SVM Classification Based on Supervised Subset Density Clustering

Yong Sun *,**, ZhenChao Sun*, Ran Zhan*, WeiDong Feng***, Geng Zhang****, ShiDong Liu****

* School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China

** Beijing Key Laboratory of Network System Architecture and Convergence, China

*** State Grid Hubei electric power company, China

**** China Electric Power Research Institute, China

sunyong@bupt.edu.cn

***Abstract*—A way of combining SVM(Support Vector Machine) with Supervised Subset Density Clustering is proposed in this paper. How to minimize the training set of SVM by means of clustering is researched. Original center positions are of great importance to clustering accuracy. However the traditional clustering center choosing algorithm doesn't work properly when the same kind of samples aren't closely-spaced or the shape of the sample distribution isn't regular, an self-adaptive multiple centers choosing method is proposed to solve the problem. Another problem addressed in the paper is that there are areas that are covered by multi-class samples which is of great difficulty for traditional clustering to deal with, so a supervised method for the improved density clustering is designed to make out such areas and referring the samples to SVM. The experimental results show that the algorithm reduces the iteration time of the whole training process without compromising the accuracy and generalization capacity of the algorithm obviously.**

***Keywords*—Self-adaptive, Center Choosing, Supervised Subset Density Clustering, SSDC-SVM, Classification**

I. INTRODUCTION

Support Vector Machine (SVM) [1] is a machine learning method put forward by Vapnik, which is based on the statistical learning theory (SLT) and structural risk minimization (SRM) principle [2] to solve binary classification problem. In the case of small samples, SVM algorithm has a strong portability and high classification accuracy. And it has become a useful tool in machine learning field and has been widely used in many fields, such as classification and pattern cognition. But when the scale of the sample set increases greatly, SVM training will take an unbearable long time. To solve this problem, many scholars have proposed improved algorithms, of which a popular one is diminishing the scale of the training set by clustering [3-6].

Clustering belongs to the field of unsupervised machine learning, which clusters samples by similarity. K-means is a clustering algorithm based on partition and has been widely applied in many fields. But it's performance is sensitive to the initial center positions. Most improved K-means algorithms rely on sample density to select initial centers [7, 8]. That is to select high-dense point as centers and to guarantee the Euclidean distance among centers is large enough. It proved effective, but the center will be fault if there exists more than one cluster in the dense region. Also, clustering is sensitive to noise points and the distribution shape of samples. Because the performance of clustering will affect the accuracy of SVM classification, therefore it's important to improve clustering performance.

In order to reduce the scale of SVM training set and guarantee the accuracy at the same time, this paper proposes SSDC-SVM (Supervised Subset Density Clustering-Support Vector Machine) algorithm. SSDC (Supervised subset density clustering) algorithm removes some samples and SVM trains the rest. The idea of SSDC algorithm is inspired by density clustering algorithm. There are two main innovations, one is adaptive center choosing to reduce the impact of sample distribution, and the other is supervised subset clustering to avoid getting centers from aliasing area. 3 datasets from UCI [10] data base and one

generated Gaussian-distributed datasets were tested and analysed. The classification result shows it's satisfactory for both accuracy and scale of training.

II. ALGORITHM DESCRIPTION

A. Algorithm Analysis

In this paper, how to find an effective way to diminish the scale of the training set is researched. The key part is the Supervised Subset Density Clustering algorithm (SSDC). There are several apparent improvements which have been made in contrast with the traditional clustering.

When the sample set is too large, it has been proved feasible by former researchers to select a relatively small number of samples to represent the features of the whole sample set, thus the complexity of cluster center choosing can be scaled down without scarifying the accuracy obviously.

The traditional K-means clustering isn't suitable for center choosing of classes with irregular shapes, in other words, when the samples of the same class aren't closely-spaced, the K-means algorithm can't find a center good enough, which may lead to the inaccuracy of the clustering. To solve this problem, change is made that in the clustering period number of centers isn't fixed, both number and position of centers can change to adapt to the distribution of the training set, and the strategy proved to be efficient in our experiment.

