A new data classification improvement approach based on kernel clustering

Bingsen Guo*

The College of Intelligent Equipment, Shandong University of Science and Technology, Tai’an, Shandong, 271000

*(guobingsen0909@163.com)

Abstract—Data classification is one of the most critical issues in data mining with a large number of real-life applications. In many practical classification issues, there are various forms of anomalies in the real dataset. For example, the training set contains outliers, often enough to confuse the classifier and reduce its ability to learn from the data. In this paper, we propose a new data classification improvement approach based on kernel clustering. The proposed method can improve the classification performance by optimizing the training set. We first use the existing kernel clustering method to cluster the training set and optimize it based on the similarity between the training samples in each class and the corresponding class center. Then, the optimized reliable training set is trained to the standard classifier in the kernel space to classify each query sample. Extensive performance analysis shows that the proposed method achieves high performance, thus improving the classifier's effectiveness.

1. INTRODUCTION

Classification is an important data mining technology with a wide range of applications. It can classify almost all types of data in any field in our life, such as skin lesion classification [1], customer loss prediction and classification [2], credit card fraud detection and classification [3], text classification [4] and other fields. At present, many mature classification algorithms have been proposed, such as Support Vector Machine (SVM) [5], K-nearest Neighbor (KNN) [6], Naive Bayes (NB) [7][8], Decision Tree (DT) [9] and so on. These methods for a more complete data set can achieve a better classification effect.

However, there are common shortcomings of the traditional classification methods; when there are outliers in the training set, the existing methods are difficult to obtain satisfactory results. Because of such problems, scholars put forward many improved methods. In the reference [10], a coarse-grained parallel genetic algorithm is proposed to make the feature subset and parameters of support vector machine better, but it takes much time to train SVM with this method. In the literature [11], a heuristic genetic algorithm is used to optimize SVM kernel parameters. The heuristic strategy is adopted to dynamically adjust the genetic operator and take the classification accuracy of the model as the objective function to realize the parameter optimization of SVM classification model based on Gaussian kernel, but the influence of feature weighting on the accuracy of SVM is not considered in the method.

In the reference [12], an intrusion detection method based on the wavelet kernel least square method is developed, which can enhance the detection ability of SVM in complex nonlinear systems, but the computational complexity of this method is high. K-nearest neighbor algorithm (KNN) is simple to implement and has significant classification performance, but the biggest disadvantage of the traditional KNN method is that the K value is difficult to determine. In the reference [13], an enhanced fuzzy KNN...
A way is proposed to adaptively specify the optimal K value through particle swarm optimization. In the literature [14], the optimum value of K is proposed by combining two widely used empirical methods, rules summary and example learning. In [15], the number of locally adjusted nearest neighbors is submitted based on confidence degree. In [16], a probabilistic nearest neighbor method is proposed to infer the number of nearest neighbors, namely the optimal K value.

This paper proposes a new data classification improvement approach based on kernel clustering, which reduces the influence of outliers on the training set to a certain extent. We first use the existing kernel clustering method to cluster the training set and get different cluster centers. The reliable training set is optimized by calculating the similarity between the center of each class and the corresponding sample. The reliable training set is used to train the conventional classifier and classify the test samples. It is worth noting that the proposed method is based on the kernel method. The kernel function maps the data that need clustering and classification from the original space to the high-dimensional feature space. By nonlinear mapping to the high-dimensional space, the differences between data samples are amplified, and data partitioning is improved to obtain better classification results.

In section 2, we first briefly review FCM, SVM, and KNN algorithms. In section 3, we introduce the detailed steps of applying the proposed method to data classification. For proving this method is effective, we carried out experiments on five real data sets, and the results are given in section 4. Finally, the conclusion and discussion are given in section 5.

2. BACKGROUND KNOWLEDGE

2.1 Fuzzy c-means Clustering(FCM)

The Fuzzy C-means (FCM) algorithm [1] is a famous fuzzy clustering algorithm widely used in pattern recognition, image segmentation, data sample analysis, and other fields. FCM continuously updates the membership matrix and clustering center to minimize the objective function, obtain the membership of each sample point to all class centers, and determine the class of sample points according to the principle of maximum membership. The class of sample points is determined according to the maximum membership degree principle. KFCM applies the kernel method to the classical fuzzy C-means clustering algorithm and adopts the kernel induction measure to replace the Euclidean norm measure in FCM. Previous studies have shown that the performance of KFCM is better than that of FCM [17].

