CORRELATION BASED FEATURE SELECTION (CFS) TECHNIQUE TO PREDICT STUDENT PERFROMANCE

Mital Doshi [1], Dr.Setu K Chaturvedi, Ph.D [2]

[1]Mtech. Research Scholar
Technocrats Institute of Technology Bhopal, India

[2]Professor & HOD (Dept. of CSE)
Technocrats Institute of Technology Bhopal, India

ABSTRACT

Education data mining is an emerging stream which helps in mining academic data for solving various types of problems. One of the problems is the selection of a proper academic track. The admission of a student in engineering college depends on many factors. In this paper we have tried to implement a classification technique to assist students in predicting their success in admission in an engineering stream.We have analyzed the data set containing information about student's academic as well as socio-demographic variables, with attributes such as family pressure, interest, gender, XII marks and CET rank in entrance examinations and historical data of previous batch of students. Feature selection is a process for removing irrelevant and redundant features which will help improve the predictive accuracy of classifiers. In this paper first we have used feature selection attribute algorithms Chi-square.InfoGain, and GainRatio to predict the relevant features. Then we have applied fast correlation base filter on given features. Later classification is done using NBTree, MultilayerPerceptron, NaiveBayes and Instance based –K- nearest neighbor. Results showed reduction in computational cost and time and increase in predictive accuracy for the student model

KEYWORDS

Chi-square, Correlation feature selection, IBK, Infogain, Gainratio, Multilayer perceptron, NaiveBayes, NBTree

1. INTRODUCTION

Feature selection is a preprocessing step in machine learning. We have three main categories wrapper, filter and embedded .algorithms [1]. The filter model selects some features without the help of any learning algorithm. In the wrapper model we use some predetermined learning algorithm to find out the relevant features and test them.Wrapper model is more expensive than filter one because it requires more computations so when generally there are large number of features we prefer filter model. In this paper, we have tried to use the filter model and our aim is to improve the accuracy of recommending the stream to the student to help him develop a bright future according to his choice by predicting the success at the earliest. Fast correlation base filter is an algorithm which is much successful in removing the redundant and irrelevant features from the dataset so that computation time is decreased and predictive accuracy is increased.

DOI : 10.5121/ijcnc.2014.6315

2.CLASSFICATIONTECHNIQUES

2.1 NBTree

NBTree is a hybrid algorithm with Decision Tree and Naïve-Bayes. In this algorithm the basic concept of recursive partitioning of the schemes remains the same but here the difference is that the leaf nodes are naïve Bayes categorizers and will not have nodes predicting a single class. [2]

2.2 Naïve Bayes

The Naïve Bayes classifier technique is used when dimensionality of the inputs is high. This is a simple algorithm but gives good output than others. We are using this to predict the dropout of students by calculating the probability of each input for a predictable state. It trains the weighted training data and also helps prevent over fitting.

2.3 Instance-based-k-nearest neighbor

In this technique a new item is classified by comparing the memorized data items using a distance measure. For this we require storing of a dataset. Matching of items is done by putting them close to original item. Nearest neighbors can be done by using cross-validation either automatically or manually.

2.4 Multilayer Perceptron

It is one of the most widely used and popular neural networks. Its network consists of a set of sensory elements which forms the input layer, one or more hidden layers of processing elements, and the output layer is of the processing elements. The back propagation algorithm ANN can be used for predicting both continuous and discrete data. ANN Algorithm represents each cluster by a neuron based on the neural structure of the brain. Here each connection has an associated weight, which is calculated adaptively during learning. The only point about ANN is that it takes long training times and is therefore more suitable for applications where long training is feasible. Here we have used Multilayer Perceptron technique of ANN. [3]

3. RELATED WORK

Pumpuang [4]had proposed the classifier algorithm for building Course Registration Planning Model from historical dataset.The model used four classifiers including Bayesian Network, C4.5, Decision Forest and NBTree. Results showed that NBTree seemed to be the best for prediction of GPA of the student.

Tanna[5] has implemented a decision support system for admission in engineering colleges which is based on entrance exam marks. Results show it will return colleges and streams categorized as Ambitious, Best Bargain and Safe using an offset value.