The traditional density clustering has an advantage over k-means in the aspect of center choosing, it determines the centers according to the density of samples. Centers are set in the areas where the points are closely-spaced, but it can't distinguish the high-density conditions where different kinds of samples are spaced closely, leading to incorrectly choosing centers. So the idea of supervision center choosing is adopted. When the potential centers are found by density clustering, each one will be checked if it's correctly chosen. Thus the accuracy of clustering is improved.

After all the improvements made above have been applied, the clustering process starts to discard the points irrelevant to SVM training. The points within a certain distance from the centers are regarded as inner part of the class and won't do influence to the construction of hyper-plane so these points are discarded in this period. Then the training process of SVM adopts the regular training algorithm since scale of the training set has diminished apparently.

B. Terminology definition

In clustering process, to distinguish the importance of different features, w (weight value) of every feature for each sample is defined as:

$$w_r = \frac{\sum_{i=1}^{n} x_{ir}}{\sum_{i=1}^{n}\sum_{r=1}^{m} x_{ir}}, w = (w_1, w_2, ..., w_m) \quad (1)$$

to strengthen the important factors and to weaken the inferior, thus making the clustering more reasonable.

Note that n is the number of samples and m is the dimension of the feature.

Then wd (the weighted Euclidean distance) is defined as:

$$wd(x_i, x_j) = \sqrt{\sum_{r=1}^{m} w_r (x_{ir} - x_{jr})^2} \quad (2)$$

awd (average weighted euclidean distance) is defined as

$$awd = \frac{1}{n^2} \sum_{i,j=1}^{m} wd(x_i, x_j) \quad (3)$$

Considering the mixed areas can seriously violate the centers choosing, dp (density parameter) of each point is defined in the process accordingly to avoid the inaccuracy. It is the number of same-kind points one sample has around it within a certain distance of $\alpha \cdot \mathrm{awd}$ in the hyperspace.

$$dp(x_i) = \sum_{j=1}^{n} \mathrm{u}(\alpha \cdot \mathrm{awd} - \mathrm{d}(\mathrm{x}_i, \mathrm{x}_j))$$
$$\mathrm{u}(\mathrm{x}) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}, 0 < \alpha < 1 \quad (4)$$

Note that there is a new coefficient α in the formulas (4), it is an adjusting coefficient to adapt to different datasets.

Then adp (average density parameter) is defined to find the best threshold.

$$adp = \frac{1}{n} \sum_{i=1}^{n} dc(x_i) \quad (5)$$

Drop Rate is defined as:

$$\text{drop rate} = \frac{\text{number of samples dropped by clustering}}{\text{number of the total samples}} \quad (6)$$

to evaluate the performance of the clustering scaling downing the training set.

With all the definition above it's possible to adopt the Supervised Subset Density Clustering algorithm called SSDC to diminish scale of the training set efficiently.

All the variables defined above are essential to our SSDC-SVM algorithm, of which w is special defined according to the SSDC algorithm.

C. Algorithm description

1)Supervised Density Clustering:

The algorithm is described as follows,

1) Select samples from the original dataset D randomly to make a smaller training set, then classify the points manually and mark their classes.

Suppose the number of the samples is n and that of features is m.

2) Calculate the weight of every feature and then calculate wd between every two sample points accordingly, and furthermore, average distance of the whole set is computed as (3).

3) Compute dp for every point and adp of the training set according to (4) and (5). Then mark the samples with points of different kinds around it.

4) Check dp of every point and mark the ones whose dp is lower than the threshold $\beta * adp(0 < \beta < \alpha)$ as $label(x_i) = -1$, then put all the points $label(x_i) \neq -1$ into set Cc.

5) Find the point Xi with the largest dp in Cc, then check the number Nc of Xi, if there is only one, then the point Xi found is the center and is marked as Xc, move it to the center set C, else if Nc>1, then calculate the average position Xav of the points found, check if the point Xi is within the distance $\gamma \times awd, (0 < \gamma < \alpha)$ from the center, leave alone the points out of the range, then recalculate the center using the points left and discard them afterwards, put the newly chosen center in C and mark the center's class.

6) Discard the points in Cc which is within the distance of awd from the center just chosen.

7) Repeat 5) and 6) until Cc is null.

