A NOVEL HYBRID FEATURE SELECTION APPROACH FOR THE PREDICTION OF LEARNING DISABILITIES IN SCHOOL-AGED CHILDREN

Sabu M.K
Department of Computer Applications, M.E.S College, Marampally, Aluva, Kerala, India

ABSTRACT

Feature selection is one of the most fundamental steps in machine learning. It is closely related to dimensionality reduction. A commonly used approach in feature selection is ranking the individual features according to some criteria and then search for an optimal feature subset based on an evaluation criterion to test the optimality. 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. The process of feature ranking follows a method of calculating the significance or priority of each symptoms of LD as per their contribution in representing the knowledge contained in the dataset. Each symptoms significance or priority values reflect its relative importance to predict LD among the various cases. 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. For comparative analysis and to establish the importance of rough set theory in feature selection, the backward feature elimination algorithm is combined with two state-of-the-art filter based feature ranking techniques viz. information gain and gain ratio. The experimental results show the proposed feature selection approach outperforms the other two in terms of the data reduction. Also, the proposed method eliminates all the redundant attributes efficiently from the LD dataset without sacrificing the classification performance.

KEYWORDS

Rough Set Theory, Data Mining, Feature Selection, Learning Disability, Reduct, Information gain, Gain ratio.

1. INTRODUCTION

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. A possible approach to build computer assisted systems to handle LD is: collect a large repository of data consisting of the signs and symptoms of LD, design data mining algorithms to identify the significant symptoms of LD and build classification models based on the collected data to classify new unseen cases. Feature selection is an important data mining task which can be effectively utilized to develop knowledge based tools in LD prediction. Feature selection process not only reduces the dimensionality of the dataset by preserving the significant features but also improves the generalization ability of the learning algorithms.

Data mining, especially feature selection is an exemplary field of application where Rough Set Theory (RST) has demonstrated its usefulness. RST can be utilized in this area as a tool to discover data dependencies and reduce the number of attributes of a dataset without considering

DOI : 10.5121/ijaia.2015.6205

any prior knowledge and using only the information contained within the dataset alone [2]. In this work, RST is employed as a feature selection tool to select most significant features which will improve the diagnostic accuracy by Support Vector Machine (SVM). For this purpose, a popular Rough Set based feature ranking algorithm is implemented to rank various symptoms of the LD dataset. Then by integrating this feature ranking technique with backward feature elimination [15], a new hybrid feature selection technique is proposed. 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.

The rest of the paper is organized as follows. In section 2, a review of feature selection procedures are described. An overview of information gain, gain ratio and rough set based feature ranking processes are given in section 3. A brief description on Learning Disability dataset is presented in Section 4. Section 5 introduces the proposed approach of feature selection process. Experimental analysis and results comparison of the proposed feature selection approach are highlighted in Section 6. A discussion of the experimental results is given in Section 7. The last section concludes this research work.

2. FEATURE SELECTION

The Feature selection is a search process that selects a subset of significant features from a data domain for building efficient learning models. Feature selection is closely related to dimensionality reduction. Most of the dataset contain relevant as well as irrelevant and redundant features. Irrelevant and redundant features do not contribute anything to determine the target class and at the same time deteriorates the quality of the results of the intended data mining task. The process of eliminating these types of features from a dataset is referred to as feature selection. In a decision table, if a particular feature is highly correlated with decision feature, then it is relevant and if it is highly correlated with others, it is redundant. Hence the search for a good feature subset involves finding those features that are highly correlated with the decision feature but uncorrelated with each other [1]. Feature selection process reduces the dimensionality of the dataset and the goal of dimensionality reduction is to map a set of observations from a high dimensional space M into a low dimensional space m ($m<<M$) by preserving the semantics of the original high dimensional dataset. Let $I = (U, A)$ be an information system (dataset), where $U = \{x_1, x_2, \ldots, x_n\}$ be the set of objects and $A = \{a_1, a_2, \ldots, a_M\}$ be the set of attributes used to characterize each object in I. Hence each object x_i in the information system can be represented as an M dimension vector $[a_1(x_i), a_2(x_i), \ldots, a_M(x_i)]$, where $a_j(x_i)$ yields the j^{th} ($j = 1, 2, 3, \ldots, M$) attribute value of the i^{th} ($i = 1, 2, 3, \ldots., n$) data object. Dimensionality reduction techniques transform the given dataset I of size $n \times M$ into a new low dimensional dataset Y of size $n \times m$.
While constructing a feature selection method, two different factors namely search strategies and evaluating measures [2] are to be considered. Commonly used search strategies are complete or exhaustive [3], heuristic [4] and random [5][6]. In general feature selection methods are based on some exhaustive approaches which are quite impractical in many cases, especially for high dimensional datasets, due to the high computational cost involved in the searching process [25]. To reduce this complexity, as an alternate solution strategy, heuristic or random search methods are employed in modern feature selection algorithms.

