Handwritten dynamics classification of Parkinson’s disease through support vector machine and principal component analysis

Zhifei Xu1, Zhaohui Zhu2*

1Department of Information Technology and Management, Zhejiang Police Vocational Academy, Hangzhou, Zhejiang Province, 310018, China
2School of Medicine, Hangzhou Normal University, Hangzhou, Zhejiang Province, 311121, China
*Corresponding author’s e-mail: zhuzhaohui@stu.hznu.edu.cn

Abstract. Parkinson's disease is a chronic neurodegenerative disease that affects the daily lives of tens of thousands of middle-aged and elderly people. The intelligent classification method of Parkinson's disease has received extensive attention in recent years. The paper proposes a new auxiliary classification model of Parkinson's disease based on principal component analysis and support vector machine. The model first samples and preprocesses the collected handwritten sensor data, then performs dimensionality reduction by principal component analysis, and finally inputs the dimensionality reduction data into a linear kernel support vector machine for Parkinson's disease classification and prediction. The experiment uses 5-fold cross-validation for dataset segmentation and performance verification. The average performance results obtained on the Meander r dataset are: accuracy is 70.86%, specificity is 67.23%, sensitivity is 75.98%, and F1-Score is 69.72%, and the average performance results obtained on the Spiral dataset are: accuracy is 77.45%, specificity is 70.26%, sensitivity is 85.58%, and F1-Score is 77.10%.

1. Introduction
Parkinson's disease (PD) is a progressive neurodegenerative disease that causes motor and non-motor symptoms, and is the second most common neurological disease [1]. The motor symptoms of Parkinson's disease include tremor, stiffness, postural instability and motor retardation, which are the direct result of insufficient dopamine signal due to the selective degeneration of dopamine producing neurons in the substantia nigra of the midbrain. In addition, more diverse and heterogeneous non-motor symptoms including autonomic dysfunction, hypofunction, cognitive decline, depression and sleep dysfunction. The exact cause of Parkinson’s disease is not yet known, but it’s thought to involve the complex interaction between genetics, biology and the environment.

The early diagnosis of Parkinson's disease is very important for the later health management and delay the development of the disease. In recent years, with the development of machine learning, computer-aided diagnosis of Parkinson's disease has received extensive attention from many scholars [2]. The data sources of computer-aided diagnosis are mainly reflected in: (1) voice data of Parkinson's patients. (2) Gait, handwritten dynamics data, etc.; (3) clinical data, such as clinical scale data, MRI images, DaTSCAN images.

Difficulty or sound disturbance is one of the earliest symptoms of Parkinson's disease. Among patients with Parkinson's disease, about 70% to 90% of the people suffer from Parkinson's disease
speech disorders [3]. People with PD have low voice and no expression, which makes themselves difficult to be understood due to voice tremors, sudden bursts, long pauses, and other problems [4]. Sharma et al. [5] proposed an improved grey wolf optimization method to identify patients with Parkinson's disease, which has been verified on both voice and language data sets. Experiments show that their method can maximize the accuracy and minimize the number of features. Sakar et al. [6] first studied the feature extraction of speech signals from patients with Parkinson's disease using Q-factor adjustable wavelet transform. In their study, 252 subjects' recordings were collected, and multiple feature recording subsets were extracted. Feature subsets are fed to multiple classifiers, and the classifiers are combined with ensemble learning method. The results show that the performance of Q-factor adjustable wavelet transform is better or comparable to the latest speech signal processing technology used in Parkinson's disease classification.

There are also many researches on intelligent diagnosis methods in gait and handwriting dynamics testing. Patients with Parkinson's disease have obvious frozen gait in the middle and late stage. The appearance of frozen gait reduces the patient's quality of life and leads to falls. These movement characteristics are collected by scholars to help identify Parkinson's disease. Zhao et al. [7] proposed a machine learning-based method that can automatically identify patients and assess their severity based on gait information (especially the sequential data of vertical ground reaction force recorded by foot sensors). A dual channel model combining long-term short-term memory (LSTM) and convolutional neural network (CNN) are developed to learn the spatial-temporal patterns behind gait data, which is verified on three public datasets. Experimental simulation proves the effectiveness of this method.

