Comparative Study for Prediction of Dyscalculia Using Smo and Naïve Bayes Classifier

Computer Science

KEYWORDS : SVM, Naïve Bayes, dyscalculia, classification

Ms. Sampada Margaj Department of Computer Science, University of Mumbai, India

Dr. Seema Purohit Department of Mathematics, Kirti College Mumbai, India

ABSTRACT

Aim of this work is to study the two classification methods, Sequential Minimal Optimization (SMO) and Naïve Bayes, for the prediction of Dyscalculia in school-age children. Prediction of any of the categories of learning disability is not an easy task. Same is the case of dyscalculia. Detail knowledge of the subject is mandatory in accurate prediction of dyscalculia in any child. A sooner the detection faster we can overcome it which will help the child for bright future. The above mentioned classifiers gives us satisfactory results. This study will also reflect on determining the best classification method for our specific domain.

INTRODUCTION

In our country there has not been much research and work in the field of learning disabilities. Also there isn't enough awareness or rehabilitation measures available for children with learning problems. Learning disability is broadly classified as dyslexia, dysgraphia and Dyscalculia. In this work we have focused on dyscalculia which is nothing but the inability of children for performing mathematical operations. Since international research reports that 5 - 8% of school-age children experience difficulties that interfere with their acquisition of mathematical concepts or procedures [Geary et al.,2004][Fuchs et al., 2002] and that an average of 3.6 – 6.5% have severe difficulties with acquiring numeracy and mathematics [Lewis et al., 1994] an increasing interest has been shown towards the subject by international professionals.

Thus this study aims at targeting the children who are under tremendous pressure due to their bad performance in the school examinations. We have considered schools in and around Mumbai, which do not have computer facilities and other amenities unlike the private schools, where implementation and detection of dyscalculia is much easier. This paper evaluates performance of two classification algorithm Sequential Minimal optimization and Naïve Bayes. Also compares which classifier gives more accurate results.

Author [Zan Huang et al., 2004] has compared the Naïve Bayes, Logistic function, RBF Network, Decision Table, SMO function algorithms, and the results showed that the Naïve Bayes algorithm outperforms with 86% accuracy in 0 second, this comparative study was applied to determine the most effective techniques that are capable for the detection of heart valve disease with a high accuracy. A comparative study of corporate credit rating analysis using support vector machines and back propagation neural network were analysed [Zan Huang et al., 2004]. In this study the performances of three models were compared with the Naive Bayes classifier, tree augmented Naive Bayes, the SVM, C4.5 and the nearest neighbour classifier and the obtained results demonstrated that the proposed models could significantly improve the performance of the naive Bayes classifier, yet at the same time maintain its simple structure[Taheri et al., 2013].Julie M., determined the relevance of various data preprocessing methods in classification it was done along with dimensionality reduction for the long list of attributes. The results obtained from this study was illustrated that the data pre- processing method gives good results in prediction system and can be used to improve the performance of classifiers [Julie M. et al., 2011].

SIGNIFICANCE OF THE STUDY

In our country there has not been much research and work in the field of dyscalculia. As it is always be confused with mathematical difficulty. Also there isn't enough awareness or rehabilitation measures available for children with learning problems. Thus this study aims at targeting the children who are under tremendous pressure due to their bad performance in the school examinations. We have considered schools in and around Mumbai, which do not have computer facilities and other amenities unlike the private schools.

Data Representation and Pre Processing

In this study we used 237 real world datasets from schools in and around Mumbai. The population is from primary section and all sample falls under same age group. The final sheet is arranged with names and the scores of all the subtests which are our attributes in this case. Below Table1 has the list of attributes with their descriptions based on which we are going to predict the student is affecting with dyscalculia or not. For final counseling the result based on these attributes, child's earlier history and general observations by parents and teachers are also considered.

TABLE I. LIST OF ATTRIBUTES

Sr.No.	Attribute	Description
1	DSR	Difficulty with Shape Recognition
2	DSD	Difficulty with Size Discrimination
3	DNA	Difficulty with Number Arrangements
4	DGS	Difficulty with Grouping Sets
5	DPV	Difficulty with Place Values
6	DNC	Difficulty with Numeric Calculations
7	DVA	Difficulty with Verbal Analysis
8	DCC	Difficulty with Counting Concepts

METHODOLOGY USED

In this paper, two different classifiers namely, Sequential Minimal Optimization (SMO) and Naïve Bayes Classifier are used for efficiency comparison of prediction of dyscalculia.