2.2 Support Vector Machine(SVM)

Support Vector Machines is a new machine learning method based on statistical learning theory. SVM is used to construct one or more high-dimensional hyperplanes to divide the sample data, namely the classification boundary between the samples. The kernel function is an essential part of SVM, which is extended to nonlinear separable problems. The biggest feature of SVM is the introduction of kernel tricks. By mapping to samples and mapping to high-dimensional feature space, samples can be divided into high-dimensional space. Its generalization ability largely depends on the choice of the kernel function. With a solid theoretical foundation, SVM has become an excellent machine learning algorithm under its many advantages, and in most cases, SVM can provide better performance than traditional machine learning [18].

2.3 K-nearest Neighbor(KNN)

The core idea of the traditional KNN algorithm is to find K nearest neighbor training samples for the test sample and classify them according to the categories of K nearest neighbor samples. The distance measurement is usually taken, such as Euclidean distance. Use the kernel approach to redefine the distance metric by mapping $\phi : R^n \rightarrow H$. Samples are mapped to higher-dimensional Hilbert space, which improves the separability of samples in higher-dimensional space. Then the kernel function
is \( K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \) and we can calculate the distance of the sample in the high-dimensional space only by calculating the inner product [19].

3. THE PROPOSED METHOD
In this part, the proposed method is theoretically analyzed. The proposed method can be divided into two steps. The first step is to optimize the training set through kernel clustering, and the second step is to train the classifier with the optimized training set and then classify the classified samples.

3.1 Optimized training set
In this part, we optimize the training set. The optimization method is to obtain the center by kernel clustering of the training set, and then set a distance threshold to select reliable training samples by calculating the Euclidean distance between the center of each class of the training set and the corresponding class samples of the training set.

The most important is that the kernel method can improve the partitioning between data, make the clustering and classification results more accurate, realize efficient calculation, and effectively deal with high-dimensional input.

Consider a training set \( X = \{x_1, x_2, \ldots, x_N\} \), \( x_j \in \mathbb{R}^r \) and a test set \( Y = \{y_1, y_2, \ldots, y_M\} \), \( y_j \in \mathbb{R}^r \). Firstly, a nonlinear mapping \( \Phi : X \rightarrow H \) is introduced to map data to a higher-dimensional feature space. \( X \) represents the original data space and \( H \) describes the high-dimensional feature space. The dot product of any two \( s \)-dimensional vectors \( a \) and \( b \) in the high-dimensional feature space \( H \) is:

\[
K(a, b) = \langle \phi(a) \cdot \phi(b) \rangle = \phi(a)^T \phi(b)
\]  

(1)

The square of Euclidean distance is:

\[
\|\phi(a) - \phi(b)\|^2 = K(a, a) - 2K(a, b) + K(b, b)
\]  

(2)

Gaussian function as a kernel function is adopted, and its form is: \( K(a, b) = \exp(-\|a - b\|^2 / 2\sigma^2) \) and then we can figure out \( K(a, a) = K(b, b) = 1 \). Then equation (2) becomes:

\[
\|\phi(a) - \phi(b)\|^2 = 2 - 2K(a, b)
\]  

(3)

The objective function of the KFCM algorithm is:

\[
J_{\phi, m}(U, V) = \sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^m \|\phi(x_j) - \phi(v_i)\|^2
\]  

(4)

c is the true number of categories in the sample, \( \mu_{ij} \) represents the membership degree of \( x_j \) to cluster center \( v_i \). \( V \) is the cluster center matrix, \( m \) is the fuzzy system, and the general value is 2.

Calculate the class center according to equation (5)

\[
v_i = \frac{\sum_{j=1}^{N} \mu_{ij}^m K(x_j, \omega_i) x_j}{\sum_{j=1}^{N} \mu_{ij}^m K(x_j, \omega_i)}, \forall i = 1, 2 \cdots, c
\]  

(5)
After labeling the cluster center, the Euclidean distance of the data in each class is calculated, and \( x_{ik} \) represents the kth data in class i:

\[
d(x_{ik}, v_i) = \| \phi(x_{ik}) - \phi(v_i) \| = \sqrt{2 - 2K(x_{ik}, v_i)}
\]  

(6)

Set a distance threshold \( \delta_i \) for each class, we take the average of the Euclidean distance of data \( x_{ik} \) corresponding to class center \( v_i \) in each class as the threshold:

\[
\delta_i = \frac{1}{N_i} \sum_{k=1}^{N_i} d(x_{ik}, v_i)
\]  

(7)

Threshold \( \delta_i \) is used to select an optimized training set, and if the distance between sample \( x_{ik} \) and class center \( v_i \) is less than \( \delta_i \), the sample is included in the new training set, otherwise discarded. Define the optimized data set as \( X_{new} \):

\[
X_{new} = \{ x_{ik} \mid x_{ik} - v_i \| \leq \delta_i, 1 \leq k \leq N_i, 1 \leq i \leq c \}
\]  

(8)

Discussion: In the above optimization of the training set using the kernel method, we chose the KFCM algorithm. Of course, other kernel-based clustering methods can be used as potential choices. For the selection of the kernel function, can choose the commonly used linear kernel function and polynomial kernel function, the radial basis kernel function, and so on, due to low dimension, high dimension and small sample, large sample, and so on apply RBF kernel, convergence domain-wide, wide application and is an ideal classification function [20]. Hence, the kernel function all adopts the RBF kernel algorithm in this paper. By optimizing the training set, the influence of abnormal data can be reduced.

