A classification method based on linear discriminant analysis and multivariate adaptive spline

Zhihui Li, Jiaxin Liu
School of Computer Science and Technology Harbin Engineering University
Harbin, China
lizhihui@hrbeu.edu.cn

Yufei , Huang, Jia Xu
Beijing Institute of Spacecraft System Engineering
Beijing, China
13072417213@163. com

Abstract—**In this paper, a data classification method based on linear discriminant analysis and multivariate adaptive splines is proposed, and this method is a combination of dimension reduction and classification. Firstly, the most effective classification features are determined by linear discriminant analysis. Then, the input variables are divided into intervals by multiple adaptive regression splines (MARS), and the non-linear classification is transformed into the linear classification problem. Finally, classification is realized by perceptron. Experiment results prove that the method is effective both in classification performance and predicting speed.**

Keywords—machine learning; data dimension reduction; classification; multiple adaptive regression splines.

I. Introduction

Data dimension reduction and classification technology in machine learning is the core technology of artificial intelligence which has a wide range of applications. In the traditional segmentation algorithm ,support vector machine(SVM) as in [1], decision tree(random forest) as in [2] and deep learning as in [3] have good classification effect, support vector machine as in [4] and decision tree need to extract features, and their classification performance depends on the validity of features, , support vector machine uses inner product kernel function to replace the non-linear mapping to high-dimensional space, but it is difficult to implement for large-scale training samples, when the sample size is large, the computation of data will consume a lot of memory and time, and there is no suitable method to solve the kernel function. Decision tree is suitable for high-dimensional data with less computation, however, it is easy to over-fit data with inconsistent sample sizes. Deep learning is the best classifier at present, it enables computers to learn pattern features automatically and feature learning is integrated into the process of model building, thus, the incompleteness caused by artificial design features is reduced, however, deep learning can not estimate the rules of data without bias. In order to achieve better accuracy, a large number of training samples and great hardware support are needed.

II. Method

For the classification of some high-dimensional samples, such as soil classification based on infrared remote sensing images, in this paper, a method combining dimension reduction and classification is proposed, firstly, the most effective classification features are determined by linear discriminant analysis method, and then the interval segmentation of input variables is realized by multiple adaptive regression splines as in [5], and then it converting the non-linear classification problem to the linear classification problem, finally, the perceptron is used to realize classification as in [6]. The algorithm has the characteristics of fast and accurate.

The classification algorithm in this paper includes two processes: model building and classification prediction, and model building includes segment model building and classification model building. The specific steps of the algorithm in this paper are as follows:

The flow chart of the proposed method shows in Fig. 1.

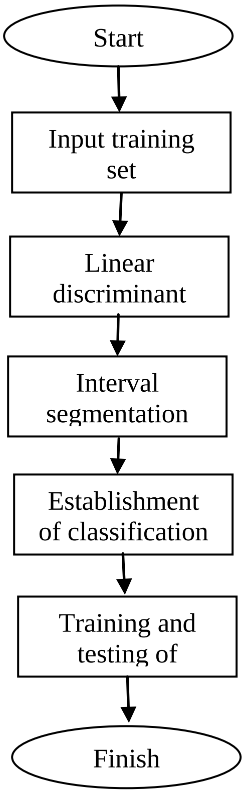

Fig. 1 Flow chart of the method

A. Establishment Of Model

1) Linear Discriminant Analysis

For M-dimensional vector x_o to be classified k is the number of categories, X_j is the class j sample set, μ_j is the mean vector of class j sample. First, linear discriminant analysis of X is carried out, and the dimension after dimension reduction is pre-set to d, the specific steps of dimension reduction are as follows:

a) Calculating intra-class divergence matrix S_w *:*

$$S_w = \sum_{j=1}^{k} \sum_{x \in X_j} (x_o - \mu_j)(x_o - \mu_j)^T \tag{1}$$

b) Computation of inter-class divergence matrix S_b *:*

$$S_b = \sum_{j=1}^{k} N_j (\mu_j - \mu)(\mu_j - \mu)^T \tag{2}$$

$N_j (j = 1, 2, \cdots, k)$ is the number of samples in category j.

c) Computational matrix $S_w^{-1} S_b$ *.*

d) Computing the largest d eigenvalues and corresponding d eigenvectors of $S_w^{-1} S_b$ *by matrix similarity diagonalization, get the projection matrix W,* w_1-w_d *is the column vector of W and W is the matrix of M column of d row.*

For each sample feature x_o in the sample set, transforming the new sample $x = W^T x_o$, x is the sample reduced to d-dimension.