2)SSDC-SVM

With the points irrelevant to the SVM hyper-plane discarded by SSDC, the training process of SVM becomes much easier. The rest part of the whole algorithm is:

1) Calculate distance of all points in the original sample set D to the centers chosen by SSDC $d_w(X_i, X_c)$

2) Discard the point if $\min(\mathrm{d_w}) < \eta \times \mathrm{avd_w}, 0 < \eta \leq \alpha$

3) Train the left points with the traditional SVM training method.

4) Test the classifier, and optimize the parameters until the classifier is steady.

III.EXPERIMENT AND SIMULATION

The simulation is based on the dataset Wine, Iris and Skin from UCI database and dataset[10] generated which follows Gaussian distribution, of which dataset wine consists of 178 samples with 13 features, and iris consists of 150 samples with 4 features, for these two datasets, choose 100 sample points from each one randomly as the input of clustering. While skin consists of 245057 samples with 3 features, choose 5000/10000/20000 points randomly as the training sample, and the generated dataset consists of 23000 points with 2 features, choosing 5000/10000/20000 points as the input of supervised subset density clustering. The software used in the experiment is Matlab (R2010b) and Libsvm-3.18 [9].

This section consists of two parts, the first is about the simulation of SSDC and analysis of its performance, the second is about the simulation of SSDC-SVM and its analysis.

For each dataset, traditional K-means center choosing, DC (Density center choosing Clustering) and SSDC are all carried out to compare the efficiency.

A. SSDC

In the SSDC process, the points correctly discarded in the clustering period make hardly any difference to the distribution of the support vectors in SVM training progress, so we care mainly about how many points can be diminished by the SSDC process and the rate of points being discarded correctly. The result is as shown in Fig. 1 and Fig. 2. In Fig. 1, the accuracy and points dropping rate of the algorithm on Wine are tested. In Fig. 2, accuracy of the clustering is set to 100% to compare the drop rate of different algorithms.

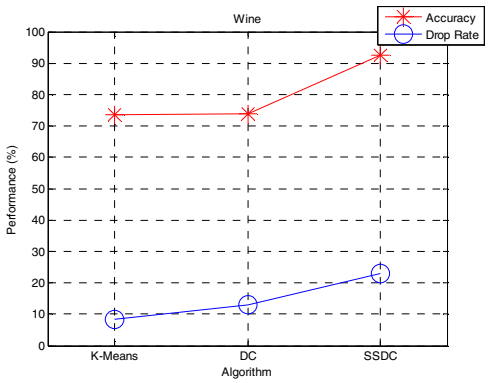

Fig.1 Clustering Performance on Wine

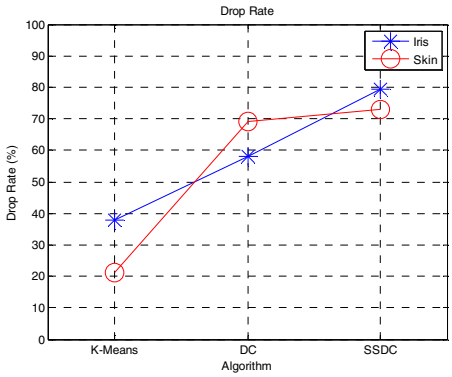

Fig.2 Drop Rate of three Algorithms

To make the result more convincing, number of classes isn't assigned for DC, and in each experiment for DC, more than one center is chosen, thus the influence caused by the distribution shape of samples is balanced. However the traditional DC cannot make out the crowded areas covered by many different kinds of points while the SSDC is capable to erase the influence which is critical to improve the clustering accuracy, and in addition the SSDC deals with only a part of the whole training set, reducing the total time of training and reduce the difficulty of marking the samples. In brief, SSDC has huge advantages in clustering accuracy and drop rate over the traditional algorithm.

