Machine learning algorithm for improving performance on 3 AQ-screening classification

Taftazani Ghazi Pratama*, Rudy Hartanto, Noor Akhmad Setiawan

Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Jl. Grafika No.2, Yogyakarta 55281, Indonesia

Article history:
Received: 24 May 2019 / Received in revised form: 24 July 2019 / Accepted: 18 September 2019

Abstract

Autism Spectrum Disorder (ASD) classification using machine learning can help parents, caregivers, psychiatrists, and patients to obtain the results of early detection of ASD. In this study, the dataset used is the autism-spectrum quotient for child, adolescent and adult, namely AQ-child, AQ-adolescent, AQ-adult. This study aims to improve the sensitivity and specificity of previous studies so that the classification results of ASD are better characterized by the reduced misclassification. The algorithm applied in this study: support vector machine (SVM), random forest (RF), artificial neural network (ANN). The evaluation results using 10-fold cross validation showed that RF succeeded in producing higher adult AQ sensitivity, which was 87.89%. The increase in the specificity level of AQ-Adolescents is better produced using an SVM of 86.33%.

Keywords: Classification, ASD, machine learning, sensitivity, specificity, SVM, RF, ANN

1. Introduction

ASD (Autism Spectrum Disorder) is a neurological development disorder in individuals that can be identified through characteristics such as repetitive behaviour, limited social interaction and communication [1]. The causes of the emergence of ASD are unknown and clear [2]. Some factors that influence the emergence of ASD are: genetic, environmental, and immunologists of an individual. Children aged 2-9 years assessment of diagnosis involves professional experts [3]. According to the Centers for Disease Control and Prevention (CDC), an individual can be diagnosed with ASD by going through two stages: 1. Initial screening / detection 2. Interview with psychiatrists with children or parents and observing behaviour by pediatricians and neurologists [4].

At this time the instruments used for screening ASD are Autism Behavior Checklist (ABC), Autism Spectrum Quotient (AQ), Child Behavior Checklist (CBCL), and many others [5–7]. Inside the instrument there are many items that represent questions related to the characteristics of ASD. The use of the instrument is carried out by parents, teachers, caregivers, and patients themselves. However, with many answers that must be filled, it makes the time taken longer and inefficient. Therefore, fewer items are needed in order to speed up the ASD screening process. The solution provided by C. Allison et al. [7] produces screening instruments with fewer items. For example, AQ that is applied at the age of toddler to adult from the beginning has a number of items of 50 and then decreases to 10 items. The items in the ASD screening instrument are given to the patient to be filled in or answered in order to find out the condition of the patient regarding the characteristics that exist whether they meet the ASD criteria or not. The ASD classification in particular using Autism-Quotient has guidelines provided by the National Institute for Health and Care Excellence (NICE) [8,9] to assess each item during the screening process. In addition to the existing guidelines, ASD classification can also be perform using machine learning. Machine learning can classify individuals as indicated by ASD by learning from previous data. Previous studies related to ASD classification using the AQ-Screening methods (AQ-child, AQ-adolescent, AQ-adult) have been carried out by Thabitah et al. [10] by means of a feature selection method: Variable analysis (Va). The feature selection results from Va obtained the following subset of features: AQ-child obtain 8 features, AQ-adolescent obtain 8 features, and AQ-adult obtain 6 features. The subsets of features obtained along with the existing instances are then classified using the classification algorithm: C4.5 and Ripper.

However, the classification results using C4.5 and Ripper showed a misclassification that caused a low sensitivity level on AQ-adult and a low specificity level on AQ-Adolescent. Hence, it leads to poor performance of the prediction results of patients who are being identified as an ASD. With this in mind, several studies has been conducted to increase the sensitivity and specificity by proposing better classification characterized algorithms to reduce the misclassification. Some of them are successfully producing higher sensitivity and specificity algorithms, especially algorithms that related to ASD classification and other medical fields [11–15]. Previous studies conducted by Komisicky et al. [11], autistic detection using SVM, Naïve Bayes, decision tree variants, RF. Some of these algorithms, SVM and RF can produce higher sensitivity
and specificity than other algorithms. The dataset used is the Autism Diagnostic Observation Schedule (ADOS) in module 2 and module 3 obtained from 5 autism organizations, namely: Autism Consortium (AC), Autism Genetic Resource Exchange (AGRE), National Autism Research Database (NDAR), Simons Simplex Collection (SSC), Simons Variance in Individual Projects (SSVIP).

