Application of Feature Selection Methods in Educational Data Mining

Anal Acharya
Department of Computer Science,
St Xavier's College, Kolkata,
India.

Devadatta Sinha
Department of Computer Science and Engineering,
University of Calcutta, Kolkata,
India.

ABSTRACT
In the recent years, web based learning has emerged as a new field of research due to growth of network and communication technology. These learning systems generate a large volume of student data. Data mining algorithms may be applied on this data set to study interesting patterns. As an example, student enrollment data and his past examination records could be used to predict his grades in the term end examination. However this prediction could mean examining a lot of features of the student data resulting in creation of a model with high computational complexity. In this context this work first defines a student data set with 309 records and 14 features collected by a survey from various graduation level students majoring in Computer Science under University of Calcutta. Different feature selection algorithms are applied on this data set. The best results are obtained by Correlation Based Feature Selection algorithm with 8 features. Subsequently classification algorithms may be applied on this feature subset for predicting student grades.

General Terms
E-Learning, Feature Selection, Data Mining.

Keywords
Educational Data Mining (EDM), Kappa Statistic, F-measure, Prediction Accuracy, College Education, WEKA.

1. INTRODUCTION
In the recent years the most important innovation that has changed the face of modern education is the Internet [1]. It has been used to provide computer aided instruction to any location as well as any platform. This change has lead to the development of a variety of web based educational systems offering varied learning goals based on learning needs. These systems generate a huge amount of student data based on various types of Learning Objects accessed, student learning pattern, and their assessment results to name a few [2]. In this context, Educational Data Mining (EDM) has emerged as a new field of research to analyze educational data in order to resolve various educational research issues [3,4].

As stated above, an web based educational system generates a lot of student data. These could be used for a variety of analytical purposes. As an example, student "past records" could be used to predict his degree of success in the "final exams". This would require the application of certain classification and prediction models of data mining. However "past records" could mean huge amount of data with a large number of features derived from system logs. Applying classification and prediction algorithms on such a large data and feature set would result in creation of a model with high computational complexity. In this context, the objectives of this work are as follows:

(i) Defining a data set for a group of students graduating in various colleges under University of Calcutta majoring in Computer Science.
(ii) Identifying the most relevant minimal Feature Sub Set (FSS) for achieving a acceptable degree of accuracy for predicting their final grades in semester end examinations using various Filter and Wrapper based Feature Selection algorithms.
(iii) Evaluate the goodness of FSS thus found by applying certain baseline classifiers using Kappa Statistic and F-Measure parameters.

The organization of the paper is as follows: The next section discusses briefly about Feature Selection along with related literature on it in context of EDM. The details of the data set used in experimentation are then discussed along with the Feature Selection algorithms applied on them. The details of the experiments are then discussed. Finally conclusions are drawn by deriving the most relevant FSS and Feature Selection algorithm.

2. FEATURE SELECTION
Feature Selection (FS) is the task of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept [5] which in this case is classification. The subset is chosen on the assumption that input data contains a large number of irrelevant features the removal of which would not affect the accuracy of the classification model [6]. There are several benefits behind this: reduced computational complexity, better model interpretability and reduced data over fitting by enhancing generalization [7]. Keeping classification as a target two types of FS algorithms are used here: Filter Based FS Algorithms (FBFSA) and Wrapper Based FS Algorithms (WBFSA) [8]. FBFSA ranks the features based on inherent relationships between data. Some examples of these are Correlation Based FS (CBFS), Chi-Square Based Feature Evaluation (CSBFE), Information Gain Attribute Evaluation (IGATE) etc [9]. WBFSA evaluates the importance of the features based on learning algorithms. Several popular machine learning algorithms [10] may be used for this purpose.

