ASM Sci. J., Special Issue 6, 2019 for SKSM26, 22-27

Accuracy Improvement for Parkinson's Disease Classification Using Fuzzy Cluster and Forward Feature Selection with Support Vector Machine

Roselina Sallehuddin*, Tay Tze Loong, Nor Haizan Mohd Radzi, Azlan Mohd Zain, Yusliza Yusoff and Zuriahati Mohd Yunos

School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81300 Skudai Johor

Improving the predictive accuracy of Parkinson Disease classification has been an important task and become a hot topic these days since early diagnosis can assist the patients improve and maintain their quality of life. However, the performance of Parkinson classification is constrained by the inter-correlation between features and insignificant features in the Parkinson dataset. The objective of this paper is to combine SVM classifier with clustering algorithm and forward feature selection for obtaining accurate Parkinson classification while reducing the computational cost. Fuzzy C-Means (FCM) algorithm solves the correlation problem by subdividing dataset into several clusters while forward feature selection extracts a significant subset of features from the dataset by calculating the features weight according to a ranking order. Here, forward feature selection is implemented through the combination of PCA and wrapper approach and it is used to identify the optimum significant features that affect the SVM classification performance. The removal of insignificant features also can hasten the training time for model development because the Parkinson dataset can be represented with fewer features. Experimental results show that the combination of Fuzzy C-Means and forward feature selection has improved the classification ability of SVM in determining the Parkinson's disease. As a result, it can help the medical expert in making decision for providing better treatment to the patients.

Keywords: Parkinson classification; correlation; clustering; feature selection; Support Vector Machine; fuzzy c-means; forward feature selection

I. INTRODUCTION

Parkinson's disease (PD) becomes a common age-related disease after Alzheimer that affected many older adults worldwide. PD is a chronic deteriorating neurodegenerative disease caused by the death of neuron cells in the brain which releases dopamine. The reduction of dopamine content in the brain affects mainly the coordination of muscle tissues of the whole body. The patients who are suffering from Parkinson disease do not have the same symptoms. There are no definite tests for the diagnosis of Parkinson that can differentiate Parkinson disease from other disorders that have analogous clinical presentations. Parkinson disease is an idiopathic disease. Patients who suffered from Parkinson disease are often associated with cardinal symptoms such as bradykinesia, hypokinesia, stiffness or rigidity, rest tremor and posture instability (Khan, 2017; Rustempasic and Can, 2013). Apart from that, one may also suffer from psychiatric problems like depression and dementia (Shubham *et al.*, 2015).

Based on the recent studies, although previous researchers had used various methods including supervised learning method, unsupervised learning method, ensemble clustering method and hybridization of various methods, the accuracy of the performance is still remaining unsatisfactory (Saloni *et al.*, 2015; Froelich *et al.*, 2014). The performance of Parkinson classification is influenced by many factors. Two main issues in Parkinson classification are accuracy and computational time. The accuracy of existing classifiers is constrained by

*Corresponding author's e-mail: roselina@utm.my

uninformative and ambiguous variables that have been used as input variables for classifier during the development of the model. The Parkinson dataset contains both prominent and insignificant features which lead to data redundancy as well as misclassification on the Parkinson dataset. The data redundancy affects the classification accuracy and the training time for constructing the model. The ambiguous samples in dataset will lead to inappropriate output label and will affect the decision making. So, they need to be identified and eliminated from the training phase of the classifier. Therefore, an accurate classifier for Parkinson classification is required so that possible diagnosis errors made will be at minimal level and computational time for training the model can be reduced.

The objective of this study is to propose a hybrid model named FCM-PCA-SVM that combine classifier with clustering algorithm and forward feature selection method for obtaining accurate classification for Parkinson detection.

A. Literature review

Researchers had used different methods including univariate classification and multivariate clustering method to increase the accuracy of classification results on diagnosis of Parkinson's disease. The classifiers often used by researchers for Parkinson classification include K-Nearest Neighbor (KNN), Naïve Bayesian, Bagging, Boosting, Random Forest, Artificial Neural Network (ANN) as well as Support Vector Machine (SVM), (Subham *et al.*, 2015). Based on previous studies, ANN and SVM provided a better performance in decision making system and hence were being applied widely for Parkinson classification (Sriram *et al.*, 2016; Prashanth *et al.*, 2014; Gharehchopogh and Mohammadi, 2013).

However, the above methods do not have any mechanism to identify the relevant inputs and to reduce the ambiguous relations between the input and output. Therefore, in this study, a new classification model is proposed to handle these problems.