Based on the procedures used for evaluating the scalability of the generated subset, heuristic or random search methods are further classified into three – classifier specific or wrapper methods [7][8][9][10][11], classifier independent or filter methods [12][13][14] and hybrid models [15] which combines both filter and wrapper approach to achieve better classification performance. In a classifier specific feature selection method, the quality of the selected features is evaluated with the help of a learning algorithm and the corresponding classification accuracy is determined. If it satisfies the desired accuracy, the selected feature subset is considered as optimal; otherwise it is modified and the process is repeated for a better one. The process of feature selection using wrapper (classifier specific) approach is depicted in Figure 1. Even though the wrapper method may produce better results, it is computationally expensive and can encounter problems while dealing with huge dataset.

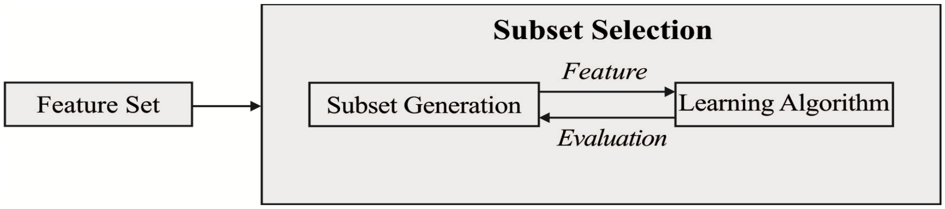

Figure 1. Wrapper approach to feature selection

In the case of classifier independent method, to evaluate the significance of selected features one or more of classifier independent measures such as inter class distance [12], mutual information [16][17] and dependence measure [13][18] are employed. In this approach, the process of feature selection is treated as a completely independent pre-processing operation. As an outcome of this pre-processing, irrelevant/noisy attributes are filtered. All filter based methods use heuristics based on general characteristics of the data rather than a learning algorithm to evaluate the optimality of feature subsets. As a result, filter methods are generally much faster than wrapper methods. Since this method does not depend on any particular learning algorithm, it is more suitable in managing high dimensionality of the data.

In the case of hybrid model, as a first step, features are ranked using some distance criterion or similarity measure and then with the help of a wrapper model an optimal feature subset is generated. The method usually starts with an initial subset of features heuristically selected beforehand. Then features are added (forward selection) or removed (backward elimination) iteratively until an optimal feature subset is obtained.

3. FEATURE RANKING METHODS

Feature ranking is one of the most fundamental steps in hybrid feature selection approaches. This section reviews two conventional feature ranking methods followed by a detailed discussion of a rough set based feature ranking approach.

3.1. Information Gain

Information gain is a well known measure used for attribute selection which is based on the concept of entropy. This attribute selection measure provides a ranking for the set of conditional attributes A of the decision table $T= \{U, A, d\}$. The attribute having the higher information gain is the one which minimizes the information needed to classify the tuples in the resulting partitions of T and reflects the least randomness in these partitions [27][28][29].

Let $a \in A$, be a conditional attribute and d be the decision attribute of the decision table T. Also, assume that both are discrete attributes taking values $\{a_1, a_2,..., a_l\}$ and $\{d_1, d_2,...,d_k\}$ respectively. Then the entropy of d, $H(d)$, is defined as:

$$H(d) = -\sum_{i=1}^{k} P(d = d_i)\log_2 P(d = d_i) \qquad 1$$

where $P(d=d_i)$ is the probability that an arbitrary tuple in T belongs to the i^{th} class d_i.

The information gain of a given attribute $a\in A$ with respect to the decision attribute d is the reduction in uncertainty about the value of d when the value of a is known. The uncertainty about d given the value of a is measured by the conditional entropy $H(d/a)$. Then the information gain of a with respect to d, $IG(a)$ is defined as:

$$IG(a) = H(d) - H(d/a) \qquad 2$$

The conditional entropy of d given a, $H(d/a)$ is:

$$H\left(d\middle/a\right) = -\sum_{j=1}^{l} P\left(a = a_j\right)H(d/a = a_j) \qquad 3$$

3.2. Gain Ratio

A limitation of using information gain measure in attribute selection is it prefers to select attributes having a large number of values over the attributes with fewer values even though the latter is more informative [28]. In order to overcome this limitation, an extension of information gain, known as gain ratio, is used in attribute ranking. To obtain the gain ratio of a conditional attribute $a\in A$, a kind of normalization is performed to the $IG(a)$ using a split information value, say $Splt_a(U)$, of the set of objects U with respect to a [28]. $Splt_a(U)$ is computed by splitting U into l partitions corresponding to the l values of the conditional attribute a. Hence, the gain ratio of the attribute $a\in A$, $GR(a)$ is define as:

$$GR(a) = \frac{IG(a)}{Splt_a(U)} \qquad 4$$

The split information value, $Split_a(U)$, is given by:

$$Splt_a(U) = -\sum_{j=1}^{l} \frac{\left|\{a = a_j\}\right|}{|U|} log_2\left(\frac{\left|\{a = a_j\}\right|}{|U|}\right) \qquad 5$$

With the help of this gain ratio, it is possible to rank the conditional attributes of the decision table T.

