Detecting Parkinson Disease Using Machine Learing Models

Bin Wang
Faculty of Science and Technology .
Univerisyt of Canberra
Canberra, Australia
u3199441@uni.canberra.edu.au

Xiaotong Yang
Faculty of Science and Technology .
Univerisyt of Canberra
Canberra, Australia
u3179117@uni.canberra.edu.au

Song Wang
Faculty of Science and Technology .
Univerisyt of Canberra
Canberra, Australia
u3194938@uni.canberra.edu.au

Yolanda Zhang
Faculty of Science and Technology .
Univerisyt of Canberra
Canberra, Australia
u3200644@uni.canberra.edu.au

***Abstract*—This project will employ three machine learning models to analyze Parkinson's Disease (PD) based on the patients' voice information. The dataset will be collected from the open source website, and further processed before imported into different models. The health status of the patients will be classified as health or PD positive. Different score measurements will be applied to demonstrate the efficiency of the proposed models. Different result comparisons will be conducted to find the best model to detect PD health problem.**

Keywords—machine learning, Parkinson's disease, voice information

I. Literature Review

PD is a long-term neurodegenerative disorder, affecting 2-3% of the population who are aged 65 or older [1]. It is caused by neuronal loss, resulting intracellular inclusions and dopamine deficiency. Not only it is a disorder of movement, it can also induce autonomic dysfunction, sleep deprivation, and cognitive impairment. PD Mortality increases after first ten years infection. One of the earliest indications of being infected by PD is vocal impairment, and nearly 90% of PD patients exhibit certain kind of vocal impairment. It is discovered that non-standard methods in combination with traditional harmonic-to-noise ratios can separate healthy from PD positive [2].

Traditionally, the diagnosis of PD is concluded based on patients' health record and physical examination [3]. However, along with the development of computer technology, Artificial Intelligence (AI) and machine learning models has become innovative approaches to detect PD through voice information or facial images. Decision tree is one of the most practical and powerful algorithms in data mining and is widely recognized in both classification and regression tasks [4]. Naïve Bayesian is based on the Probability Theory, which requires less training data. Its classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes [5]

Therefore, with the help of AI technologies, different models will be compared and analyzed in this paper to detect PD problems, providing references to the academic and medical communities.

II. Motivation and Research Problem

A. Research Motivation

PD is a disorder of the central nervous system, causing movement disability. It will cause server damage to public health, especially to the elderly. Even though scientists from different fields are devoting their researches to this problem, there is no effective cures. However, identifying the patients can allow early medical treatments.

There are different machine learning models available across different industrials for image recognition, spam emails filtering, speech recognition, traffic prediction, and so on. Each model has its own strength and weakness. Furthermore, data used for machine learning modeling also varies. Images, numbers, audio and descriptive texts are the most common form of data used in AI modelling. Therefore, it is essential to compare different models used for the targeted data to verify the efficiency and accuracy.

B. Description of Research Prolem Solved

In order to predict the Parkinson's disease, dataset has been selected and processed before being introduced to the different models. Based on the nature of the data, three models were chosen. What is more, for each model, part of data will be trained before the models were established, leaving the rest of the data tested for verification. The modelling process was implemented by using Python libraries. Finally, the results of using different models to detecting Parkinson's disease were visually presented.

III. Machine Learning Modelling and Implementation

Based on the nature of our dataset, we select the following classification models for training our data

- Naïve Bayesian
- Decision Tree
- Support Vector Machine (SVM)

A. Data Selection

The dataset was selected from UCI website at https://archive.ics.uci.edu/ml/datasets/Parkinsons, which is shared by Max Little in 2008. The data type is limited to numbers.

It has 197 instances with 23 attributes, among which one attribute indicates the health status as either health (0) or PD positive (1). One attribute is the name of each data record.

The other 21 attributes are the health indicators, including vocal fundamental frequency, variation in fundamental frequency, variation in amplitude, ratio of noise to tonal components in the voice, nonlinear dynamical complexity, and signal fractal scaling exponent.

There are no noise and missing data in the dataset. The health status reveals the modelling task as classification rather than regression.

