ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 181

4th International Conference on Social Science and Higher Education (ICSSHE 2018)

The Design and Implementation of the Data Classification System

Dandan Xue, Zengguo Sun*, Yang Liu, Rui Shi, Jie Ding

School of Computer Science

Shaanxi Normal University

Xi'an, China

Corresponding Author: Zengguo Sun, duffer2000@163.com.

***Abstract*—The artificial data classification is time-consuming and inefficient. To solve this problem, the data classification system is developed to classify the data automatically and helps users to find the inconsistent points easily. Firstly, the requirement analysis, the summary design, the detailed design and the code design are given in this paper. Secondly, the MATLAB R2016 is used to implement this system. The data classification system has four algorithms: Fisher classifier, clustering analysis classifier, Bayes classifier, and linear SVM classifier. Users can choose different algorithms to classify the data and study the effect of different classification algorithms.**

Keywords—data classification; requirement analysis; summary design; detailed design; MATLAB development tool

I. INTRODUCTION

The society is a big data society. Data classification becomes a very essential part of the data processing. However, the artificial classification is time-consuming and error-prone. This paper develops a data classification system, including four classifiers: Fisher classifier, clustering analysis classifier, Bayes classifier and Linear SVM classifier. With this system, the data can be classified conveniently [1]. Among them, the Fisher classifier uses Fisher discriminant method. The main idea of it is to project n class and m dimensional data sets into one direction of a straight line as far as possible, so that the different classes can be separated as far as possible [2]. The clustering algorithm classifier is divided into two methods: top-down and bottom-up. The former is to treat all the samples as a class, and then continuously divided the class into small classes until they can no longer be separated. The latter, on the contrary, all samples are a class of their own and then continuously merge into two classes, until eventually a few categories are formed. The Bayes classifier is based on Bayes theorem. It uses the priori probability of an object to calculate the priori probability by Bayes formula, which is the probability that the object belongs to a certain class, and then selects the class with the maximum priori probability as the class to which the object belongs [3]. The linear SVM classifier is implemented based on the SVM theory that provides a way to avoid the complexity of high dimensional space, and directly uses the inner product function (the kernel function) of this space, then the decision problem of the corresponding high dimensional space is solved directly by using the solution method in the linear separable case. When the kernel function is confirmed, the problem of high dimensional space is easier. At the same time, SVM is based on small sample statistical theory, and has better generalization ability than neural network.

The Fisher classification algorithm, the clustering classification algorithm, the Bayes classification algorithm and the linear SVM classification algorithm are applied in many fields. For example: The bus forecast of the citizen when they are in travel, the identification of the species of microorganism, precision marketing of the big data, the image mining of users and so on. The different data classification algorithms have their own advantages [4]. It is necessary to compare different data classification algorithms and select the most appropriate algorithm according to the requirements of the applied object. The data classification system can help researchers to choose the better data classification algorithms.

II. REQUIREMENT ANALYSIS

The data classification system needs to implement the selection of the data classifier, the classification of data, the display of the data classification and other functions. The function of data classification with different data classifiers is the core of the system. The system can classify the data imported by users, and show the classification results to users by graphical representation according to different classification algorithms. The classifier provided by the system contains four kinds: Fisher classifier, clustering analysis classifier, Bayes classifier and linear SVM classifier. The data classification system has good man-machine interaction and can satisfy actual demand of the users.

III. SUMMARY DESIGN

This section gives the summary design of the data classification system based on the requirement analysis, and determines the framework and the functional modules of the system [5]. The structure diagram of the data classification system is shown in Fig. 1. This structure diagram contains all the functions of the data classification system.

This work is supported by the fourth batch of Information Demonstration Course Construction Project of Shaanxi Normal University, All English Teaching Demonstration Course Construction Project of Shaanxi Normal University in 2017, Research Project on the Reform of Graduate Education and Teaching (Research Project of Degree and Graduate Education Reform) of Shaanxi Normal University in 2018 (Grant No. GERP—18—57), and "Teaching Model Innovation and Practice Research" Special Foundation of Shaanxi Normal University in 2018 (Grant No. JSJX2018L212).

