A Comparative Study of Face Recognition Classification Algorithms

Wang Changyuan
School of computer science and engineering
Xi'an Technological University
Xi'an, China
E-mail: cyw901@163.com

Xue Pengxiang
School of computer science and engineering
Xi'an Technological University
Xi'an, China
E-mail: xuepx@xatu.edu.cn

Li Guang
Northwest Institutes of Advanced Technology
Xi'an Technological University
Xi'an, China
E-mail: 865413666@qq.com

Wu Qiyou
School of computer science and engineering
Xi'an Technological University
Xi'an, China
E-mail: 314650592@qq.com

***Abstract*— Due to the different classification effects and accuracy of different classification algorithms in machine learning, it is inconvenient for scientific researchers to choose which classification algorithm to use. This paper uses the face data published by Cambridge University as an experiment. The experiment first reduces the dimensionality of the data through the principal component analysis (PCA) algorithm, extracts the main features of the data, and then respectively through linear logic classification, linear discrimination LDA, nearest neighbor algorithm KNN, support vector machine SVM and the integrated algorithm Adaboost are used for classification. By comparing the advantages and disadvantages of the classification performance and complexity of different algorithms, the final review reviews accuracy, recall, f1-score, and AUC as evaluation indicators.**

Keywords-Classification Algorithm; Machine Learning; Face Recognition; Model Evaluation

I. INTRODUCTION

With the rise of artificial intelligence and machine learning, face recognition technology is widely used in life, such as station security, time and attendance punching, and secure payment [1-3], but different face recognition devices use different algorithms. Therefore, this paper analyzes and compares the commonly used classification algorithms in face recognition. The data set in this paper uses the ORL face data set published by Cambridge University in the United Kingdom. The methods used involve linear logistic regression, linear discriminant analysis (LDA), K-Nearest Neighbor (KNN), support vector machine (SVM), Na ïve Bayes (NB) and other methods. The definition and advantages and disadvantages of the act are briefly explained. Finally, the five methods are compared and analyzed according to the evaluation indicators such as the accuracy rate, recall rate, F1-score, and AUC area commonly used in machine learning.

DOI: 10.21307/ijanmc-2020-024

II. Related Work

A. *Principal component analysis PCA data dimensionality reduction*

The data set contains a total of 400 photos. We use the machine learning library Scikit-learn provided by python to process the data, and display part of the data set pictures as shown in Figure1.

Figure 1. ORL partial face image

The experimental data has 4096 features per picture. Since the number of features is much greater than the number of samples, it is easy to regenerate overfitting during training. Therefore, a principal component analysis algorithm is required to reduce the dimensions of the features and select K main features as the input of the data set. . The main idea of PCA [4] uses the covariance matrix to calculate the degree of dispersion of samples in different directions, and selects the direction with the largest variance as the main direction of the sample set. Processing process:

a) Data preprocessing normalizes and scales the data first. Normalization makes the mean value of data features 0, and scaling is to solve the case where feature values are different by an order of magnitude.

b) Calculate the covariance matrix and eigenvectors of the processed data. The eigenvectors can be obtained by singular value decomposition.

c) Retain the feature vector and feature value corresponding to the largest first K feature values to form an orthogonal basis.

d) Project the sample into a low-dimensional space, and the acquired dimensionality-reduced data can represent the original sample approximately, but with a certain degree of distortion.

This paper use formula (1) to calculate the distortion of PCA.

$$X = \frac{\frac{1}{m}\sum_{i-1}^{m}\left\|X^{(i)} - X^{(i)}_{approx}\right\|^2}{\frac{1}{m}\sum_{1}^{m}\left\|X^{(i)}\right\|^2} \qquad (1)$$

Through scikit-learn processing, the reduction rate after dimensionality reduction is obtained from the PCA model diagram is shown in Figure 2. It can be seen from the figure that the larger the value of k, the smaller the distortion rate. As k continues to increase, the data reduction rate will approach 1 Using this rule, choose between 10 and 300, and perform sampling calculation every 15th. Under the k features of all samples, the reduction rate is obtained after processing by the PCA algorithm. We select the reduction ratios at 98%, 90%, 80%, and 70%, and the corresponding k values are 195, 75, 32, and 20. The corresponding pictures after PCA processing are displayed. The first line of the image is the original image, and then each column corresponds to the It is a picture at different reduction rates. The lower the reduction rate, the more blurred the image is shown in Figure 3.

