Effect of Feature Selection to Improve Accuracy and decrease execution time with predicating Learning Disabilities in school going children

Margaret Mary. T [1], Hanumanthappa. M [2]

[1]Research Scholar, Dept., of CS, Bharathiar University Coimbatore-641 046, and Assistant Professor in MCA Dept., Sambhram Academy of Management Studies, Bangalore-97
[2]Professor, Department of Computer Science & Applications, Bangalore University, Bangalore-56

ABSTRACT

Learning disability in school children is the representation of brain disorder which includes several disorders in which school going child faces the difficulties. The evaluation of learning disability is a crucial and important task in the field of educational field. This process can be accomplished by using data mining approaches. The efficiency of this approach is based on the feature selection while performing the prediction of the learning disabilities. In paper mainly aims on the efficient method of feature selection to improve the accuracy of prediction and classification in school going children. Feature selection is a process to collect the small subset of the features from huge dataset. 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 evaluation criterion to test the optimality. In the Wrapper model we use some predetermined learning algorithm to find out the relevant features and test them. It requires more computations, so if there are large numbers of features we prefer to filter. In this paper first we have used feature selection attribute algorithms Chi-square. Info Gain, and Gain Ratio to predict the relevant features. Then we have applied fast correlation base filter algorithm on given features. Later classification is done using KNN and SVM. Results showed reduction in computational cost and time and increase in predictive accuracy for the student model. The objective of this work is to predict the presence of Learning Disability (LD) in school-aged children more accurately and help them to develop a bright future according to his choice by predicting the success at the earliest.

Keywords: Data Mining, Learning Disabilities, feature selection, KNN and SVM.

1. INTRODUCTION

Learning disability is a crucial neurological disorder found in the school going children which causes the difficulty in learning and various skills i.e. speaking, reading, listening and writing etc. There are various approaches have been proposed to build a computer based system to handle the learning disability which follows collection of huge datasets which contains the symptoms and signs of the learning disability. These kinds of approaches use data mining techniques to build the efficient interface for prediction of learning disability and development of classification model. For improved prediction of the learning disability, feature selection is a key parameter. It is an important parameter which helps us to build the efficient model of the learning disability prediction. By using this process the dimension of the data is reduced and by preserving the key features of the data. It comes under pre-processing stage in machine learning methodology. This process of feature selection is classified into three subcategories (i) Wrapper method, (ii) Filtering and (iii) Embedding method.

Wrapper method is performed to by using predetermined learning techniques to find out the relevant features and to perform testing on them. Main drawback of this is that it is computation complexity is more when compared to others. Next method is embedding technique which combines the feature selection and classification stage together [12]. For this purpose, feature selection approach is proposed by integrating a popular SVM and KNN based feature ranking process with a probability. The approach follows a ranking of the symptoms of LD according to their importance in the data domain. 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.[12] The experimental results shows the success of the proposed method in removing redundant attributes efficiently from the LD dataset without sacrificing the classification performance.

2. DATA MINING

Data mining is an approach which dispense an intermixture of technique to identify a block of data or decision making knowledge in the database and eradicating these data in such a way that they can be put to use in decision support, forecasting and estimation [11]. Educational data mining concerns with developing methods for discovering knowledge from data that come from educational environment.[15] In this paper we used educational data mining to analyse learning behavior. The characteristics of learning disabilities are identified early in the child's development; a program of cognitive training can be used to give that child the right skills for normal and successful life. Data mining techniques are the result of long process of research and product development. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis [7]. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events of real world problems. Support Vector Machines, a powerful algorithm based on statistical. learning theory. It implements Support Vector Machines for classification, regression, and anomaly detection. It also provides the scalability and usability that are needed in production quality data mining system.[16] This paper introduces and analyses SVM and KNN supervised algorithm, which will help to understand the tuning, diagnostics & data preparation process and advantages of SVM in Data Mining. SVM can model complex, real-world problems such as text and image classification, hand-writing recognition, and bio-informatics and bio-sequence analysis [18]. "k-nearest neighbor" use small neighborhoods in the attribute space to determine classification and regression predictions. These predictions can be greatly skewed by redundant attributes. The remaining paper is organized as follows. Section 2 Learning disabilities and dataset; we then provide a brief explanation of Feature selection and step of feature selection in section 3. The next section i.e. section 4 overview of conventional describes the basics of KNN and SVM diagnostic technique used for handling this problem, followed by actual implementation methodology and contains the test result in Section5. Finally section 6 Discussion and the conclusion of the paper and feature analysis.

