Research of Incremental Learning Algorithm for SVM Based on Class Center Diameter

Jinfeng Li\textsuperscript{1, a}, Wenhao Xie*\textsuperscript{2, b}

\textsuperscript{1}School of Computer Science and Technology Xidian University Xi’an, China
\textsuperscript{2}School of Science Xi’an Shiyou University Xi’an, China
\textsuperscript{b} Corresponding author: xwhaoxwhao@163.com

Abstract—In the learning process based on ISVM, how to effectively retain the history information and selectively discard some new training data, so as to maintain the classification accuracy and save the storage space after adding new samples every time, which is the key of the current ISVM classification algorithm. This thesis proposes a new incremental learning algorithm, that is, SVM incremental learning algorithm based on cluster diameter (CD-ISVM). This algorithm firstly calculates the two centers of the positive and negative samples, and then to build the boundary vector set by the coordinates of the two class centers. Moreover, KKT conditions are combined to filter incremental data and boundary vector set, and feedback information is provided to adjust the position of the boundary vector set. Then, the union of SV set and boundary vector set is taken as the increment set. After several increments, a classifier with strong generalization is finally trained.

1. Introduction

With the rapid development of information technology, the sample data collected in the field of industrial production often with dynamic change, constant update and other characteristics. With the accumulation of new samples, the traditional SVM classification algorithm based on batch data will continuously increase the training time and storage space, and it cannot dynamically update the classification model based on new samples while retaining the original classification information. In order to solve the above problems, an incremental learning algorithm based on Support Vector Machine — ISVM was proposed\cite{1-4}. Compared with traditional machine learning methods, the incremental learning algorithm makes full use of the previous results in the new training, can effectively retain the history information, and at the same time learn the newly added data information in real time, form a gradual and continuous learning process. In addition, incremental learning loses worthless samples, reduces the training time after incremental learning, and it does not need to maintain the initial training results, greatly reducing the storage space, and has important practical significance.

At present, many domestic and foreign researches on ISVM algorithm mainly focus on the optimization of support vector (SV), which are divided into two categories: one is incremental learning algorithm based on sample selection strategy\cite{5}; the other is incremental learning strategy based on KKT condition\cite{6}. For example, ISVM algorithm based on generalized KKT condition optimization support vector set proposed in literature\cite{7} and ISVM algorithm based on cluster preselection vector proposed in literature, etc. In order to adapt to the precise classification of huge amounts of data, literature\cite{8} proposes a SVM incremental learning algorithm for SVM based on combined reserved set. This algorithm is based on building the reserved set. Certain numbers of the original samples and the
incremental samples which satisfy the KKT conditions are added to the reserved set, used to build the new training set for the next incremental learning.

It can be seen from the above studies that, in incremental learning, with the addition of new samples, how to select a new support vector set so that useful information will not be discarded but the original training results can be retained has become an important content in the construction of ISVM learning model.

2. Relationship Between KKT Conditions And Sample Distribution

In the learning process of ISVM, KKT conditions can be used to verify the impact of new data objects on historical training results. The nonlinear classification model based on SVM can be formulated as convex semidefinite programs:

\[
\begin{array}{l}
\min \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{l} \alpha_i \\
\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \\
0 \leq \alpha_i \leq C, i = 1, \ldots, l
\end{array}
\]

(1)

Solving the above problem to obtain the classification function: \( y = \text{sgn} \, f(x) \), where \( f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \) is the decision function, if and only if each sample point in the training set satisfies the KKT condition below, that is:

\[
\begin{align*}
\alpha_i = 0 & \Rightarrow y_i f(x_i) \geq 1 \\
0 < \alpha_i < C & \Rightarrow y_i f(x_i) = 1 \\
\alpha_i = C & \Rightarrow y_i f(x_i) < 1
\end{align*}
\]

(2)

\( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l) \) is the optimal solution to the above programming problem.