3.3. Rough Set based Attribute Ranking – The Proportional Rough Set Relevance Method

RST introduced by Z. Pawlak in the early 1980's provides a methodology for data analysis based on the approximation of imprecise or vague concepts in information systems [30][31]. The philosophy of RST follows a mathematical approach suitable to intelligent data analysis and data mining. The rough set approach considers data acquired from experience as explicit facts about reality, which is represented as an information system I= (U, A) where U is a finite non empty set of objects and A is a finite non empty set of primitive attributes. The classification of I according to various attributes in A represent the explicit knowledge obtained from the collected data. If

$P \subseteq A$ and $P \neq \phi$, then $\cap P$ gives the P-basic knowledge about U in I. The partition induced by P is denoted as IND(P). The set of all indiscernible (similar) objects belongs to $IND(P)$ is called an elementary set or a category and forms a basic granule (atom) of the knowledge P. The indiscernibility relation generated in this way is the mathematical basis of RST [18].

In RST, a dataset is always termed as a decision table. A decision table presents some basic facts about the universe along with the decisions (actions) taken by the experts based on the given facts. An important issue in data analysis is whether the complete set of attributes given in the decision table are necessary to define the knowledge involved in the equivalence class structure induced by the decision attribute. This problem arises in many real life situations and referred to as knowledge reduction [18]. A real fact is, the whole knowledge available in the collected dataset is not always necessary to define our interested categories represented in the dataset. This motivates the need for efficient automated feature selection processes in the area of data mining. With the help of RST, we can eliminate all superfluous attributes from the dataset preserving only the indispensable attributes [18]. In reduction of knowledge, the basic roles played by two fundamental concepts in RST are *reduct* and *core*. A reduct is a subset of the set of attributes which by itself can fully characterize the knowledge in the given decision table. A reduct keeps essential information of the original decision table. In a decision table there may exist more than one reduct. The set of attributes which is common to all reducts is called the core [18]. The core may be thought of as the set of indispensable attributes which cannot be eliminated while reducing the knowledge involved in the decision table. Elimination of a core attribute from the dataset causes collapse of the category structure given by the original decision table. In the following section, a popular reduct based feature ranking approach known as PRS relevance method [19] is presented. In this method, the ranking is done with the help of relevance of each attribute/feature calculated by considering its frequency of occurrence in various reducts generated from the dataset.

The Proportional Rough Set (PRS) relevance method is an effective Rough Set based approach for attribute ranking proposed by Maria Salamó and López-Sánchez [19]. The concept of reducts is used as the basic idea for the implementation of this approach. The same idea is also used by Li and Cercone to rank the decision rules generated from a rule mining algorithm [20][21][22][23]. There exist multiple reduct for a dataset. Each reduct is a representative of the original data. Most data mining operations require only a single reduct for decision making purposes. But selecting any one reduct leads to the elimination of representative information contained in all other reducts. The main idea behind this reduct based feature ranking approach is the following: the more frequent a conditional attribute appears in the reducts and the more relevant will be the attribute. Hence the number of times an attribute appears in all reducts and the total number of reducts determines the significance (priority) of each attribute in representing the knowledge contained in the dataset. This idea is used for measuring the significance of various features in PRS relevance feature ranking approach [19]. These priority values provide a ranking for the conditional features available in the dataset.

4. LEARNING DISABILITY DATASET

Learning disability (LD) is a neurological condition that affects the child's brain resulting in difficulty in learning and using certain skills such as reading, writing, listening, speaking and reasoning. Learning disabilities affect children both academically and socially and about 10% of children enrolled in schools are affected with this problem. With the right assessment and remediation, children with learning disabilities can learn successfully. As nature and symptoms of LD may vary from child to child, an early diagnosis of LD is critically difficult. Identifying students with LD and assessing the nature and depth of LD is essential for helping them to get around LD. By integrating soft computing techniques with machine learning it is possible to increase the diagnostic accuracy of LD prediction. The proposed methodology of feature selection is helpful to identify the presence and degree of LD in any child at an early stage.

To apply the proposed methodology, a dataset consisting of the signs and symptoms of the learning disabilities in school age children is selected. It is collected from various sources which include a child care clinic providing assistance for handling learning disability in children and three different schools conducting similar studies. This dataset is helpful to determine the existence of LD in a suspected child. It is selected with a view to provide tools for researchers and physicians handling learning disabilities to analyze the data and to facilitate the decision making process.

The dataset contains 500 student records with 16 conditional attributes as signs and symptoms of LD and the existence of LD in a child as decision attribute. Various signs and symptoms collected includes the information regarding whether the child has any difficulty in reading (DR), any difficulty with spelling (DS), any difficulty with handwriting (DH) and so on. There are no missing values or inconsistency exists in the dataset. Table 1 gives a portion of the dataset used for the experiment. In this table *t* represents the attribute value true and *f* represents the attribute value false. Table 2 gives key used for representing the symptoms and its abbreviations.