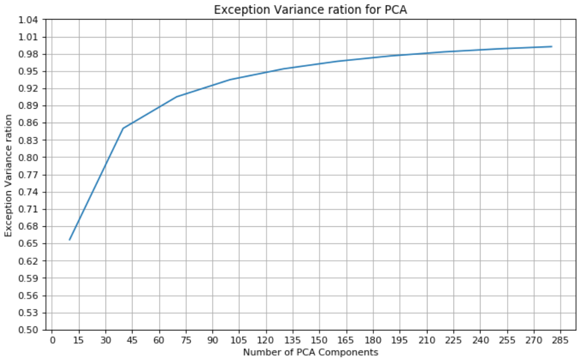

Figure 2. Relationship between reduction rate and K characteristics

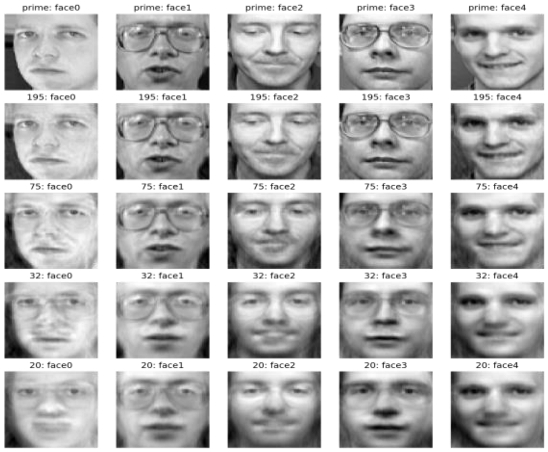

Figure 3. Relationship between reduction rate and K characteristics

B. Research on classification method

1) Logistic Regression

Supervised learning is the most widely used branch of machine learning in industry. Classification and regression are the main methods in supervised learning. Linear classifier is the most basic and commonly used machine learning model. This paper uses linear logistic regression to classify and recognize faces.

The prediction function of linear regression is:

$$\mathrm{h}_\theta(x) = [\theta_0, \theta_1 \ldots. \theta_n] \begin{bmatrix} X_0 \\ X_1 \\ . \\ . \\ . \\ X_n \end{bmatrix} = \theta^T X \tag{2}$$

Where is the prediction function, x is the feature vector, To handle the classification problem, this paper hope that the value of the function is [0,1], This paper introduce the Sigmoid function:

$$\mathrm{g}(z) = \frac{1}{1+e^{-z}} \tag{3}$$

Combined with linear regression prediction function:

$$\mathrm{h}_\theta(\mathrm{x}) = \mathrm{g}(z) = \mathrm{g}(\theta^T X) = \frac{1}{1+e^{-\theta^T X}} \tag{4}$$

If there is a simple binary classification of class A or class B, then this paper can use Sigmoid as the probability density function of the sample, and the result of the input classification can be expressed by probability:

$$P(y=1 \mid x, \theta) + P(y=0 \mid x, \theta) = 1 \tag{5}$$

The cost function is a function that describes the difference between the predicted value and the true value of the model. If there are multiple data samples, the average value of the replacement price function is obtained. It is expressed by J(θ). Close to, based on the maximum likelihood estimate available cost function J (θ).