Set as the new data object, then the samples which breach the KKT conditions generally have the following three situations [9]:

1. \((x_i, y_i)\) is located in the margin, and on the same side of the classification boundary with the samples of the same type, it can be correctly classified by original classifier, satisfies \( 0 \leq y_i f(x_i) < 1 \);

2. \((x_i, y_i)\) is located in the margin, and on the opposite side of the classification samples of the same type, it cannot be correctly classified by original classifier, satisfies that \( -1 \leq y_i f(x_i) \leq 0 \);

3. \((x_i, y_i)\) is outside the margin, and on the opposite side of the classification boundary with the samples of the same type, it can’t be correctly classified by original classifier, satisfies that \( y_i f(x_i) < -1 \);

In summary, the sample points that break the KKT conditions satisfy that \( y_i f(x_i) < 1 \).

Reference[10-11] tells us when refactoring support vector set, we can not only consider the new samples that violated the KKT conditions as the training samples of classification model, but also consider the nonsupport vectors near the border of the classification hyperplane because they could be the new support vectors.

3. SVM Incremental Learning Algorithm based on Cluster Center Diameter (CD-ISVM Algorithm)

3.1. Summary of CD-ISVM algorithm

Due to the disadvantages of traditional SVM-based incremental learning algorithm such as low efficiency, memory limitation and poor generalization, many scholars have further optimized and improved it. Based on previous studies, this paper makes some improvements on ISVM, and proposes
a new incremental learning algorithm (CD-ISVM) based on cluster diameter, which in some extent improves the classification accuracy and classification efficiency.

This paper first calculates the centers of the two classes of samples, extracts the two data dimensions with maximum classification information, the line of the two centers of the positive and negative samples is taken as the diameter, and uses the samples falling into the boundary domain as the initial training samples of SVM to obtain the initial support vector set. For the newly added samples, first uses the KKT conditions for screening to determine whether they violate the KKT conditions. For the samples that violate the KKT conditions, through the geometric knowledge which carried by the samples in to modify the scope of the elliptic domain, that is averages the class centers of the positive and negative of the new samples with the previous class centers, obtains the updated class centers’ coordinates, and according to the weighted sum of the ratio of these points to the diameter of the positive and negative class centers to determine the diameter of the new boundary vector. From this, it obtains a new round of incremental learning set, and then together with the previous round of SV set to carry out a new round of SVM training, repeats the process until all samples are trained.

3.2. The selection of the boundary vectors
The CD-ISVM algorithm first calculates the average distance between the samples to get the positive and negative class centers, for the positive or negative data, each class center is generally defined as:

\[ C_i = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \quad (i = 1, 2) \]  \hspace{1cm} (3)

For any 2 points of dimensional sample points, in the original vector space, use Euclidean distance to define the distance between them:

\[ d(x_i, x_j) = \|x_i - x_j\| = \sqrt{\sum_{i=1}^{d} (x_i^j - x_j^j)^2} \]  \hspace{1cm} (4)

Therefore, we can get the coordinates of the centers of the positive and negative class data as:

\[ C_+ = \frac{1}{n} \sum x^+, \quad C_- = \frac{1}{n} \sum x^- \]  \hspace{1cm} (5)

According to the above method, we can obtain the two center coordinates of the positive and negative samples. Taking the distance between these two centers coordinates as the diameter; the elliptic domain is drawn in the feature space as the boundary vector range.

Define parameters \( \sigma, \mu \) (the initial value \( \sigma = 1 \), the initial \( \mu \) value is 0) as the parameters to divide the size of boundary vector set. According to different distribution of data set, dynamically adjust the size and position of boundary vectors range, so that as many data points with classification feature information falls in this region as possible.

**Definition 1:** According to the coordinates operation of the vectors in Euclidean space, it obtains the coordinates of the center point of the boundary vector set:

\[ x_{center} = \frac{1 - 0.1 \mu}{2} (C_+ + C_-) \]  \hspace{1cm} (6)

Among them, parameter \( \mu \) is used to correct the position of the center point, so that as many effective data points fall into the set of boundary vectors as possible, and its definition is shown in (12).