2) Establishment of interval segmentation and classification model

a) Interval segmentation

Interval segmentation is a cyclic iterative process, the purpose is to represent the new data sample x in the form of MARS(Multivariate adaptive regression splines) basis function $B_m(x)$.

$$B_m(x) = \prod_{k=1}^{K_m} [s_{km}(x_{km} - t_{km})]_+ \tag{3}$$

For the m-dimensional component x_m of x, t_{km} is the node of the k-th straight line segment, $s_{km} \in \{-1, 1\}$, $[\cdot]_+$ denotes that the vector in square brackets takes only the part greater than 0, x_{km} is x_m after k-th linear piecewise. MARS divides x_{km} into two segments by t_{km}, k-th segmentation by recursive segmentation, the obtained basis function $B_m(x)$ is x_m after piecewise operation, however, it is necessary to subtract the value of node t_{km} before participating in the calculation of linear classification model. Among them, $\{m, s_{km}, t_{km}\}$ is the parameter of the basis function and $B_m(x)$ is the vector of the basis function, the specific steps are as follows:

- Adding the base function composed of each node t_{km} to the parameter set *basicFunctionList* of the current base function, *basicFunctionList* includes dimension dim of x to be partitioned, node t_{km} , the *basicFunctionLi-st{i}.dim*-dimension of x is divided into two dimensions greater than and less than the node value *basicFunctionList{i}.* t_{km} and added to the vector set *basisTmp* of temporary basis functions, establishing temporary classification model based on current basis function and calculating the error of current model
- The basis function vector corresponding to the minimum error is added to the set of basis function vectors *basisSet.*
- Establishment of classification model based on vector set *basisSet* of current basis function set.
- If the current error is less than the error threshold, the iteration is withdrawn

basisSet is $B_m(x)$ in equation (3), which is a new vector for classification and the original x is changed from d dimension to K_m dimension, and the whole $B(x)$ dimension is expressed in K_m dimension.

b) Establishment of classification model

The obtained $B(x)$ can be used as input variable of classifier to establish classification model, then the classification model is solved by perceptron method:

- Make $z = B(x)$, It is necessary to add one-dimensional constant term with value of one to z and the $K_M + 1$ -dimensional vector v whose number is c is initialized to a zero matrix.
- For each type of cycle, in the i loop, select another class j and calculate the probability difference e according to equation (4) for all samples z_i belonging to class i:

$$e = \langle v_i \cdot z_i \rangle - \langle v_j \cdot z_j \rangle \tag{4}$$

Select sample z_m with the minimum and negative value of parameter e in all sample z_i , When all $\langle v_i \cdot z_i \rangle > \langle v_j \cdot z_j \rangle$, calculate $v_i = v_i + z_m$, $v_j = v_j - z_m$, or exit the iteration when the number of iterations reaches the limit. The obtained model is $(K_M + 1) \times c$ -dimensional matrix, expressed by V.

B. Classified Prediction

For M -dimensional vector x_0 to be classified, c -dimensional real vector is obtained P_n , and classified prediction is achieved through the following five steps:

a) The vector x_o to be classified is multiplied by the projection matrix W obtained in linear discriminant analysis and then get a new sample $x = W^T x_o$; x is the sample reduced to d dimension.

b) According to the model parameter $\{m, s_{km}, t_{km}\}$, each one-dimensional component x_m in the model is segmented to form the vector $B(x)$ after the segmentation.

c) Let the first dimension of z be a full 1 vector, the second dimension and more than the second dimension be $B(x)$.

d) Calculating N c -Dimensional Real Vectors P_n according to $P_n = Vz$.

e) According to the maximum value of each P_n, its class number can be obtained.

III. RESULT

In this paper, three different types of high-dimensional small sample datasets are selected: ORL face data set as in [7], soil classification data set infrared images and infrared image fire detection data set. ORL face data set is a test set for classification and judgment criteria and the data set of infrared soil and fire detection was selected from the project "Application of infrared remote sensing in environmental protection, atmospheric and geological anomalies".

ORL face data set is a standard database created by Olivetti Research Laboratory in Cambridge, UK, it contains forty people, each of whom has ten pictures, each of which is eleven thousand two hundred and ninety-two in size, that is the dimension of each sample is ten thousand three hundred and four.

The soil data set（doi:10.3972/heihe.00134.2016.db）was downloaded from the Heihe Planned Data Management Center. The infrared image was selected using LANSAT8 data, and the LANSAT8 images were selected at different times, each image included eleven bands. The three thousand one hundred and thirty-one neighborhoods of the corresponding points of each soil data were selected to form a row, the dimension of each sample is ten thousand five hundred and seventy-one. There are one hundred and twenty-six samples and four types of soil.

Fire point data set is a self-made data set, which marks the known fire point information on the image, and LANSAT8 infrared remote sensing image is also used fire point data set is a self-made data set, which marks the known fire point information on the image, and LANSAT8 infrared remote sensing image is also used. A total of 110 fire spots are included in 18 remote sensing images. There are only two categories of fire detection: there are yes(fire points) and no (fire points). Because each remote sensing image is 7701×7821 in size and each corresponding point has eleven bands, which is equivalent to eleven large images, the 31×31 neighborhoods of corresponding points of fire point data are represented as a row, and a sample data is composed of two hundred and fifty-six samples, which are classified, that is, each sample has nine thousand and nine hundred dimensions and there are 110 samples in the data set. The location marker of the fire point is 1, the location marker of the non-fire point is 0.

