TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 3, June 2020, pp. 1433~1438
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i3.14837

Prediction schizophrenia using random forest

Zuherman Rustam, Glori Stephani Saragih
Department of Mathematics, Faculty of Mathematics and Science, Universitas Indonesia, Indonesia

Article Info

Article history:

Received Aug 15, 2019
Revised Feb 12, 2020
Accepted Feb 21, 2020

Keywords:

Classification
Machine learning
Random forest
Schizophrenia

ABSTRACT

Schizophrenia is a mental illness with a very bad impact on sufferers, attacking the part of human brain that disables the ability to think clearly. In 2018, Rustam and Rampisela classified Schizophrenia by using Northwestern University Schizophrenia Data, based on 66 variables consisting of group, demographic, and questionnaires statistics, based on the scale for the assessment of negative symptoms (SANS), and scale for the assessment of positive symptoms (SAS), and then classifiers that used are SVM with Gaussian kernel and Twin SVM with linear and Gaussian kernel. Furthermore, this research is novel based on the use of random forest as a classifier, in order to predict Schizophrenia. The result obtained is reported in percentage of accuracy, both in training and testing of random forest, which was 100%. This classification, therefore, shows the best value in contrast with prior methods, even though only 40% of training data set was used. This is very important, especially in the cases of rare disease, including schizophrenia.

Corresponding Author:

Zuherman Rustam,
Department of Mathematics,
Faculty of Mathematics and Science,
Universitas Indonesia,
Margonda Raya St., Pondok Cina, Beji, Depok, Jawa Barat 16424, Indonesia.
Email: rustam@ui.ac.id

1. INTRODUCTION

Schizophrenia is a mental illness that has a very bad impact on sufferers, based on its ability to attack parts of the human brain, thus disabling the persons ability to think clearly [1]. Generally, patients experience a change either behaviorally or on the mind, which subsequently affects reality. This disease possesses the propensity to attack everyone at any age, the average attacks starting at the age of 20s in men, while in women, it was observed at the end of 20s [1]. It is, therefore, important to proficiently identify the symptoms on time, in order to prevent the disease from occuring in earnest. Schizophrenia is generally divided with three symptoms, including 1) the positive, presented with unnecessitated extra brain activities, including hallucinations, 2) the negative, indicated by the loss of brain activities, 3) the cognitive, which is exhibited as challenges with ability to remember and think. These are severe facts that confer potential negative effects on the life of sufferers, and their mortality rates are 2 to 2.5 times higher than the general population [2], 10% commited suicide and 20-40% attempted suicide at least once [3]. Then, the cause of schizophrenia has not been determined for sure [4] and this illness has been evaluated to get worse due to the incompetence of early detection, hence, the need to identify other approaches for detection, and one of which involves the use of computational methods, comprising of machine learning.

However, several papers have attempted utilizing this technique in the diagnosis of schizophrenia, including SVM with Gaussian kernel, Twin SVM with linear and Gaussian kernel [5], linear discriminant

analysis and k-Nearest neighbor [6], fisher linear discriminant analysis [7], Elastic Net, as well as least absolute shrinkage and selection operator [8]. This current research involved the use of random forest as a classify, although it has widely been used in various studies, including the prediction of bank financial failures, with accuracy of 93% [9], diabetes mellitus at 80.8% [10], automated diagnosis of heart disease at 83.6%, applying the weighted random forest [11], classify prostate cancer [12], chronic kidney disease [13] and osteoarthritis disease [14]. Hence, the selected approach has proven the capability of this classifier to apply to any problem, exhibiting good model performance in the process. This research is organized as follows: section 1 provides background of details, while the second specifies data and research method used. In addition, result and analysis were discussed in section 3, and finally, conclusions are included in section 4.

2. DATA AND RESEARCH METHOD

2.1. Data

Information from the database of Northwestern University Schizophrenia Data was used in this study [1]. Furthermore, there were 392 observations divided into 4 groups, with distributions as seen in Table 1. The study followed the grouping used by Rustam and Rampisela in the paper entitled "Support vector machines and twin support vector machines in the classification of schizophrenia data" [5]. This was because of the similarity in research focus, which was based on the categorization into schizophrenics and non-schizophrenics only, which served as the group variable [2]. Non-schizophrenics group consists of healthy siblings of the patients, the control, and siblings of control. A total of 66 data variables were collected, including the group, and demographics, consisting of gender, dominant hand, race, ethnicity, and age, using questionnaires statistics of scale for the assessment of negative symptoms (SANS) [15] and scale for the assessment of negative symptoms (SAPS) [16], as shown in Table 2.

