Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

RELEVANT TEST SET USING FEATURE SELECTION ALGORITHM FOR EARLY DETECTION OF DYSLEXIA

S. N. W. Shamsuddin*, N. S. F. N. Mat and M. Makhtar

Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, Terengganu, Malaysia

Published online: 10 November 2017

ABSTRACT

The objective of feature selection is to find the most relevant features for classification. Thus, the dimensionality of the information will be reduced and may improve classification's accuracy. This paper proposed a minimum set of relevant questions that can be used for early detection of dyslexia. In this research, we investigated and proposed a feature selection algorithm that is correlation based feature selection (CFS) and generate classification models based on five different classifiers namely Bayes Net, Simple Logistic and Decision Table. This paper used dataset collected from a computer based screening test developed consists of 50 questions. The result shows that the new set of question suggested from the feature selection algorithm was significantly achieved 100% accuracy of classification and less time was taken for conducting screening test among students.

Keywords: feature selection; dyslexic children; computer based screening test.

Author Correspondence, e-mail: syadiah@unisza.edu.my

doi: http://dx.doi.org/10.4314/jfas.v9i6s.66

1. INTRODUCTION

Generally, dyslexia screening tests can be defined as psychological assessments that which identify with consistent application of some criteria [1]. Learning support officers or teachers

can be trained to screen children for dyslexia through paper-based screening test or computer-based screening test. Research in [2] explained those screening tests are designed to be used on very large numbers of individuals. It is also to narrow down the group of individuals who might need a more complete test for possible dyslexia.

Screening test can therefore be implemented in the form of questionnaires referring to the symptoms of dyslexia whereby present challenges that dyslexic normally struggles. The probability of the disease will be sort out through the performing tests, examinations or other procedures of screening which is the presumptive detection of unrecognized disease. After that, further action will be taken by psychologist with positive or suspicious result that needs further diagnosis and necessary treatment [3]. Therefore, the efficiency screening test to be used should be well-established by means of a prospective validation study whereby should be carried out in the absence of intervention.

Computer-based screening test as said in [3] is more precise in measurement. A computer-based screening tool should be designed in such a way that is more attractive, efficient, fun and interesting so as to motivate and promote positive feeling of the user. Singleton agreed that there are various advantages of computer-based assessments over conventional assessments including being reportedly more efficient and cost effective to administer [1].

The use of computer-based in screening dyslexic helped in formulating an appropriate solution to overcome the limitation of traditional method. More important, the application that is both efficient and effective can give precise result of screening. From our study, children have to answer number of questions before they can be diagnosed as dyslexia. This process takes a lot of time and requires a person to assist the screening process.

This study targets to evaluate the effect of feature selection methods towards early detection of dyslexic children using computer based screening test developed in previous study called i-Dyslex. The purpose of this paper is to find the most relevant questions that can be used for early detection of dyslexic children. Thus, the question can be reduced and less time is required for screening test. We proposed correlation based feature selection (CFS) as a feature selection algorithm available in Weka.

Attributes or variables are also known as features that describe the nature of the knowledge

that are valuable for machine learning. Data mining is part of Knowledge Data Discovery (KDD). In KDD, features are the main components that contribute to the machine learning. However, not all features in the data set are significant for predictive modelling. Thus, by selecting the most relevant features for data mining will give a great improvement on the classification accuracy.

1.1. Computer Based Screening Test (i-Dyslex)

The assessment activities structure design for this research consist of five different modules that are Module Reading ("Membaca"), Spelling ("Mengeja"), Sorting ("Menyusun"), Hearing ("Mendengar") and IQ ("Berfikir"). Fig. 1 shows total of 50 questions that are available from 5 different modules of i-Dyslex.