2. Methodology

This paper proposes a handwriting dynamics classification and prediction model for Parkinson's disease based on support vector machine and principal component analysis. The methods mainly include data acquisition, feature dimensionality reduction, classifiers modeling and Parkinsonism prediction, as shown in Figure 1. The main innovations of this paper are as follows: (1) A new handwriting dynamics classification model for Parkinson's disease is proposed; (2) The classification prediction model uses cross-validation for data segmentation and performance evaluation.
2.1. Data acquisition
This paper uses the internationally public available Parkinson's disease handwriting dynamics data set NewHandPD for modeling and prediction [12, 13]. The data set consists of 66 individuals, 35 healthy people and 31 sick people. Each individual uses a smart pen to perform 12 handwriting dynamics tests, which are: draw 4 spirals, 4 Meanders, 2 circular motions (1 is to draw a circle in the air, the other is to draw a circle on paper) and alternate movement of left and right hands. During the test, the handwriting dynamics of each individual was recorded through a smart pen, and each individual had 12 sensor signals. This paper only uses the spiral sensor signal and the zigzag line sensor signal recorded by the smart pen to construct the Parkinson's disease classifier. The population is described as follows: (1) The healthy group is composed of 18 men and 17 women, with an average age of 44.05±14.88; (2) The patient group is composed of 21 men and 10 women, with an average age of 57.83±7.85. The spiral dataset and Meander dataset used in this paper include 264 samples respectively.

2.2. Preprocessing
The smart pen for handwriting test is equipped with 6 different sensors to capture individual handwriting signals. Since the time for each individual to complete the spiral or meander test is different, we need to down-sample the signal of each sensor to obtain a fixed length n for further processing. It’s easy to get that the dimension of a sensor signal is (6, n). When n=1000, the handwriting test completed by the individual and the corresponding sampled sensor signal are shown in Figure 2. Figure 2(a) shows a spiral drawn by an individual with Parkinson's disease and the corresponding 6-channel sensor signal. Figure 2(b) shows the meander drawn by a normal individual and the corresponding 6-channel sensor signal.

Figure 2 (a). A spiral drawn by a sick individual and its corresponding sensor signal.
2.3. **feature dimensionality reduction**

Principal component analysis is a feature dimensionality reduction technique, which is widely used in many fields [12, 14]. It maps n-dimensional features to k-dimensional features through orthogonal transformation (k<=n). The obtained k-dimensional features are called principal components, and the variables are linearly independent and orthogonal with most of the original information retained. In short, principal component analysis is essentially a basis transformation, so that the transformed data has the largest variance, that is, utilizing the rotation of the coordinate axis and the translation of the coordinate origin get the smallest variance between one of the axes (main axis) and the data point. The orthogonal axis with high variance is removed after coordinate transformation to obtain a dimensionality reduction data set. Through principal component analysis, the original space is effectively compressed, and the high-dimensional data space is mapped to the low-dimensional space.

The data in this article is 6-channel sensor data, which can be directly input to the support vector machine for classification and prediction model construction after preprocessing. In order to improve the prediction accuracy, the data obtained by preprocessing is first reduced by principal component analysis, and then the reduced data is input into the support vector machine for classification model construction. The main steps of principal component analysis include:

1. Obtain the sample data matrix X;
2. Calculate the covariance matrix of the sample data matrix;
3. Calculate the eigenvalues and eigenvectors of the covariance matrix;
4. The feature vector is sorted according to the feature value from large to small;
5. Calculate the cumulative contribution rate of the first k feature vectors;
6. Obtain the principal components after dimensionality reduction based on the original data and feature vectors.