SEQUENTIAL MINIMAL OPTIMIZATION

SMO is basically a new form of SVM (Support Vector Machine). Because SMO spends most of its time evaluating the decision function, rather than performing Quadratic Programming, it can exploit data sets which contain a substantial number of zero elements. SMO does particularly well for sparse data sets, with either binary or non-binary input data [John C. et al., 2000]. In this study the data sets which we have collected is matches with the same structure

NAÏVE BAYES

Naïve Bayes is fast, easy to implement with the simple structure, and effective. It is basically useful for high di-

Volume : 5 | Issue : 8 | August 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

mensional data as estimating the probability of each feature independently is a challenged [Taheri et al., 2013]. Wu. Et al (2008) is also listed Naïve Bayes as one of the 10 top algorithms in data mining. In Naive Bayes (NB), features are conditionally independent given the class. It means that each feature has the class as an only parent. The efficiency of NB is proved in many different areas in real world applications such as medical diagnosis, recommender systems, email filtering, web page perfecting and fraud detection [Kononenkon et al., 2001][Crawford et al., 2002][Miyahara et al., 2000].

MEASURES USED FOR PERFORMANCE EVALUATION

There are several different measures are used but we have considered whoever are appropriate for our dataset. Using these measures the efficiency of classifiers are evaluated.

Classification Accuracy

Classification results could have an error rate and it may fail to classify correctly. Classification accuracy can be calculated as follows. Accuracy = (Instances Correctly Classified / Total Number of Instances)*100 % [Tian-Shyug Lee et al., 2002]

Mean Absolute Error

It is the average of difference between predicted and actual value in all test cases. The formula for calculating MAE is given in equation shown below: MAE = (|a1 – c1| + |a2 – c2| + … +|an – cn|) / n

Here, "a◎ is the actual output and "c◎ is the expected output [Tian-Shyug Lee et al., 2002].

Root Mean Square Error

It is used to measure differences between values actually observed and the values predicted by the model. It is calculated by taking the square root of the mean square error as shown in equation given below:

RMSE = [√ ((a1 – c1)2 + (a2 – c2)2+ … + (an – cn)2)] / n.

Here, "a◎ is the actual output and c is the expected output. The mean-squared error is the commonly used measure for numeric prediction [Tian-Shyug Lee et al., 2002].

Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. [Tian-Shyug Lee et al., 2002]

RESULTS AND DISCUSSION

In our study for the prediction of Dyscalculia and evaluation of the performances of both Sequential Minimal Optimization and Naïve Bayes we have used Waikato Environment for Knowledge Analysis (Weka).It is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. Here we have checked the performance using the Training set itself and using different Cross Validation and Percentage Split methods. The class (Dyscalculic and Non-Dyscalculic) is obtained by considering the values of all the specified attributes.

Performance of Sequential Minimal Optimization (SMO) Classifier

The overall evaluation summary of Sequential Minimal Optimization (SMO) Classifier using training set and different cross validation methods is given in Table XIV. The classification summary of SMO Classifier for different percentage split is given in Table XV. The confusion matrix for each different test mode is given in Table XVI to Table XXV. The chart showing the performance of SMO Classifier with respect to Correctly Classified Instances and Classification Accuracy with different type of test modes are depicted in Fig. 4, Fig. 5 and Fig. 6. SMO classifier gives 99% for training data sets. On average, SMO Classifier gives around 95% of accuracy in prediction of Dyscalculia.