Table 1**.** Learning Disability (LD) dataset

DR	DS	DH	DWE	DBA	DHA	DA	ED	DM	LM	DSS	DNS	DLL	DLS	STL	RG	LD
t	t	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t
t	t	f	t	f	t	f	t	t	t	t	f	t	f	t	f	t
t	t	f	t	f	t	f	t	t	t	t	f	t	f	t	f	t
t	t	f	f	f	f	t	t	t	t	f	f	f	f	f	f	t
f	f	f	t	t	f	f	f	f	f	f	f	f	f	f	f	f
f	f	f	f	f	f	t	t	t	f	f	f	f	f	f	f	f
t	t	t	t	t	f	t	t	t	t	f	f	f	f	t	f	t
f	f	f	f	f	f	f	f	f	t	f	f	t	f	t	f	f
t	t	f	t	f	f	f	f	f	f	f	f	t	f	f	f	t
t	t	f	t	f	t	t	t	t	t	t	f	t	t	t	f	t
t	t	f	t	f	t	t	t	t	t	t	f	f	f	t	f	t
f	f	f	t	f	f	t	f	f	f	f	f	f	f	f	f	f
t	t	f	t	f	t	f	t	f	t	t	f	t	f	t	f	t
f	f	f	f	f	t	f	t	f	f	f	f	f	f	f	f	f

Table 2**.** Key used for representing the symptoms of LD

Key/ Abbreviations	**Symptoms**	**Key/ Abbreviations**	**Symptoms**
DR	Difficulty with Reading	LM	Lack of Motivation
DS	Difficulty with	DSS	Difficulty with Study
DH	Difficulty with Handwriting	DNS	Does Not like School
DWE	Difficulty with Written Expression	DLL	Difficulty in Learning a Language
DBA	Difficulty with Basic Arithmetic	DLS	Difficulty in Learning a Subject
DHA	Difficulty with Higher Arithmetic skills	STL	Is Slow To Learn
DA	Difficulty with Attention	RG	Repeated a Grade
ED	Easily Distracted	LD	Learning Disability
DM	Difficulty with		

5. Proposed Approach

The proposed method of feature selection follows a hybrid approach which utilizes the complementary strength of wrapper and filter approaches. 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 [26]. To calculate the significance and to rank various features of the LD dataset, in this work, PRS relevance approach is used. To explain the feature ranking process, consider a decision table $T = \{U, A, d\}$, where U is the non-empty finite set of objects called the universe, $A = \{a_{1,} a_{2, \dots,} a_n\}$ be the non-empty finite set of conditional attributes/features and d is the decision attribute. Let $\{r_1, r_{2,\dots,} r_p\}$ be the set of reducts generated from T. Then, for each conditional attribute $a_i \in A$, reduct based attribute priority/significance $\beta(a_i)$ is defined as [19][20][21]:

$$\beta(a_i) = \frac{\left|\{r_j | a_i \in r_j, j = 1,2,3,\dots,p\}\right|}{p}, i = 1,2,3,\dots,n \qquad 6$$

where the numerator of the Eq. 6 gives the occurrence frequency of the attribute a_i in various reducts.

From Eq. 6 it is clear that an attribute a not appearing in any of the reducts has priority value $\beta(a) = 0$. For an attribute a, which is a member of core of the decision table has a priority value $\beta(a) = 1$. For the remaining attributes the priority values are proportional to the number of reducts in which the attribute appear as a member. These reduct based priority values will provide a ranking for the considered features.

After ranking the features, search process start with all available features and successfully remove least significant features one by one (backward elimination) after evaluating the influence of this feature in the classification accuracy until the selected feature subset gives a better classification performance. When a certain feature is eliminated, if there is no change in the current best classification accuracy the considered feature is redundant. If the classification accuracy is increased as a result of elimination, the removed feature is considered as a feature with negative influence on the classification accuracy. In these two cases, the selected feature is permanently removed from the feature subset; otherwise it is retained. Feature evaluation starts by considering the classification accuracy obtained from all available features as the current best accuracy. The search terminates when no single attribute deletion contributes any improvement in the current best classification accuracy. At this stage, the remaining feature subset is considered as optimal. For classification, Sequential Minimal Optimization (SMO) algorithm using the polynomial kernel is used in this work. It is implemented through Weka data mining toolkit [24]. This algorithm is used for the prediction of LD because it is simple, easy to implement and generally faster. The proposed feature selection algorithm *FeaSel* is presented below. The algorithm accepts the ranked set of features obtained from the PRS relevance approach as input and generates an optimal feature subset consisting of the significant features as output. The overall feature selection process is represented in figure 2.

```
Algorithm FeaSel($F_n$, Y, n, $X_n$)
//$F_n$ ={$f_1$, $f_2$,...,$f_n$}– Set of features obtained from PRS relevance approach ranked in
  descending order of their significance.
//Y – class; n – total number of features.
// $X_n$ – The optimal feature subset.
{
    $X_n$=$F_n$;
```

```
        max_acc=acc(F_n,Y); //acc() returns the classification accuracy given by the classifier
        for (i=n to 1 step -1) do
          {
            F_n=F_n-{f_i};
            curr_acc=acc(F_n, Y);
            if (curr_acc==max_acc)
                X_n=F_n;
            else if (curr_acc>max_acc)
                {
                X_n=F_n;
                 max_acc=curr_acc;
                }
                else
                     X_n=F_n∪{f_i};
            F_n=X_n;
          }
      return(X_n, max_acc);
}
```