Table 1. Distribution of group

Group	Number of observations
Schizophrenics	171
Siblings of Schizophrenics	44
Control	111
Siblings of Control	66

Table 2. The variable of schizophrenia data

nth Variable	Data Group	Variable	Description
1 - 34	Questionnaires of SAPS	SAPSi, $i = 1, \dots, 34$	SAPS is used to evaluate the positive symptoms of schizophrenia, SAPS is divided into 4 main sections containing 34 different symptoms, namely hallucination, delusion, bizarre behavior and thought disorder [16]. The data is in scale (0.5)
35 - 60	Questionnaires of SANS	SANSj, $j = 35, \dots, 60$	SANS is used to evaluate the negative symptoms of schizophrenia, SANS is divided into 5 main sections containing 25 different symptoms, namely emotional reaction decline, alogia, avolition and apathy, anhedonia and asociality, and attention [15]. The data is in scale (0.5)
61	Demographic	Gender	Gender is divided into 2 categories: 1) Male and 2) Female
62	Demographic	Dominant Hand	Dominant Hand is divided into 2 categories: 1) Left and 2) right
63	Demographic	Race	
64	Demographic	Ethnicity	Ethnicity is divided into 3 categories: 1) Kaukasia, 2) Africa-America, and 3) others
65	Demographic	Age	Integer (13.66)
66	Group	Group	Group is divided into 2 classes: 0) Non-Schizophrenics and 1) Schizophrenics

2.2. Research method

2.2.1. Bootstrap

In 1948, Queneiville introduced Jackknife as a resampling method, while Bradley Efron initiated bootstrap as its revolution in 1979, serving as a resampling method with replacement [17]. This allows for the creation of new data sets from the original, through repeatedly sampling the observations, as this is more feasible, in contrast with the method that required obtaining data from the population frequently [18]. This approach is a training data set, denoted as B, which contains n observations, randomly selected to

produce the new set, B^{*1}, indicating a similarity is size with the original. In addition, sampling was performed with replacement, which signifies the propensity of same observations appearing more than once [19]. The process is conducted continuously to the point where T new data set, T is capable of setting personally (e.g $T = 100$) or tune by an independent validation data set.

2.2.2. Random forest

In 2001, Breiman introduced random forest as a categorization tool, consisting of a collection of tree-structured classifiers $\{h(\boldsymbol{x}, \Theta_k), k = 1, \ldots\}$, where $\{\Theta_k\}$ are identical independent distributed random vectors, where each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$ [20].The approach used is aimed at improving stability and accuracy of the decision tree, through the creation of numerous units from existing training data, using the bootstrap method [21]. Random forest is capable of improving accuracy through randomization and voting methods, and it is also able to reduce the correlation between trees, without significantly reducing the strength of each [22]. Therefore, when overfitting is observed in a particular training data, others do not behave in the same manner [20]. Based on Breiman's paper, the process building of numerous trees does not create an overfit, although it produces a generalization error that converges to a value [20].

Algorithm random forest for classification [9].

1. Given the training data set, with n and m as observations and variables, respectively
2. For $b = 1$ to B
   a. Draw a bootstrap sample with n number of observations from the training (original) data set
   b. Build the decision tree T^b from each new result derived, where individual nodes are chosen at random.
      i. Select p variable at random from m, with $p \leq m$, where $p = 1,2, \ldots or\ less\ than\ \sqrt{m}$ [23].
      ii. Choose the best feature that provides satisfactory Information Gain or Gini Index [24].
      iii. Split the node
   c. Grow each without pruning
3. Output the ensemble of decision trees $\{T^b\}_1^B$
4. Conduct voting, i.e., if $\hat{C}_b(x)$ is the class prediction of the bth random forest tree, then $\hat{C}_{rf}^B(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$

The algorithm above can be representative with Figure 1.