3.2 Classified test set

After the optimized training set is obtained, the classifier is trained. In this part, many classifiers can be selected. It is worth noting that the method proposed in this paper is based on the kernel method, so we prefer to use the same kernel-based classifier. Kernel support vector machine is a potent model and performs well on all kinds of data sets. The KNN does not require much tuning to get good performance. This algorithm is a good benchmark to try out before considering more advanced techniques. Therefore, here we present two kernel-based classification methods, one is kernel-based SVM, the other is kernel-based KNN. The classification results of the \( y_k \) sample in the test set are given by:

\[
P_k = Y_{1 or 2}(y_k \mid X_{new})
\]  

(9)

Here, \( Y_1 \) stands for SVM, \( Y_2 \) stands for KNN.

The pseudo-code of an improved data classification method based on kernel clustering is given below
Input:
Training set $X$ and test set $Y$ in $R^d$

Parameters:
- $\hat{Y}$: classifier
- $K(\cdot, \cdot)$: Kernel function

for
- The training set $X$ was clustered by kernel clustering method.
- The cluster center was obtained, and the tag was matched.
- The distance between the center of each class and the corresponding class sample is calculated by Eq.(7)
- Set the distance threshold $\delta$ by Eq.(8).
- Reliable training samples were selected according to Equation (9)
- Reliably training samples constitute the optimized training set $X_{new}$
- The classifier is trained with the optimized training set
end

4. EXPERIMENTAL DESIGN AND RESULTS

4.1 The experiment design

We use five data sets in UCI Database, WDBC, Handshake, Database on Vowel Recognition, Synthetic Control Chart Time Series and Landsat Satellite. For each dataset, 70% of it is divided into training sets and 30% into test sets. In order to eliminate the influence of dimensional and singular samples and shorten the training time, the data set is normalized. KFCM was used to cluster the training set. The clustering center was calculated by equation (5), the threshold was calculated by equation (8), and reliable training samples were selected by equation (9) to form the optimized training set. In order to better reflect the feasibility of this method and its improvement over traditional classification methods, SVM and KKNN were used as classifiers for experiments, and the RBF kernel function was selected for all kernel functions. The classification accuracy, precision, recall and F1-measure were used as evaluation indexes to compare the four indexes for classification after training the classifier directly with the initial training set and the classification method proposed in this paper.

4.2 The experimental results

Table 1 shows the average accuracy of different data sets under different methods. We can see that the proposed method achieves better performance than other methods because the proposed method can deal with the distribution of nonlinear data set well by introducing kernel function and effectively avoiding the influence of outliers by optimizing the training set.

| Dataset  | SVM          | The proposed method (SVM) | KKNN        | The proposed method (KKNN) |
|----------|--------------|---------------------------|-------------|----------------------------|
|          | Dataset1     | Dataset2                  | Dataset3    | Dataset4                   | Dataset5     |
| SVM      | 94.69%       | 82.63%                    | 72.97%      | 86.64%                     | 67.97%       |
| The proposed method (SVM) | 97.33%       | 91.22%                    | 85.19%      | 95.31%                     | 79.16%       |
| KKNN     | 94.91%       | 81.69%                    | 75.15%      | 88.77%                     | 66.15%       |
| The proposed method (KKNN) | 96.2%        | 84.39%                    | 86.03%      | 94.61%                     | 78.19%       |

Figures (a), (b), (c), (d), (e) and (f) show the average precision, average recall and average F1-measure of different data sets under different methods. Kernel clustering is used to optimize the
training set so that the classifier can better learn the useful feature information of the training set. Especially in the face of a large amount of data, the proposed method has obvious advantages.
Figure 1. (a) Precision comparison between direct classification using SVM as a classifier and classification employing the way put forward in this article. (b) Recall comparison between direct classification using SVM as a classifier and classification via the proposed method in this article. (c) F1 comparison between direct classification using SVM as a classifier and classification via the proposed method in this work. (d) Precision comparison between direct classification using KKNN as a classifier and classification via the proposed method in this article. (e) Recall comparison between direct classification using KKNN as a classifier and classification via the proposed method in this work. (f) F1 comparison between direct classification using KKNN as a classifier and classification employing the way put forward in this article.