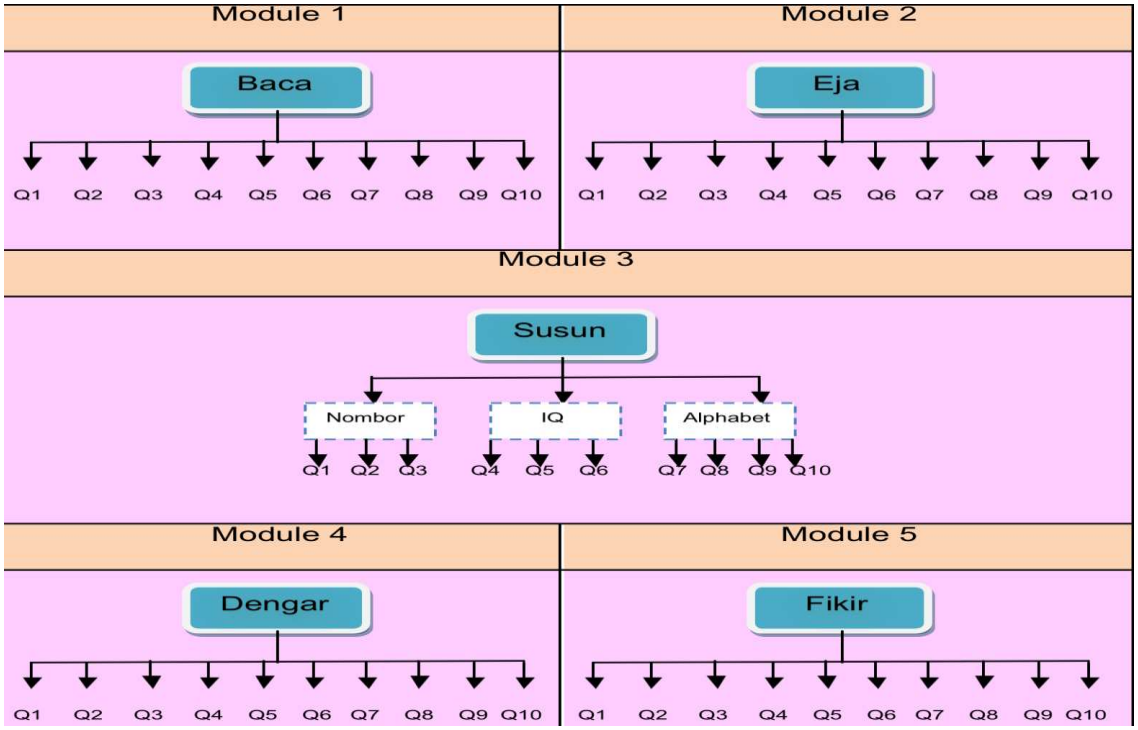

Fig.1. I-Dyslex modules

1.2. Module Reading

One of the important assessment is by screening the reading ability. Studies have shown that reading difficulties is one of the symptoms of dyslexia. Research in [4] showed that dyslexics have difficulties in their phonological decoding process, which causes phonological deficit [4-6, 20]. When accurate and fluent word reading develops very incompletely, it is proof that dyslexia face reading disability and it is due to the symptoms in dyslexia.

This module consists of word recognition activities which are focused on phonics, phonemes and reading fluency. There are ten questions with 20 words to test out where these words are the commonly confused by the dyslexic.

1.3. Module Sorting

Children that have difficulties in mathematics may also be dyslexic [7]. A report from previous researches showed that 60% of dyslexic students would typically face difficulties in learning mathematics [8].

The activities in this module applied the drags and drop approaches. It focused on remembering, understanding, manipulating number and number facts. The questions are related to numbers that regularly confused by the dyslexic. It showed further evidence that the difficulties faced by dyslexics in learning mathematics compared to normal students were due to their disabilities in mathematics skills.

1.4. Module Hearing

Reports from previous researches proved that dyslexics' reading difficulties were due to visual and auditory deficits [4, 9]. Auditory Processing Disorder is widely diagnosed in the USA and Australia with more reports from the UK and elsewhere [10]. This clearly shows that problems regarding auditory deficit is serious.

This module proposed in this research is to test hearing abilities among dyslexic children. It focused on selecting the correct answer as instructed by the sound.

1.5. Module Spelling

Studies have found that dyslexics have difficulties in identifying phonemes which occur during the spelling process due to exchanging of letters such as the letters ‘b-d’, ‘u-n’, ‘m- w’, ‘g-q’, ‘p-q’ and ‘b-p’ [11]. The main feature of dyslexia is a problem with word decoding, where it impacts the development of reading fluency and spelling performance [12].

This module focused on spelling the correct words as instructed by the sound. It emphasized on spelling abilities that consist of phonemes method.

1.6. Module IQ

One of the characteristics of dyslexia is memory loss among dyslexics [13-14]. Previous studies have found that individuals with dyslexia may suffer from memory loss which is related to and categorized as neurological deficit [9].

This module will test IQ skills among students. The module task examined the short term memory of dyslexic children by matching the objects and rapid memorizing the numbers.

2. METHODOLOGY

Feature selection is a technique to identify the most relevant features or attributes, which are used to generate predictive models on a training data set [15]. By using a raw data set (with no feature selection), the model will have to learn from all the features available. For data sets that have hundreds of features, the accuracy of the models may be lower because most of the features have no relationship to target classes and the accuracy is improved when the feature selection algorithms are applied [16-17]. It is because the model learns better about the data using the relevant attributes selected using a feature selection algorithm, while irrelevant features do not enter noise anymore during the learning stage [18-19].