B. Model Selction

There are numerous effective machine learning models, widely applied in medical science. Apart from the models we selected, there are models available like the Logistic Regressions, K-Nearest Neighbour, Convolutional Neural Networks and Clustering. The comparison between different models and reasons for selecting our models are as follows,

- SVM is supervised learning models, used for both classification and regression analysis. SVM Classifiers provide accuracy and is fast in prediction. It has been widely used in medical engineering, classifying genes and patients for their biological problems. The dataset we selected only consists of discrete numerical data, which also suits the SVM model. It also suits binary classification which fits the dataset.
- The aim of our models is to detect if the person has Parkinson disease. Therefore, regressions, predicting instead of classifying the results, is not appliable.
- CNN is widely used for imaging detection where each neuron receives input from every element of the previous layer, extracting features from images.
- Clustering is not suitable for our dataset since it is mostly used to recognize activities, identify pictures into different categories or collecting spam information. It involves the grouping of data where data in different groups should have significantly dissimilar properties.
- Naïve Bayesian is based on the Probability Theory, which needs less training data. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This feature suits our data.
- Because of the simplicity and integrity, the decision tree is widely recognized in both classification and regression tasks. Discrete variables are allowed for this model, and this model has been used for the diagnosis of medical conditions from the pattern of symptoms..

C. Implementation

1) Feature selection

The dataset has only 197 instances and 23 attributes. It is computable. All the information is organized and presented as numbers. Therefore, it is not needed for pre-processing. Also, there is no need for further sampling and the whole dataset is selected for training and testing. Because the attribute status is domain dependent and is prior information, it is selected as the feature.

2) Choice of parameters

80 Percent of the data will be used for training, leaving 20 percent of the data for testing. Beside using attribute of status as the feature and the attribute of name, the other 21 attributes are used as labels.

The default parameters are used for the Decision Tree model as

```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'presort': 'deprecated',
 'random_state': None,
 'splitter': 'best'}
```

The default parameters ae used for SVM model, as

```
{'C': 1.0,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 'scale',
 'kernel': 'rbf',
 'max_iter': -1,
 'probability': False,
 'random_state': None,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```

The default parameters ae used for Naive Bayesian model, as

```
{'priors': None,
 'var_smoothing': 1e-09}
```

3) Tuning

GridSearch CV is adopted for tuning the parameters for the three models. Tuning can be computationally expensive, if there are too many choices for the parameters. Therefore, their choice range are chosen accordingly. For Decision tree model, the possible tuning options are as follow

```
'max_depth':range(1,10)
'criterion':['gini','entropy']
'max_features':['auto','sqrt','log2',None]
'max_leaf_nodes':[1000,200,30,14,15,16,17,None]
```

For SVM model, the possible tuning options are as follow:

```
'C': [0.5, 0.1, 1, 5, 10]
'kernel': ['linear', 'poly', 'rbf']
class_weight': ['balanced', None]
gamma': [0.0001, 0.001, 0.01, 0.1, 0.005, 0.05,0.5]
```

Tuning for the Naive Bayesian model is not wildly recognized in application since there are only two parameters. The tuning options is listed as 'var_smoothing':[1e-010,1e-09,1e-09,1e-08]

IV. Model Evaluation

A. Results

A testing schedule is adopted where different tuning strategies are used to obtain the best results possible without sacrificing the computing consumption. The overall accuracy is used to verify the results. The results for the Decision Tree model are listed as

TABLE I. Decision Tree Tuning Testing

Tuning Parameters	Best Combination	Accuracy
No tuning	Null	0.897435897
max_features	sqrt	0.769230770
max_features, max_depth	None, 4	0.923076923
max_leaf_nodes	None	0.897435897
max_leaf_nodes, criterion, max_depth	1000, gini, 9	0.923076923

The tuning results for SVM model are listed in Table 2, as

TABLE II. SVM Tuning Testing

Tuning Parameters	Best Combination	Accuracy
No tuning	Null	0.846153846
kernel	linear	0.871794871
C	5	0.820512821
gamma, class_weight	0. 001, balanced	0.794871795
C, class_weight, kernel	1, balanced, linear	0.923076923

The tuning results for Naïve Bayesian model are listed in Table 2, as

TABLE III. Naïve Bayesian Testing

Tuning Parameters	Best Combination	Accuracy
No tuning	Null	0.692307692
No tuning with MultinomialNB	Null	Error occurred
var_smoothing	1e-08	0.769230770

B. Algorithm Analysis

From the three tuning results shown in tables, it can be illustrated that:

- All three models can successful predict the Parkinson's disease using patients' vocal information.
- The best accuracy obtained is 0.923076923, which can be achieved by either Decision Tree model or SVM model.
- Tuning can improve the accuracy. However, tuning more parameters with larger range does not necessary ensure better results. It is a process of testing and adjusting.
- MultinationalNB algorithm does not fit the dataset, since there are negative values. Instead, GaussianNB was employed. It has much lower accuracy than the others.
- Between the Decision Tree model and SVM model, Decision Tree model is much easier to reach the best accuracy. The general results for different parameter combinations are also better.

Furthermore, without tuning, the modelling results of accuracy and deviation is calculated in Fig. 1. It can also demonstrate the advantages of Decision Tree model over the other two models. The mean accuracy is higher and deviation is smaller.

Therefore, based on results, the Decision Tree Model is recommended for detecting this health problem and other problem with similar data structure and data type.

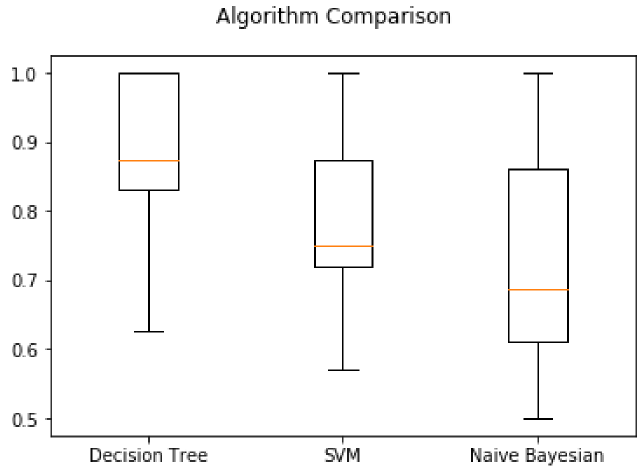

Fig. 1. Algorithm Comparison Between Different Models without Tuning

More measurement in analysing the best modelling results produced by the Decision Tree model are shown as

Best Hyper Parameters: {'criterion': 'gini', 'max_depth': 6, 'max_leaf_nodes': 1000}
Accuracy: 0.9230769230769231
Precision: 0.9203574203574204
Recall: 0.9230769230769231
F1-score: 0.9207100591715977
Matrix: [[5 2]
[1 31]]

	precision	recall	f1-score	support
0	0.83	0.71	0.77	7
1	0.94	0.97	0.95	32
accuracy			0.92	39
macro avg	0.89	0.84	0.86	39
weighted avg	0.92	0.92	0.92	39

It can be further demonstrated that, among 39 testing instance, only 3 instances were predicted incorrectly. F1-score, taking both the false positive and false negative into account, can reach as high as 0.92071006.

V. Major Contribution

This research provides an academic reference in diagnose Parkinson's disease using machine learning models. A testing schedule is adopted where different tuning strategies are used to obtain the best results possible without sacrificing the computing consumption. The overall accuracy is used to verify the results. The result comparison between different models proves that the best model to detect Parkinson's disease and other diseases with similar data structure is the Decision Tree Model.

However, this research has limits regarding data variance and research depth. In the future, more work should be carried to add more datasets in the proposed models. Patients' facial images can be also modelled. Furthermore, theoretical theories behind the models should be further researched to improve the models' efficiency and accuracy.

VI. References

[1] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A. E. Schrag, and A. E. Lang, "Parkinson disease. Nature reviews", Disease primers, vol. 3, no. 1, pp. 1-21, March 2017.

[2] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease". Nature Precedings, pp.1-1, 2008.

[3] J. M. Savitt, V. L. Dawson, and T. M. Dawson, "Diagnosis and treatment of Parkinson disease: molecules to medicine, " The Journal of clinical investigation, vol. 116, no. 7, pp.1744-1754. July 2006.

[4] Y. Y. Song, and L. U. Ying, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, vol. 27, no. 2, p.130, April 2015.

[5] I. Rish, "August. An empirical study of the naive Bayes classifier," In IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, No. 22, pp. 41-46. August 2001.