In [6] Malaya used a knowledge based decision technique will guide the student for admission in proper branch of engineering. They used two algorithms decision tree algorithm and ANN to find out which one is more accurate for decision making. Results showed that accuracy of MLP algorithm has proved to be better for training partition size 50 & testing partition size 50 upto 86%

Al-Radaideh [7] proposed in his paper a simple classification model to provide a guideline to help students and school management to choose the right track of study for a student. Decision tree using the C4.5 algorithm (J48 in WEKA), was built by selecting the best attributes using the information gain measure. The classification rules to find were based on more than one factor such as the Ratio and the average of student mark in the 10th class (AVERAGE), and the average of the student mark in 8^{th}, 9^{th}, and 10^{th} classes (AVG89_10). Results show that accuracy of the model was 87.9% where 218 students were correctly classified out of 248 students.

Hany et al. [8] applied six classifiers on ASSISTments dataset having 15 features. They used VF1,IBK,NaiveBayes Updateable, ONER, j48 and k means clustering classifiers to rank the features. Results showed that k means clustering was the best in giving ranks to features and Naïve Bayes was better in giving prediction accuracy.

Lei Yu [9] in their work proposed a feature selection algorithm which is specially used for high dimensional data which is called as fast correlation base filter. This algorithm is for removing irrelevant and redundant data. They applied FCBF, ReliefF, CorrF, and ConSF on four datasets and recorded the running time and number of features selected. Then they applied C4.5 and NBC classification on the data.

Bharadwaj and Pal [10] conducted experiment to predict the performance at the end of semester using student's data like attendance, class test, seminar and assignment marks from the student's previous database results

Hijazi and Naqvi [11] conducted a study on student performance on 300 students from group of colleges of Punjab University. Results showed that student's attitude towards attendance in class are dependent on the time they spend in college for study after college hours. Other factors such as mother's age and education are related with student's performance found by simple linear regression analysis.

Khan [12] conducted an experiment on 200 boys and 200 girls of Secondary school of Aligarh Muslim University. Their main aim was to find out variables which determine the success in higher education in science stream. So they used demographic variables, personality measures as an input. They had used cluster sampling technique for division into groups or clusters and a random sample of cluster was used for further analysis. Results showed that girls with high socio-economic status had relatively higher academic achievement in science whereas boys with low socio-economic status had higher academic achievement in general.

Z. J. Kovacic [13] presented a case study on educational data mining to identify up to what extent enrolment data can be used to predict student's success. They had used CHAID and CART on students of diploma college of New Zealand. They got two decision trees in their results and accuracy of classifiers obtained was 59.4 and 60.5.

Al-Radaideh [7] proposed in his paper a simple classification model to provide a guideline to help students and school management to choose the right track of study for a student. Decision tree using the C4.5 algorithm (J48 in WEKA), was built by selecting the best attributes using the information gain measure. The classification rules to find were based on more than one factor such as the Ratio and the average of student mark in the 10th class (AVERAGE), and the average of the student mark in 8^{th}, 9^{th}, and 10^{th} classes (AVG89_10). Results show that accuracy of the model was 87.9% where 218 students were correctly classified out of 248 students.

Hany et al. [8] applied six classifiers on ASSISTments dataset having 15 features. They used VF1,IBK,NaiveBayes Updateable, ONER, j48 and k means clustering classifiers to rank the features. Results showed that k means clustering was the best in giving ranks to features and Naïve Bayes was better in giving prediction accuracy.

Lei Yu [9] in their work proposed a feature selection algorithm which is specially used for high dimensional data which is called as fast correlation base filter. This algorithm is for removing irrelevant and redundant data. They applied FCBF, ReliefF, CorrF, and ConSF on four datasets and recorded the running time and number of features selected. Then they applied C4.5 and NBC classification on the data.

Bharadwaj and Pal [10] conducted experiment to predict the performance at the end of semester using student's data like attendance, class test, seminar and assignment marks from the student's previous database results

Hijazi and Naqvi [11] conducted a study on student performance on 300 students from group of colleges of Punjab University. Results showed that student's attitude towards attendance in class are dependent on the time they spend in college for study after college hours. Other factors such as mother's age and education are related with student's performance found by simple linear regression analysis.