*Corresponding Author: Zengguo Sun, duffer2000@163.com.

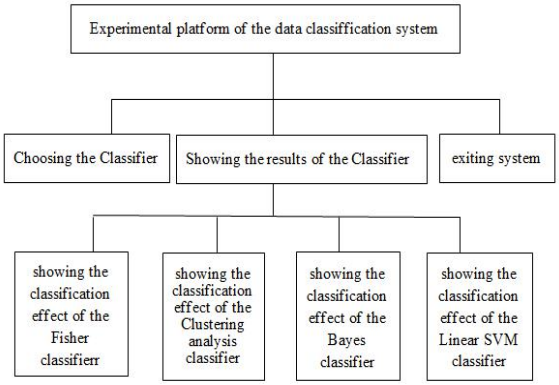

Fig. 1. Structure diagram of the data classification system

IV. Detailed Design

This section gives the detailed design of the data classification system, the specific implementation process of each function module is determined in this section. The basic operation flow chart of the data classification system is shown in Fig. 2. Users can select the type of operations after entering the main interface of the system. If users choose the Fisher classification function, the Fisher discriminant method is used to classify the data and the classification results are displayed in the graphical format. If users choose the clustering classification function, the K-means clustering algorithm is applied and the classification results are displayed in the graphical format. If users choose the Bayes classification function, the Bayesian algorithm is carried out and the classification results are displayed in the graphical format. If users choose the linear SVM classification function, the SVM algorithm is used and the classification results are displayed in the graphical format. If users choose the exit function that the system will exit.

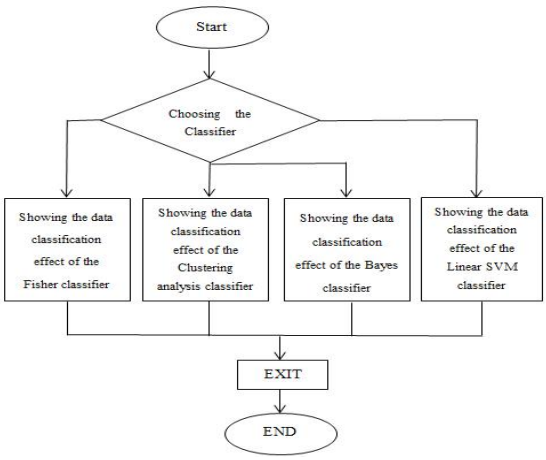

Fig. 2. The basic operation flow chart of the data classification system

V. Code Design

According to the results of the detailed design, the code design of the data classification system can come true. MATLAB R2016 is a business mathematics software that used for algorithm development, data visualization, data analysis and numerical calculation [6,7]. This section mainly introduces the implementation of the Fisher classifier (Fisher linear discriminant) in the data classification system. The following is the implementation of the Fisher classification algorithm.

```
%Calculating the mean of the sample
m1=mean(w1)';
m2=mean(w2)';
%s1: the class inner discrete matrix of the first class of samples.
%s2: the class inner discrete matrix of the second class of samples.
s1=zeros(2);
[row1,colum1]=size(w1);
for i=1:row1
    s1 = s1 + (w1(i,:)'-m1)*(w1(i,:)'-m1)';
end;
s2=zeros(2);
[row2,colum2]=size(w2);
for i=1:row2
    s2 = s2 + (w2(i,:)' - m2)*(w2(i,:)' - m2)';
end;
%Calculating the discrete matrix of the total class Sw.
Sw=s1+s2;
%Calculating the solution of maximum value of Fisher criterion function w.
w=inv(Sw)*(m1-m2);
%Calculating the threshold w0.
ave_m1 = w'*m1;
ave_m2 = w'*m2;
w0 = (ave_m1+ave_m2)/2;
%Drawing two types of training sample pointsv
axes(handles.axes2)
%Drawing two types of sample points
plot(X1,Y1,'.r',X2,Y2,'.b');
hold on;grid;
%Drawing the solution when taking the maximum value w.
x = [-40:0.1:40];
y = x*w(2)/w(1);
axes(handles.axes2);
plot(x,y,'g');
```

VI. Interface Design

This section shows the main interface and the function test results of the system. The test data is imported into the system ahead of time; four data classifier results are displayed respectively after entering the system. The effect of the data classification is displayed in graphics. At the same time, the system has a brief introduction of all kinds of algorithms.