Due to the characteristic of the problem, the eigenvalues and eigenvectors of the covariance matrix can be estimated by the singular value decomposition method [15]. Assuming there is s×n dimensional data samples X, with s samples in total and n-dimensionality each row, the s×n matrix can be decomposed into:

\[ X = U \Sigma V^T \]  \hspace{1cm} (1)

Here, the dimension of the orthogonal matrix U is p×n, and the dimension of the orthogonal matrix V is n×n (satisfies: \( U^T U = V^T V = I \)), which is an n×n diagonal matrix. According to singular value decomposition, the data after dimensionality reduction can be expressed as:

\[ Y_k = U \Sigma_k \]  \hspace{1cm} (2)

This paper sets the number of features k equal to 200 after dimensionality reduction. After calculation, its cumulative contribution rate is greater than 99.0%, that is, the principal component after dimensionality reduction retains most of the information of the original data.
2.4. Support Vector Machine

The support vector machine (SVM) method, as a kind of machine learning, is first used to solve the binary classification problem and transform the classification problem into a convex quadratic programming problem under inequality constraints. Compared to neural networks and fuzzy logic methods, it is supported by more reliable mathematical principles. In supervised classification problems, SVM has been widely used in many fields as an effective classifier. The SVM method is mainly to find a hyperplane to divide the sample, separate the positive and negative examples in the sample with the hyperplane, and try the best to maximize the interval between the positive and negative examples.

Given a sample label pair \((p_i, q_i), i=1, 2, ..., s\), where \(s\) is the number of samples, \(p_i\) is a sample vector, and \(q_i \in \{1, -1\}\) represents one of the two classes category. The SVM is used for classification to solve an optimization problem, the formula is as follows:

\[
\begin{align*}
\min_{w, b, \xi} & \quad \frac{1}{2} w^\top w + C \sum_{i=1}^{s} \xi_i \\
\text{subject to} & \quad q_i (w^\top \phi(p_i) + b) \geq 1 - \xi_i,
\end{align*}
\]

Among them, the feature vector \(p_i\) is mapped to a high-dimensional space through the function \(\Phi\), and \(C>0\) is the penalty coefficient. \(w\) is the weight vector, \(b\) is the bias term, and \(\xi\) is the slack variable. The final optimal weight \(w\) and the classification decision equation can be defined as:

\[
\begin{align*}
w &= \sum_{i=1}^{n} q_i \alpha_i \phi(p_i) & (4) \\
\text{sgn}(w^\top \phi(p_i) + b) &= \text{sgn}\left(\sum_{i=1}^{n} q_i \alpha_i K(p, p_i) + b\right) & (5)
\end{align*}
\]

In this paper, due to the particularity of the problem, a linear kernel function is selected, which is defined as:

\[
K(p, p_i) = p^\top p_i
\]

2.5. Performance evaluation

This paper uses 5-fold cross-validation to perform data set segmentation and classifier performance evaluation. The relationship between the predicted category and the truth category of the classifier on the test set is shown in Table 1. The specific evaluation indicators of the classifier include accuracy, sensitivity, specificity and AUC area. The first three indicators are shown in equations (7)-(9).

| Predicted category | Truth category |
|--------------------|----------------|
| Positive (individuals with Parkinson's disease) | Negative (normal individuals) |
| Positive | True Positive (TP) | False Positive (FP) |
| Negative | False Negative (FN) | True Negative (TN) |

\[
\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

\[
\text{Sensitivity} = \frac{TP}{FN + TP}
\]
3. Experiment results

The experiment in this paper is based on the Windows 10 platform, the programming language is Python 3.7 and the machine learning library is Scikit-learn. The hardware configuration is CPU= Intel® Core™ i7-7500U with RAM=8.0GB. We chose Logistic regression (LR) classifier and extreme gradient boosting tree (XGBoost) classifier [16, 17] for comparison experiments, and analyzed from different sampling frequencies and four different performance indicators. The LR classifier parameters use the default parameters in the machine learning library Scikit-learn, the number of trees in the XGBoost classifier is 100, and the remaining parameters are set to the default parameters in Scikit-learn. The SVM classifier uses a linear kernel, and the remaining parameters are default parameters. For the two datasets: Meander and Spiral, when n=1000 and n=2000, any classifier performed10 times of 5-fold cross-validation, and the average performance indicators obtained are shown in Table 2 and Table 3. Before the data is input to each classifier, PCA dimensionality reduction technology is used to reduce the dimensionality of the features.