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} (y^{(i)} \log(h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right] \tag{6}$$

2) Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [5，6], also known as Fisher Linear Discriminant (FLD), was introduced into the field of machine learning by Belhumeur. LDA is a dimensionality reduction technique in supervised learning, which can not only reduce dimensionality but also classify, and mainly project data features from high latitude to low latitude space. The core idea is that after projection, the projection points of the same category of data should be as close as possible, and the distance between the category centers of different categories of data should be increased as much as possible [5]. If the data set has two data sets, for the center of the two classes

Then within-class scatter matrix (within-class scatter matrix):

$$S_w = \sum 1 + \sum 2 \\ = \sum_{x \in x_1} (x-u_1)(x-u_1)^T + \sum_{x \in x_2} (x-u_2)(x-u_2)^T \qquad (7)$$

Between-class scatter matrix:

$$S_b = (\mathrm{u}_1 - u2)(\mathrm{u}_1 - u2)^T \qquad (8)$$

3) KNN

Among N training samples, find the k nearest neighbors of the test sample x. Suppose there are m training samples in the data set, and there are c categories, namely $\{\varpi_1, .. \varpi_c\}$, and the test sample is x. Then the KNN algorithm can be described as: Find k neighbors of x in m training samples, among which the number of samples belonging to category w_i in k neighbors of x are $k_1, k_2, ..kc$, then the discriminant function is

$$g_i(x) = k_i, i = 1,2,..c \qquad (9)$$

The core idea of K-nearest neighbor algorithm [7] is to calculate the distance between unlabeled data samples and each sample in the data set, take the K nearest samples, and then K neighbors vote to decide the type of unlabeled samples.

KNN classification steps:

a) Prepare the training sample set X, which contains n training samples, and select an appropriate distance measurement method according to specific requirements. This paper use dis(xa,xb) to represent the distance between ax and bx in the sample set.

b) For the test sample x, use the distance measurement formula to calculate the distance between the test sample x and n samples to obtain the distance set Dis, where

$$Dis = \{dis(x, x_1), dis(x, x_2), \ldots, dis(x, x_n)\}.$$

c) Sort the distance set, select the smallest k elements from it, and get k samples corresponding to k elements.

d) Count the categories of these k samples, and obtain the final classification results by voting.

Assuming that x_test is an unlabeled sample and x_train is a labeled sample, the algorithm is as follows:

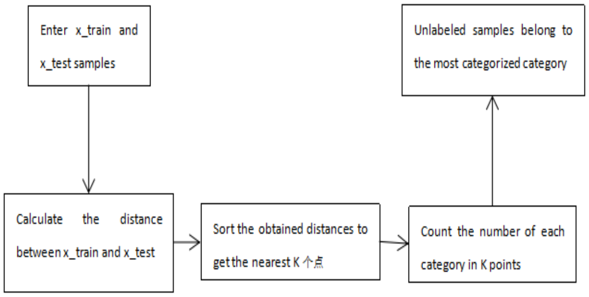
Enter x_train and x_test samples
Calculate the distance between x_train and x_test
Sort the obtained distances to get the nearest K 个点
Count the number of each category in K points
Unlabeled samples belong to the most categorized category

Figure 4. K nearest neighbor algorithm

For distance measurement, Euclidean distance, Manhattan distance, Chebyshev distance, etc. are usually used. Generally, Euclidean distance is mostly used, such as the distance between two points and two points in N-dimensional Euclidean space.

$$\mathrm{d}_{ab} = \sqrt{\sum_{i=1}^{N} (x_{1i} - x_{2i})} \qquad (10)$$

4) SVM

Support Vector Machine (SVM)[7] for short is a very important and extensive machine learning algorithm. Its starting point is to find the optimal decision boundary as far as possible, which is the farthest from the two types of data points. Furthermore, is the farthest from the boundary of the two types of data points, so the data point closest to the boundary is defined as a support vector. Finally, our goal becomes to find such a straight line (multi-dimensional called hyperplane), which has the largest distance from the support vector. Make the generalization ability of the model as good as possible, so SVM prediction of future data is also more accurate, as shown in Figure 5 below. Find the best Dahua margin.