Table XIV

SMO Classifier Overall Evaluation Summary

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model(sec)
Training Set	236	1	99.58%	0.9817	0.0042	0.065	0.07
5 Fold CV	226	11	95.36%	0.7873	0.0464	0.2154	0.15
10 Fold CV	225	12	94.94%	0.7712	0.0506	0.225	0.06
15 Fold CV	226	11	95.36%	0.7873	0.0464	0.2154	0.03
20 Fold CV	225	12	94.94%	0.7712	0.0506	0.225	0.05
50 Fold CV	225	12	94.94%	0.7712	0.0506	0.225	0.05

Table XV

SMO Classifier Percentage Split Overall Evaluation Summary

Test Mode	Total Test Instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model(sec)
66% Split	81	77	4	95.06%	0.7722	0.0494	0.2222	0.05
33% Split	159	152	7	95.60%	0.7755	0.044	0.2098	0.07
75% Split	59	54	5	91.53%	0.6203	0.0847	0.2911	0.05
80% Split	47	46	1	97.87%	0.8773	0.0213	0.1459	0.05

Table XVI

Confusion Matrix - SMO on Training Dataset

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	205	0	205
Dyscalculic	1	31	32
Predicted (Total)	206	31	237

Table XVII

Confusion Matrix - SMO for 5 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	202	3	205
Dyscalculic	8	24	32
Predicted (Total)	210	27	237

Table XVIII

Confusion Matrix - SMO for 10 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	201	4	205
Dyscalculic	8	24	32
Predicted (Total)	209	28	237

Table XIX

Confusion Matrix - SMO for 15 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	202	3	205
Dyscalculic	8	24	32
Predicted (Total)	210	27	237

Table XX

Confusion Matrix - SMO for 20 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	201	4	205
Dyscalculic	8	24	32

Table XXI

Confusion Matrix - SMO for 50 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	201	4	205
Dyscalculic	8	24	32

Table XXII

Confusion Matrix - SMO for 66% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	69	3	72
Dyscalculic	1	8	9
Predicted (Total)	70	11	81

Table XXII

Confusion Matrix - SMO for 33% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	138	2	140
Dyscalculic	5	14	19
Predicted (Total)	143	16	159

Table XXIII

Confusion Matrix - SMO for 75% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	49	1	50
Dyscalculic	4	5	9
Predicted (Total)	53	6	59

Table XXIV

Confusion Matrix - SMO for 80% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	42	0	42
Dyscalculic	1	4	5
Predicted (Total)	43	4	47

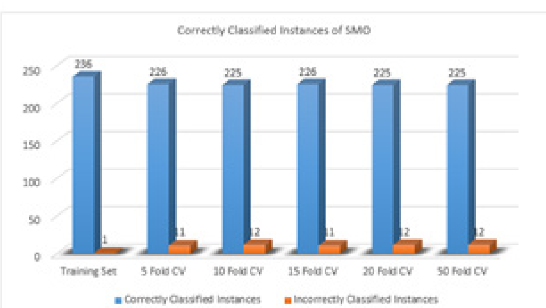

Fig 4. Correctly Classified Instances of SMO Classifier

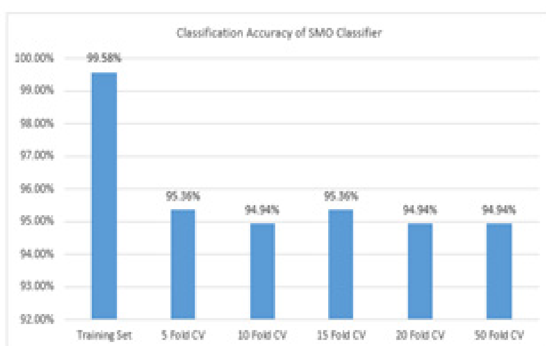

Fig.5 Classification Accuracy of SMO Classifier

Performance of Naïve Bayes Classifier

The overall evaluation summary of Naïve Bayes Classifier using training set and different cross validation methods is given in Table XXV. The classification summary of Naïve Bayes Classifier for different percentage split is given in Table XXVI. The confusion matrix for each different test mode is given in Table XXVII to Table XXXVII. The chart showing the performance of Naïve Bayes Classifier with respect to correctly classified instances and classification accuracy with different type of test modes are depicted in Fig. 7 and Fig. 8. Naïve bayes classifier gives 94.51% for training data sets. On average, Naïve Bayes Classifier gives around 94% of accuracy in prediction of Dyscalculia.

