Improved SVM classification algorithm based on KFCM and LDA

Xiaoyan ZHANG, Mengjuan WANG

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, 710000, China
wmengjuan6359@163.com
zhangxy@xust.edu.cn

Abstract. To address the problem that SVM is sensitive to outliers and noise points, in order to improve the classification accuracy of SVM, this paper introduces fuzzy theory and intra-class dispersion theory, proposes an improved SVM classification algorithm, uses KFCM and LDA to filter the data set, and selects reasonable training samples, thereby reducing the number of wild points and noise points in the training sample, and thus reducing its impact on the classification effect of the classification model. Compared with the traditional SVM, the algorithm in this paper considers the impact of training samples on the classification effect, introduces fuzzy theory and intra-class dispersion, and eliminates the wild points and noise points in the training samples that affect the classification accuracy of the classification model. Experimental verification shows that the classification accuracy of the SVM classification model trained by the filtered training samples is higher than that of the SVM classification model without the trained training samples.

1. Introduction

SVM is an effective classification method that provides ideal decision making between two or more categories [1]. SVM algorithm has many unique advantages in solving pattern recognition problems such as small samples, nonlinear and high dimensions. However, SVM also has disadvantages, such as sensitivity to wild points and noise points, and low training speed and classification speed when the number of samples is large [2].

At present, a number of scholars have adopted a variety of methods to optimize SVM to improve the classification accuracy of big data. Among them, some scholars have improved SVM parameters through optimization algorithm. Chiang, H[3] et al. proposed a decentralized artificial bee colony food source optimization algorithm and used it to optimize the kernel parameters of the support vector machine model, thus creating a new mixture and further improving the classification accuracy. Gao, Y. [4] et al. proposed a TWSVM algorithm based on an improved artificial fish swarm algorithm, which solved the TWSVM parameter selection problem through the improved artificial fish swarm algorithm. At the same time, some scholars improve the classification accuracy of SVM by removing redundant support vectors. Wang Yu et al. [5] and Zhao Xiaojian et al. [6] improved the training speed and classification effect of SVM by reducing the scale of training set and reducing redundant support vector.

This paper proposes an improved SVM classification algorithm based on KFCM and LDA, aiming at the impact of the outliers and noise points in the training samples on the SVM classification effect.
The algorithm takes into account the data structure of the training data set, introduces intra-class dispersion on the basis of KFCM, filters the training data, reduces the influence of noise points and outliers in the training samples on the classification algorithm, and thus constructs an efficient classification algorithm. By comparing and analyzing the performance of various algorithms, the results show that the improved SVM classification algorithm based on KFCM and LDA has a high classification accuracy.

2. Theoretical basis

2.1. SVM basic theory

The main idea of SVM is to establish a classification hyperplane as the decision surface to maximize the margin between positive and negative examples [7]. In other words, when the sample is linearly separable, an optimal hyperplane can be found to separate the points of different categories, and the optimal classification plane obtained correctly separates the two categories, the classification interval is also the largest. Let the sample set be \( \{x_i, y_i\}_{i=1}^n \), \( y_i \in \{+1, -1\} \), \( x_i \) is the \( i \) eigenvector, \( y_i \) is the class marker, \( y_i = +1 \) is the positive example, \( y_i = -1 \) is the negative example.

Let the hyperplane equation obtained is \( w^T x + b = 0 \), \( w \) is the weight normal vector, \( b \) is the deviation displacement term, and the linear equation be normalized, so that the above equation can satisfy:

\[
y_i \left( w^T x_i + b \right) \geq 1 \quad i=1,2,\cdots,n.
\]

In formula (1), the classification interval is \( \frac{2}{\|w\|} \), to make the interval maximum is to make \( \|w\| \) minimum. At this point, the optimization problem can be expressed as: under the constraint of formula (1), find the minimum value of \( A \). Therefore, the solving problem of SVM can be transformed as follows:

\[
\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad y_i \left( w^T x_i + b \right) \geq 1 \quad i=1,2,\cdots,n.
\]