3. LEARNING DISABILITIES

In this section we discuss about the learning disability. Learning disability is a general term which denotes the various neurological disorders which can cause learning disorders. This can be classified into writing, speaking, reading, reasoning, listening and doing math etc. According to a survey by National Health Interview in 2004, it was examined the 8% children between 3-16 years of age had learning disability in the United States. It was found that Dyslexia is primarily learning disability which affects the writing, reading and speaking difficulties in the children. "Generally learning disability is a disorder of the neurological system which affects the process of receiving, storing the information in brain and responding to the query or information". It is not an intellectual disability. It occurs at all levels of intelligence, sub average, average, above average and highly gifted [6]. According to conventional techniques, a sever between intellectual ability and academic achievement has to exist in one or more of these academic areas: Oral expression, Listening comprehension, Writing expression, Basic Reading skills, Reading comprehension and Mathematical calculations.[22] These aforementioned parameters are commonly examined to evaluate whether a child is learning disabilities or not. Developmental brain disorders represent one of the most interesting and challenging research areas in neuroscience. In this research project, we have tried to adopt a K Nearest Neighbor and Support Vector machine to identify Learning disabilities problems at an early stage.

A. Dataset

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 and schools providing assistance for handling learning disability in children and two 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 680 student records with 16 conditional attributes as signs and symptoms of LD and the existence of LD in a child as decision attribute. 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 1 gives key used for representing the symptoms and its abbreviations.

Table- 1. List of attributes

Sl. No.	Attribute	Signs & Symptoms of LD
1	DR	Difficulty with Reading
2	DS	Difficulty with Spelling
3	DH	Difficulty with Handwriting
4	DWE	Difficulty with Written Expression

5	DBA	Difficulty with Basic Arithmetic skills
6	DHA	Difficulty with Higher Arithmetic skills
7	DA	Difficulty with Attention
8	ED	Easily Distracted
9	DM	Difficulty with Memory
10	LM	Lack of Motivation
11	DSS	Difficulty with Study Skills
12	DNS	Does Not like School
13	DLL	Difficulty in Learning a Language
14	DLS	Difficulty in Learning a Subject
15	STL	Slow To Learn
16	RG	Repeated a Grade

4. FEATURE SELECTION

Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore feature selection techniques needs to be applied before any kind of mining algorithm is applied. [10] The main objectives of feature selection are to avoid over fitting and improve model performance and to provide faster and more cost-effective models. 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 $U = \{x_1, x_2, …, x_n\}$ be the set of objects and A = {a1, a2… aM} be the set of attributes used to characterize each object in I. Hence each object xi in the information system can be represented as an M dimension vector $[a_1(x_i), a_2(x_i), …, a_M(x_i)]$,where $a_i(x_i)$ yields the ith ($i = 1, 2, 3, …, M$) attribute value of the ith ($i = 1, 2, 3, …., n$) data object. Dimensionality reduction techniques transform the given dataset I of size n × M into a new low dimensional dataset Y of size $n \times m$.[12]. Feature Selection or attribute selection is a process by which automatically search for the best subset of attributes in dataset. The notion of "best" is relative to the problem are trying to solve, but typically means highest accuracy. [14] A useful way to think about the problem of selecting attributes is state-space search. The search space is discrete and consists of all possible combinations of attributes could choose from the dataset. The objective is to navigate through the search space and locate the best or a good enough combination that improves performance over selecting all attributes.[1]

The simplest is feature selection, in which the number of LD in an experiment is reduced by selecting only the most significant according to some criterion such as high levels of activity. Feature selection algorithms are separated into three categories [10]:

1. The filters which extract features from the data without any learning involved.
2. The wrappers that use learning techniques to evaluate which features are useful.
3. The embedded techniques which combine the feature selection step and the classifier construction

Three key benefits of performing feature selection on data are:

1. Reduces Over fitting: Less redundant data means less opportunity to make decisions based on noise.
2. Improves Accuracy: Less misleading data means modeling accuracy improves.
3. Reduces Training Time: Less data means that algorithms train faster.

The feature selection is a search process that select a subset of significant features from a data domain for building efficient learning models.[12]. KNN and SVM have been applied to an increasing number of real world problems of considerable complexity. The most important advantage is that KNN and SVM are capable of solving problems that are too complex for conventional technologies. These two techniques used for identifying significant attribute and to improve accuracy level.