Khan [12] conducted an experiment on 200 boys and 200 girls of Secondary school of Aligarh Muslim University. Their main aim was to find out variables which determine the success in higher education in science stream. So they used demographic variables, personality measures as an input. They had used cluster sampling technique for division into groups or clusters and a random sample of cluster was used for further analysis. Results showed that girls with high socio-economic status had relatively higher academic achievement in science whereas boys with low socio-economic status had higher academic achievement in general.

Z. J. Kovacic [13] presented a case study on educational data mining to identify up to what extent enrolment data can be used to predict student's success. They had used CHAID and CART on students of diploma college of New Zealand. They got two decision trees in their results and accuracy of classifiers obtained was 59.4 and 60.5.

4.CORRELATIONFEATURE SELECTION

Feature selection is a preprocessing step to machine learning which is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. [14]

4.1 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. [15]

Steps:

a. Subset generation: We have used four 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 5 features of the subset are selected.
d. Result validation: We have used 10-fold cross validation method for testing each classifier's accuracy.

4.2CORRELATION-BASED MEASURES

Here we shall discuss the measures used to find the goodness of a feature for classification. We find a feature to be good if it is more relevant to the class and not redundant to any other features of the class. So in short a feature should be highly correlated to the class and not much correlated to any other feature of the class. For this we have used information theory based on entropy. Entropy is a measure of uncertainty of a random variable. It can be defined by the following equation 1 as

$$H(X) = - \sum_i P(x_i) \log2 (P(x_i) \quad (1)$$

And the entropy of X after observing values of another variable Y is defined in equation 2 as

$$H(X/Y) = - \sum_j P(y_j) \sum_i P(x_i/y_j) \log2 (P(x_i/y_j)) \quad (2)$$

Here, $P(x_i)$ is the prior probabilities for all values of X, and $P(x_i/y_j)$ is the posterior probabilities of X when values of Y are given. The amount by which the entropy of X decreases reflects additional information about X provided by Y is called information gain given the equation 3 as

$$IG(X/Y) = H(X)-H(X/Y) \quad (3)$$

We can conclude that feature Y is regarded to be more correlated to feature X than to feature Z, if $IG(X/Y) > IG(Z/Y)$.
We have one more measure symmetrical uncertainty which shows correlation between features defined by equation 4 as

$$SU(X, Y) = 2 [IG(X/Y) / H(X) + H(Y)] \quad (4)$$

SU compensates information gain's bias toward features with more values and normalizes its value to range of [0,1] with 1 showing that knowledge of either one completely predicts the value of other and 0 shows that X and Y are independent. It considers pair of features symmetrically. Entropy based measures require nominal features, but they can be applied to measure correlations between continuous features as well if they are discretized properly.

5. ALGORITHM

Based on the methodology presentedbefore, wehave used the following algorithm,named FCBF(FastCorrelation- Based Filter). [9]

Input: S (F_1,F_2, F_N , C) // training data set

δ // predefined threshold value

Output: S_{best} // an optimal subset

1 begin

2 for i = 1 to N do begin

3 calculate $SU_{i,c}$ for F_i ;

4 if ($SU_{i,c} \geq \delta$)

5 append F_i to S'_{list} ;

```
6       end;
7       order S'list in descending SUi,c value;
8       Fp = getFirstElement(S'list)
9       do begin
10  Fq =getNextElement(S'list,Fp)
11          if (Fq<>  NULL)
12              do begin
13                  F 'q = Fq ;
14                 if (SUp,q ≥ SUq,c)
15                  remove Fq from S'list
16.      Fq = getNextElement(S'list F'q);
17.  else Fq  = getNextElement(S'list, Fq);
18              end until (Fq  = = NULL);
19          Fp  = getNextElement(S'list,Fp);
20      end until (FP = = NULL);
21    Sbest = S'list    ;
22   end;
```

6. PROPOSED SYSTEM

6.1 DATA PREPARATIONS

We have collected students data from a Mumbai college going to enroll in 2014 which is a training dataset consisting of information about students admitted to the first year. Data is in the excel format and has details of students personal and academic record. It has details such a s student's name, admission type, sex, marks in 12th standard, marks in math, physics, chemistry, average of all, common entrance test marks, and personal details as father's occupation, qualification, mother's qualification and occupation, interest of student.