Several researchers have performed FS before applying classification algorithms for reasons specified earlier. Ramaswami and Bhaskaran [11] have used CSBFE technique to determine high potential variables for higher secondary examination grade prediction for a set of 772 students in

Tamil Nadu, India. These features were used to construct a CHAID tree classification model. Overall prediction accuracy of the model was 44.69% and results found to be quite satisfactory when compared to other prediction model. Nguyen et al [12] has used IGATE technique for predicting academic performance of undergraduate and post graduate students in two universities in Thailand. Decision trees and Bayesian Network algorithms were then applied to develop the prediction model. The sample size used for developing the prediction model was 20492 and 936 respectively. These predictions were found to be very good in identifying and assisting failing students. Kovacic [13] has again used CSBFE technique to find out the importance of the dependent variables for predicting student dropout at the Open Polytechnique of New Zealand. The number of students sampled was 450. Classification and Regression Trees (CART) and CHAID were used for prediction model construction. He concluded that classification done on enrollment data alone is not quite good in separating successful from unsuccessful students. Osmanbegovic and Suljic [14] has used four FS algorithms namely, CSBFE, OneR, IGATE and Gain ratio based attribute evaluation and computed the average ranking of these attributes for predicting student performance in the University of Tuzla, Uganda. Naïve Bayes(NB), Multi Layer Perceptron (MLP) and C4.5 algorithms were then used for developing the prediction models. However all these papers use FS only as a prerequisite to classification.

3. DATA SET AND FEATURE SELECTION ALGORITHMS USED

When a student takes admission to a college for graduation he is basically at academic cross roads. Students from various backgrounds join him; some of the factors which create this variation is difference in family income, medium of instruction at the higher secondary level, the caste and religion of the student to name a few. These differences lead to their varying performance in semester end examinations. This is particularly true in a semester system where the students are inducted in the examination system within six months of admission. The aim of this study is to develop a prediction model to identify the set of students who may perform poorly in the semester end examination. Once this is done some sort of remedial lesson plan may be developed for them. The current study is based on data collected for 309 students who are attending first semester classes majoring in computer science in some undergraduate colleges under University of Calcutta. The data has been collected just after the announcement of mid semester examination results. A detailed questionnaire was prepared with help from the following sources: (i) Vice Principal's office of the related colleges (ii) Faculty members and students of the respective departments (iii) Related literature [11,12,13,14].

The attributes which has been used to develop the prediction model is now discussed. It has been widely observed that there is a difference in study pattern of boys and girls especially at college level and thus gender plays a role in determining student results. Although India is a secular country, caste and religion play a vital role in academic development of students over the years. It has generally been observed that students of higher castes have performed better in academic arena. Family size is also related to academic performance as more the number of children parents have the lesser is the degree of attention given by them to their academic performance. The type of board the student study in and the medium of instruction at higher secondary level affect the performance of the student as in college the medium of instruction is English. Students from rural or municipal background also may have difficulty in understanding classes for this reason as it is virtually impossible to teach computer science in regional languages. Kolkata being cosmopolitan city students from various states take admission to city colleges. In addition to academics these students have to take the additional burden of fooding and lodging which may affect their academic performance adversely. In graduation level most of the students take private tuition from various college teachers in their domain and computer science is no exception. It has been seen that quality private tuition enhances student performance. Needless to say that a student has to pay a large amount of fees to the private tutors for this purpose. Thus a student's family income has a huge bearing on the student's performance. It has been noted that students who have performed well in the board exams generally performs well in graduation exams. Another indicator in this regard is the performance of the student in mid semester exams, which indicates the degree of student's grasp over the subject. The number of hours studied by the student also determines the quality of his academic performance as does his attendance in class. All these attributes are used for predicting student's grades in semester end examinations. Related literature [11,12,13,14] suggests that this set is exhaustive for predicting student's grades. The model developed attempts to predict student grades as a 7-class problem. For example, if a student secures between 90 and 100% he is assigned a grade 'O', if he secures between 89 and 80% he secures a grade 'E' and so on. Continuing in this way the grades A, B, C, D and F are defined. The student attributes, prediction variable and their corresponding domain are shown in Table 1 for easy reference.

As indicated earlier the aim of this work is to identify a optimal set of features that would reduce computational complexity without sacrificing predictive accuracy. For this three FBFSA along with three popular machine learning algorithms are identified to be used with WBFSA. These algorithms along with their search methods are shown in Table 2.