B. Steps of Feature Selection

A feature of a subset is good if it is highly correlated with the class but not much correlated with other features of the class. [21]Steps to be taken are as follows:

a. Subset generation: We have used two classifiers to rank all the features of the data set.
Then we have used top 3, 4, and 5 features for classification.
b. Subset evaluation: Each classifier is applied to generated subset.

c. Stopping criterion: Testing process continues until 8 features of the subset are selected.
d. Result validation: We have used 10-fold cross validation method for testing each classifier's accuracy.
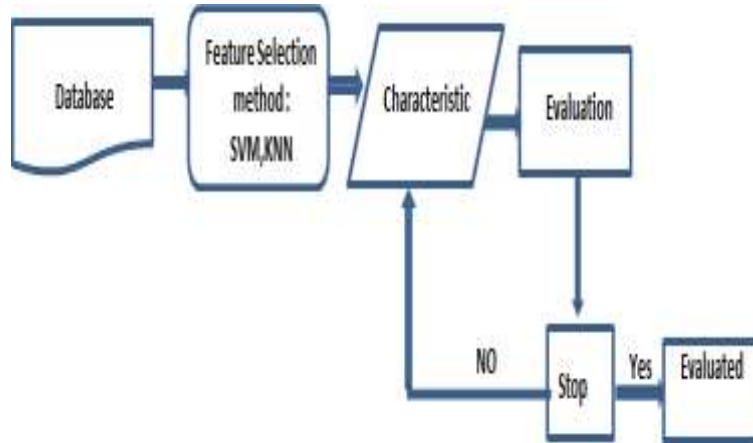
5. IMPLEMENTATION OF MODEL

This work is implemented using MATLAB tool. MATLAB (matrix laboratory) is multi-paradigm numerical computing environment and fourth-generation programming language, developed by Math Works. MATLAB permits matrix manipulations, employment of algorithms, inception of user interfaces, plotting of functions and data and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python [17]. The classify panel is used for classification, to estimate the accuracy of resulting predictive model, visualize erroneous predictions, or the model itself. Net Beans is used to implement FCBF. For good results we need to know the weight age of each variable necessary for identifying learning disabilities in student. So we have used feature selection algorithms tests such as Info gain, Chi squared and gain Ratio. The following table 2 shows the features ranked according to the algorithm.

Table II. Rank of Feature and Average Rank

| Sl. No	Feature	Gain Ratio	Info Gain	Chi squared
1	DR	0.145887	0.145817	120.5478
2	DS	0.104722	0.104627	87.6333
3	DLS	0.089254	0.089037	75.0805
4	DHA	0.064551	0.063383	52.5809
5	DA	0.045683	0.044441	38.2477
6	DH	0.033437	0.033412	28.5096
7	DWE	0.026997	0.026656	22.5457
8	DM	0.021703	0.021311	18.0414
9	DNS	0.007888	0.00788	6.7659
10	LM	0.006284	0.006275	5.3769
11	DLL	0.005614	0.005603	4.8017
12	STL	0.002997	0.002992	2.5718
13	DB	0.002674	0.002649	2.2799
14	DSS	0.001382	0.001381	1.187
15	DBA	0.000582	0.000581	0.4989
16	ED	0.000408	0.00359	0.3093
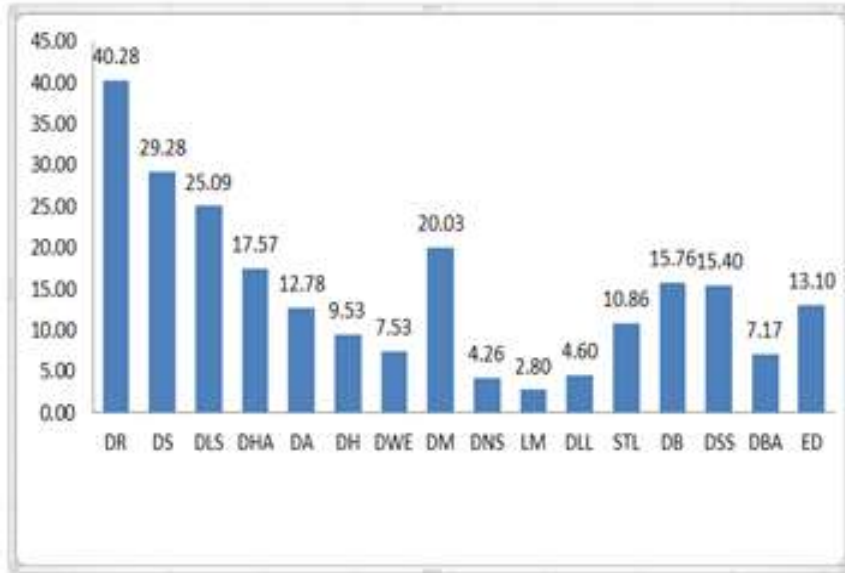
Figure 2 Average Feature Rank