The aim of feature selection is to find relevant features that have the most discriminating information from the original feature set. Since there are number of questions in the screening test, we want to determine the most substantial questions for early detection of dyslexic children. This paper aims to investigate a feature selection algorithm which is correlation based Feature Selection (CFS) and built classification models based on 5 different classifiers namely ZeroR, MultiClassClassifer, J48, Bagging and Bayes Net.

Following are the functions used for attribute evaluation (feature selection) within this research:

- Correlation Based Feature Selection (CFS)-Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
- Classifier Subset Evaluator-Evaluates attribute subsets on training data or a separate hold out testing set.
- Consistency Subset Evaluator-Evaluates the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes.

All the attributes were searched using these algorithms:

- BestFirst-Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.

- Genetic Search-Performs a search using the simple genetic algorithm.
- PaGreedy Step Wise-Performs a greedy forward or backward search through the space of attribute subsets.

3. RESULTS AND DISCUSSION

The classification models were generated using Weka with 10-fold cross validation for all feature selection algorithm and classifiers. Feature selection was used to find sets of attributes that are highly correlated with the target classes. From the experiments, the most significant features were comes from CFS with 10 relevant features. The features were selected form modules hearing (3 questions), spelling (3 questions), reading (1 question), IQ (1 question) and sorting (2 questions). The output generated for relevant features can be seen as Fig. 2 using CFS Subset Evaluator.

From the Table 1, the highest accuracy is 100% using CFS as a feature selection algorithm and Bayes Net as the classifier. The CFS returns 10 significant features meaning that we can simplify the screening test using only 10 questions. The questions are not just the normal question. For the dyslexic children, the design and color of the questions itself contribute to the process of screening test. We already proposed our own design in previous study.