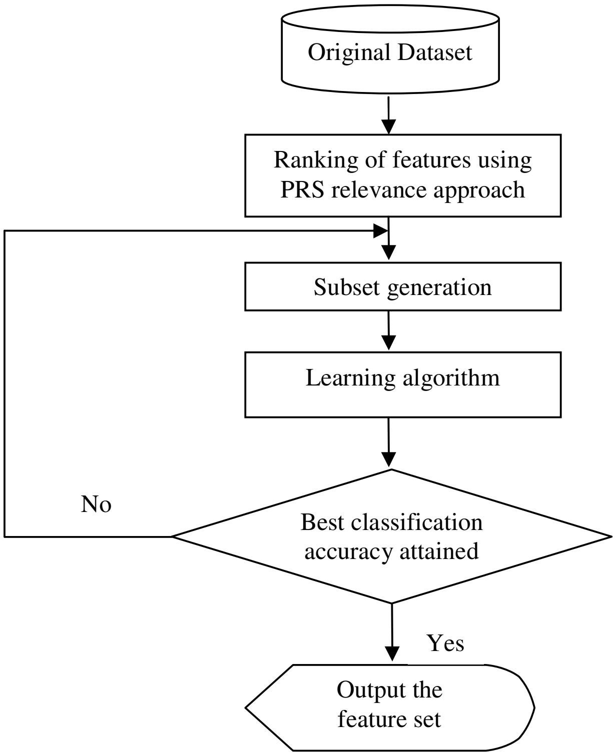
Original Dataset
Ranking of features using PRS relevance approach
Subset generation
Learning algorithm
No
Best classification accuracy attained
Yes
Output the feature set

Figure 2. Block diagram of the feature selection process

6. EXPERIMENTAL ANALYSIS AND RESULTS

In order to implement the PRS relevance approach to rank the features, as a first step of the process, various reducts are generated from the LD dataset. For this purpose, the discernibility matrix approach of Rough Sets Data Explorer software package ROSE2 is used which generates 63 reducts from the original LD dataset. Then frequencies of various features occurring in these reducts are computed. These frequencies are given in Table 3. Based on these frequencies and by applying Eq. 6, the priority/significance values of various features are calculated. Ranked features as per their significance are shown in Table 4.

Table 3. Frequencies of various symptoms in reducts

Feature	Frequency	Feature	Frequency
DR	63	DSS	18
DS	34	DNS	23
DWE	32	DHA	21
DBA	41	DH	16
DA	44	DLL	50
ED	63	DLS	27
DM	63	RG	36
LM	41	STL	27

Table 4. Symptoms with priority values

Rank	Feature	Significance	Rank	Feature	Significance
1	DR	1	9	DS	0.5397
2	ED	1	10	DWE	0.5079
3	DM	1	11	DLS	0.4286
4	DLL	0.7937	12	STL	0.4286
5	DA	0.6984	13	DNS	0.3651
6	LM	0.6508	14	DHA	0.3333
7	DBA	0.6508	15	DSS	0.2857
8	RG	0.5714	16	DH	0.2540

For feature selection using the proposed algorithm, the classification accuracy of the whole LD dataset with all available features is determined first. In the feature selection algorithm the construction of the best feature subset is mainly based on this value. Then, the set of features ranked using PRS relevance approach is given to the proposed feature selection algorithm *FeaSel*. Since the features are ranked in decreasing order of significance, features with lower ranks gets eliminated during initial stages. The algorithm starts with all features of LD and in the first iteration the algorithm selects lowest ranked feature DH as a test feature. Since there is no change occurs in the original classification accuracy while eliminating this feature, it is designated as redundant and hence it is permanently removed from the feature set. The same situation continues for the features DSS, DHA, DNS, STL, and DLS selected in order from right to left from the ranked feature set and hence all these features are removed from the feature set. But when selecting the next feature DWE, there is a reduction in the classification accuracy which signifies the dependence of this feature with the class attribute LD and hence this feature is retained in the feature set. The process is continued until all features are evaluated. The influence of various symptoms of LD in classification during the proposed feature selection process is depicted in figure 3.