4. RESULTS AND DISCUSSION

The experimental methodology for FBFSA is first discussed and the same methodology is followed for WBFSA. Initially the three FBFSA are applied to the student data to obtain a set of attributes in descending order of their importance. The predictive accuracy of this set is then to be evaluated using some baseline classifier. Here C4.5 Decision Tree (DT) is used as the baseline classifier since it does not take long training time and has a high degree of predictive accuracy [15]. Two parameters [18] are used to determine the degree of predictive accuracy. The Kappa statistic measures the agreement of prediction with true class; a value of 1.0 signifies complete agreement.

Table 1. Student attributes

Attribute Number	Attribute Name	Description	Domain
1	Gender	Student's sex	m-male; f-female
2	Caste	Student's caste	1-General;2-SC;3-ST;4-OBC
3	Religion	Student's religion	1-Hindu;2-Muslim;3-Crhistian;4-Others
4	Fsize	Student's family size	Numeric
5	Board	Student's board at higher secondary level	1-ISC;2-CBSE;3-State board;4-other boards
6	Sorigin	Student's state of origin	1-West Bengal;0-Other States
7	Income	Student's family income	1- <=10K;2- >10K but <=20K;3- >20K but <=30K;4- >30K but <=40K;5- >50K
8	Boardmarks	Student's aggregate % of marks at higher secondary level	1- >=90%; 2-89% to 80%;3-79% to 70%;4-69% to 60%;5-59% to 50%; 6-<50%
9	Hday	Average number of hours studied by the student per day	1-<3 hours ;2-3 to 6 hours;3-7 to 9 hours;4- >9 hours
10	Atten	Student's % attendance in class for the semester	1- >=90%; 2-89% to 80%;3-79% to 70%;4-69% to 60%;5-59% to 50%; 6-<50%
11	Midsem	Student's % marks in mid semester exam	1- >=90%; 2-89% to 80%;3-79% to 70%;4-69% to 60%;5-59% to 50%; 6-<50%
12	Medium	Student's medium of study at higher secondary level	1-English;2-Bengali;3-Hindi;4-Others
13	School	Student's type of school at higher secondary level	1-Urban;2-Minucipal;3-Rural
14	Ptution	Whether the student has taken private tution	1-Yes;0-No
15	Grade (Dependent Variable)	Grade secured by the student in semester end examination	O-91% to 100%;E-81% to 90%;A-71%-80%;B-61% to 70%;C-51% to 60%;D-41%-50%;F<=40%

Table 2. FS Algorithms along with corresponding search methods in this study

Type of FS Algorithm	Algorithms	Search Methods
Filter Based Algorithms	Correlation Based FS(CBFS)	Rank Search
	Chi-Square Based Feature Evaluation(CSBFE)	Ranker
	Information Gain Attribute Evaluation(IGATE)	Ranker
Wrapper Based Learning Algorithms	C 4.5	Rank Search
	Naïve Bayes(NB)	Rank Search
	1-Nearest Neighbor(1-NN)	Rank Search

Two other parameters Precision and Recall have been used by several researchers to determine predictive accuracy. However F-measure which is a combination of Precision and Recall is used here. It is to be noted that Receiver Operating Characteristic (ROC) parameter is not used here since it applies to a 2-class problem. The experiments for this paper have been conducted in WEKA (Waikato Environment for Knowledge Analysis) which is a free software [17] written in Java for data analysis and predictive modeling. The steps used for computing the maximum Kappa statistic for the CBFS algorithm is shown in Table 3.

This process of computing the maximum Kappa statistic is continued for CSBFE and IGATE algorithms and the results shown in Table 4. The above experiments are repeated for determining the minimal FSS with maximum F-measure for CBFS, CSFE and IGATE algorithms and the results shown in Table 5. Based on Kappa Statistic and F-measure values IGATE and CBFS algorithm generated FSS are selected for further investigation (Table 6).