6.2 DATA PROCESSING

Student data warehouse contains details as follows. It contains 380 instances with 32 attributes. From this list we have selected 17 attributes which we felt as relevant related to our work. Following table 1 is the list of reduced number of attributes.

Table 1: List of Attributes

1. Cat	caste	{ST,SC,OBC,OPEN}
2. A_overall	All india rank	{ 1- 1500}
3. CET	Common entrance test marks	{ 1-200}
4. Interest	Interest In taking engineering field	{YES,NO}
5. Twe_per	Percentage in 12th std	>40
6. Pcm	Total Marks in physics chemistry and maths	<=300
7. Maths	Marks in maths	<=100
8. Ad_type	Type of admission	CAP,Direct
9. Medium	Medium of study	{ENGLISH,HINDI}
10. Hostel	Students stays in hostel	{YES,NO}
11. Family size	Members in family	{1-4}
12. Family_ income	Annual income of family	{BPL,POOR,MEDUIM,HIGH}
13. Fqual	Qualification of father	{PHD,PG,UG,SECONDARY,ELEMENTARY, NO EDUCATION,NA}
14. Mqual	Qualification of mother	{PHD,PG,UG,SECONDARY, ELEMENTARY,NO EDUCATION,NA}
15. Foccu	Occupation of father	{SERVICE,BUISNESS,AGRICULTURE,RETIRED}
16. Moccu	Occupation of mother	{HOUSEWIFE,SERVICE,RETIRED}
17. Family_pr	Family pressure	{HIGH,LOW}

6.3 IMPLEMENTATION OF MODEL

WEKA is open source software which is freely available for mining data and implements a large collection of mining algorithms. It can accept data in various formats and also has converter supported with it.So we have converted the student dataset into arff file. The file was loaded into WEKA explorer. 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 weightage of each variable necessary for the success of admission of student in engineering. So we have used feature selection algorithms tests such as Info gain, Chi squared, gain Ratio. The following table 2 shows the features ranked according to the algorithm.

Table 2: Rank of features and Average rank.

	Feature	Gain ratio	Info gain	Chi-squared	Avg.
1.	Cat	0.038546	0.018912	2.449	0.835486
2.	A_overall	0	0	0	0
3.	CET	0	0	0	0
4.	Interest	0.002441	0.002329	0.3886	0.131123
5.	Twe_per	0.179664	0.143336	23.4341	7.919033
6.	Pcm	0	0	0	0
7.	Maths	0	0	0	0
8.	Ad_type	0.000685	0.000685	0.1139	0.038423
9.	Medium	0	0	0	0
10.	Hostel	0.004073	0.004007	0.6677	0.22526
11.	Family size	0	0	0	0
12.	Family_income	0.003327	0.006361	1.0603	0.356663
13.	Fqual	0.030984	0.060847	10.0359	3.37591
14.	Mqual	0.01206	0.02314	3.1626	1.065933
15.	Foccu	0.016929	0.028436	3.2978	1.114388
16.	Moccu	0.002683	0.003823	0.645	0.217169
17.	Family_pr	0.012732	0.008976	0	0.007236

So instead of trusting on any one attribute selector we have taken the average of their ranks and selected the features. So the ranking is 17,8,4,16,10. From the above table we conclude that family pressure is the most important factor for prediction of admission in engineering which is followed by admission_type, interest of student, mother's occupation, and residence in hostel. Next we have applied classification algorithms NBTree, MultilayerPerceptron, Naïve Bayes and IBK on the selected features. For this we take the subset of 3 features and then add on feature to see the accuracy of the algorithms. The below table 3 shows the evaluation criteria of features classified.

TABLE 3: EVALUATION OF CLASSIFIERS USING SUBSET OF 3, 4, 5 FEATURES (PA-Predictive Accuracy)

No. of Features	NBTree		MLP		Naïve Bayes		IBK	
	PA	Time	PA	Time	PA	Time	PA	Time
3	65	0.22	62.5	0.37	58.33	0.01	65	0.01
4	61.66	0.13	65.83	1.31	61.66	0	75	0
5	63.33	0.28	64.16	0.51	58.33	0	69.16	0
MAX	**65**		**65.83**		**61.66**		**75**	

From the table we can see that highest PA of NBTree is 65% with three features. For MLP we get highest PA of 65.83% with four features gives highest accuracy with 3 features. For Naïve Bayes we get highest accuracy of 61.66% with four features. And for IBK we get PA of 75% also with four features. Also amongst all the classifiers we conclude that IBK is the best classifier amongst all with minimum time.