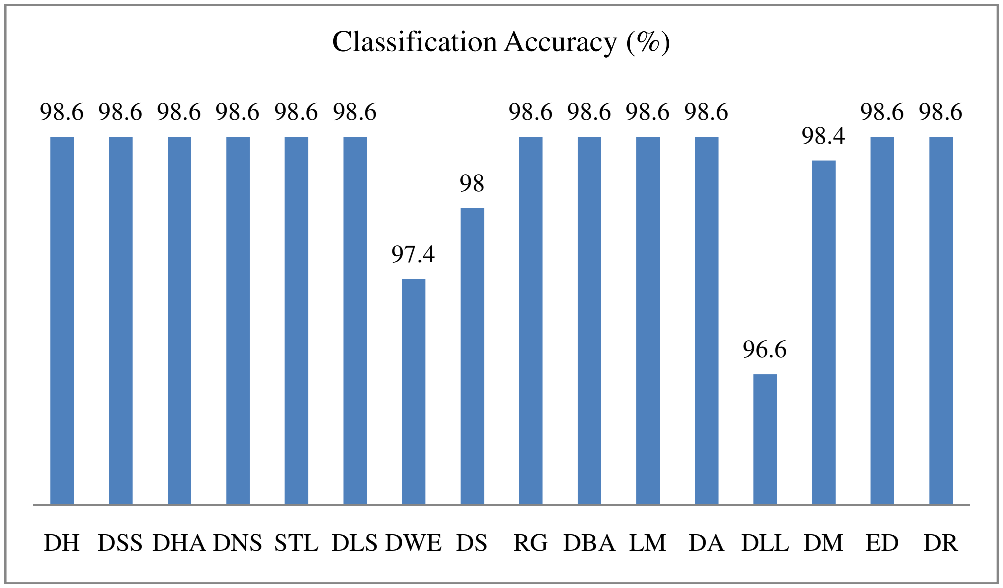

Figure 3. Influence of various symptoms in proposed approach

After evaluating all features of the LD dataset, the algorithm retains the set of features {DWE, DS, DLL, DM}. These four features are significant because all other features can be removed from the LD dataset without affecting the classification performance. Table 5 shows the results obtained from the classifier before and after the feature selection process. To determine the accuracy 10 fold cross validation is used.

Table 5. Classification results given by SMO

Various cases	Dataset prior to perform feature selection	Dataset reduced using the proposed approach
No. of features	16	4
Classification accuracy (%)	98.6	98.6
Time taken to build the model (Sec.)	0.11	0.01

6.1 Comparison of feature selection with Information Gain and Gain Ratio

The proposed feature selection with PRS relevance approach is compared with two similar hybrid feature selection approaches viz. feature selection with information gain and feature selection with gain ratio. For comparison the same LD dataset is used. In the first method, information gain is used as a measure to rank various symptoms of LD dataset. In the second method gain ratio is used for ranking the symptoms. The same modified backward feature selection algorithm, *FeaSel* is used for selecting the significant symptoms in both these methods. The symptoms ranked using information gain measure are given in Table 6 and the performance of various symptoms of LD in feature selection process is presented in figure 4. Similarly, Table 7 gives the symptoms ranked using gain ratio measure and figure 5 presents symptom's performance in classification. Using feature selection with information gain, it is possible to remove eleven features (symptoms) and using feature selection with gain ratio ten features can be eliminated without affecting the classification performance.

Table 6. Symptoms ranked using information gain

Rank	Feature	Significance	Rank	Feature	Significance
1	DR	0.65294	9	DNS	0.10687
2	DS	0.59359	10	DH	0.10169
3	DSS	0.34597	11	ED	0.09946
4	LM	0.27086	12	DHA	0.08743
5	DWE	0.2685	13	DLS	0.0626
6	STL	0.17331	14	DBA	0.02944
7	DM	0.14574	15	DA	0.00933
8	DLL	0.12301	16	RG	0.00473

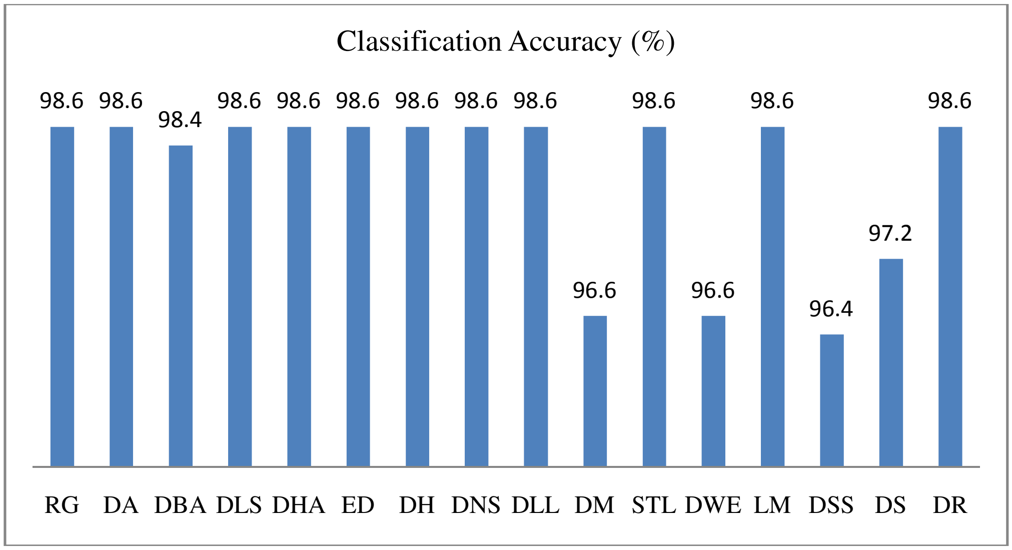

Figure 4. Influence of symptoms in feature selection with information gain

Table 7. Symptoms ranked using gain ratio

Rank	Feature	Significance	Rank	Feature	Significance
1	DR	0.70899	9	DH	0.14246
2	DS	0.62968	10	DLS	0.1317
3	DSS	0.34637	11	DLL	0.12422
4	DWE	0.28914	12	ED	0.09996
5	LM	0.272	13	DHA	0.09622
6	STL	0.17711	14	DBA	0.03837
7	DNS	0.16249	15	DA	0.00944
8	DM	0.14664	16	RG	0.00695