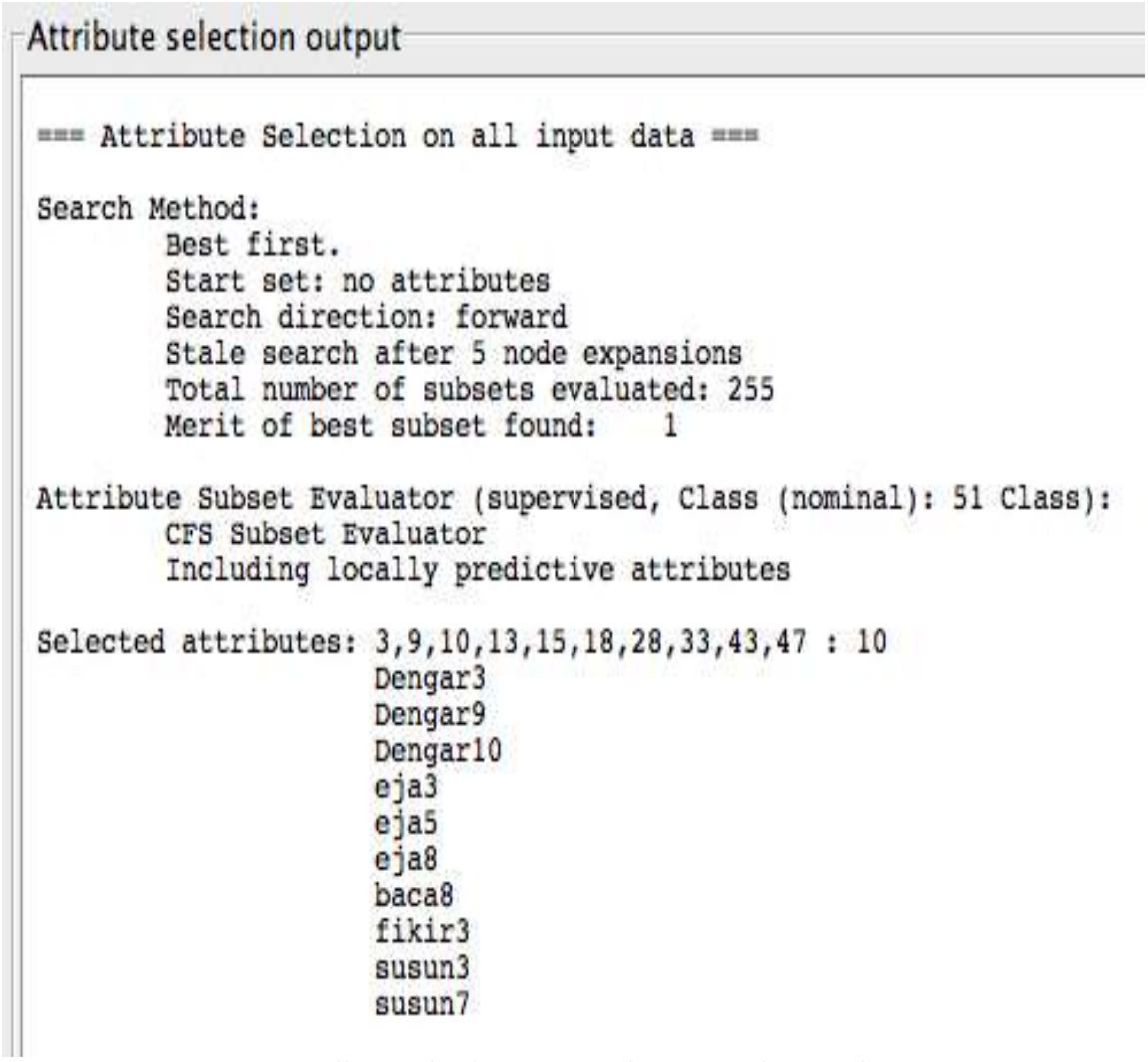

Fig.2. Attribute selection output using CFS subset evaluator

After we have all the relevant features, we apply those features to generate classification models. All the models were validated using 10-Folds Cross Validation. From the following result (see Table 1), it can be concluded that by applying CFS methods, the classification accuracy improved for Bayes Net.

Table 1. The accuracy of classification models with and without feature selection method

Classifiers	Accuracy for All Features	Accuracy Using 10 Selected Features (CFS)
ZeroR	85.71	85.71
MultiClassClassifier	96.91	95.91
J48	97.95	97.95
Bagging	97.95	97.95
Bayes Net	91.83	100.00

Classifier output

```
Bayes Network Classifier
not using ADTree
#attributes=11 #classindex=10
Network structure (nodes followed by parents)
Dengar3(1): Class
Dengar9(2): Class
Dengar10(2): Class
eja3(2): Class
eja5(2): Class
eja8(2): Class
baca8(2): Class
fikir3(1): Class
susun3(2): Class
susun7(1): Class
Class(2):
LogScore Bayes: -176.7929961297162
LogScore BDeu: -188.81829921842495
LogScore MDL: -199.972871141118
LogScore ENTROPY: -170.78421890528833
LogScore AIC: -185.7842189052883


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          49               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0188
Root mean squared error                  0.0799
Relative absolute error                  7.3185 %
Root relative squared error             22.6392 %
Total Number of Instances               49

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        no
                 1         0          1         1         1          1        yes
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b   <-- classified as
  7  0 |  a = no
  0 42 |  b = yes
```

Fig.3. The performance of Bayes Net classification model using 10 relevant features

Based on the results (see Table 1), it is proved that feature selection algorithm is able to improve the prediction accuracy and at the same time producing lower error rates. Furthermore, it also improved the overall performance of the classifiers by reducing the time taken to complete the prediction. In the long run, it is beneficial to the school and teachers who are going to handle the computer based screening test with less time consuming using minimum set of questions.

Following is the minimum relevant set of questions from the 5 modules that can be used to determine whether a student is dyslexic or not.

1. Module Hearing

   Hearing the words island (“pulau”), ask (“tanya”) and quiet (“sunyi”). The print screen of the test is as Fig. 4 to Fig. 6.

2. Module Spelling

   Spelling the words anchor (“sauh”), pole (“tiang”) and voice (“suara”). The print screen of the test is as Fig. 7.

3. Module Reading

   Reading the both words of pillow (“bantal”) and scrub (“sental”). The print screen of the test is as Fig. 8.

4. Module Thinking

   Select the best icon from the sound given. The print screen of the test is as Fig. 9.

5. Module Sorting

   Sort the numbers and alphabet in ascending and descending order. The print screen of the test is as Fig. 10 and Fig. 11.

Fig.4. Listening module of word “island”

Fig.5. Listening module of word “ask”

Fig.6. Listening module of word "quiet"

Fig.7. Spelling module of words "anchor", "pole" and "voice"

Fig.8. Reading module of words "pillow" and "scrub"

Fig.9. Thinking module of a sound

Fig.10. Sorting module of a set of numbers in ascending order

Fig.11. Sorting module of a set of alphabet in descending order

4. CONCLUSION

In this paper, feature selection algorithms have been reviewed. The results provided proved that the dataset with feature selection gives higher classification accuracy. The classifier is able to achieve 100% accuracy produced by the Correlation based Feature Selection (CFS) method with the Bayes Net classifier. Overall, this study recommends the use of feature reduction algorithms in the context of selection questions for computer based screening test in order to improve accuracy and performance of classification of dyslexic children. Feature selection techniques show that more information is not always good in machine learning applications. As a conclusion, the developed screening tool can be a good alternative for early detection of dyslexic children.