Feature selection Algorithm:

1. Input feature set, KNN SVM
2. Output Final feature subset
3. Normalize features subset using Epsilon algorithm
4. If (Normalized feature subset = =0)
▪ Feature subset = =1
▪ Else
▪ Feature subset = feature subset
▪ End
5. Compute R value
6. If (R -> 1)
▪ Optimal feature = feature subset
▪ Else if
▪ R < 0.5 (step 3)
▪ Reevaluation
▪ End

.**Classification techniques**

There are different combinations of variables in the real world, but we cannot always determine correlations between the variables. This project seems to be one of the first to attempt the use of KNN and SVM for addressing this challenging problem that has drawn the attention of many medical researchers. In actual application, therefore, the "Feature Selection Method" is often adopted to delete unnecessary feature variables and reduce the number of variables included. This simplifies problems and improves accuracy in problem classification.[21] In our study students with LD are evaluated. The method we adopted in our study is the Wrapper-based feature selection based on SVM (Support Vector Machine) and KNN (K Nearest Neighbor) algorithm, were selected as the embedded algorithms to test the optimal combination of feature selection

5.1 Support Vector Machine

Vladimir Vapnik invented Support Vector Machine in 1979. SVM has received more and more attention from researchers in recent year, and its greatest feature is the capability of dealing with linear and non-linear questions and is free from the limitation of data size. [13] In other words, training with few parameters can be conducted without having sufficient data or premises. In classification, the primary concept of SVM is to construct an optimal hyper-plane that serves as an interface for classification decisions. [20, 21] Through this interface we can effectively separate positive examples and negative examples.

Classification of SVM is shown on Figure. 3.
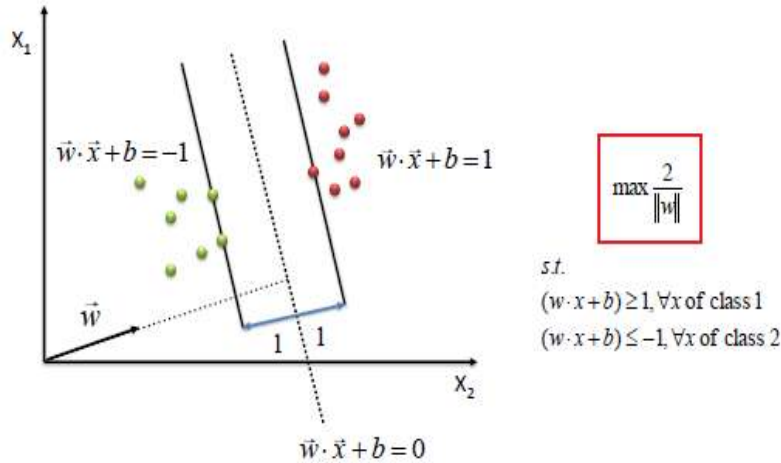
The details accuracy by class and confusion matrix in respect of SVM as shown in Table III below Detailed accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.973	0.854	0.933	0.969	0.965	0.953	1
0.854	0.974	0.949	0.854	0.657	0.895	0

Table IV. Confusion matrix

A	B	Classified as
419	13	a = t
6	242	b = f

5.2 K- Nearest Neighbor

Feature weights are easily incorporated into learning algorithms such as k nearest neighbor, but the advantage of feature weighting over feature selection is minimal at best, due to the increased chance of over fitting the data. In general, feature weighting does not reduce the dimensionality of the original data.[4]. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic techniques. The special case where the class is predicted to be the class of the closest training sample (i.e. when k = 1) is called the nearest neighbor algorithm. [6]

The new instance is assigned to the retrieved instance's class. Equation 1.1 shows the distance metric employed by IB1.

$$D(x, y) = \sqrt{\sum_{j=1}^{n} f(x_j, y_j)} \qquad (1)$$

Equation 1 gives the distance between two instances x and y; x_i and y_j refer to the jth feature value of instance x and y, respectively. For numeric valued attributes $f(x_j, y_j) = (x_i - y_j)^2$; for symbolic valued attributes $f(x, y) = 0$, if the feature values x_j and y_j are the same, and 1 if they differ.