B. SSDC-SVM

The process consists of two periods, firstly diminish the training set with clustering algorithm, secondly train the data left with the SVM algorithm. After the above two steps the hyper-plane is constructed and the SVM classifier can do the classification on the data left to verify the performance. The SVM adopts the RBF kernel [11], this experiment tested the algorithm with the skin dataset and generated dataset, and the parameters C and g were optimized by cross-validation. The result is shown as follows:

TABLE 1. PERFORMANCE OF CLASSIFICATION ON SKIN

Algorithm	Performance			points
	Accuracy (%)	Drop Rate (%)	SV	
SVM	96.77	0.00	41	5000
DC-SVM	92.03	56.46	21	
SSDC-SVM	95.93	55.56	37	
SVM	97.07	0.00	65	10000
DC-SVM	88.13	59.03	25	
SSDC-SVM	97.30	53.83	58	
SVM	99.63	0.00	181	20000
DC-SVM	91.50	55.69	149	
SSDC-SVM	99.53	51.4	173	

TABLE 2. AVERAGE PERFORMANCE OF CLASSIFICATION

Algorithm	Skin		Gaussian	
	Accuracy	Drop Rate	Accuracy	Drop Rate
SVM	97.82	0.00	84.60	0.00
DC-SVM	90.55	57.09	84.25	13.93
SSDC-SVM	97.58	53.60	84.45	14.70

TABLE 3. PERFORMANCE OF CLASSIFICATION ON GAUSSIAN DISTRIBUTION

Algorithm	Performance			Points
	Accuracy (%)	Drop Rate (%)	SV	
SVM	84.5	0.00	140	5000
DC-SVM	84.63	14.2	127	
SSDC-SVM	84.5	15.28	140	
SVM	84.6	0.00	267	10000
DC-SVM	83.53	13.56	247	
SSDC-SVM	84.57	13.6	266	
SVM	84.7	0.00	485	20000
DC-SVM	84.6	14.03	446	
SSDC-SVM	84.73	15.24	486	

Table I shows that performance of SSDC-SVM has an apparent advantage in accuracy and drop rate on the skin dataset. When the scale of the training set is larger, the accuracy of SVM and SSDC-SVM is improved, but the scale of training set of SSDC-SVM is much smaller than that of SVM because of SSDC dropping irrelevant points. Taking the number of support vectors into account, both SSDC-SVM and DC-SVM have a loss, SV number of SSDC-SVM declined by 4.88% on average compared with SVM, while the ratio is 32.06% for DC-SVM, which is the main reason why the accuracy of SSDC-SVM declines and accuracy of DC-SVM is the worst of the three.

Table II is the comparison of performance among the three algorithms on the Skin set and generated Gaussian set. As is shown in the table, the SSDC-SVM performs better than DC-SVM and SVM, especial in practical conditions.

Table III shows that on the generated dataset that follows Gaussian distribution both DC-SVM and SSDC-SVM have an improvement in diminishing the scale of training set, but note that there are points wrongly dropped in the process of DC while in SSDC all the points are discarded correctly, which means the robustness of SSDC-SVM is better than DC-SVM.

The distribution of the data of the Skin set in different stages is shown in the following figures to interpret the process and advantage of SSDC-SVM.

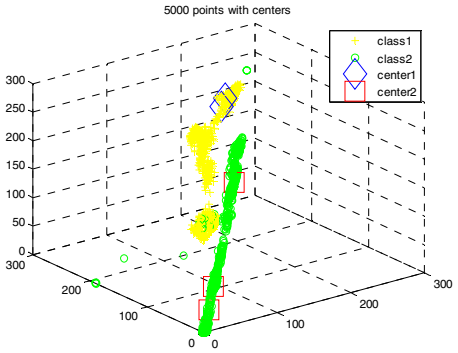

Fig.3 Distribution of 5000 samples and centers

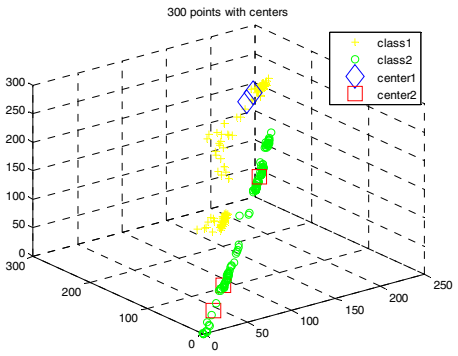

Fig.4 300 randomly chosen samples with centers

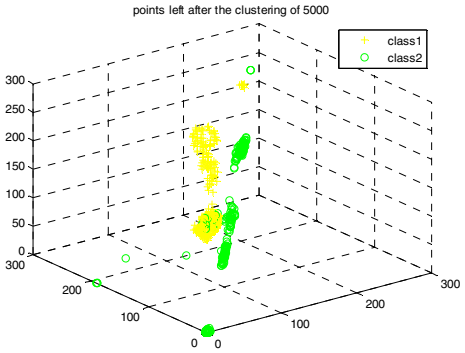

Fig.5 Points left after SSDC

Fig. 3 is the 5000 points chosen as the original training set, Fig. 4 is the randomly chosen 300 points that is applied in the clustering process. Fig. 5 is the samples left for the SVM training.