Another study conducted by Maenner et al. [12] using RF for ASD classification through words or sentences written with supervision of various clinical and school supervision. A text-based dataset obtained from Georgia ADDM (The Autism and Developmental Disabilities Monitoring). In addition to ASD classification through data, ASD classification can also be done through brain image as done by Zhang et al. [13] using SVM. Image-based datasets that describe brain images was obtained from the Autism Research Center through Children's Hospital of Philadelphia.

In other medical fields, in the study [14] which classified chronic kidney disease using SVM and ANN. The higher levels of sensitivity are generated by ANN than SVM. The dataset is obtained from Apollo Hospitals, Tamilnadu, India.

In research by [15], the authors compared to machine learning algorithms: RF, Logistic Regression (LR), ANN (MLP) for predictions of fatty liver disease. The results of optimal sensitivity and specificity is generated by RF. While ANN is third after LR. However, these previous studies did not yet apply their algorithms on ASD datasets in this study. So, this study will use SVM, RF, and ANN to acquire a higher sensitivity and specificity algorithms using the same algorithms as the [11–15]. The aim of this study is to increase the level of sensitivity and specificity in previous studies [10] for ASD classification on 3 AQ-Screening (AQ-Child, AQ-Adolescent, and AQ-Adult).

2. Materials and Method

2.1. Dataset

The dataset used in this study was obtained from the UCI Machine Learning Repository [16–18] consisting of 3 age-related ASD screening datasets from children to adults. AQ-Children, AQ-Adolescent, AQ-Adult have the same number of features or attributes: 21. The number of instances of the 3 datasets is different because there are different proportions between ASD and non-ASD classes. The number of instances contained in the three datasets amounts to the following: 292, 104, and 704.

Features and Description are shown in Table 1. In Table 1, the values contained in A1-A10 attributes in each dataset are coded with 0 or 1 depending on the answers given by the participant during the ASD screening process.

| Feature | Type | Description |
|---------|------|-------------|
| Age     | Number | Participant age |
| Gender  | String | Male or female |
| Ethnic  | String | Ethnic participant |
| Born with | Boolean | Does jaundice appear at birth |
| jaundice | (yes or no) | |
| A native of | Boolean | Whether families have history of autism |
| autistic family | (yes or no) | |
| Who does the | String | Parent, self, caregiver, medical staff, clinician, etc |
| screening | |
| Country | String | List of participant countries |
| Have already | Boolean | Has the participant ever screened |
| screened through | (yes or no) | through the previous application? |
| the application | |
| Description of age | String | The age range of children, adolescents and adults |
| First to Tenth | Binary (0, 1) | The question can be seen on Table 2 |
| question answer (A1-A10) | |
| Seening Score | Integer | The final result of screening through the app with interval rate 1-10 |

In Table 2, 10 questions in one of the AQ-Screening Methods are presented, namely AQ-10 Child. Each question represents the categories in ASD screening. Giving a value of 0 or 1 for each dataset varies. For AQ-Child on questions A1, A5, A7, A10 coded with value 1 if the participant chooses answer definitely or slightly agree and value 0 if the participant chooses answer definitely or slightly disagree. In questions A2, A3, A4, A6, A8, A9 is coded with a value of 1 if the participant chooses the answer definitely or slightly agree. While the value of 0 if the participant chooses the answer definitely or slightly agree. For AQ-Adolescent, the value 1 is set for questions A1, A5, A8, and A10 if the participant answers definitely or slightly agree.

| Feature | The Questions on AQ-10 Child |
|---------|-----------------------------|
| A1      | He or she often noticed small voices while others didn't |
| A2      | He or she usually concentrates on the whole picture rather than the small details |
| A3      | In social groups, conversations of different people can easily be tracked |
| A4      | He or she felt it was easy to go back and forth between various activities |
| A5      | He or she doesn't know how to maintain conversations with friends |
| A6      | He or she is good at social chatting |
| A7      | When he or she reads a story, he or she finds it difficult to know the character's intentions or feelings |
| A8      | When he or she was in preschool, playing a game pretending he or she always enjoys |
| A9      | He or she found it easy to know what someone was thinking or feeling just by looking at their faces |
| A10     | He or she found it difficult to make new friends |

2.2. Information Gain

Information Gain is a measure used to calculate how much information is obtained from a feature on a class [19]. The
formula for Information Gain can be seen in Equation (1).

$$Gain(X, A) = E(X) - \sum_{i=1}^{n} \frac{|X_i|}{|X|} \cdot E(X_i)$$  

(1)

where $X$ is the set of cases, $E$ is Entropy, $n$ is the number of partitions of attribute $A$, $|X_i|$ is the number of cases of partition $I$, and $|X|$ is the number of cases in $X$.