It is important to remove redundant and irrelevant attributes from dataset before evaluating algorithms. This task should be tackled in the Prepare Data step of the applied machine learning process. [26]. The Search Method is the structured way in which the search space of possible attribute subsets is navigated based on the subset evaluation. Baseline methods include Random Search and Exhaustive Search, although graph search algorithms are popular such as Best First Search. The details accuracy by class and confusion matrix in respect of KNN as shown in Table V below

Table V Detailed accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.967	0.975	0.985	0.969	0.966	0.977	0
0.975	0.969	0.949	0.949	0.968	0.962	1

Table VI. Confusion matrix

A	B	Classified as
247	8	a = t
9	416	b = f

5.3 Wrappers Method

A number of deterministic investigations have been used to examine LD such as a combination of a wrapper and sequential forward selection. SFS is a deterministic feature selection method that works by using hill-climbing search to add all possible single-attribute expansions to the current subset and evaluate them. It starts from an empty subset of LD attributes and sequentially selects attributes, one at a time, until no further improvement is achieved in the evaluation function. The feature that leads to the best score is added permanently [23]. For classification, support vector machines (SVMs), k-nearest neighbors, and probabilistic were used in an attempt to classify between LD and non LD [24]. Very accurate results were achieved using SVMs techniques. Two methods based on SVMs are very widely used in disabilities datasets: After removing the illogical entries, the remaining data underwent feature selection with SVM, feature selection with KNN, and to improve the accuracy result. The wrapper feature selector is as follows:
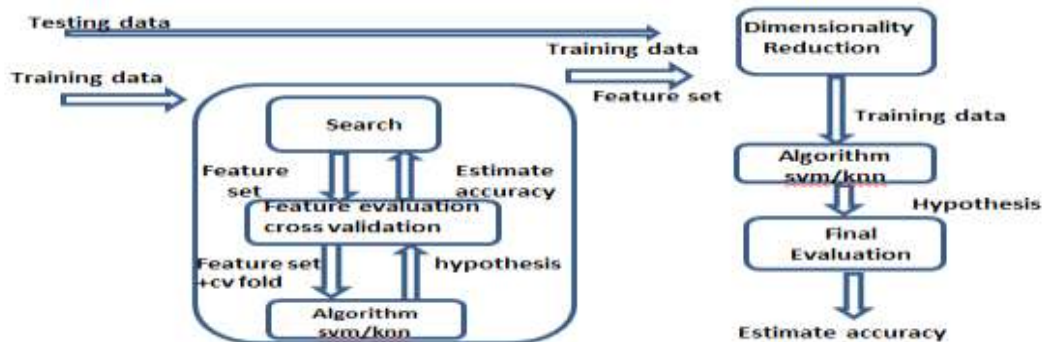

Figure 4. Wrapper feature selector

The experimental comparison of classification algorithms are done based on the performance measures of classification accuracy rate and execution time. After evaluating all features of the LD dataset, the algorithm retains the set of features $\{DWE, DS, DLL, DM, DLS, DA, DH, STL\}$. These Eight features are significant because all other features can be removed from the LD dataset without affecting the classification performance. Table VII 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 VII: Classification results given by SVM/KNN

Comparison of Features Selection Classification Results				
Various cases	Dataset prior to perform feature selection		Dataset reduced using the proposed approach	
	KNN	SVM	KNN	SVM
No of feature	16	16	7	8
Correctly Classified Instances (In percentage)	92%	94%	92%	97%

Incorrectly Classified Instances (In Percentage)	8%	6%	8%	3%
Time Taken to build a model (in seconds)	0.98	0.40	0.98	0.13

6. DISCUSSION

From the experimental results presented in Table VII it is clear that, in case of the proposed approach 50% reduction in the dataset, it affected only 3% the classification accuracy. It follows that the original dataset contains about 50% redundant attributes and the feature selection approach presented is efficient in removing these redundant attributes affecting the classification accuracy. From the comparison of 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 in SVM. 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.