**Definition 2:** For the axial length sum of the set of boundary vectors \( a \) and \( b \), it can be defined as:

\[ b = \sigma d(C_+, C_-) \]  \hspace{1cm} (7)
\[ a = 2b \]  \hspace{1cm} (8)

Among them, \( \sigma \) represents the adaptive parameter of the set of boundary vector, according to the different distribution characteristics of the data set, it can adjust and correct the width of the boundary vector set, \( d(C_+, C_-) \) represents the distance between the center of the positive sample and the center of the negative sample.
Definition 3: define the angle $\theta_e$ between the direction of axis $b$ and the axis $x$, take a basis vector in space $e_i = (1,0)$ with the direction vector of the set of boundary vectors as inner product, it can obtain the included angle expression $\theta_e$ :

$$\theta_e = \arccos(e_i \cdot (C_+ - C_-) / d(C_+ , C_-))$$

Among them, $(x_i - x_c)$ represents the direction vector of the set of boundary vectors, that is, the difference between the coordinates of positive and negative center points.

According to the above parameters, it can obtain the elliptic domain as shown in Fig 1. The data set located in this ellipse is defined as the set of boundary vectors, the star symbols indicate the central coordinates of the positive and negative classes, the red range represents the set of boundary vectors. In the CD-SVM algorithm, every time add new samples, the boundary vector set will replace the SV set in the traditional ISVM algorithm as the training samples to train the classifier again.

![Figure 1 The set of boundary vectors](image)

3.3. The determination of the incremental data sets

Based on the description of it, this paper selects the set of boundary vectors to train the classifier again.

(1) For the newly added sample points $x_i (i = 1,2,3\cdots)$, CD-ISVM algorithm calculates the positive and negative class center coordinates and combines with the class center coordinates obtained from the last operation to obtain the new class center coordinates. That is:

$$x_+ = \frac{1}{n_+ + n_n^+} (n_+^+ \cdot C_+ + n_n^+ \cdot C_+^{add})$$

$$x_- = \frac{1}{n_+ + n_n^-} (n_+^- \cdot C_- + n_n^- \cdot C_-^{add})$$

Among them, $C_+^{add}$, $C_-^{add}$ respectively represent the positive and negative class centers of the new data set, and $C_+, C_-$ represent the positive and negative class centers of the original data set, $n_+^+, n_+^-$ respectively represent the number of the positive class samples of the original data set and the new data set, and $n_n^+, n_n^-$ respectively represent the number of the negative class samples of the original and the new data set.

At the same time, it calculates the distance ratio between the newly added $x_i$ from the center of positive class $C_+$ and the center of negative class cluster $C_-$, denoted as $S_i$, and can be defined as:

$$S_i = \frac{d_+(x_i, C_+)}{d_-(x_i, C_-)}$$

(11)
It can use $S_i$ to quantify the intensity of the relationship between the newly added test sample $x_i$ and the positive and negative data centers, that is when $S_i > 1$ indicates that the sample is close to the negative class, when $S_i < 1$ indicates that the sample is close to the positive class, for all newly added samples, define as:

$$
\mu = \frac{1}{n_n} \sum_{i=1}^{n_n} 1 - S_i
$$

Among them $n_n$ represents the number of new added samples, parameter $\mu$ can describe the degree to which the distribution of new samples is close to the center of the positive class and the negative class, therefore, the boundary vector region is moved to the class with fewer newly added samples, so that the positive and negative data are more balanced in the training.

(2) Using the above parameters, construct a new set of boundary vectors, make it more adaptable to the new added data, and then determine whether the incremental data $x_i$ belongs to the set of boundary vectors. For the data falling in the boundary vector region, added it into the boundary vector set for the next calculation.

(3) Update the set of boundary vectors. In order to accurately describe the boundary condition of the data, all the support vectors of the previously trained support vector machine were found and added it into the set of boundary vectors.

Through the above three steps, the boundary vector set can be selected as the training set for ISVM classifier.

3.4. The steps of the CD-ISVM algorithm

Based on the above analysis, this paper proposes a new incremental learning method of SVM—CD-ISVM algorithm. This algorithm tries its best to retain some data that may be converted into support vectors, and abandons the useless samples for classifier training, so as to ensure the algorithm accuracy and improve the algorithm classification efficiency.

The specific process of this algorithm promoted by this paper is described as follows:

Premise: Suppose that the initial sample set is:

$$
T = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}, x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, 2, \cdots, n
$$

and the new sample set is $T^\pm$.