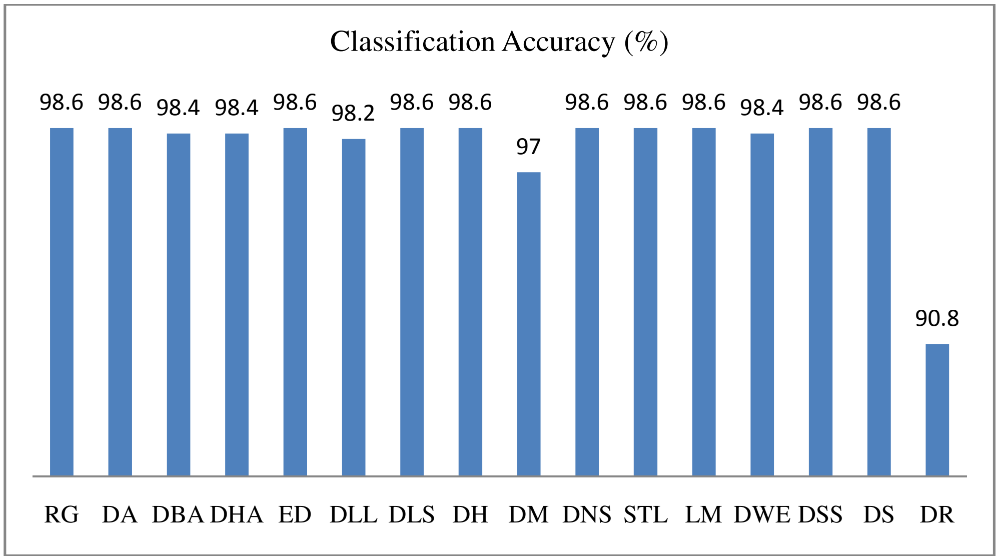

Figure 5. Influence of symptoms in feature selection with gain ratio

Table 8 shows the percentage of data reduction obtained during the proposed feature selection process, feature selection with information gain and feature selection with gain ratio.

Table 8. Comparison of feature selection for LD dataset

Various cases	Feature selection using the proposed approach	Feature selection with information gain	Feature selection using gain ratio
No. of features selected	4	5	6
Data reduction (%)	75	68.75	62.5

7. DISCUSSION

From the experimental results presented in Table 5 it is clear that, in the case of the proposed approach a 75% reduction in the dataset does not affect the classification accuracy. It follows that the original dataset contains about 75% redundant attributes and the feature selection approach presented is efficient in removing these redundant attributes without affecting the classification accuracy. From the presented results, it can be seen that when using the selected significant features for classification, the time taken to build the learning model is also greatly improved. This shows that in an information system there are some non-relevant features and identifying and removing these features will enable learning algorithms to operate faster. In other words, increasing the number of features in a dataset may not be always helpful to increase the classification performance of the data. Increasing the number of features progressively may result in reduction of classification rate after a peak. This is known as peaking phenomenon.

The comparison results presented in Table 8 shows that, in terms of data reduction, the proposed method of feature selection is efficient compared to the feature selection with information gain and feature selection with gain ratio.

8. CONCLUSION

In this paper, a novel hybrid feature selection approach is proposed to predict the Learning Disability in a cost effective way. The approach follows 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 reflect its relative importance to predict LD among the various cases. By ranking these symptoms in the decreasing order of their significance, 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 feature selection with information gain and feature selection with gain ratio.

ACKNOWLEDGEMENTS

This work was supported by University Grants Commission (UGC), New Delhi, India under the Minor Research Project program (MRP(S)-344/12-13/KLMG061/UGC-SWRO dated 29-03-13).

REFERENCES

[1] Richard Jensen (2005) Combining rough and fuzzy sets for feature selection, Ph.D thesis from Internet.

[2] Yumin Chen, Duoqian Miao & Ruizhi Wang, (2010) "A Rough Set approach to feature selection based on ant colony optimization", Pattern Recognition Letters, Vol. 31, pp. 226-233.

[3] Petr Somol, Pavel Pudil & Josef Kittler, (2004) "Fast Branch & Bound Algorithms for Optimal Feature Selection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 7, pp. 900-912.

[4] Ning Zhong, Juzhen Dong & Setsuo Ohsuga, (2001) "Using Rough Sets with heuristics for feature selection", Journal of Intelligence Information systems, Vol. 16, pp.199-214.

[5] Raymer M L,Punch W F, Goodman E D,Kuhn L A & Jain A K, (2000) "Dimensionality Reduction Using Genetic Algorithms", IEEE Trans. Evolutionary Computation, Vol.4, No.2, pp. 164-171.