II. MATERIALS AND METHODS

A. Parkinson Dataset

The dataset is obtained from the University of California at Irvine (UCI) machine learning repository, prepared by Little *et al.*, in 2008. The Parkinson disease dataset contained samples of 31 patients with 23 patients diagnosed with Parkinson disease and the sample dataset consisted of 195 voice recordings with 23 attributes respectively. In the dataset, 48 of the subjects are healthy and 147 of the subjects were analyzed with Parkinson disease. Six phonations were recorded from each individual. The voice recordings analyzed the variation in vocal fold vibration frequency, measures of variation in amplitude, ratio of noise to tonal components, nonlinear dynamical complexity measure in the dataset for Parkinson classification. The model will discriminate the test subjects with output label as "1" as Parkinson patients and "0" as healthy person.

B. Proposed Hybrid Method

The research framework for developing FCM-PCA-SVM is shown in Figure 1. There are three phases involve namely Cluster analysis, Forward feature selection and SVM classifier.

The fuzzy C-Means clustering algorithm (FCM) helps to solve inter-correlation between features and identify misclassification cases in the Parkinson dataset. Cluster analysis plays an important role on amendment of mismatched data. After the cluster analysis, the results were extracted and the data which have mismatched cases to the clusters were eliminated manually. In other words, FCM will correct the mismatch data by clustering the data to appropriate cluster: 1 (Parkinson) and cluster 2 (healthy).

Next, both clusters are combined and PCA is used to rank the features based on weighted score before the irrelevant features are identified using forward feature selection.

Finally, the selected optimum features are fed into SVM classifier after data partitioning. Data division of Parkinson dataset was applied to split the whole dataset into training and testing set by a specific ratio. The classification performance of Support Vector Machine

was validated and analyzed using statistical errors.

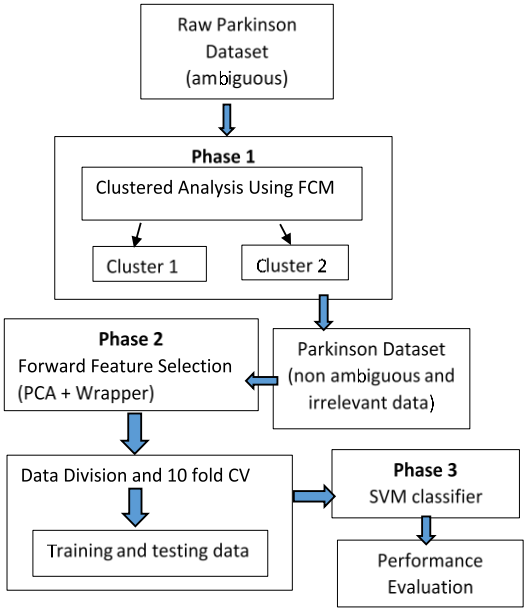

Figure 1. Framework for FCM-PCA-SVM model

C. Cluster Analysis Using Fuzzy C-Means (Fcm) Algorithm

Clustering algorithms such as Fuzzy C-Means are implemented to divide data objects into smaller sets of clusters so that each cluster demonstrates a high degree of intra-cluster similarity and inter-cluster dissimilarity. Based on previous studies, Fuzzy C-Means algorithm provided a better performance than K-Means algorithm because Fuzzy C-Means analyse the data based on the distance between various input data points which allows the data points to be belongs to more than one cluster (Lawrence *et al.*, 2013). The degree of membership value was determined by the distance from data points to various cluster centers. Cluster analysis aims to cluster the highly correlated features into clusters with similarity within the data points in the cluster and inter-dissimilarity between clusters and solves data redundancy. In the Parkinson dataset, the irrelevant data, correlation between features and misclassification of data affects the performance for Parkinson classification. Cluster analysis help to cluster data points which exhibits similar properties in the features in order to ease the differentiation of data points in the classifier. Besides, cluster analysis aids to identify the misclustered or misclassified data in the dataset. The misclassified data is caused by either noises or outliers and this might affect the accuracy of the classification. In order to handle the misclassified data, amendment of data is implemented via swapping data between clusters, changing cluster output or removal of misclustered data.

The main objective of fuzzy c-means algorithm is to minimize:

$$J(Z,U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m \left\|x_i - v_j\right\|^2 \qquad (1)$$

where $\left\|x_i - v_j\right\|$ is the Euclidean distance between x_i and v_j, μ_{ij} is the degree of membership of x_i data point in jth cluster center, x_i is the ith of d-dimensional measured data, v_j is the d-dimension center of the cluster, m is the fuzziness index where m is any real number between 1 and infinity [1,∞], n is the number of data points and c is the number of cluster center.