Algorithm target: Finding the SVM classification based on $T \cup T^\pm$.

Step1. Calculate the boundary vector set $B$ according to the positive and negative class center;

Step2. Perform SVM training on set $B$ to obtain SV set, and the classifier is denoted as $S$;

Step3. Check whether the increment set exists. If not, the algorithm will end and $S$ is the final classifier. Otherwise, enter into Step4;

Step4. For the newly added data, first judge whether there are samples violating KKT condition. If there are no such samples, the algorithm will end and $S$ is the final classifier. Otherwise, the samples violating KKT condition will be added to the vector set;

Step5. Calculate the positive and negative class center of the samples in $T^\pm$ and average with the original class center to get the new positive and negative class center;

Step6. Calculate the positive and negative distance center ratio from the new class center, and then update the parameter $\mu$;

Step7. Construct a new boundary vector set $B^*$ by the class center coordinates and parameter $\mu$, and add the SV set into the set $B^*$together;

Step 8. Set $B = B^* \cup T^*$ and go to Step 2.

4. Experiment And The Analysis Of The Experimental Results

In order to verify the effectiveness of the proposed method, we have a numerical simulation experiment on PC. We use different number of data sets with different distribution characteristics to simulate and
analyze the experimental results of the incremental learning algorithm, and analyze the advantages and disadvantages of the algorithm and its reliability.

4.1. Experimental data
This paper selects some classical dichotomous data sets in UCI [12] to carry on the experiment. The data sets of UCI used in this paper and their data characteristics are shown in Table I.

| Data set   | Train sample size | Test sample size |
|------------|-------------------|------------------|
| breast_cancer | 2000              | 770              |
| heart      | 4250              | 2500             |
| mushroom  | 5644              | 1000             |
| splice     | 10875             | 5000             |
| thyroid    | 2800              | 1500             |
| titanic    | 10255             | 750              |

4.2. Experimental results
In order to verify whether the CD-ISVM algorithm proposed in this paper is effective, the simulation experiment in the PyCharm is carried out using the data set in table 1, and using each data sample, experiments are carried out for CD-ISVM, Simple-ISVM[13], GGKKT-ISVM[14] and CSV-ISVM[15] one by one. Then use classification accuracy, training time and other indicators of each algorithm to evaluate the performance of it.

We use Gaussian RBF kernel function in this experiment and the value of parameter C = 200 and gamma = 10.

4.2.1. Experimental results:
This experiment compares the training results of the four incremental learning algorithms include Simple-ISVM, GGKKT-ISVM, CSV-ISVM and CD-ISVM in the above six data sets. The initial training data sets for several algorithms are 500, and 500 data are added each time for incremental learning. The results obtained are shown in Table II.

| Data set   | breast_cancer | heart | mushroom |
|------------|---------------|-------|----------|
| Algorith   | Simple        | GGKK  | CSV      | CD       |
| Acc        | 0.7613        | 0.7699| 0.7656   | 0.7898   |
| Time       | 141.645       | 179.824| 323.506 | 170.079  |
| APOS       | 0.373         | 0.438 | 0.429    | 0.521    |
| Data set   | splice        | thyroid| titanic |
| Acc        | 0.6119        | 0.6188| 0.6173   | 0.6331   |
| Time       | 104.888       | 23.929| 15.972   | 209.95   |
| APOS       | 0.794         | 0.722 | 0.811    | 0.788    |

4.2.2. Result analysis:
According to the results of the Table 2, we make a comparison chart of the average accuracy, total training time and average proportion of SV for the four algorithms, as shown in Figs. 2 and 3.
It can be seen from Fig. 2 that it is higher than the overall average accuracy of the CD-ISVM algorithm, and it has better classification accuracy in different data sets. Compared with the Simple-ISVM algorithm, the accuracy of CD-ISVM has been greatly improved. Compared the GKKT-ISVM and CSV-ISVM algorithms, it also has certain advantages in most data sets. Compared the run time of several algorithms, it can also be seen that the running time of the CD-ISVM algorithm is basically the same as that of the GKKT-ISVM algorithm, which is slightly higher than the Simple-ISVM algorithm. Therefore, CD-ISVM improves classification accuracy without significantly increasing the training time.