5. CONCLUSION
A new data classification improvement method based on kernel clustering is proposed. The method first obtains the corresponding class centers based on the kernel clustering method and then optimizes the reliable training set by calculating the similarity of samples in the related classes of each class center. By doing so, we can avoid membership point confusion and reduce the learning ability of the classifier. Extensive analysis shows that compared with the traditional direct training classifier, this method has better usability and improves the validity of classification. In future work, we intend to extend the proposed approach to the classification of incomplete data [21]-[22].

REFERENCES
[1] Hameed N, Shabut A M, Ghosh M K, et al. Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques[J]. Expert Systems with Applications, 2020, 141: 112961.
[2] Liu J, An F P. Image classification algorithm based on deep learning-kernel function[J]. Scientific programming, 2020, 2020.
[3] de Sá A G C, Pereira A C M, Pappa G L. A customized classification algorithm for credit card fraud detection[J]. Engineering Applications of Artificial Intelligence, 2018, 72: 21-29.
[4] Wang Z, Qu Z. Research on Web text classification algorithm based on improved CNN and SVM[C]//2017 IEEE 17th International Conference on Communication Technology (ICCT). IEEE, 2017: 1958-1961.

[5] Mygdalis, V., Tefas, A., Pitas, I.: Learning multi-graph regularization for svm classification. In: ICIP, pp. 1608–1612 (2018)

[6] S. Malhotra, V. Bali, and K. K. Paliwal, “Genetic programming andk-nearest neighbour based intrusion detection model,” in Proc. IEEE 7th Int. Conf. Cloud Comput., Data Sci. Amp. Eng. Conf., Jan. 2017, pp. 42–46.

[7] Domingos, P., and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29:103-130.

[8] Li, Y. H., and A. K. Jain. 1998. Classification of text documents. Comput. J. 41:537-546.

[9] Mu Y, Liu X, Wang L. A Pearson’s correlation coefficient based decision tree and its parallel implementation[J]. Information Sciences, 2018, 435: 40-58.

[10] Z. Chen, T. Lin, N. Tang, and X. Xia, “A parallel genetic algorithm based feature selection and parameter optimization for support vector machine,” Sci. Program., vol. 2016, Jun. 2016, Art. no. 1

[11] T. Yerong, S. Sai, X. Ke, and L. Zhe, “Intrusion detection based on support vector machine using heuristic genetic algorithm,” in Proc. IEEE 4th Int.Conf. Commun. Syst. Netw. Technol. (CSNT), Apr. 2014, pp. 681–684.

[12] Y. Guang and N. Min, “Anomaly intrusion detection based on wavelet kernel LS-SVM,” in Proc. IEEE 3rd Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT), Oct. 2013, pp. 434–437.

[13] D.-Y. Liu, H.-L. Chen, B. Yang, X.-E. Lv, L.-N. Li, and J. Liu, “Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnosis system for thyroid disease,” J. Med. Syst., vol. 36, no. 5, pp. 3243–3254, 2012

[14] G. Góra and A. Wojna, “RIONA: A classifier combining rule induction and k-NN method with automated selection of optimal neighbourhood,” in Proc. ECML, 2002, pp. 111–123

[15] J. Wang, P. Neskovic, and L. N. Cooper, “Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence,” Pattern Recognit., vol. 39, no. 3, pp. 417–423, 2006.

[16] S. Manocha and M. Girolami, “An empirical analysis of the probabilistic k-nearest neighbour classifier,” Pattern Recognit. Lett., vol. 28, no. 13, pp. 1818–1824, 2007

[17] Da-Qiang Zhang, Song-Can Chen. Clustering Incomplete Data Using Kernel-Based Fuzzy C-means Algorithm[J]. Neural Processing Letters, 2003, 18(3):155-162.

[18] Burges C J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discov er y, 1998, 2(2): 121–167

[19] Kai Yu,Liang Ji,Xuegong Zhang. Kernel Nearest-Neighbor Algorithm[J]. Neural Processing Letters, 2002, 15(2):147-156.

[20] KR Müller, Mika S , G Rätsch, et al. An introduction to kernel-based learning algorithms[J]. IEEE Transactions on Neural Networks, 2001, 12(2):181.

[21] Z. F. Ma, Z. Liu, Y. Zhang, L. Song, J. He, Credal Transfer Learning With Multi-Estimation for Missing Data. IEEE Access, 2020, 8, 70316-70328.

[22] Cao T T , Zhang M , Andreae P , et al. An effective and efficient approach to classification with incomplete data[J]. Knowledge-Based Systems, 2018, 154(15):1-16.