Lagrange multiplier is introduced into formula (2), after solution, the optimal solution \( \alpha \) can be obtained. According to the obtained \( \alpha \), \( w \) and \( b \) can be calculated. Meanwhile, the optimal classification hyperplane is \( w^T x + b = 0 \), and the optimal classification function is:

\[
f(x) = w^T x + b = \sum_{i=1}^n \alpha_i y_i x_i^T x + b
\]

In linear inseparability, SVM introduces kernel technique to map the linear inseparability data in the input space into a high-dimensional feature space, so that the data can be divided in the high-dimensional feature space [8]. We use \( \phi(x) \) to represent the eigenvector after \( x \) mapping, then the objective function becomes:

\[
f(x) = w^T x + b \quad \text{s.t.} \quad y_i \left( w^T \phi(x_i) + b \right) \geq 1 \quad \forall i=1,2,\cdots,n.
\]

Lagrange multiplier is introduced to the above objective function, and kernel function \( k(x_i, x_j) \) is introduced. Its corresponding classification function becomes:

\[
f(x) = \sum_{i=1}^n \alpha_i y_i k(x_i, x_j) + b
\]

2.2. Kernel fuzzy C-means clustering algorithm

Based on C-means clustering algorithm, KFCM algorithm introduces fuzzy membership degree to spread "hard" clustering and "soft" clustering, thus transforming the clustering problem into the fuzzy division problem of logarithmic data points [9]. The objective function of FCM is:
2.3. Linear Discriminant Analysis

LDA is a classical Discriminant method. The idea of this method is as follows: given a training set sample, in a feature projection space, the interval between data samples of the same category is the smallest, and the center interval between two different categories is the largest [11].

Let dataset \( D = \{(x_i, y_i)\}_{i=1}^m \), where dataset sample \( x_i \) is an \( N \)-dimensional vector, \( y_i \in \{0, 1\} \). If the projection line is vector \( \omega \), for any sample \( x_i \), its projection on line \( \omega \) is \( \omega^T x_i \). Therefore, the projection points of the same type of data should be as close as possible, meanwhile, the projection points of samples of different categories should be as far as possible. The center of the two classes, \( \mu_0 \) and \( \mu_1 \), just have to make \( \|\omega^T \mu_0 - \omega^T \mu_1\|^2 \) as big as possible.

In-class divergence matrix \( S_o \) is defined as:
\[
S_o = \sum_{x \in x_0} (x - \mu_0)(x - \mu_0)^T + \sum_{x \in x_1} (x - \mu_1)(x - \mu_1)^T
\]  
(10)

The inter-class divergence matrix \( S_b \) is defined as:
\[
S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T
\]  
(11)

The main purpose of LDA algorithm is to find a reasonable \( \omega \), to minimize \( S_o \) and maximize \( S_b \):
\[
\max J(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_o \omega}
\]  
(12)

3. Improved SVM classification algorithm

KFCM calculates the membership degree of each sample point to all class centers by optimizing the objective function. The membership degree is any number within the interval of \([0,1]\), and the sum of membership degree is 1. The data point belongs to a class is determined by the membership function,
so that the membership of each point does not directly belong to a single clustering center. For binary classification, if a data point to two clustering center membership degree value is calculated by the hour, the data points is hard to judge what kind of the data points is likely to affect the classification effect of classification model, so you can according to calculate the membership degree values of the training sample to carry on the preliminary screening, eliminate the noise points.

The LDA algorithm finds a line based on a given sample and projects the sample onto the line. In other words, the projection of the same kind of samples on the straight line is as close and dense as possible, and the projection points of different kinds are as far away as possible. Discriminant analysis was carried out on the training samples using LDA algorithm, get the data distribution of the projection, when the data distribution is more closely, can think these data points is a main characteristic of the class, therefore, can according to the projection of the distribution of the sample points, the training sample screening, get features evident in the data set as the training set, training the classification model.

In the SVM classification algorithm, the field points and noise points in the training samples have a great negative impact on the classification effect of the SVM classification model. Therefore, this paper considers the data structure of the training data set, introduces fuzzy theory and in-class dispersion, and selects the training samples. The improved SVM classification algorithm based on KFCM and LDA is named KFLDPSO_SVM.