![Figure 2 Comparison of accuracy and time of each algorithm](image)

We all know that it is the support vector set that the main contribution to the classification hyperplane, so in each training how many samples will be used as the SV set is an important indicator that affects the performance of the algorithm. From Fig 3, we can see that the average SV ratio of the CD-ISVM algorithm is higher than that of other algorithms on data sets with different classification characteristics. This also shows that the data set is more effective when selected by the CD-ISVM algorithm in each round of learning, carries more classification feature information, and improves the efficiency of learning.

![Figure 3 The average SV proportion curve of each algorithm under data sets](image)

5. Conclusion

It can be seen from the experimental results that the CD-ISVM algorithm maintains a high classification accuracy in most data sets. The algorithm filters the data set by constructing the boundary vector set, and adds part of the data set with important classification features to the training, thus reducing the impact of some noise points on the training result, and raising the accuracy of the Simple-ISVM that discards all points except the SV set. In addition, the algorithm dynamically adjusts the size and position of the boundary vector set according to the geometric characteristics of the data set itself in each round of incremental learning, which reduces the classification instability caused by the static boundary vector constructed by the CSV-ISVM algorithm. At the same time, the algorithm did not forget the results of the previous training, but added the SV set obtained from the previous training to the new round of training, so it is less susceptible to interference from noise points than the GKKKT-ISVM algorithm.
References
[1] Yongmin Li. On incremental and robust subspace learning [J]. Pattern Recognition, 2004, 37(7):1509-1518.
[2] Shouxian Cheng, Frank Y. Shih. An improved incremental training algorithm for support vector machines using active query [J]. Pattern Recognition, 2007, 40(3):964-971.
[3] J. Wang, W. Zhang, B. Qin, et al. Product Quality Modelling Based on Incremental Support Vector Machine [J]. Journal of Physics: Conference Series, 2012, 364:012101.
[4] Nadeem Ahmed Syed, Syed Huan, Liu Kah. Incremental Learning with Support Vector Machines[C]. IEEE International Conference on Data Mining. IEEE Computer Society. Piscataway:IEEE, 1999:1104-1105.
[5] Rong Xiao, Jicheng Wang, Zhengxing Sun, etc. A SVM incremental learning algorithm [J]. Journal of Nanjing University: Natural Science Edition, 2002, 38(2): 152-157.
[6] Wenjian Wang. A redundant incremental learning algorithm for SVM[C]. Machine Learning and Cybernetics, 2008 International Conference on. Piscataway:IEEE, 2008, 2:734-738.
[7] Ertekin, Seyda, Bottou, Leon, Giles, C. Lee. Nonconvex Online Support Vector Machines [J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(2):368-381.
[8] Yanfang Li, Bo SU, Gongshen L . An Incremental Learning Algorithm for SVM Based on Combined Reserved Set[J]. Journal of Shanghai Jiaotong University, 2016, 50(7):1054-1059.
[9] Wenhua Zeng, Jian Ma. A new incremental learning algorithm for support vector machines. Journal of Xiamen University (Natural Science Edition), 2002, 41(6), 687-691.
[10] Weida Zhou, Li Zhang, Licheng Jiao. Analysis on the promotion ability of support vector machine[J]. Chinese Journal of Electronics, 2001, 29(5): 590-594.
[11] Xiaodan Wang, Chunying Zheng, Chongming Wu. A new SVM peer-to-peer incremental learning algorithm [J]. Computer Applications, 2006. 26 (10): 2440-2443.
[12] http://archive. ics. uci. edu/ml/, 2020-10.
[13] Pan S. C. (2015) Researches on Incremental Support Vector Machine. Shanxi University. pp. 13-32.
[14] Wenhao Xie, Gongqian Liang, Pengcheng Yuan. Research on the incremental learning SVM algorithm based on the improved generalized KKT condition. 2019, 1237(2).
[15] Roshan Chitrakar, Chuanhe Huang. Selection of Candidate Support Vectors in incremental SVM for network intrusion detection. 2014, 45:231-241.