class of samples, then write down the Dmin at this time, the step will be changed to 1/2 at this time, repeat step 3, 4.

6. Repeat steps 3, 4, and 5, until the sample is divided into two categories, one for the left subclass, and one for the right subclass.

7. Repeat the steps in each subclass until the sample is completely separated.

By combining this method with two binary tree support vector machine, forming a combination of partial two binary tree and binary tree two completely unbalanced multiple classification methods, the structure of the two fork tree is determined by the spatial position of the sample, it can be completely two forks tree, can also be a partial two binary tree, the more is between two. This structure combines the advantages of the two, which is relatively simple structure, high accuracy, no unclassifiable regions, at the same time, the structure will be the largest sample separability was divided, the classification error from the root node to reduce the error caused by pre classification error accumulation. The structure is formed by the machine, has a strong ability to promote.

REFERENCES

1. Vapnik V. The nature of statistical learning theory. New York: Springer, 1995.
2. Zhi-jie L, Chun-yan Y. An Improved Support Vector Machines with Balanced Binary Decision Tree for Multi-class Classification[J]. Journal of Chinese Computer Systems. 2014;(05).
3. Yang W, You-li C, Wait LJ. A Binary Tree-support Vector Machine Multi-category Classification Based on Similarity Direction[J]. Journal of Sichuan Normal University(Natural Science). 2008;(06).
4. U-Pair Wise K. Classification and support vector machines. Advances in Kernel Methods Supprot Vector Learning, Cambrige. 1999; 255-268
5. Botton L, Cortes C, Denker J. Comparison of classifier methods:a case study in handw riting digit recognition.Proc of the 12thIAPＲ Int Conf on Pattern Ｒecongniton, Jerusalerm. 1994;77-82.
6. Lei X, Guang-zhou Z, Hong G. Novel one-vs-rest classifier based on SVM and multi-spheres. Journal of Zhejiang University. 2009;43(2):303-308.
7. Platt JC, Cristianini N, Shawe-Taylor J. Large margin DAGs formulticlass. Advances in Neural Information Processing Sys-tems, 2000;547-553.
8. Cheng Z, Zhao-feng M, Xin-xin N. A method basedon dynamic sequence analysis of an API and DAG-SVM multi-class support vector machines of unknow n virus detection. Journal of Chinese Computer Systems, 2012;33(12) : 2724-2728.
9. Madzarov G, Gjorgjevikj D, Chorbev I. A multi-class SVM classifier utilizing binary decision tree. Informatics, 2009;33(2):233-241.
10. Zhi-hua D, Chun-jiang Z, Xin-yu G. A new SVM multi-class classification algorithm based on balance decision tree. Control and Decision. 2011;26(1):149-156.
11. Vural V, Dy J. A hierarchical method for multi-class support vector machines. In: Proceedings of the 21st international conference of ICML. 2004, Banff, Alberta, Canada. 2004;105-113.
12. Sheng-fa Y, Fu-lei C. Support Vector Machines And Its Applications In Machine Fault Diagnosis. Journal of Vibration and Shock. 2007;(11).
13. Si-yu X, Hong P, Li-zuo J. Multi-class SVM method based on non-balanced binary tree. Computer Engineering and Applications. 2009;(07).
14. Zhi-qiang X, Li G, Jing Y. Improved multi-class classification incremental learning algorithm based on sphere structured SVM. Journal of Harbin Engineering University. 2009;(09).
15. Xin Z, Guo-hua Z, Xiang-fen Wait Z. Study on Transformer Fault Diagnosis Based on Euclidean Distance of Binary Tree SVM. Electrical Measurement & Instrumentation. 2013;(06).
16. Shuang L, Peng C. Improved hyper-sphere Support Vector Machine. Computer Engineering and Applications. 2009;(16).
17. Lin ZM, Dong LX, Fu CS. Solving the Problem of Multi-class Pattern Recognition with Sphere-structured Support Vector Machines. Journal of Nanjing University (Natural Sciences). 2003;(02).
18. Haop Y, Liny H. A new multiclass support vector machine with multi-sphere in the feature space. Lecture Notes in Computer Science. 2007.
19. Aiwu Z, Guojin W, Dandan C. An improved binary tree points support vector machine (SVM) method. Microcomputer & Its Applications. 2011;(06).

20. Zhi-qiang X, Li G, Jing Y. SVM multi-class classification algorithm based on full-binary tree of sphere-structured. Application Research of Computers. 2008;(11).
21. Tong Z, Jo F. Text categorization based on regularized linear classification methods. Information Retrieval. 2001;4(1):5-31.