2.3. Chi-Square

Chi-Square is a method for calculating correlations between 2 variables. Non-class variables with class variables where the data types of the two variables are nominal [20]. The formula for Chi-Square can be seen in Equation (2).

$$Chi - Square = \frac{N. (AD - BC)^2}{(A+B)x(C+D)x(A+C)x(B+D)}$$  

(2)

where $N$ is the number of instances in training data, $A$ is the frequency of non-class variables with class variables, $B$ is the frequency of non-class variables without class variables, $C$ is the frequency of class variables without non-class variables, and $D$ is the frequency of unbound instances on non-class variables and class variables.

2.4. Variable analysis (Va)

Variable analysis is a filter-based feature selection method made by Thabtah et al. [10] by combining Information Gain and Chi-Square to get fewer features in the dataset. Va produces a score for selected features that are optimal by not decrease the performance level of the classification model badly.

Va combines the IG and Chi-Square scores with the following steps: 1. Calculate the selected attribute or feature score using IG and Chi-Square; 2. Normalize the score of each feature from the calculation of IG and Chi-Square respectively; 3. Calculate the magnitude of score from the previous normalization process by combining the two scores of each attribute of the IG calculation and Chi-Square.

The result of Va is a new vector based on the IG and Chi-Square score each feature and calculates the magnitude of the vector which is also called $M_{score}$. $M_{Score}$ in Va is using as a strong measure in choosing optimal features. Mathematical formulations are shown in Equations (3) - (6).

$$\text{ChiSQv} = \frac{\text{ChiSQ}}{\text{ChiSQmax}}$$  

(3)

where $\text{ChiSQv}$ is the result of normalization of $\text{ChiSQ}$ scores for each attribute with the maximum $\text{ChiSQ}$ score of features, $\text{ChiSQ}$ is the $\text{ChiSQ}$ score generated on each feature, and $\text{ChiSQmax}$ is the maximum $\text{ChiSQ}$ score.

$$\text{IGv} = \frac{\text{IG}}{\text{IGmax}}$$  

(4)

where $\text{IGv}$ is the result of normalization of IG scores for each attribute with maximum IG score of features, IG is the Information Gain score generated on each feature, and $\text{IGmax}$ is the maximum IG score.

The score vector of each feature(v) can be defined to be

$$SVy = \left[ \frac{\text{IGv}}{\text{IGmax}} \right]$$  

(5)

The magnitude of the score vector is calculated by summing ChiSQv and IGv from each attribute and then squaring it then rooting it. Formula for vector size is shown in Equation (5).

$$|SVy| = \sqrt{(IGv)^2 + (\text{ChiSQv})^2}$$  

(6)

2.5. Support Vector Machine (SVM)

SVM is a machine learning algorithm made by Vapnik in 1995 which is used for classification problems and along with the development of this classifier can solve regression problems. SVM has the task of separating 2 classes linearly and non-linearly. The dividing line in SVM is called Hyperplane.

Hyperplane can be said to be the best separator between two classes by measuring the hyperplane margin and looking for the maximum point. Margin is the distance between the hyperplane and the closest data point in each class. Support vector is a line that determines the closest data point. SVM has a way to handle classification for data that cannot be separated linearly. This method is called the kernel trick. The types of kernels that are often used when using SVM include: linear, polynomial, Radial Basis Function (RBF), sigmoid [21].

| Kernel Function | Expression |
|----------------|------------|
| Linear | $k(x,y) = xy$ |
| Polynomial | $k(x,y) = (x.y + 1)^d$ |
| RBF | $k(x,y) = \exp(-d|u-v|^2)$ |
| Sigmoid | $k(x,y) = \tan(x.y + b)$ |

2.6. Random Forest (RF)

Random Forest is a machine learning algorithm developed by Breiman for classification and regression problems. Random Forest has the task of making multiple random trees based on training data and randomly selecting the root node with the same distribution per sample. Prediction results were chosen based on a majority vote.