The comparison method is the traditional principal component analysis (PCA) combined with SVM classifier, that is, PCA as in [8] is first used to reduce the dimension of sample data, and then SVM classifier is used to discriminate. In this paper, the comparison index is the accuracy of classification, and the verification method is K-Fold cross-validation as in [9]. The method of this paper is to reduce the dimension first, and then segment the reduced dimension by MARS. The purpose of this paper is to transform the non-linear classification problem into the linear classification problem, and the dimension after the segmentation is improved to a certain extent. Although the dimension of segmentation is improved, the speed is still very fast because of the linear classifier.

Table 1 lists the dimensionality reduction of PCA and the method in this paper, and K-Fold cross-validation on each data set and where K in the table is K in K-Fold cross validation.

Table 1 List of experimental parameters for each data set

Data Set Name	Sample dimension	PCA dimensionality reduction	Dimension reduction in this paper	K
ORL Face Data Set	10304	199	585	4
Soil Data Set	10571	125	600	3
Firepoint Detection Data Set	9900	255	105	3

The classification results of SVM classification method and this paper classification method on three data sets are shown in Table 2-Table 4.

Table 2 Soil data classification test results

Image Number	Image Data	Detection Rate of SVM	Detection rate of this method
1	20140426	82.4%	82.5%
2	20140917	83.5%	84.2%
3	20150413	82.2%	83.1%
4	20150718	79.2%	81.0%
5	20151022	81.5%	81.6%
6	20151225	77.4%	78.1%
Average detection rate		81.0 %	81.8%

Table 3 The results of classification and detection of fire detection data

Image Number	Image Data	Detection Rate of SVM	Detection rate of this method
1	20140612	81.3%	82.8%
2	20140725	85.1%	84.9%
3	20140813	86.9%	87.3%
4	20140910	89.9%	89.2%
5	20141024	82.7%	83.1%
6	20141125	89.7%	89.3%
Average detection rate		85.9 %	86.1%

Table 4 ORI Face Data Classification Results

Classification accuracy of SVM	The classification accuracy of the method in this paper
81.5%	81.8%

The average prediction me of each sample is 2.5 ms, 2.8 ms and 1.6 ms respectively. The prediction time of SVM is 15 ms, 10 ms and 7.8 ms. In summary, the method studied in this paper has obvious advantages in time.

Statistical results show that for three different types of data with high-dimensional and small samples, the classification accuracy of this algorithm is higher and more stable than that of traditional support vector machine, and the prediction time is faster than that of traditional support vector machine as in [10].

IV. CONCLUSIONS

For some high-dimensional small sample classification problems, such as soil classification based on infrared remote sensing images, in this paper, a method combining dimension reduction with classification is proposed. Firstly, the most effective classification features are determined by the linear discriminant analysis method. Then, the input variables are divided into intervals by MARS, and the non-linear classification is transformed into the linear classification problem. Finally, classification is realized by perceptron. The experimental results on face, soil and firepoint data sets show that the proposed method is superior to the combination of PCA and SVM in classification performance, and the prediction time is shorter.

ACKNOWLEDGMENT

The project is a subproject of “Application research on infrared remote sensing in environmental, atmospheric and geological abnormality” supported by “Beijing Institute of Spacecraft System Engineering”

REFERENCES

[1] Hongxin Cao, Research on Network Intrusion Detection Based on SVM[D].Nanjing: Nanjing University of Technology,2004.

[2] ZHANG X H, LIN B G. Research on Internet of Things Security Based on Support Vector Machines with Balanced Binary Decision Tree[J]. Netinfo Security, 2015, (8):20-25.

[3] SunZhi-jun，XueLei，Xu Yang-ming，eta1．Overview of deep learning[J]．Applicatin Research of Computers，2012，29(8)：2806—2810.

[4] M. Palaniswami, (2002), “Machine learning using support vector machines”, IEEE Signal Processing Society - North Melbourne.

[5] Zhang W, Goh ATC, Zhang Y, Chen Y, Xiao Y (2015) Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. Eng Geol.

[6] V. Cherkassky, and Y. Ma, (2002), “Practical Selection of SVM Parameters and Noise Estimation for SVM Regression”, submitted to Neurocomputing, special issue on SVM 2002.

[7] Li Kangshun, Li Kai, Zhang Wensheng. A PCA face recognition algorithm based on improved BP neural network [J]. Computer application and software, 2014, 31 (1): 158-161.

[8] Yuhong Wu, Xiaohong Tian, Yan’an Tong. 2010. Soil based on principal component analysis Evaluation of Comprehensive Index of Soil Fertility. Journal of Ecology, 29 (1): 173-180..

[9] Rodriguez, J. D., Perez, A., and Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3), 569-575.

[10] G. Nalbantov, R. Bauer and I. Sprinkhuizen-Kuyper, (2004), “Equity Style Timing Using Support Vector Regressions”, Social Science Research Network, Tomorrow’s Research Today.