Table 3. Feature Subset corresponding to maximum Kappa statistic for CBFS

Step	Attribute Set in order of decreasing importance	Kappa Statistic
Initialization	7,5,12,8,3,11,13,2,1,14,10,6,9,4	0.64
Prune 4	7,5,12,8,3,11,13,2,1,14,10,6,9	0.70
Prune 9	7,5,12,8,3,11,13,2,1,14,10,6	0.69
Prune 6	7,5,12,8,3,11,13,2,1,14,10	0.68
Prune 10	7,5,12,8,3,11,13,2,1,14	0.69
Prune 14	7,5,12,8,3,11,13,2,1	0.69
Prune 1	7,5,12,8,3,11,13,2	0.71
Prune 2	7,5,12,8,3,11,13	0.67
Prune 13	7,5,12,8,3,11	0.66

Table 4. FSS maximizing Kappa Statistic for FBFSAs.

Algorithm	CBFS	CSFE	IGATE
Number of attributes	8	8	10
Attribute Set	7,5,12,8,3,11,13,2	7,5,12,8,3,11,13,2	7,11,8,5,3,10,2,13,12,9
Kappa Statistic	0.73	0.71	0.84

Table 5. FSS maximizing F-Measure Value for FBFSAs

Algorithm	CBFS	CSFE	IGATE
Number of attributes	6	7	7
Attribute Set	7,5,12,8,3,11,13,2	7,8,11,5,7,10,3	7,11,8,5,3,10,2
F Measure	0.73	0.70	0.71

Table 6. FBFSAs and FSS selected for further analysis

Measure	Algorithm	Number of attributes	Attribute Set
Kappa Statistic	IGATE	10	7,11,8,5,3,10,2,13,12,9
F measure	CBFS	8	7,5,12,8,3,11,13,2

Three classes of popular machine learning algorithms are next identified for use in WBFSA. Since DT can classify both categorical and numerical data C4.5 is chosen as a representative of this class. Bayesian Learning algorithms theoretically have minimum error rate compared to other classifiers. They also provide a theoretical justification for other classifiers that do not use Bayes theorem. Naïve Bayes algorithm is chosen as a representative of this class. A typical example of lazy learning algorithm is k-nearest neighbor classifier. This method is computationally inexpensive when the training set is small. In this case k=1 is taken. Following a process similar to FBFSA the FSS maximizing Kappa Statistic (Table 7) and F-measure (Table 8) for the above mentioned machine learning algorithms is obtained. NB and 1NN algorithms are selected for further analysis.

Table 7. Maximum Kappa Statistic for three learning algorithms

Algorithm	C4.5	NB	1-NN
Number of attributes	8	7	8
Attribute Set	2,3,5,7,8,11,12,13	7,5,12,11,8,13,2	7,5,12,8,3,11,13,2
Kappa Statistic	0.71	0.72	0.71

Table 8. Maximum F-Measure measure for three learning algorithms

Algorithm	C4.5	NB	1-NN
Number of attributes	8	6	8
Attribute Set	2,3,5,7,8,11,12,13	7,5,12,11,8,13	7,5,12,8,3,11,13,2
F Measure	0.73	0.69	0.74

Table 9. Learning algorithms and FSS selected for further analysis

Measure	Algorithm	Number of attributes	Attribute Set
Kappa Statistic	NB	7	7,5,12,11,8,13,2
F-Measure	1NN	8	7,5,12,8,3,11,13,2

The four algorithms and feature subsets (Table 6 and Table 9) thus selected are renamed as CBFS8, IGATE10, NB7, 1NN8 for convenience. The best FS algorithm and FSS is then selected by finding the predictive accuracy of these

Table 10. Average F-Measure for each feature subset

Feature Subset	F-Measure				Average F-Measure
	CART	NB	MLP	1NN	
CBFS8	0.72	0.69	0.72	0.73	0.7150
IGATE10	0.76	0.68	0.71	0.68	0.7075
NB7	0.75	0.69	0.74	0.67	0.7125
1NN8	0.76	0.68	0.71	0.68	0.7075
Without FS	0.67	0.72	0.72	0.76	0.7175

FSS on four classes of machine learning algorithms: (i) CART from decision tree (ii) NB (from Bayesian Learning) (iii) MLP (from Artificial Neural Networks) (iv) 1NN(from Lazy learning). Machine Learning algorithms from four classes are taken so as to introduce a certain degree of variation among

the classifiers, i.e. they should not make identical or correlated errors [9]. Average F-measure of these is then computed to determine their predictive accuracy. Ten fold cross validation was used for this purpose. The results are summarized in Table 10. The average F measure obtained without performing FS is also shown.