Comparison of SMO and Naïve Bayes Classifiers

We can now clearly compare the two classifiers, Sequential Minimal Optimization (SMO) and Naive Bayes with the following graphical charts. These are depicted in the Fig.9, Fig.10 and Fig.11 in terms of classification accuracy and correctly classified instances. The overall ranking is done based on the classification accuracy, correctly classified instances, MAE and RMSE values and other statistics found using Training set results, Percentage Splits and Cross Validation Techniques. Based on these results, it is observed that SMO performs better than Naïve Bayes classifier.

Table XXV

Naïve Bayes Classifier Overall Evaluation Summary

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model(sec)
Training Set	224	13	94.51%	0.7901	0.0619	0.1939	0.00
5 Fold CV	224	13	94.51%	0.7901	0.0619	0.1939	0.00
10 Fold CV	223	14	94.09%	0.7712	0.0833	0.2315	0.00
15 Fold CV	223	14	94.09%	0.7712	0.0831	0.2293	0.00
20 Fold CV	222	15	93.67%	0.7578	0.0841	0.2339	0.00
50 Fold CV	223	14	94.09%	0.7712	0.0842	0.2322	0.00

Table XXVI

Naïve bayes Classifier Percentage Split Overall Evaluation Summary

Test Mode	Total Test Instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model(sec)
33% Split	159	148	11	93.08%	0.7048	0.0741	0.2191	0.01
66% Split	81	76	5	93.83%	0.7486	0.0873	0.2336	0
75% Split	59	57	2	96.61%	0.8798	0.0612	0.1704	0
80% Split	47	46	1	97.87%	0.8972	0.0439	0.1446	0

Table XXVII

Confusion Matrix - NB on Training Dataset

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	194	11	205
Dyscalculic	2	30	32
Predicted (Total)	196	41	237

Table XXVIII

Confusion Matrix - NB For 5 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	194	11	205
Dyscalculic	2	30	32
Predicted (Total)	196	41	237

Volume : 5 | Issue : 8 | August 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

Table XXIX

Confusion Matrix - NB For 10 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	194	11	205
Dyscalculic	3	29	32
Predicted (Total)	197	40	237

Table XXX

Confusion Matrix - NB For 15 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	194	11	205
Dyscalculic	3	29	32
Predicted (Total)	197	40	237

Table XXXI

Confusion Matrix - NB For 20 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	193	12	205
Dyscalculic	3	29	32
Predicted (Total)	196	41	237

Table XXXII

Confusion Matrix - NB For 50 Fold CV

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	194	11	205
Dyscalculic	3	29	32
Predicted (Total)	197	40	237

Table XXXIV

Confusion Matrix - NB for 66% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	67	5	72
Dyscalculic	0	9	9
Predicted (Total)	67	14	81

Table XXV

Confusion Matrix - NB for 33% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	132	3	135
Dyscalculic	8	16	24
Predicted (Total)	140	19	159

Table XXXVI

Confusion Matrix - NB for 75% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	48	2	50
Dyscalculic	0	9	9
Predicted (Total)	48	11	59

Table XXXVII

Confusion Matrix - NB for 80% Split

Class	Non Dyscalculic	Dyscalculic	Actual (Total)
Non Dyscalculic	41	1	42
Dyscalculic	0	5	5
Predicted (Total)	41	6	47

Fig. 9: Comparison of correctly classified instances between NB and SMO

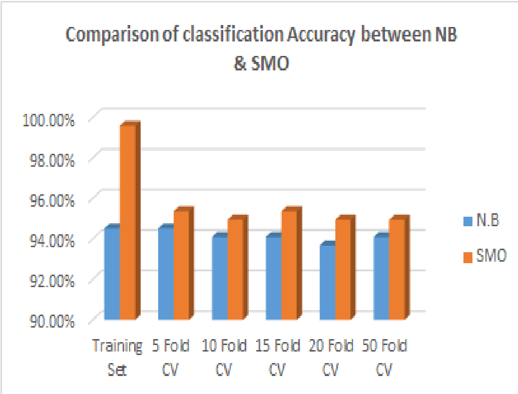

Fig. 10: Comparison of classification Accuracy between NB and SMO

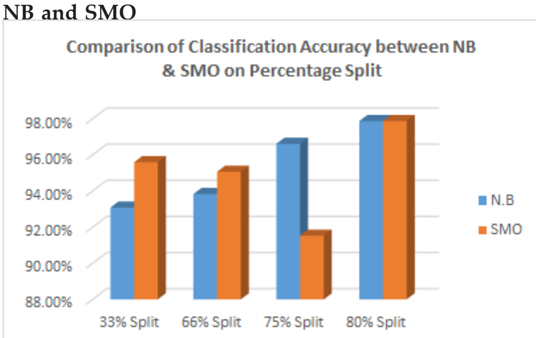

Figure 11: Comparison of classification Accuracy between NB and SMO on percentage split