The Majority votes are calculation proportion of the number of classes in the model which decision trees are counted more. Calculation most classes are chosen to predict a case. If a majority vote > 0 then the result of the prediction will result in higher accuracy. If the majority vote < 0 then the result predictions will have a high error rate [22].

2.7. Artificial Neural Network (ANN)

ANN is a machine learning algorithm that adopted the human brain system in producing information. ANN is often referred to as a neuron which consists of several elements such as input layer, hidden layer, weight, and output layer. Input layer functions as a signal receiver, hidden layer to hide the desired prediction, weight functioned as a liaison between
the input layer and the hidden layer, the output layer functioned as another nerve.

2.8. Methodology

Fig. 1 shows the paths carried out in this study. The dataset used in this study is related to ASD screening obtained from the UCI Machine Learning Repository which consists of 3 datasets namely: AQ-Screening Child, AQ-Screening Adolescent, AQ-Screening Adult [16−18].

Before classifying ASD using a machine learning algorithm, all features in the dataset are selected using the Va method to obtain a subset of relevant features in the ASD screening process and transform non-numerical labels to numerical labels. After the feature subset is obtained and label have been transformed, the dataset is separated into training data and data testing using 10-fold cross validation. 10-fold cross validation is one of the validation technique that functions as divider the dataset into 10 parts. In every 1 part, there are 90% training data and 10% testing. The distribution of training data and testing is done randomly 10 times. The process of modelling is done by using training data by involving machine learning algorithms such as SVM, RF, and ANN.

Then the model is tested on testing data that is not involved in the training process. The results of the performance of each machine learning algorithm are compared and measured using performance metrics such as sensitivity and specificity.

Table 4. Confusion Matrix for ASD Classification

|                | Predicted ASD | Predicted No ASD |
|----------------|---------------|------------------|
| Actual ASD     | True Positive (TP) | False Negative (FN) |
| Actual No ASD  | False Positive (FP) | True Negative (TN) |

Table 5 shows the results of evaluating the average sensitivity using 10-fold cross validation from each classifier in the ASD classification on the three datasets. In the previous study [10], the lowest level of sensitivity with the C4.5
classifier occurred at Adult AQ: 80.95%. This happens because of ASD misclassification. Of the 189 instances which in the case of real positive ASD, 36 instances were classified incorrectly negatively ASD (FN). 36 instances in real cases should be positive ASD. In this study, classifiers who succeeded in increasing the sensitivity level were higher at AQ-Adult namely: RF with a sensitivity of 87.89%. This result was obtained because it successfully reduced ASD misclassification.

| Dataset       | Classifier | C4.5 (This study) | Ripper [10] (This study) | SVM (This study) | RF (This study) | ANN (This study) |
|---------------|------------|-------------------|--------------------------|-----------------|----------------|-----------------|
| AQ-Child      | 85.82      | 82.98             | 95.61                    | 91.21           | 92.78          |
| AQ-Adolescent | 90.48      | 87.30             | 92.74                    | 92.74           | 97.32          |
| AQ-Adult      | 80.95      | 82.54             | 86.81                    | 87.89           | 83.14          |

Of the 189 instances which in the case of real positive ASD, only 24 instances were incorrectly classified as negative ASD (FN). In the overall comparison of five classifiers in the classification of the three datasets, the classifier that produced the highest level of sensitivity was: ANN of 97.32% in the AQ-Adolescent dataset.

Table 6 shows the results of evaluating the average specificity using 10-fold cross validation from each classifier in the ASD classification on the three datasets. In the previous study, the lowest specificity level with the C4.5 classifier occurred at AQ-Adolescent: 78.00%. This happens because of ASD misclassification. Of the 41 instances which in real negative ASD cases, 9 instances were classified incorrectly positive ASD (FP). 9 instances in real cases should be negative ASD.

| Dataset       | Classifier | C4.5 (This study) | Ripper [10] (This study) | SVM (This study) | RF (This study) | ANN (This study) |
|---------------|------------|-------------------|--------------------------|-----------------|----------------|-----------------|
| AQ-Child      | 83.40      | 80.10             | 91.85                    | 91.12           | 91.67          |
| AQ-Adolescent | 78.00      | 80.50             | 86.33                    | 83.81           | 73.89          |
| AQ-Adult      | 92.00      | 91.18             | 92.41                    | 92.53           | 94.27          |

In this study, the classifier that succeeded in increasing the level of specificity was higher at AQ-Adolescent, namely: SVM with a specificity of 86.33%. This result was obtained because it successfully reduced ASD misclassification. Of the 41 instances which in real negative ASD cases, only 6 instances were classified incorrectly positive ASD (FP). In the overall comparison of five classifiers in the classification of the three datasets, the classifier which produced the highest level of sensitivity was: ANN of 94.27% on the AQ-Adult Dataset.