The results clearly show that CBFS gives the best result on the given data set with 8 attributes. Also the most important attributes in order as determined by CBFS is 7,5,12,8,3,11,13,2. Thus a student's family income is the most important factor in determining his end semester grades. This fact is also supported by IGATE, 1NN and NB algorithms. As seen in Table 10, the average F-measure obtained without FS is just higher than the average F-Measure obtained after FS. In terms of performance, there is no difference between FBFSA and WBFSA; it is noticeable that both of them perform equally well on given data set.

5. CONCLUSION

In this paper a data set is first defined for a group of 309 students majoring in computer science in various colleges under University of Calcutta. This data set was used for predicting the grades of the students in semester end examinations. However, this data set contains 16 features and applying prediction algorithms on these would increase the computational complexity. Thus FBFSA and WBFSA algorithms have been applied to perform FS. C4.5 algorithm has been used as a baseline classifier to derive the optimal FSS based on F measure and Kappa statistic for FBFSA and WBFSA.

Final conclusions about the best Feature Subset are drawn by averaging the F-measure obtained from applying four classifiers on the features subsets obtained previously. CBFS with 8 features give the highest average F-measure value. This value is only marginally less than the average F measure obtained without FS. Thus FSS has been reduced from 14 to 8 without sacrificing predictive accuracy.

Future work involves predicting student grades using these 8 features. Various machine learning algorithms may be used for this purpose. A set of decision rules may then be derived from these algorithms to predict student performance in semester end examinations. The predictive accuracy obtained using these algorithms may be further increased by Combining Multiple Classifiers (CMC) using Ensemble Methods. The utility of genetic algorithms in enhancing classification accuracy may also be examined.

6. REFERENCES

[1] Castro, F., Vellido, A., Nebot A., and Francisco M,. Applying Data Mining Techniques to e-Learning Problems.

[2] Minei-Bidgoli, B., Kashy, D. A.,,Kortemeyer, G, Punch, W F., 2003, Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA. 33rd ASEE/IEEE Frontiers in Education Conference

[3] Romero, C., Ventura, S.,2007, Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications

[4] Romero, C., Ventura. S., Educational Data Mining: A Review of the State-of-the-Art. IEEE Transactions on Systems, Man, and Cybernatics—PART C: Applications and Reviews

[5] Kira, K., Rendell, L.A., 1992. A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning. pp. 249–256.

[6] Selwyn, Piramuthu.,2004, Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research 156 p 483–494.

[7] Mitra, P., Murthy, C. A., Pal, S K.,2002, Unsupervised Feature Selection using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence", Vol 24, No 4.

[8] Dash, M., Liu, H.,1997, Feature Selection for Classification.Intelligent Data Analysis.

[9] Stefanowski, J,. An Experimental Study of Methods Combining Multiple Classifiers - Diversified both by Feature Selection and Bootstrap Sampling.

[10] Kotsiantis, S., Piarrekeas, C., Pintelas P, 2004, Predicting Students' performance in Distance Learning using Machine Learning Techniques, Applied Artificial Intelligence, Volume 18, pp 411-426.

[11] Ramaswami, M., Bhaskaran,, R.,2010, A CHAID Based Performance Prediction Model in Educational Data Mining. International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1.

[12] Nguyen, N. N., Janeck, P., Haddawy, P. A ,2007, Comparative Analysis of Techniques for Predicting Academic Performance. 37th ASEE/IEEE Frontiers in Education Conference.

[13] Kovacic, Z. J.,2010, Early Prediction of Student Success: Mining Students Enrolment Data. Informing Science & IT Education Conference.

[14] Oladukun, V. O., Adebanjo, A. T., Charles-Obawa, O. E., Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course.

[15] Zhao, Y. and Zhang Y., Comparison of decision tree methods for finding active objects.

[16] Vafaie, H. and Imam, I. F,. Feature Selection Methods:Genetic Algorithms vs. Greedy-like Search.

[17] WEKA Manual for Version 3-6-10.

[18] Haan, J., Kamber, M,.2011, Data Mining-Concepts and Techniques. Third Edition, Elsivier.