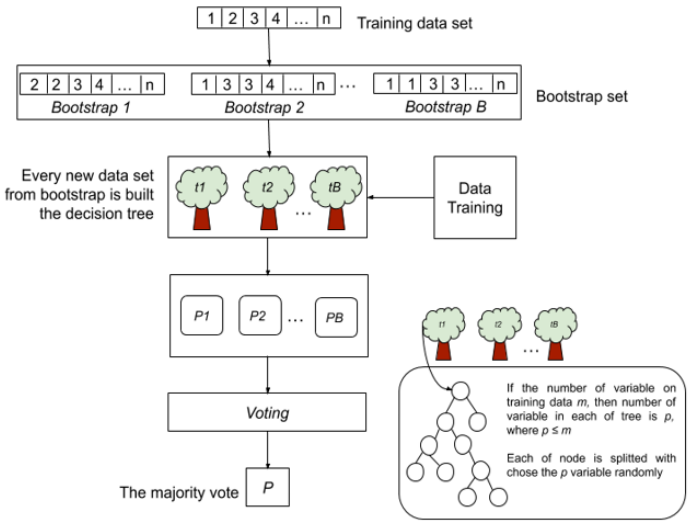

Figure 1. Flow of random forest

2.2.3. Evaluation of model performance

The evaluation of model performance is important, due to its ability to provides knowledge on the tools' efficiency of classifying data. This was assessed through the measurement of accuracy, obtained from the result of model with confusion matrix, where a high value indicates a good condition of the classification model [21]. Basically, it is known to contain comparable information with the result made by the model, including the binary type of classification, which indicates the presence of 2 output class, encompassing schizophrenics and non-schizophrenics. However, there are 4 parts to this confusion matrix:

1) TP: True Positive, observed on instances where schizophrenia is detected as Schizophrenics
2) TN: True Negative, is when non-schizophrenia is detected as non-schizophrenics
3) FP: False Positive, is seen in cases where non-schizophrenia is detected as schizophrenics
4) FN: False Negative, is when schizophrenia is detected as non-schizophrenics

Table 3 shows the confusion matrix used in this reserch.

Table 3. Confusion matrix

		Actual Class	
		Schizophrenics	Non-Schizophrenics
Predicted Class	Schizophrenics	TP	FN
	Non-Schizophrenics	FP	TN

Therefore, based on the value of TP, TN, FP, and FN from the matrix, it is possible to obtain the valued accuracy with the following formula:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP + TN + FP + FN}} \quad (1)$$

3. RESULTS AND ANALYSIS

3.1. Preprocessing data

There 66 variables in schizophrenia Data, therefore, feature selection was conducted before fitting the model, in order to improve accuracy. This was based on the percentage of missing data from variables less than 10%, thus, the feature is selected. Then, If the individual variance is more than that in the data collected from group, then choices are made based on those conditions, with 60 used in the model. Another means of promoting accuracy is by tuning the hyperparameters, which include those that affect the model structure and also the result output, therefore, there is need to identify its optimal set [25]. This is obtained by learning various algorithms with different sets, and subsequently comparing the results of eachs' performance, also known as tuning the model. Furthermore, the paper by Saragih and Rustam was followed for the use criterion (Gini and Entropy) [9]. The number of estimators/trees serves as hyperparameters to improve the accuracy, which collectively with the criterion is a function that measures the equality of a split, entropy for the Information Gain and Gini index [9].

3.2. Result and analysis

A study [5] applied 4 types of SVM on the same schizophrenia data, therefore, the main goal in this research is novel, through the use of random forest, in order to enhance predictability. This research required that the algorithm was run 10 times, and the repetition was performed due to the presence of element random in this experiment. As mention in section 3.1, the model tuning was conducted with the use of 2 hyperparameter combinations, and Table 4 provides the result of classification, using random forest with entropy, while gini was the criterion in Table 5. In this research, we used *scikit-learn* library. From Table 4, random forest with entropy in accuracy of training data was able to correctly classify schizophrenia data, with a 100% accuracy level for all compositions of training data set, and number of trees. This occurs on instances where the number of trees is 50 and 100 for 50-80% of the composition training data. From Table 5, the gini in accuracy of training data for random forest, provides the same result as entropy in the classification of schizophrenia data, which is 100% for all data set, and number of tree. Therefore, if this percentage was obtainable with entropy for testing in 50-80% training data set, then the result of gini in 40-80%, with the number of trees is 100, is correct.