CONCLUSIONS

Past studies in different domain have shown a very good contribution of data mining classifiers. In this study we analyzed the efficiency of two different classifiers namely, Sequential Minimal Optimization (SMO) and Naïve Bayes for the prediction of Dyscalculia among primary school children. Comparison of both the classifiers has been done by considering different measures of performance evaluation. After all the research and study of the different aspects in this work, it is observed that Sequential Minimal Optimization (SMO) performs better than the Naïve Bayes for the prediction of Dyscalculia. In future study can be done to get more accurate results.

REFERENCES:

[1] Durgadevi. A[1], Prof. Saravanapriya.K[2] ,[1,2]PG and Research Department of Computer Applications Sacred Heart College, Tirupattur, Vellore ."Comparative Study of Data Mining Classification Algorithms in Heart Disease Prediction",International Journal of Recent Research in Mathematics Computer Science and Information Technology Vol. 2, Issue 2, pp: (22-26), Month: October 2015 – March 2016, Available at: www.paperpublications.org

[2] Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," Decision Support Systems (Elsevier) 37 (2004) pp. 543– 558.

[3] Sona Taheri⊚, Musa Mammadov⊚,⊚⊚ ⊚ Centre for Informatics and Applied Optimization School of Science, Information Technology and Engineering, University of Ballarat, Victoria 3353, Australia "LEARNING THE NAIVE BAYES CLASSIFIER WITH OPTIMIZATION MODELS"sonataheri@students.ballarat.edu.au,sona.taheri@unisa.edu.au Victoria Research Laboratory National ICT Australia, Victoria 3010, Australia e-mail: m.mammadov@ballarat.edu.au,musa.mammadov@nicta.com.au, Int. J. Appl. Math. Comput. Sci., 2013, Vol. 23, No. 4, 787–795 DOI: 10.2478/amcs-2013-0059.

[4] Julie M. David et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 , 2011, 829-835,"Prediction of Learning Disabilities in School Age Children using SVM and Decision Tree".

[5] Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen, "Credit scoring using the hybrid neural discriminant technique," Expert Systems with Applications (Elsevier) 23 (2002), pp. 245– 254

[6] John C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," 2000

[7] Wu, X., Vipin Kumar, J., Quinlan, R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, J., Ng, A., Liu, B., Yu, P. S., Zhou, Z., Steinbach, M., Hand, D. J. and Steinberg, D. (2008). Top 10 algorithms in data mining, Knowledge and Information Systems 14: 1–37. Yatsko,

[8] S. O. N. A. Taheri, "Learning the naive bayes classifier with optimization models," vol. 23, no. 4, pp. 787–795, 2013.

[9] Geary, D. C. (2004). Mathematics and learning disabilities. Journal of Learning Disabilities, 37, 4-15.

[10] Fuchs, L.S., & Fuchs, D. (2002). Mathematical problem-solving profiles of students with mathematical disabilities with or without comorbid reading disabilities. Journal of Learning Disabilities, 35, 563-573.

[11] Lewis, C., Hitch, G., & Walker, P. (1994). The prevalence of specific arithmetic and specific reading difficulties in 9- and 10-year old boys and girls. Journal of Pupil Psychology and Psychiatry, 35, 283-292

[12] Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective, Artificial Intelli- gence in Medicine 23: 89–109.

[13] Crawford, E., Kay, J. and Eric, M. (2002). The intelligent email sorter, Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia, pp. 83–90

[14] Miyahara, K. and Pazzani, M.J. (2000). Collaborative filtering with the simple Bayesian classifier, Proceedings of the 6th Pacific Rim International Conference on Artificial Intelli- gence, Melbourne, Australia, pp. 679–689.