A. Program Main Interface

The main interface of the system is in Fig. 3. The interface text introduces the benefits of the classification. The four buttons respectively indicate the four data classifier from the top to the bottom on the left: Fisher classifier, clustering analysis classifier, Bayes classifier, linear SVM classifier. The two coordinates on the right are used to show the classification effect's diagram.

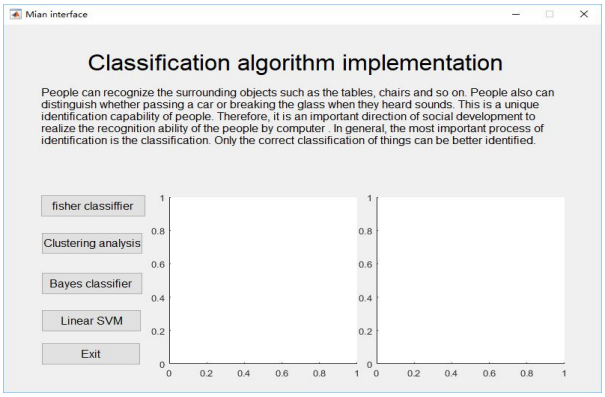

Fig. 3. Program main interface

B. *Fisher Classifier*

Clicking the Fisher button, the classification effect of the Fisher classifier is shown in Fig. 4. The interface text is a brief introduction of the Fisher classification algorithm.

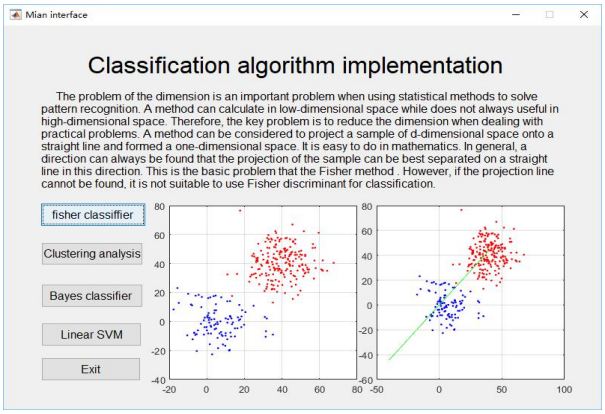

Fig. 4. The classification effect of the Fisher classifier

C. *Clustering Analysis Classifier*

Clicking the button of clustering analysis, the classification effect of the clustering analysis classifier is shown in Fig. 5. The interface text is a brief introduction of the clustering analysis classification algorithm.

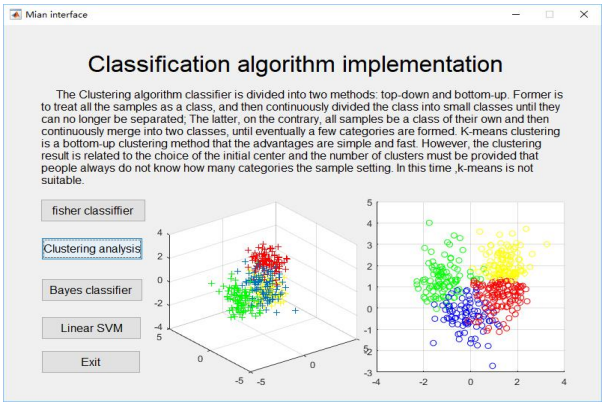

Fig. 5. The classification effect of the clustering analysis classifier

D. *Bayes Classifier*

Clicking the Bayes button, the classification effect of the Bayes classifier is shown in Fig. 6. The interface text is a brief introduction of the Bayes classification algorithm.