The steps of the improved SVM classification algorithm based on KFCM and LDA are as follows:

Step1: Pre-screen the training data, and define the membership range of the pre-training set of the screening as: greater than the minimum membership value and less than the average membership value. KFCM was used to cluster the data to obtain the minimum membership minF and the average membership AVG. For the data within the membership range of the pre-training set, it was screened out and put into ftrain.csv as the pre-training set. For the data outside the range, it was put into test.

Step2: The LDA algorithm is used to perform linear discriminative analysis on the pre-training data set. According to the data distribution after projection, the closely distributed data is put into the new data file train.csv as the training set, and the scattered data is put into test.csv as the test set.

Step3: The improved PSO is used to optimize the SVM algorithm. In PSO-SVM, the population size is 60, the number of iterations is 30, and the fitness function is a 3 fold crossover calibration verification function. Adaptive inertia weight, $\omega_{\text{max}} = 0.9, \omega_{\text{min}} = 0.4$, asynchronous learning factor, $c_{1, \text{start}} = c_{2, \text{end}} = 2.5, c_{1, \text{end}} = c_{2, \text{start}} = 0.5$.

Step4: build a SVM classification model according to the optimized SVM parameters, and classify the data. Among them, the kernel function is RBF function, and test the classification effect of the model with train.csv training classification model and test.csv as test set. Finally, the classification results are output.

4. Experimental verification of algorithm

4.1. Relevant data extraction

In order to verify the classification effect of the new model, the experimental data selected in this paper are from the data provided by the LibSVM official website [12]: German, Heart, Australian. Table 1 shows the attribute information of the data in turn according to the size of the experimental data.

| Data set name | Data source          | Experimental data | Data dimension |
|---------------|----------------------|-------------------|----------------|
| German        | Statlog / German     | 1000              | 24             |
| Australian    | Statlog / Australian | 690               | 14             |
| Heart         | Statlog / Heart      | 270               | 13             |
4.2. Verification and analysis

Firstly, the experimental data set is input, KFCM algorithm is used to cluster the data set, and the membership degree of each data point is calculated, the data points with membership value within the specified range are screened out. Then, the LDA algorithm is used to perform linear discriminant analysis on the filtered data and eliminate outliers that affect the efficiency of the classification model. After the LDA discriminates the data, the following data distribution can be obtained:

![Diagram](a) German data set (b) Australian data set

Fig.1 Distribution of pre-training samples after LDA projection

In the diagram above, the y axis is the tag data classes, the x axis is received after the projection on the original data values, in figure 2 (a) the graph is German data set after by KFCM screening of the preliminary training samples by LDA projection distribution, we can see that the German advance training sample after the LDA for projection, negative class are mainly distributed in the range of (0.004, 0.011), is class are mainly distributed in the range of (0.058, 0.012); FIG. 2 (b) shows the distribution of pre-training samples of Data set Australian projected by LDA. It can be seen from the figure that after the pre-training samples of Australian were projected by LDA, the negative classes were mainly distributed within the range of (-0.022,-0.005), while the positive classes were mainly distributed within the range of (-0.017,-0.003). The densely distributed data points in the pre-training sample were put into train.cvs and the sparsely distributed points were put into test.cvs.

The PSO-SVM classification model was trained with the training set obtained by screening, and the remaining data was used as the test set to verify the model. At the same time, PSO-SVM and GA-SVM classification models are compared with the model proposed in this paper, where the ratio of training set to test set is 3:7.

| Data set  | PSO-SVM | GA-SVM | KFLDPSO_SVM |
|-----------|---------|--------|-------------|
| german    | 75%     | 72%    | 75.9%       |
| heart     | 65.4%   | 64.2%  | 71.9%       |
| australian| 67.6%   | 61.8%  | 82.6%       |