The average performance of different classifiers for the Meander dataset is shown in Table 2. When n=1000, XGBoost has the best classification performance, with an average accuracy of 70.99% and an F1-score value of 73.24%. Followed by the linear kernel SVM algorithm, the average accuracy rate is 70.86%, and the F1-score value is 69.72%. When n=2000, the classification performance of XGBoost drops significantly, while the LR classifier has an average accuracy of 72.16% and an average F1-score value of 71.03%. In general, the linear kernel support vector machine classifier used in this paper has relatively stable performance on the zigzag data set.

Table 2. Comparison of the classification performance of different classifiers on the Meander dataset

| Classifier   | n=1000        | n=2000        |
|--------------|---------------|---------------|
|              | Accuracy  | Sensitivity | Specificity | F1-score | Accuracy  | Sensitivity | Specificity | F1-score |
| LR           | 70.18%    | 73.55%      | 66.61%      | 69.40%    | 72.16%    | 78.63%      | 67.38%      | 71.03%    |
| XGBoost      | 70.99%    | 74.91%      | 68.37%      | 73.24%    | 69.15%    | 72.34%      | 67.65%      | 69.33%    |
| SVM_linear   | 70.86%    | 75.98%      | 67.23%      | 69.72%    | 70.38%    | 74.54%      | 66.49%      | 69.10%    |

For the Spiral dataset, the average performance of different classifiers is shown in Table 3. When n=1000, the linear kernel SVM in this paper achieves the best performance, with an average accuracy of 77.45%, an average sensitivity of 85.58%, and an average F1-score of 77.10%. All three performance indicators are the highest. When n=2000, this method has the highest average accuracy rate of 76.17%. In general, the XGBoost classifier has the lowest performance on this data set, and the SVM has the best performance.

Table 3. Comparison of classification performance of different classifiers on the Spiral dataset

| Classifier   | n=1000        | n=2000        |
|--------------|---------------|---------------|
|              | Accuracy  | Sensitivity | Specificity | F1-score | Accuracy  | Sensitivity | Specificity | F1-score |
| LR           | 75.52%    | 82.86%      | 70.26%      | 74.92%    | 75.20%    | 89.52%      | 66.20%      | 75.53%    |
| XGBoost      | 74.22%    | 77.67%      | 73.27%      | 75.42%    | 71.62%    | 74.28%      | 71.04%      | 70.31%    |
| SVM_linear   | 77.45%    | 85.58%      | 70.26%      | 77.10%    | 76.17%    | 83.47%      | 70.46%      | 74.94%    |

From Table 2 and Table 3, it can be found that the classification performance of each classifier on the spiral data set is better than that on the zigzag data set. For example, the accuracy of LR on the spiral data set is above 75.20%, while the accuracy on the zigzag data set is lower than 72.16%. For another example, the accuracy of the SVM on the spiral data set is above 76.17%, while the accuracy on the Meander data set is lower than 70.86%.
4. Discussion

This paper proposes a method combining support vector machine and principal component analysis to realize a new model of Parkinson's disease handwritten signal classification. On the Meander and Spiral data sets, after sampling preprocessing, feature dimensionality reduction, classification modeling, and Parkinson's disease prediction, the data set segmentation and performance evaluation are realized based on 5-fold cross-validation. The proposed method has the characteristics of few adjustable parameters and easy implementation.

This paper proposes a new solution for Parkinson's disease classification and prediction, but it also has certain limitations. First of all, the number of records in the public data set used is small. Secondly, the preprocessing and classifier modeling in this article only consider the spatial feature information, but not fully consider the time series information. Finally, the classifier modeling in this paper is carried out on the respective sensor data sets, and the fusion of multiple sensor information is not fully considered.