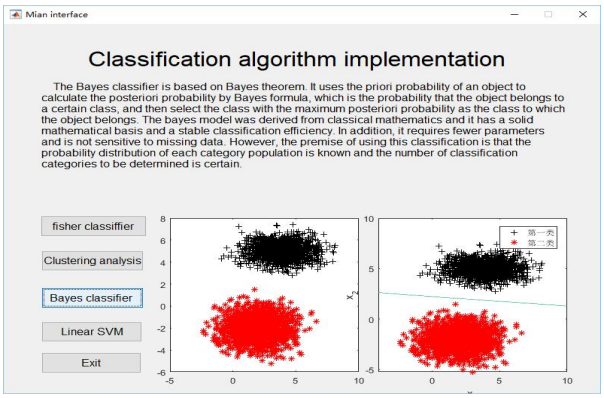

Fig. 6. The classification effect of the Bayes classifier

E. *Linear SVM Classifier*

Clicking the linear SVM button, the effect of the linear SVM is shown in Fig. 7. The interface text is a brief introduction of the SVM classification algorithm.

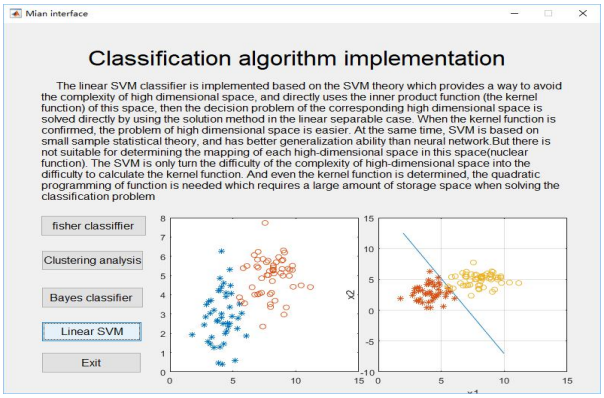

Fig. 7. The classification effect of the Linear SVM classifier

VII. SUMMARY

The artificial data classification is time-consuming and inefficient. To solve this problem, this paper develops a data classification system and gives the requirement analysis, the summary design, the detailed design and the code design. The system uses the MATLAB R2016 development tool to implement the system. This paper gives the part of the implementing code of algorithms and shows the running interfaces of the system. It also introduces the corresponding algorithm in the system [8]. The system has the function of Fisher classifier, clustering analysis classifier, Bayes classifier, linear SVM classifier and so on. The system has beautiful interfaces, simple operation and strong practicability. The graphical display of the data classification results facilitates observation and analysis of the experimental data. So the system can be used as a platform of the data classification, and

can be a comparison platform for the classification effects of Fisher classifier, clustering analysis classifier, Bayes classifier, and linear SVM classifier.

REFERENCES

[1] Zhang X, Ding S, Xue Y. An improved multiple birth support vector machine for pattern classification [J]. Neurocomputing, 2016, 225.

[2] Dong L, Wesseloo J, Potvin Y, et al. Discrimi- nation of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression [J]. Rock Mechanics & Rock Engineering, 2016, 49(1): 183-211.

[3] Taheri S, Mammadov M. Learning the naive Bayes classifier with optimization models [J]. International Journal of Applied Mathe- matics & Computer Science, 2013, 23(4): 787-795.

[4] Fuangkhon P. Multiclass Contour-Preserving Classification with Support Vector Machine (SVM) [J]. Journal of Intelligent Systems, 2017, 26(2): 323-334.

[5] Saeedi J, Faez K. A classification and fuzzy-based approach for digital multi-focus image fusion [J]. Pattern Analysis & Applications, 2013, 16(3): 365-379.

[6] Wilson H B, Halpern D, Turcotte L H. Advanced Mathematics and Mechanics Applications Using MATLAB, Third Edition [M]. Taylor and Francis, 2002.

[7] Hahn B, Valentine D. Essential MATLAB for Engineers and Scientists (Sixth Edition) [M]. Netherlands: Elsevier Ltd, 2017.

[8] Zhang G, Zhang C, Zhang H. Improved K-means Algorithm Based on Density Canopy [J]. Knowledge-Based Systems, 2018.