CONCLUSION

In this work we have proposed a new approach of feature selection for classification of learning disability by using data mining approach. In order to achieve the better classification results we have used feature selection algorithms which are Chi-square, Info Gain and Gain Ratio which results in the prediction of the relevant features. Later the optimization of selected feature is performed by using fast correlation based filtering approach. Finally KNN and SVM classifiers are utilized to show the classification accuracy. From the results, it can be concluded that the SVM achieves increased classification performance, yields results that are accurate, hence it is considered as best classifier when compared with K nearest Neighbour classifier algorithm. Perhaps, K nearest Neighbour classifier classifies the data with minimum execution time. In future other feature selection techniques can be applied on the dataset.

REFERENCES

[1]   Ladha L. and Deepa T., "Feature Selection Methods and Algorithms", International Journal on Computer Science and Engineering (IJCSE), 2011.
[2]   H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In Proceedings of the Ninth National Conference on Artificial Intelligence, pages 547–542. MIT Press, 1991.
[3]   Translated by Kung, F., 1997. DSM-IV. Taipei, Ho-chi Book Publishing Co., Ltd..
[4]   Huang, T., 2005. A study on the traits demonstrated by students with LD when taking the WISC-III test. Unpublished master's thesis, Department of Special Education, NHCUE, Hsinchu.
[5]   Hung, L., 2005.Discussion and suggestions regarding LD identification: Discussing the "Identification of Students with LD Chart" implemented in Taiwan. Li-yu Hung's Website Date acquired: 2006.10.16,From: http://192.192.250.54/liyuhung/chineses/modules/datashare//ld_identification_taiwan2005.pdf.
[6]   https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#Feature_extraction.
[7]   Y. Ma, B. Liu, C.K. Wong, P.S. Yu, and S.M. Lee, "Targeting the Right Students Using Data Mining", Proceedings of KDD, International Conference on Knowledge discovery and Data Mining, Boston, USA, 2000, pp. 457-464.
[8]   Jai Ruby & K. David, "A study model on the impact of various indicators in the performance of students in higher education", IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, pp.750-755, May 2014.
[9]   Monika Goyal & Rajan Vohra, "Applications of Data Mining in Higher Education" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, pp.130-120, March 2012.
[10]   M.A Hall, and L.A Smith, 1998, selection for machine learning, In Proceedings of the Australian Computer Science Conference, pp181.
[11]   Sk. Althaf Hussain Basha, A Govardhan, S. Viswanadha Raju, Nayeen Sultana, 2010, A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University European Journal of Scientific Research, Vol 46, No 2, pp.186-193.
[12]   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.
[13]   Gualtieri. J. A, Chettri. S. R, Cromp. R. F and Johnson.L. F, (1999) Support vector machine classifiers as applied to AVIRIS data, in Summaries 8th JPL Airborne Earth Science Workshop, JPL Pub. 99-17, pp. 217–227.
[14]   Ron Kohavi & Dan Sommerfield, (1995) "Feature subset selection using the wrapper method: Overfitting and dynamic search space topology", Proceedings of the First International Conference onKnowledge Discovery and Data Mining, pp. 192-197.
[15]   Ian H. Witten and Eibe Frank.(2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition.
[16]   Neumann J, Schnörr C & Steidl G, (2005) "Combined SVM based feature selection and classification", Machine Learning, Vol.61, pp.129-150.
[17]   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.

[18]   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.

[19]   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.

[20]   Fadzilah Siraj, Mansour Ali Abdoulha, (2011). Mining Enrollment Data Using Descriptive and Predictive'

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

[22]   Baker R.S.J.D., & Yacef K, "The state of educational data mining in 2009: A review and future vision", Journal of Educational Data Mining, I, pg. 3-17,2009.

[23]   Rowley, J.,"Is higher education ready for knowledge management?",International Journal of Educational Management, vol. 14(7), pp. 325–333, 2000.

[24]   Lubega, J. T., Omona, W., & Weide, T. V. D., "Knowledge management technologies and higher education processes_: approach to integration for performance improvement", International Journal of Computing and ICT Research, vol. 5(Special Issue), pp. 55–68, 2011.

[25]   Hung, L., 1999. Education for individuals with learning disabilities. Taipei, Psychological Publishing Co., Ltd.

[26]   Mark A. Hall The thesis of Correlation-based Feature Selection for Machine Learning,