The clustering algorithm of Fuzzy C-Means can be described in the following steps:

1. Consider a dataset of n sample to be clustered, x_i.
2. Choose the number of clusters c, where $2 \leq c < n$.
3. Set the number of membership degree (fuzziness index), where $m \in R > 1$.
4. Initialize the $(n\ x\ c)$ sized membership matrix μ to random values, where $\mu_{ij} \in [0,1]\ and\ \sum \mu_{ij} = 1$.
5. Calculate cluster centers v_j defined by
6. $v_j = \frac{\sum_{i=1}^{n} {\mu_{ij}}^m x_i}{\sum_{i=1}^{n} {\mu_{ij}}^m}$ (2)

   for $j = 1, 2, \ldots, c$
7. Calculate the distance measure using Euclidean distance

   $d_{ij} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{c}\left(x_{ij} - v_{ij}\right)^2}$ (3)

   for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, c$.
8. Update the fuzzy membership matrix μ
9. $\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$ (4)
10. Stop if $\left\|\mu^{(k+1)} - \mu^k\right\| < \varepsilon$, where $\varepsilon \in [0,1]$. Otherwise repeat from step 5.

The cluster analysis is conducted using Fuzzy C-Means algorithm. The dataset sample is subdivided into 2 clusters by using Fuzzy C-Means algorithm. Based on the previous researchers, the membership's degree (m) equal to 2 is used as an ideal fuzzy exponent (Rustempasic and Can, 2013). The maximum number of iterations was set to 100 and the minimum gain ($\mathcal{E}$) was set to 0.001.

D. Forward Feature Selection

In this study, forward feature selection is implemented through combination of PCA and wrapper approach and it is used to identify the optimum significant features that affect the SVM classification performance. Principal Component Analysis (PCA) is applied to remove insignificant features and noise data. After the cluster analysis, forward feature selection ranking method has been applied.

Forward selection in feature selection refers to a feature addition which starts with the most significant feature to the least significant feature in the dataset based on the ranking order. In forward selection method, the most significant feature is examined and the performance is compared when each feature is added into the dataset based on the ranking order. The feature is added and evaluated repeatedly until the increment of feature does not provide any effect on the performance of the Parkinson classification. It indirectly affects the accuracy of the model as well as the computational speed required while learning the model. The training time for constructing the model can be speed up by using fewer features which are able to represent the whole Parkinson dataset. Forward selection has been applied with Principal Component Analysis (PCA), information gain and standard deviation. After that, the output of data pre-processing will be inserted as input data for constructing a better classifier for Support Vector Machine (SVM).

E. Development of Support Vector Machine classifier

After the data preparation process, the refined information of Parkinson dataset is used to train Support Vector Machine (SVM) classifier. The dataset is divided into training and testing dataset. Further, using 10 folds cross validation approach, the data is divided into ten equal partitions randomly. Nine of the partitions will be used as training data while the remaining partition will be used as testing data. The data division subset is applied repeatedly for 10 times in order to obtain the average result for Parkinson classification.

In SVM model, radial basis function (RBF) kernel is chosen among other 3 types of kernel functions. The gamma value of a Gaussian kernel is explored from 0.01 to 0.09 with an increment of 0.01. It is to handle non-linear classification. The C value is also explored from 0.01 to 0.09 to discover the best parameter for complex Parkinson dataset. The value of C and gamma influences the complexity of the data. 10 repetitions of 10-fold cross-validation were performed for validation by previous researchers.

F. Performance Measure

The performance of the proposed hybrid FCM-PCA-SVM is evaluated using statistical measurements and model comparison. The statistical measurements are accuracy, classification error, sensitivity and precision. To further validate the proposed hybrid model, comparative evaluations with standard SVM is carried out. The purpose of this comparison is to evaluate the effectiveness of Support Vector Machine (SVM) using Fuzzy C-Means (FCM) cluster analysis and feature selection as data pre-processing for classifier in Parkinson classification.

III. RESULTS AND DISCUSSIONS

To ensure the selected features is the best optimal features, the performance of PCA is compared with information gain and standard deviation methods. Principal Component Analysis (PCA) gives better selected features for Support Vector Machine (SVM) with least error values and a smaller number of features compared to information gain (16 features) and standard deviation feature selection (15 features) method.