[6] Carmen Lai, Marcel J.T. Reinders & Lodewyk Wessels, (2006) "Random subspace method for multivariate feature selection", Pattern Recognition letters, Vol. 27, pp. 1067-1076.

[7] Ron Kohavi & Dan Sommerfield, (1995) "Feature subset selection using the wrapper method: Over fitting and dynamic search space topology", Proceedings of the First International Conference on Knowledge Discovery and Data Mining, pp. 192-197.

[8] Isabelle Guyon, Jason Weston, Stephen Barnhill & Vladimir Vapnik, (2002) "Gene selection for cancer classification using support vector machines", Machine Learning, Kluwer Academic Publishers, Vol. 46, pp. 389-422.

[9] Neumann J, Schnörr C & Steidl G, (2005) "Combined SVM based feature selection and classification", Machine Learning, Vol.61, pp.129-150.

[10] Gasca E,Sanchez J S & Alonso R, (2006) "Eliminating redundancy and irrelevance using a new MLP based feature selection method", Pattern Recognition, Vol. 39, pp. 313-315.

[11] Zong-Xia Xie, Qing-Hua Hu & Da-Ren Yu, (2006) "Improved feature selection algorithm base on SVM and Correlation", LNCS, Vol. 3971, pp. 1373-1380.

[12] Kira K & Rendell L A, (1992) "The feature selection problem: Traditional methods and a new algorithm", Proceedings of the International conference AAAI-92, San Jose, CA, pp. 129-134.

[13] Mondrzejewski M, (1993) " Feature selection using Rough Set theory" , Proceedings of the European conference on Machine learning ECML'93, Springer-Verlag, pp. 213-226.

[14] Manoranjan Dash & Huan Liu, (2003) "Consistency based search in feature selection", Artificial Intelligence, Vpl.151, pp. 155-176.

[15] Swati Shilaskar & Ashok Ghatol. Article, (2013) “Dimensionality Reduction Techniques for Improved Diagnosis of Heart Disease”, International Journal of Computer Applications, Vol. 61, No. 5, pp. 1-8.

[16] Yao Y.Y, (2003) “Information-theoretic measures for knowledge discovery and data mining Entropy Measures, Maximum Entropy and Emerging Applications”, Springer Berlin. pp. 115-136.

[17] Miao D. Q & Hou, L, (2004) “A Comparison of Rough Set methods and representative learning algorithms”, Fundamenta Informaticae. Vol. 59, pp. 203-219.

[18] Pawlak Z, (1991) Rough Sets: Theoretical aspects of Reasoning about Data, Kluwer Academic Publishing, Dordrecht.

[19] Maria Salamo M & Lopez-Sanchez M, (2011). “Rough Set approaches to feature selection for Case-Based Reasoning Classifiers”, Pattern Recognition Letters, Vol. 32, pp. 280-292.

[20] Li J. & Cercone N, (2006) “ Discovering and Ranking Important Rules”, Proceedings of KDM Workshop, Waterloo, Canada.

[21] Li J, (2007) Rough Set Based Rule Evaluations and their Applications, Ph.D thesis from Internet.

[22] Shen Q. & Chouchoulas A, (2001) “Rough Set – Based Dimensionality Reduction for Supervised and Unsupervised Learning”, International Journal of Applied Mathematics and Computer Sciences, Vol. 11, No. 3, pp. 583-601.

[23] Jensen J (2005) Combining rough set and fuzzy sets for feature selection, Ph.D thesis from Internet.

[24] Ian H. Witten & Eibe Frank (2005) Data Mining – Practical Machine Learning Tools and Techniques. Elsevier.

[25] Alper U., Alper, M. & Ratna Babu C, (2011) “A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification”, Information Science, Vol. 181, pp. 4625-4641.

[26] Pablo Bermejo, Jose A. Gámez & Jose M. Puerta, (2011) “A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets”, Science Direct, Pattern Recognition Letters, Vol. 32, pp. 701-711.

[27] Jiawei Han, Micheline Kamber, (2006) Data Mining – Concepts and Techniques. Morgan Kaufman publishers, Elsevier.

[28] Sri Harsha Vege, (2012) Ensemble of Feature Selection Techniques for High Dimensional Data, Ph.D thesis from Internet.

[29] Margaret H. Dunham & S. Sridhar, (2006) Data Mining – Introductory and Advanced Topics, Pearson Education.

[30] Zdzislaw Pawlak. Rough sets and intelligent data analysis. Information Sciences, vol. 147, pp. 1 – 12, Elsevier, 2002.

[31] Paul S K and Skowron A(Eds.). Rough-Fuzzy Hybridization: A New Trend in Decision Making. Springer Verlag, Singapore, 1999.

Author

Sabu M K, received his Ph. D degree from Mahatma Gandhi University, Kottayam, Kerala, India in 2014. He is currently an Associate Professor and also the Head of the Department of Computer Applications in M.E.S College, Marampally, Aluva, Kerala. His research interests include data mining, rough set theory, machine learning and soft computing.