Table 4. Accuracy of schizophrenia data classification using random forest with entropy as criterion

Percentage of	Number of Tree					
Data Training	10	50	100	10	50	100
(%)	Accuracy of Testing Data (%)			Accuracy of Training Data (%)		
10	94	95	98	100	100	100
20	95	96	98	100	100	100
30	97	98	99	100	100	100
40	96	99	99	100	100	100
50	97	100	100	100	100	100
60	99	100	100	100	100	100
70	100	100	100	100	100	100
80	100	100	100	100	100	100

Table 5. Accuracy of schizophrenia data classification using random forest with gini as criterion

Percentage of Data Training (%)	Number of Tree					
	10	50	100	10	50	100
	Accuracy of Testing Data (%)			Accuracy of Training Data (%)		
10	93	96	98	100	100	100
20	95	95	98	100	100	100
30	96	98	98	100	100	100
40	96	99	100	100	100	100
50	96	99	100	100	100	100
60	97	100	100	100	100	100
70	100	100	100	100	100	100
80	100	100	100	100	100	100

Based on Table 4 and Table 5, it is seen that a greater composition used is able to produce higher accuracy, due to the presence of more data learned by the model. Hence, tuning indicates the best accuracy on instances where the criterion entropy is with number of estimators as 100. However, it is seen that the results are similar for each composition training data and number of estimators. As said in Section 2, schizophrenia data was followed that same way as in the paper written by Rustam and Rampisela [5], due to a desire to compare performance results (accuracy) of the different methods used. Table 6, therefore, shows the comparison of each methods' Testing accuracy. Table 6 shows the highest accuracy occurring at random forest, which was in contrast with other SVM methods used in the paper of Rustam and Rampisela [5]. This was recorded at a level of 100%, when the percentage of training data is 40-80%, while others only achieved 90% at 60-80%

Table 6. Performance results

Training Data (%)	Linear SVM (%)	Gaussian SVM (%)	Linear Twin SVM (%)	Gaussian Twin SVM (%)	Random Forest (%)	
					Accuracy (%)	Std deviation
10	88	88	89	89	98	0.0001
20	89	89	89	89	98	0.0002
30	89	89	89	89	98	0.0001
40	89	89	89	89	100	0
50	89	89	89	89	100	0
60	90	90	90	90	100	0
70	90	90	90	90	100	0
80	90	90	90	90	100	0

4. CONCLUSION

Classification of schizophrenia has previously been conducted by Rustam and Rampisela, using SVM models, therefore, this research was novel in the use of random forest to predict based on the information collected from the database of Northwestern University Schizophrenia Data, which was also used by Rustam and Rampisela. Therefore, it was established that random forest with entropy and gini shows similar results, although it only slightly performs better with gini and the number of estimators being 100. Subsequently, this technique was also able to predict with good accuracy, using 40% training data, which was in contrast with other methods used in prior studies which insist on the use of 80%, in order to obtain 90% accuracy. This is very important, especially in the prediction of rare disease, where data is difficult to obtain.

In comparison with past studies using the same data, random forest was observed to show better accuracy, at 100% for training, and also testing. This approach is, therefore, expected to be relevant in the medical field, especially in the prediction of schizophrenia, and subsequently in other diseases, which is currently hard in diagnosis, hence, enhancing the accuracy of the medical team in providing treatment. It is suggested that successive research use feature selection to identify the important feature assists the medical practitioners to focus on several data points, and also that the application of random forest is adopted in other dataset with updated and superior dimension.

ACKNOWLEDGMENT

This research supported financially by University of indonesia with a DRPM PUTI Q2 2020 grant scheme.

REFERENCES

[1] Wang L., et al., "Northwestern University Schizophrenia Data and Software Tool (NUSDAST)," *Frontiers in Neuroinformatics,* vol. 7, pp. 25, 2013.

[2] Yasami M. T., et al., "Living A Healthy Life with Schizophrenia: Paving the Road to Recovery," *World Federation for Mental Heatlh: World Health Organization*, 2014.

