Executive Summary of the Minor Research Project submission as part of Eleventh Plan

Prediction of Learning Disabilities in School Aged Children – A Rough Set Approach

By

Dr. SABU M.K.

(Principal Investigator)

Associate Professor, Department of Computer Applications

M.E.S. College Marampally

(Affiliated to Mahatma Gandhi University, Kottayam)

Aluva, Ernakulam, Kerala-683107

UGC Ref No. MRP(S)-344 /12-13 /KLMG061/ UGC- SWRO

Submitted to

University Grants Commission

EXECUTIVE SUMMARY

Learning Disability (LD) is a neurological disorder that affects a child's brain. It causes trouble in learning and using certain skills such as reading, writing, listening and speaking. This project mainly focuses on the effect of feature selection in the prediction of learning disability in school aged children. The process of feature selection reduces the dimensionality of the data and enables learning algorithms to operate more rapidly and effectively. In this work, a rough set based Proportional Rough Set (PRS) relevance method of feature selection is implemented as a new preprocessing approach with an objective to improve the performance of two popular supervised learning algorithms viz. Multilayer Perceptron (MLP) with back propagation and Sequential Minimal Optimization (SMO) for Support Vectors. In this approach, the features are ranked according to their significance in the data domain by constructing various reducts from the dataset. From the ranked feature set, significant features are selected by specifying a pre-defined size for the number of features to be selected or by specifying a minimum threshold value for the significance/priority of features and then fed to the learning algorithm. In order to automate the feature selection process by defining a threshold for the significance/priority of features, in this work, Selection by Threshold (ST) strategy is used. Comparative analysis performed on the LD dataset consisting of the signs and symptoms of Learning Disabilities in school aged children show that, there are some non-relevant features in the dataset and the proposed method is effective and efficient for removing these redundant features without affecting the classification performance. In the case of both the classifiers, the time taken to build the model is also promising when we consider the feature selection process.

With an objective to predict more accurately the presence of LD in school-aged children with reduced number of symptoms, a novel hybrid feature selection approach is also proposed by integrating the PRS relevance feature ranking process with a modified backward feature elimination algorithm. The approach follows a ranking of the symptoms of LD according to their importance in the data domain. Then by eliminating least significant features one by one and evaluating the feature subset at each stage of the process, an optimal feature subset is generated. A combination of four relevant symptoms is identified from the LD dataset through this approach which gives the same classification accuracy compared to the whole sixteen features. It implies that these four features were worthwhile to be taken close attention by the physicians or teachers handling LD when they conduct the diagnosis.

For comparative analysis, the feature selection algorithm is combined with two state-of-the-art filter based feature ranking techniques viz. information gain and gain ratio. The experimental results shows the proposed feature selection approach outperforms the other two in terms of the data reduction and time taken to build the learning model. Also, the proposed method eliminates all the redundant attributes efficiently from the LD dataset without sacrificing the classification performance.

Methodology

The objective of this work is to predict more accurately the presence of Learning Disability (LD) in school-aged children with reduced number of symptoms. For this purpose, a novel hybrid feature selection approach is proposed by integrating a popular Rough Set based feature ranking process with a modified backward feature elimination algorithm. Before feature selection begins, each feature is evaluated independently with respect to the class to identify its significance in the data domain. Features are then ranked in the decreasing order of their significance. To calculate the significance and to rank various features of the LD dataset, in this work, PRS relevance approach is used. Then by eliminating least significant features one by one and evaluating the feature subset at each stage of the process, an optimal feature subset is generated. The proposed method eliminates all the redundant attributes efficiently from the LD dataset without sacrificing the classification performance.

Summary

Learning disability is a neurological condition that affects a child's brain and results from impairments in one or more processes related to perceiving, thinking, remembering or learning. Feature selection is an important preprocessing technique for effective data analysis where only a subset from the original set of features is preserved after eliminating noisy and redundant or irrelevant features. Feature selection task allows reducing computational cost and improving accuracy of the data analysis process. Rough Set theory based attribute reduction provides filter based approach by which knowledge may be extracted from the domain in a concise way by preserving the semantics of the original data as far as possible. Rough Set properties such as indiscernibility, reduct, core, discernibility matrix etc. are applied effectively in attribute reduction. This results in several novel algorithms in feature selection.

In this work two different RST based feature selection approaches are discussed. In the first approach, a popular Rough Set based feature selection approach namely PRS relevance approach is implemented to identify the most significant features. The method seems to be very useful in predicting the presence of LD in a cost effective way. The main contribution of the study is a method of assigning priorities to various symptoms of the LD dataset based on the general characteristics of the data alone. Each symptoms priority values reflects its relative importance to predict LD among the various cases. The experimental results reveal the need of feature selection in classification to improve the performance of classification such as speed of learning and predictive accuracy.

Second approach follows the method of assigning priorities to various symptoms of the LD dataset and ranking these symptoms in the decreasing order of their significance. Then least significant features are eliminated one by one by considering its involvement in predicting the learning disability. The experimental result reveals the significance of rough set theory in ordering various symptoms of LD to achieve an optimal feature subset. With the help of the proposed method, redundant attributes can be removed efficiently from the LD dataset without sacrificing the classification performance. The proposed method of feature selection was also shown to perform well against the state-of-the-art feature selection strategies viz. information gain and gain ratio.

Contribution to Society

This project mainly focuses on the effect of feature selection in the prediction of learning disability in school aged children. The influence of the selected symptoms in the prediction of learning disability is studied. With an objective to predict more accurately the presence of LD in school-aged children with reduced number of symptoms, a novel hybrid feature selection approach is also proposed. The approach follows a ranking of the symptoms of LD according to their importance in the data domain. Then by eliminating least significant features one by one and evaluating the feature subset at each stage of the process, an optimal feature subset is generated. A combination of four relevant symptoms is identified from the LD dataset through this approach which gives the same classification accuracy compared to the whole sixteen features. It implies that these four features were worthwhile to be taken close attention by the physicians or teachers handling LD when they conduct the diagnosis.