The future work will be mainly carried out around the following aspects: 1. The development of deep learning is gradually applied to various fields. In the later stage, it is planned to design a suitable deep learning network model for the zigzag and spiral sensor data sets, and make full use of space and time information to distinguish individuals with Parkinson's disease. 2. To design a multi-task learning model combining multiple handwriting test data sets to make full use of the individual's multi-modal handwriting test information. 3. Collect more individual multi-dimensional feature information, comprehensively analyze the probability and possibility of individual suffering from multiple aspects.

5. Conclusion

This paper proposes a handwritten dynamics classification and prediction method for Parkinson's disease based on support vector machine and principal component analysis. This method performs classifier modeling for the spiral data set and the Meander data set, and makes full use of the six-channel sensor information to classify and identify Parkinson's disease. The method in this paper uses 5-fold cross-validation to split the data set to evaluate the prediction performance of the classification model more accurately and objectively, and achieve a reasonable accuracy rate, which can be used as a potential method for clinical diagnosis of Parkinson's disease.

References

[1] Leroy, E., Anastasopoulos, D., Konitsiotis, S., Lavedan, C., & Polymeropoulos, M. H. (1998). Deletions in the Parkin gene and genetic heterogeneity in a Greek family with early onset Parkinson's disease. Human genetics, 103(4), 424-427.

[2] Goyal, J., Khandnor, P., & Aseri, T. C. (2020). Classification, Prediction, and Monitoring of Parkinson's disease using Computer Assisted Technologies: A Comparative Analysis. Engineering Applications of Artificial Intelligence, 96, 103955.

[3] Rusz, J., Cmejla, R., Ruzickova, H., & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson’s disease. The journal of the Acoustical Society of America, 129(1), 350-367.

[4] Schulz, G. M., & Grant, M. K. (2000). Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in Parkinson's disease: a review of the literature. Journal of communication disorders, 33(1), 59-88.

[5] Sharma, P., Sundaram, S., Sharma, M., Sharma, A., & Gupta, D. (2019). Diagnosis of Parkinson’s disease using modified grey wolf optimization. Cognitive Systems Research, 54, 100-115.

[6] Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., ... & Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable Q-factor wavelet transform. Applied Soft Computing, 74, 255-263.

[7] Zhao, A., Qi, L., Li, J., Dong, J., & Yu, H. (2018). A hybrid spatio-temporal model for detection and severity rating of Parkinson’s disease from gait data. Neurocomputing, 315, 1-8.
[8] Xu, S., & Pan, Z. (2020). A novel ensemble of random forest for assisting diagnosis of Parkinson's disease on small handwritten dynamics dataset. International Journal of Medical Informatics, 144, 104283.

[9] Prashanth, R., & Roy, S. D. (2018). Early detection of Parkinson’s disease through patient questionnaire and predictive modelling. International journal of medical informatics, 119, 75-87.

[10] Sivaranjini, S., & Sujatha, C. M. (2019). Deep learning based diagnosis of Parkinson’s disease using convolutional neural network. Multimedia Tools and Applications, 1-13.

[11] Oliveira, F. P., Faria, D. B., Costa, D. C., Castelo-Branco, M., & Tavares, J. M. R. (2018). Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson’s disease based on [123 I] FP-CIT SPECT images. European journal of nuclear medicine and molecular imaging, 45(6), 1052-1062.

[12] Xu, S., & Pan, Z. (2020). A novel ensemble of random forest for assisting diagnosis of Parkinson's disease on small handwritten dynamics dataset. International Journal of Medical Informatics, 144, 104283.

[13] João Paulo. (2017) The NewHandPD dataset. http://wwwp.fc.unesp.br/~papa/pub/datasets/Handpd/.

[14] Zhou Z, Wang Y. (2009) Applied machine learning. Tsinghua University Press, Beijing

[15] Chowdhary, C. L., & Acharjya, D. P. (2018). Singular Value Decomposition–Principal Component Analysis-Based Object Recognition Approach. Bio-Inspired Computing for Image and Video Processing, 323.

[16] Xu, S., Tang, Q., Jin, L., & Pan, Z. (2019). A cascade ensemble learning model for human activity recognition with smartphones. Sensors, 19(10), 2307.

[17] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.