[3] Schizophrenia.com. *Schizophrenia symptoms and diagnosis of Schizophrenia* retrieved from http://schizophrenia.com/szfacts.htm on 22 October 2019.

[4] American Psychiatric Association. "*Diagnostic and statistical manual of mental disorders," fifth edition* (Arlington, VA: American Psychiatric Association) https://doi.org/10.1176/appi.books.9780890425596. 2013.

[5] Rustam Z., Rampisela T. V. "Support vector machines and twin support vector machines for classification of schizophrenia data," *International Journal of Engineering & Technology,* vol. 7, no. 4, pp. 6378-6877, 2018.

[6] Ahn M., Hong J. H., and Jun S. C., "Feasibility of approaches combining sensor and source features in brain-computer interface." *Journal of neuroscience methods,* vol. 204, no. 1, pp. 168–178, 2012.

[7] Neuhaus A. H., et al., "Single-subject classification of schizophrenia using event-related potentials obtained during auditory and visual oddball paradigms," *European Archives of Psychiatry and Clinical Neuroscience,* vol. 263, no. 3, pp. 241-247, 2013.

[8] Hettige N. C., et al., "Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach," *Gen Hosp Psychiatry,* vol. 47, pp. 20-28, 2017.

[9] Rustam Z., and Saragih G., "Predict bank financial failures using random forest," *Institute of Electrical and Electronics Engineers*. doi: 10.1109/IWBIS.2018.8471718. pp.81-86, 2018.

[10] Zou Q., et al., "Predicting Diabetes Mellitus with Machine Learning Techniques," *Frontiers in Genetics,* vol. 9, no. 515, 2018.

[11] Patil P. R., and Kinariwala S. A., "Automated Diagnosis og Heart Disease using Random Forest Algorithm," *International Journal of Advance Research, Ideas and Innovations in Technology.* vol. 3, no. 2, pp. 579–589, 2017.

[12] Huljanah M., et al., "*Feature Selection using Random Forest Classifier for Predicting Prostate Cancer*," IOP Conference Series: Materials Science and Engineering, vol. 546, no. 5, 2019.

[13] Rustam Z., Sudarsono E., and Sarwinda D., "*Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD),"* IOP Conference Series: Materials Science and Engineering, vol. 546, no. 5, 2019.

[14] Aprilliani U., and Rustam Z., "Osteoarthritis Disease Prediction Based on Random Forest," *Institute of Electrical and Electronics Engineers*. DOI: 10.1109/ICACSIS.2018.8618166. pp. 237-240, 2018.

[15] Andreasen N. C., "*Scale for the Assessment of Negative Symptoms* (SANS)." *conceptual and theoretical foundations. The British journal of psychiatry*, pp. 49-52, 1989.

[16] Andreasen N. C., "*Scale for the Assessment of Positive Symptoms"* (SApS. Iowa City: University of Iowa). 1984.

[17] Efron B., and Tibshirani R., "An Introduction to the Bootstra," New York: Chapman & Hall. 1993.

[18] Hastie T., Tibshirani R., and Friedman J., "The element of statistical learning data mining, inference, and prediction," California: Springer. 2009.

[19] James G., et al., "First Edition; An Introduction to Statistical Learning with Application in R," *Springer*. New York: Springer-Verlag New York, vol. 112, pp. 3-7, 2013.

[20] Breiman L., "Random forests," *Mach. Learn*, vol. 45, no. 1, pp.5–32, 2001.

[21] Mindermann S., "Random forests," Thesis. Amsterdam: Korteweg-de Vries Instituut voor Wiskunde, Universiteit van Amsterdam, 2016.

[22] Strobl C., "Statistical Issues in Machine Learning-Towards Reliable Split Selection and Variable Importance Meaasures," Munchen, 2008.

[23] Breiman L., et al., "Classification and regression trees" *Wadsworth Int*, 1984.

[24] Grabczewski K., "Meta-Learning in decision tree induction," *Torun: Springer*, vol. 1, 2014.

[25] Kohavi R. and Provost F., "On applied research in machine learning," *Editorial for the Special Issue on Application Machine Learning and the Knowledge Discovery Process, Columbia University,* vol. 30, pp. 127-132, 1998.