Can be seen from table 2, the data collection of German, the KFCM algorithm is adopted to train the data to carry on the preliminary screening, screening by LDA algorithm for pre training samples, and finally, using the selection of training samples to train the SVM classification model, classification accuracy is 75.9%, GA - SVM classification accuracy of 72%, and without screening training samples of SVM classification model of classification accuracy is 75%; For the dataset HEART, the classification accuracy of KFLDPSO_SVM algorithm is 71.9%, GA-SVM 64.2%, and PSO-SVM 65.4%. For data set Australian, the classification accuracy of KFLDPSO_SVM algorithm was 82.6%, GA-SVM 61.8%, and PSO-SVM 67.6%. It can be seen that the classification accuracy of KFLDPSO_SVM algorithm is higher than that of the other two classification models. It can be seen that the classification accuracy of KFLDPSO_SVM algorithm is higher than that of the other two classification models. We can conclude through KFCM and LDA screening of the training sample,
improve the classification accuracy of SVM classification algorithm, after KFCM and LDA selection of training samples, the category of the training sample characteristic is obvious, and the wild points and noise points in the training samples that affect the classification accuracy are effectively reduced, thus, the classification performance and accuracy of SVM classification model are improved.

5. Conclusion
In this paper, an improved classification model of support vector machine (SVM) based on KFCM and LDA is established by introducing fuzzy theory and in-class discreteness theory. By using KFCM and LDA algorithms to screen the data set, the appropriate training samples are selected, the training samples are used to train the SVM classification model, and finally the test set is used to test the classification model. Through experimental verification, we can find the SVM classification model trained by the screened training samples, its classification accuracy is higher than that of the SVM classification model with randomly selected training samples, in other words, screening training samples through KFCM and LDA can improve the classification effect of the classification algorithm. It can be seen that the number of field points and noise points in training samples is reduced after screening, and the classification accuracy of the model is improved.

References
[1] Adil Abboud, S.; Al-Wais, S.; Abdullah, S.H.; Alnajjar, F.; Al-Jumaily, A. Label Self-Advised Support Vector Machine (LSA-SVM)—Automated Classification of Foot Drop Rehabilitation Case Study. Biosensors 2019, 9, 114.
[2] Li Yitao. Improved classification algorithm based on support vector machine [J]. Computer system application,2019,28(10):145-151.
[3] Chiang, H., Sangaiah, A.K., Chen, M. et al. A Novel Artificial Bee Colony Optimization Algorithm with SVM for Bio-inspired Software-Defined Networking. Int J Parallel Prog 48, 310–328 (2020). https://doi.org/10.1007/s10766-018-0594-6.
[4] Gao, Y., Xie, L., Zhang, Z. et al. Twin support vector machine based on improved artificial fish swarm algorithm with application to flame recognition. Appl Intell 50, 2312–2327 (2020). https://doi.org/10.1007/s10489-020-01676-6.
[5] Wang Yu, MAO Yuxin. A simplified support Vector Machine model based on satellite vector [J]. Journal of Dalian University of Technology,2008(03):446-450.
[6] Zhao Xiaofang, Zhang Lu. An improved fuzzy support vector machine classification algorithm for data mining [J]. Journal of lanzhou university of technology,2017,43(05):94-99.
[7] Hu, R., Zhu, X., Zhu, Y. et al. Robust SVM with adaptive graph learning. World Wide Web 23, 1945–1968 (2020). https://doi.org/10.1007/s11280-019-00766-x.
[8] Wang, K., Liu, L., Yuan, C. et al. Software defect prediction model based on LASSO–SVM. Neural Comput & Applic (2020). https://doi.org/10.1007/s00521-020-04960-1.
[9] Kumar SVA, Harish BS, Mahanand BS, Sundararajan N, An efficient Meta-cognitive Fuzzy C-Means clustering approach. Clarivate Analytics Web of Science, 10.1016/j.asoc.2019.105838.
[10] Wang Shuwen, PI Bingkun, ZHANG Hongqiang, Ma Cong. An image classification method based on fuzzy core clustering algorithm [J]. Journal of northwest normal university (natural science edition),2016,52(05):42-45.
[11] C. Li, Y. Shao, W. Yin and M. Liu, "Robust and Sparse Linear Discriminant Analysis via an Alternating Direction Method of Multipliers," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 3, pp. 915-926, March 2020.
[12] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html.