Based on the results of this study, there are no specific factors or tuning parameters that have an impact on the performance results of the ASD classification model better than previous studies [10]. We only use the default parameters for ANN and Random Forest in improving the ASD classification performance. Whereas SVM used in ASD classification only selects kernel types, namely: linear.

4. Conclusion

This study used the feature Va selection method and three machine learning algorithms for the classification of 3 datasets (AQ-child, AQ-adolescent, AQ-adult): SVM, RF, and ANN. The experimental results show that the three classifiers could increase the level of sensitivity and specificity in previous study [10]. Specifically, the comparison with previous study [10] on AQ-Adult using RF succeeded in producing a higher sensitivity of 87.89% while the level of specificity of AQ-Adolescent could be increased using Linear SVM of 86.33%. For further research, the performance of the classification can still be improved using a classifier and other feature selection methods.

References

1. H. V Ratajczak, Theoretical aspects of autism: Causes — A review, J. Immunotoxicol. 8 (2011) 68-79.
2. C. Plauche and P. Care, Identification and evaluation of children with autism spectrum, Am. Acad. Pediat. 120 (2007) 1183-1215.
3. C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, and A. Pickles, Autism from 2 to 9 years of age, Arch. Gen. Psychiatry. 63 (2006) 694–701.
4. CDC, Screening and Diagnosis of Autism Spectrum Disorder. [Online]. Available: https://www.cdc.gov/ncbddd/autism/screening.html. [Accessed: 10-Apr-2019].
5. C. W. Rishel, C. Greeno, S. C. Marcus, M. K. Shear, and C. Anderson, Use of the child behavior checklist as a diagnostic screening tool in community mental health, Res. Soc. Work Pract. 15 (2005) 195–203.
6. P. A. DA Krug, JR Arick, Autism screening instrument for educational planning-ASIEP 2. J. Psychoeduc. Assess. 16 (1998) 280-285.
7. C. Allison, B. Auyeung, and S. Baron-cohen, Toward brief ‘red flags’ for autism screening: the short autism spectrum quotient and the short quantitative, J. Am. Acad. Child. Adolesc. Psychiatry. 51 (2012) 202–212.e7.
8. National Institute for Health and Care Excellence, Autism spectrum disorder in adults: diagnosis and management, 2012. [Online]. Available: https://www.nice.org.uk/guidance/eq142.
9. Autism Research Centre, Downloadable Tests, 2012. [Online]. Available: https://www.autismscreensresearchcentre.com/arc_tests.
10. F. Thabah, F. Kamalov, and K. Rajab, A new computational intelligence approach to detect autistic features for autism screening, International J. Med. Inf. 117 (2018) 112–124.
11. J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall, Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning, Trans. Psychiatr. 5 (2015) e514.
12. M. J. Maenner, M. Yezurin-Allsopp, K. N. Van Braun, D. L. Christensen,
13. F. Zhang, P. Savadjiev, W. Cui, Y. Song, Y. Rathi, B. Tunç, D. Parker, T. Kapur, R. T. Schultz, N. Makris, R. Verma, and L. J. O’Donnell, *Whole brain white matter connectivity analysis using machine learning: An application to autism*, Neuroimage, 172 (2017) 826–837.

14. N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, *Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study*, Comp. Biol. Med. 109 (2019) 101-111.

15. C. C. Wu, W. C. Yeh, W. D. Hsu, M. M. Islam, P. A. (Alex) Nguyen, T. N. Poly, Y. C. Wang, H. C. Yang, and Y. C. (Jack) Li, *Prediction of fatty liver disease using machine learning algorithms*, Comp. Biol. Med. 170 (2019) 23–29.

16. F. Thabtah, *Autistic Spectrum Disorder Screening Data for Children Data Set*, 2017. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++. [Accessed: 10-Apr-2019].

17. F. Thabtah, *Autistic Spectrum Disorder Screening Data for Adolescent Data Set*, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescents++. [Accessed: 10-Apr-2019].