Table 1 shows the comparative results between the proposed model and standard SVM model.

Table 1. Performance of Parkinson classification

Model	Accuracy	Classification Error	Sensitivity	Precision
FCM-PCA-SVM	85.5%	14.5%	100%	85.5%
SVM	75.37%	25.63%	95%	75.38%

The results obtained demonstrated using SVM with cluster analysis and PCA feature selection (13 features) enhance the Parkinson classification accuracy performance to 85.5% compared to 75.37% in simple SVM without cluster analysis and feature selection.

It can be seen clearly in that cluster analysis aided in the improvement of classifier models and the amendment of

cluster data can further enhance the Parkinson classification performance when comparing with the benchmark study of simple SVM without using cluster analysis and feature selection. In addition, feature selection by using fewer features aided in hastening the learning process when building the model. Based on the result obtained from the research, the combination of Fuzzy C-Means algorithm, Forward feature selection (Principal Component Analysis (PCA) with wrapper method) and Support Vector Machine (SVM) enhanced the performance for Parkinson classification.

IV. SUMMARY

Parkinson dataset contains irrelevant data, data redundancy and inappropriate output label that lead to misclassification of the data. Therefore, in order to improve the Parkinson classification accuracy, it is essential to have an efficient classifier which can handle the correlation between features and noise in the dataset. In this study, the results obtained from our proposed method FCM-PCA-SVM, showed that using cluster analysis by Fuzzy C-Means (FCM) algorithm and Principal Component Analysis (PCA) with Support Vector Machine (SVM) can enhance the Parkinson classification performance by subdividing data into correct clusters and eliminating irrelevant data.The identification of Parkinson patients in early stage is very important because it can help medical researchers to halt the disease from deteriorating and help to improve the life of the Parkinson patients.

The combination of classifier FCM-PCA-SVM model in this study has the potential of improving classification performance. It can be implemented to the classification of other diseases in medical field for detection of the disease in order to improve the life of the patients or be applied for prediction and forecasting for researchers and data analysts.

V. ACKNOWLEDGEMENT

This study is supported by GUP- tier 1 (VotNum 16H57). Authors would like to thank Research Management Centre (RMC) UniversitiTeknologi Malaysia, for the research activities and Applied Industrial Analytic Research Group (ALIAS) for the support and motivation in making this study a success.

VI. REFERENCES

Froelich, W., Wrobel, K., Porwik, P., (2014). Diagnosing Parkinson's Disease using the Classification of Speech Signals. Journal of Medical Informatics & Technologies. 23(1): 187-194.

Gharehchopogh, F.S., and Mohammadi P., (2013). A Case Study of Parkinson's Disease Diagnosis using Artificial Neural Networks. International Journal of Computer Applications. 73(19).

Khan, S.U., (2015). Classification of Parkinson's Disease Using Data Mining Techniques. Journal of Parkinson's disease & Alzheimer's disease. 2(1):4.

Lawrence, E.L., Fassola, I., Dayanidhi, S., Leclercq. C., and Valero, C.F.J., (2013). An Evaluation of Clustering Techniques to Classify Dexterous Manipulation of Individuals with and without Dysfunction. 6th Annual International IEEE EMBS Conference on Neural Engineering.

Little, M.A., McSharry, P.E., Hunter, E.J., Ramig, L.O., (2008). Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease IEEE Transactions on Biomedical Engineering. 56(4): 1015-1022.

Prashanth, R., Roy, S.D., Mandal. P.K., and Shantanu, G., (2014). Automatic Classification and Prediction Models for Early Parkinson's Disease Diagnosis from SPECT Imaging. Expert Systems with Applications. 41(7): 3333-3342.

Rustempasic, I., and Can, M., (2013). Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition. Southeast Europe Journal of Soft Computing. 2(1).

Sharma, R. K., & Gupta, A. K. (2015). Voice Analysis for Telediagnosis of Parkinson Disease Using Artificial Neural Networks and Support Vector Machines. International Journal of Intelligent Systems and Applications. 7(6): 41-47.

Shubham, B., Arvind, K.T., and Anil, K.S., (2015). A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction. International Journal of Computer Science and Information Technologies (IJCSIT). 6(2): 1648-1655.

Sriram, T.V.S., Rao, M.V., Narayana, G.V.S., and Kaladhar, D.S.V.G.K., (2016). A Comparison and Prediction Analysis for the Diagnosis of Parkinson Disease using Data Mining Techniques on Voice Datasets. International Journal of Applied Engineering Research. 11(9): 6355-6360.