Fig.3 and Fig.4 show that the distribution of the subset chosen randomly is similar to the original distribution of the set. So it's reasonable to decide the centers by the subset. As is shown in the figures, the distribution of the sample is approximately linear, so SSDC chooses several centers according to the adaptive-center choosing method. There are also areas in the figure which are covered by different kinds of points, density of these areas are large as well, but the SSDC successfully makes out the confusing area and leave them to SVM as shown in Fig. 5, which is an improvement than the traditional density clustering.

As discussed above, the algorithm are feasible in the aspect of balancing the distribution of the samples and correctly making out the centers of the same kind, thus making the points irrelevant to SVM hyper-plane construction dropped accurately as many as possible.

IV. Conclusions

The SVM training is hard to achieve when the scale of dataset is too large, the training time is too long to tolerate. So the SSDC-SVM algorithm is put forward, aiming to optimize the clustering process, and the inaccuracy in the clustering preprocessing caused by the irregular distribution of samples and samples of different kind mixing with each other is the main target for us to diminish. With all the experiment done, the algorithm is proved feasible and valid.

ACKNOWLEDGMENT

This work was supported by Science and Technology Projects of the State Grid Corporation of China (XXN17201400030), State Grid Hubei Electric Power Company, NSFC (No.61101106), and Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications.

REFERENCES

[1] Xue-Gong Zhang. "Introduction To Statistical Learn In Gtheory and Support Vector Machines". *Acta automatica sinica*. , vol.26(1).pp.32-42,2000.

[2] Vapnik V N. *Statistical learning theory*. New York：Wiley, 1998.

[3] Hua Tan, Min Zhang, Xi Tan, Yi-Dan Su. "The Optimization of Large Scale Multiple Kernel SVM Based on K-means Clustering in Kernel Space". *Internet Technology and Applications*, 2010.

[4] Shi Zhou, Bing Chen. "Intrusion Detection Method with Reduced Weighted SVM Based on Clustering and Distance Comparison". *Journal of Data Acquisition & Processing*. 2009, 24(2).pp. 232-237.

[5] Xiao-Zhang Liu, Guo-Can Feng. "Kernel Bisecting k-means Clustering for SVM Training Sample Reduction". *Pattern Recognition*, 2008 .P.8-11.

[6] Xin-Mei Tian, Xiu-Qing Wu, Li Liu."A New SVM Iterative Algorithm in Large Training Set". *Computer Engineering*. 2007, 33(8)pp.205-207

[7] Min Huang, Zhong-Shi He, Xin-Lai Xing. "A New k-means Clustering center select algorithm". *Computer Engineering and Applications*. 2011,47(35)pp.132-134

[8] Chang-Zheng Xing, Hao Gu. "K-means algorithm based on average density optimizing initial cluster centre". *Computer Engineering and Applications*.2014, 50(20)pp.135-138.

[9] Chin-Chung Chang, Chin-Jen Lin. LIBSVM: A Library for Support Vector Machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[10] UCI data set. [Online]. Available: http://archive.ics.uci.edu/ml/

[11] Wang Peng, Zhu Xiaoyan. Model Selection of SVM with RBF Kernel and its Application. Computer Engeering and Applications. 2003,39(24): pp.72-73.

Yong Sun (M'12) received the Ph.D. degree from Beiijng University of Posts Telecommunications, Beijing, China, in 2008. He is currently a Lecturer with the School of information and communication engineering, Beiijng University of Posts Telecommunications, Beijing, China. He became a Member (M) of IEEE in 2012. His current research interests include heterogeneous networks, wireless resource allocation, and network management.