5. ACKNOWLEDGEMENTS

This work is partially supported by UniSZA (Grant Code. UniSZA/2015/DPU(66)).

6. REFERENCES

[1] Brookes G, Ng V, Lim BH, Tan WP, Lukito N. The computerised-based Lucid Rapid Dyslexia Screening for the identification of children at risk of dyslexia: A Singapore study. Educational and Child Psychology, 2011, 28(2):33-51

[2] Hazawawi N A, Huang L P, Hisham S.Online reading assessment for Malaysian young adults with dyslexia.In5th International Conference on Information and Communication Technology for the Muslim World, 2014, pp. 1-6

[3] Ekhsan H M, Ahmad S Z, Halim S A, Hamid J N, Mansor N H. The implementation of interactive multimedia in early screening of dyslexia. In IEEE International Conference on Innovation Management and Technology Research, 2012, pp. 566-569

[4] Aziz F A, Husni H, Jamaludin Z. Translating interaction design guidelines for dyslexic children's reading application. In World Congress on Engineering 2013, pp. 1-4

[5] Aguilar-Alonso Á, Moreno-González V. Neuropsychological differences between samples of dyslexic and reader children by means of NEPSY. Anuario de Psicología, 2012, 42(1):35-50

[6] Mohtaram S, Pee N C, Sibgatullah A S. Mobile dyslexia screening test: A new approach through multiple deficit model mobile game to screen developmental dyslexia children. In Malaysia University Conference Engineering Technology, 2014,pp. 10

[7] Fuchs L S, Fuchs D, Stuebing K, Fletcher J M, Hamlett C L, Lambert W. Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? Journal of Educational Psychology, 2008, 100(1):30-47

[8] Ubaidullah N H, Samsudin K, Hamid J, Khan S B.Supporting dyslexic children in learning multiplication facts with a software-based scaffolding: A Malaysian experience.In Asian Conference on Education Official Conference, 2011, pp. 106-117

[9] Ismail R, Jaafar A. Interactive screen-based design for dyslexic children. In IEEE International Conference on User Science and Engineering, 2011, pp. 168-171

[10] Sirimanna T. Auditory processing disorder.In L. Peer, & G. Reid (Eds.), Special educational needs: A guide for inclusive practice. London: SAGE Publications Ltd., 2016

[11] Mohammad W M. Dyslexia in the aspect of Malay language spelling. International Journal of Academic Research in Business and Social Sciences, 2012, 2(1):308-314

[12] Snowling M J. Early identification and interventions for dyslexia: A contemporary view. Journal of Research in Special Educational Needs, 2013, 13(1):7-14

[13] Ahmad S Z, Jinon N I, Rosmani A F. MathLexic: An assistive multimedia mathematical learning aid for dyslexia children. In IEEE Business Engineering and Industrial Applications Colloquium, 2013, pp. 390-394

[14]Perera H, Shiratuddin M F, Wong K W. Review of the role of modern computational technologies in the detection of dyslexia. In K. Kim, & N. Joukov (Eds.), Information science and applications. Singapore: Springer, 2016, pp. 1465-1475

[15]Chandrashekar G, Sahin F. A survey on feature selection methods. Computers and Electrical Engineering, 2014, 40(1):16-28

[16]Nafis S, Makhtar M, Awang M K, Rahman M N, Deris M M. Feature selections and classification model for customer churn. Journal of Theoretical and Applied Information Technology, 2015, 75(3):356-365

[17]Neagu D, Craciun M, Chaudhry Q, Price N. Knowledge representation for versatile hybrid intelligent processing applied in predictive toxicology. In S. Wong, & C. S. Li (Eds.), Life science data mining. Singapore: World Scientific, 2006, pp. 213-238

[18]Ladha L, Deepa T. Feature selection methods and algorithms. International Journal on Computer Science and Engineering, 2011, 3(5):1787-1497

[19]Ramaswami M, Bhaskaran R. A study on feature selection techniques in educational data mining.Journal of Computing, 2009, 1(1):7-11

[20]Rello L, Williams K, Ali A, White N C, Bigham J P. Dytective: Towards detecting dyslexia across languages using an online game. In 13th ACM Web for All Conference, 2016, pp. 1-4

How to cite this article:
Shamsuddin S N W, Mat N S F N, Makhtar M. Relevant test set using feature selection algorithm for early detection of dyslexia. J. Fundam. Appl. Sci., 2017, *9(6S), 886-899.*