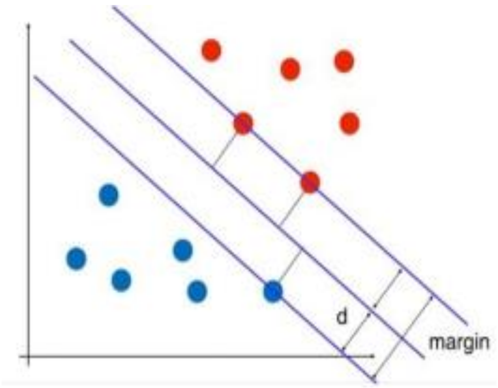

Figure 5. Classification model diagram

Let this plane be represented by g(x)=0, its normal vector is represented by w, the actual distance between a point and the plane is r point, and the distance from the plane can be measured by the absolute value of g(x) (called the function interval) .

$$\min_{w,b}\frac{1}{2}W^{T}W \\ y_i(W^{T}X_i+b)\geq 1, \\ i=1\ldots l \tag{11}$$

The penalty function is also called the penalty function, which is a kind of restriction function. For constrained nonlinear programming, its constraint function is called a penalty function, and the coefficient is called a penalty factor. The objective function of SVM (soft interval support vector machine) with penalty factor C is:

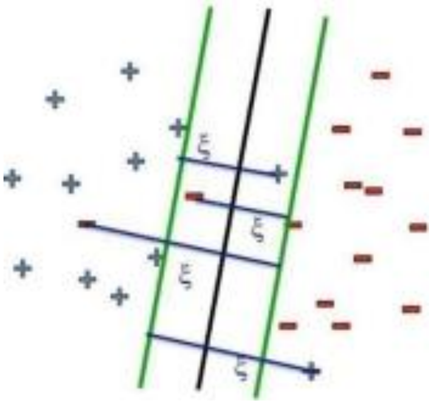

Figure 6. Introduction of C classification model diagram

$$\min_{w,b,\ \xi}\frac{1}{2}W^{T}W+C\sum_{\mathrm{i}=1}^{l}\xi \tag{12}$$

Advantages of SVM:

a) Non-linear mapping is the theoretical basis of SVM method, SVM uses inner product kernel function to replace the nonlinear mapping to high-dimensional space;

b) The optimal hyperplane to divide the feature space is the goal of SVM, and the idea of maximizing the classification margin is the core of the SVM method;

c) Support vector is the training result of SVM. It is the support vector that plays a decisive role in SVM classification decision.

d) SVM is a novel small sample learning method with a solid theoretical foundation. It basically does not involve probability measurement and the law of large numbers, so it is different from the existing statistical methods. In essence, it avoids the traditional process from induction to deduction, realizes efficient "transduction inference" from training samples to forecast samples, and greatly simplifies the usual classification and regression problems.

5) Naive Bayes

Naive Bayes [8, 9] is a conditional independence assumption, and there is no correlation between leave and leave. Suppose there is a labeled data set, where the data set has a total of categories, and for new samples, this paper predict the value. This paper use statistical methods to deal with this problem, so this paper can understand it as the probability of the category to which the sample belongs. The conditional probability formula is:

$$P(C_k \mid X) \tag{13}$$

Therefor $\mathrm{C_k} \in [\mathrm{C_1}, \mathrm{C_2} \ldots, \mathrm{C_m}]$, this paper only require the probabilities of m categories, the category belongs to the largest value, and use Bayes' theorem to solve:

$$P(C_k \mid X) = \frac{P(C_k)P(X \mid C_k)}{P(X)} \tag{14}$$

And because x has n feature vectors, it is in the determined data set， C_k , $P(x)$ are fixed values, according to the joint probability formula:

$$P(C_k)P(x|C_k) = P(C_k, x) \tag{15}$$

III. EXPERIMENTAL RESULTS

The data set in this paper uses the ORL face data set of Cambridge University, a total of 400 photos. After experimental analysis, this paper chose PCA to reduce the information rate after dimensionality reduction to 98%, and select 195 as the main features of each picture as data input. In order to make the experimental results more generalized, 80% of the data set is used as the training set and 20% is used as the test set. The 10-fold cross-validation method is used during model training. In order to ensure that the experimental results run the same data every time, a fixed random seed is set Is 7. For the test standard, this paper selected the accuracy rate (P), recall rate (R), F value, and AUC area. Compared with our prediction results, the accuracy rate indicates how many true positive samples are in the positive prediction samples. There are two possibilities for the prediction results, that is, the positive prediction is positive (TP), and the negative prediction is positive (FP), the formula is:

$$P = \frac{TP}{TP + FP} \tag{16}$$

The recall rate refers to our original sample, indicating how many positive examples in the sample were predicted correctly. There are also two possibilities, that is, the original positive class is predicted as a positive class (TP), and the other is to predict the original positive class as a negative class (FN).

$$R = \frac{TP}{TP + FN} \tag{17}$$

F1 combines the results of precision rate and recall rate. When F1 is higher, it means that the verification method is more effective.

$$F1 = \frac{2PR}{P + R} \tag{18}$$

TABLE I. THE COMPARISON OF EXPERIMENTAL RESULTS OF FIVE CLASSIFICATION ALGORITHMS

	PCA+LR	PCA+LDA	PCA+SVM	PCA+KNN	PCA+NB
Precision (%)	99	98	94	59	91
Recall (%)	99	96	91	45	85
Fl-score (%)	99	96	91	44	85

IV. CONCLUSION

Face recognition technology [10] has become one of the most popular research directions in computer vision and has made great achievements. With the development of computer technology, more and more classification methods will appear. This thesis is only through several common mainstream classification methods for experimental analysis and comparison. Through experiments, it is found that the PCA + linear logic classification method has obvious advantages in accuracy rate and recall rate. However, specific analysis should be combined with specific issues. Then I am ready to do experiments on different data sets, understand more classification algorithms, and constantly improve my results.

REFERENCES

[1] Wu Xiaotian. Research on Face Recognition Algorithm in Subway Security Inspection [D]. Dalian Jiaotong University, 2017.

[2] Chen Fuqiang. Research on invalid face filtering method in video attendance [D]. Southwest Jiaotong University, 2018.

[3] Ma Yukun. Research on key technologies of face-based secure identity authentication [D]. Beijing University of Technology, 2018.

[4] Pattern Analysis; New Pattern Analysis Findings from King Saud University Discussed (Pcapool: Unsupervised Feature Learning for Face Recognition Using Pca, Lbp, and Pyramid Pooling) [J]. Journal of Robotics &amp; Machine Learning, 2020.

[5] Zhang Yuting, Chen Junhua, Yang Xinkai, Zhang Liyan. An improved PCA+LDA face recognition algorithm [J]. Computer Knowledge and Technology, 2020, 16(03): 221-222.

[6] Guan‐Hua Huang, Chih‐Hsuan Lin, Yu‐Ren Cai, Tai‐Been Chen, Shih‐Yen Hsu, Nan‐Han Lu, Huei‐Yung Chen, Yi‐Chen Wu. Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction [J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2020, 13(5).

[7] Ou Lisong. Design and improvement of face recognition system based on SVM [J]. Network Security Technology and Application, 2019(12): 58-60. [8]Liu Jie, Song Bo. Naive Bayesian Classifier Based on Genetic Simulated Annealing Algorithm [J]. Procedia Engineering, 2011, 23.

[8] Tie Fuzhen. Application of face recognition system in hotel industry [J]. Computer Products and Circulation, 2020(07): 81+97.

[9] Jiang Ajuan, Zhang Wenjuan. A summary of face recognition [J]. Computer Knowledge and Technology, 2019, 15(02): 173-174+190.

[10] Youqiang Zhang, Guo Cao, Bisheng Wang, Xuesong Li. A novel ensemble method for k -nearest neighbor [J]. Pattern Recognition, 2019, 85.