Now we do the classification using the FCBF algorithm which is implemented in JAVA using net beans. FCBF is not supported by WEKA.

The following are the attributes which have been selected with their symmetric uncertainty values. The most important factor that we have found using this algorithm is family income followed by father qualification, all India rank in common entrance test. Now we apply the classifiers on the selected attributes. The following table 4 shows the classification using 3, 4, and 5 features.

TABLE 4: EVALUATION OF CLASSIFIERS USING FCBF ALGORITHM SUBSET OF 3, 4, 5 FEATURES (PA-Predictive Accuracy)

No. of Features	**NBTree**		**MLP**		**Naïve Bayes**		**IBK**	
	PA	time	PA	time	PA	time	PA	time
3	65.83	.05	75	.82	65.83	0	75	0
4	65.83	.08	81.6	1.64	66.6	0	100	0
5	75	.26	87.5	1.89	65.83	.01	100	0
MAX	**75**		**87.5**		**66.6**		**100**	

Results show that using FCBF we get the maximum accuracy by using the classifier IBK i.e. 100%. Other than that we see from the table that PA of NBTree is 75% and that of MLP is 87.5% and that of Naïve Bayes PA is 66.6 Also we get conclude that time is saved and accuracy is increased.

7. CONCLUSION

From the above results we conclude that feature selection techniques can improve the accuracy and efficiency of the classification algorithms by removing irrelevant and redundant features. Also by using the average of Infogain, gainratio, and Chi-square test we get the most relevant attributes. Four classifiers have been applied on the selected attributes. From the results we conclude that family pressure and interest of student are the most important factor for prediction of admission of student in engineering. So we get a predictive idea that the student should take or not admission in engineering. Also we conclude that amongst all selection techniques used FCBF gives the best output of relevancy of features. 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] R. Kohavi. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid" Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

[3] Baker, R.S.J.D. (2010). Data Mining for Education. In B. McGaw, P. Peterson, E. Baker (eds.), International Encyclopaedia of Education (3rd edition), (pp. 112-118). Oxford, UK: Elsevier

[4] Pathom Pumpuang, Anongnart Srivihok , Prasong Praneetpolgrang, "Comparisons of Classifier Algorithms: Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students", 1-4244-2384-2/08/ 2008 IEEE

[5] Miren Tanna, "Decision Support System for Admission in Engineering Colleges based on Entrance Exam Marks", IJCA(0975 – 8887) Volume 52– No.11, August 2012

[6] Malaya Dutta Borah, Rajni Jindal, Daya Gupta Ganesh Chandra Deka, "Application of knowledge based decision technique to predict student enrollment decision", 978-1-4577-0792-6/11 2011 IEEE

[7] Qasem A. Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-Shawakfa, "A classification model for predicting the suitable study track for school students", Vol8 Issue2/IJRRAS_8_2_15.pdf, August 2011

[8] Hany M. Harb1, Malaka A. Moustafa, "Selecting optimal subset of features for student performance model", IJCSI Vol. 9, Issue 5, No 1, September 2012, 1694-0814

[9] Lei Yu leiyu,Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", (ICML-2003), Washington DC, 2003.

[10] B. K. Bharadwaj and S. Pal. "Mining Educational Data to Analyze Students' Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp.63-69, 2011.

[11] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[12] Z. N. Khan, "Scholastic achievement of higher secondary students in science stream", Journalof Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[13] Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data",Proceedings of Informing Science & IT Education Conference 2010

[14] Blum & Langley, 1997; Kohavi &John, 1997

[15] Hall, M. (1999). Correlation based feature selection for machine learning. Doctoral dissertation, Universityof Waikato, Dept. of Computer Science.

[16] WEKA,http://www.cs.waikato.ac.nz/ml/weka, Last access, 8 April 2008.

Authors

Mital Mehta,B.E. in Computer engineering. Pursuing Mtech in software systems from Bhopal T.I.T College