Student Performance PredictionModeling: A Bayesian Network Approach

M. Ramaswami
Department of Computer Science
MaduraiKamarajUniversity
Madurai - 625021, India
mrswami123@gmail.com

R. Rathinasabapathy
Department of Computer Science
MaduraiKamarajUniversity
Madurai - 625021, India

Abstract

In recent times, prediction of student's academic performance at higher secondary level examination remains a difficult task. It requires vast amount of academic, socio-economic and other environmental factors for proposing an efficient prediction models. Moreover the selection of appropriate algorithm for model construction is also a non-trivial process. To this end, we explore the use of Bayesian networks for predicting academic performance of higher secondary students in India, based on values of socio-economic and other academic attributes. We present an analysis on the data obtained from the students of the higher secondary schools containing 35 attributes with 5650 data objects. The paper explains the application of the Bayesian approach in the field of education and shows that the Bayesian network classifier has a potential to be used as a tool for prediction of student's academic performance.

Keywords: Higher Secondary Education, Student Performance, Feature Selection, Simple Estimator, Search Algorithms, Bayesian Network Models

1. Introduction

Examinations serve a unique position as a measure of assessing the academic performance of a student. In fact, the performances of students in the examination mainly rely on three factors namely *demographic,academic-environment* and *socio-economic factors.*Measuring of academic performance of students is challenging tasks since student performance is a product of these three factors. Especially, accurate prediction of student performance is helpful in order to provide a student with the necessary assistance in the learning process.

In this study, we suggest an efficient predictive data-mining model using Bayesian Network (BN) approach for measuring the student academic performance at higher secondary level.

Accordingto Tuckman (1975), academic performance or excellence is used to label the observable manifestation of knowledge, skills, concepts, and understanding and ideas. In educational institutions, success of the student is measured by academic performance, or how well a student meets standards set out by an educators and the institution itself. As career competition grows ever fierce in the working world, the students are the focus of attention of parents, family members and teachers those who believe that good academic results will provide more career choices and job security. Although education is not the only road to success in the working world, much effort is made to identify, evaluate, track and encourage the progress of students in schools. By providing due considerations to these existing facts, it is felt that there is a need for devising a method of measuring academic performance of students at higher secondary level in order to create plans for improvement during their course of study.

We provide comprehensive Bayesian network models for the student performance at higher secondary level with different categories of class variable. Useful results are drawn concerning the prediction efficacy and the operational potential of BN models.

2. Bayesian Networks

A Bayesian Network [2][5] is a high-level representation of probability distribution over a set of features that are used for constructing a model of the problem domain. The benefit of the Bayesian Network representation lies in the way such a structure can be used as a compact representation for many naturally occurring and complex problem domains. More over, a Bayesian network model is constructed by explicitly determining all the direct dependencies between the features of the problem domain. In a Bayesian network each node represents one of the observable features of the problem domain, and the arcs between the nodes represent the direct dependencies between the corresponding variables. In addition, each node has to be provided with a table of conditional probabilities (CPT), where the variable in question is conditioned by its immediate predecessors in the network.

ISSN : 2349 - 6363

In recent years, there has been much interest in *learning* Bayesian networks from data. Learning such models is desirable simply because there is a wide array of off-the-shelf tools that can apply the learned models as expert systems, diagnosis engines, and decision support systems [4]. Learners also claim that adaptive Bayesian networks have advantages in their own right as a non-parametric method for density estimation, data analysis, pattern classification, and modeling. Using Bayesian network representation, we will have several advantages:Incorporation of prior knowledge, Validation and insight, learning causal interactions.

A number of Bayesian Network prediction models have been proposed for Student performance prediction in different academic environment. Nghe, Janecek, and Haddawy [6] discussed the accuracy of Decision Tree and Bayesian Network algorithms for predicting the academic performance of undergraduate and postgraduate students at two different academic institutions and they compared the accuracy for predicting student performance.

Pardos Z.A et all [7] employ the use of Bayesian Networks to model user knowledge and for prediction of student performance in 8th grade Mathematics within their ASSISTment online tutoring system. Xenos M.K [9] deploys Bayesian Networks for modeling the behaviour of the students of a bachelor course in computers in Hellenic Open University and achieves the high predictive accuracy of student behaviour.

Bekele and Menzel [1] used Bayesian networks to predict performance of high school students. Their model categorized students into three categories: below satisfactory, satisfactory, and above satisfactory.

In this paper, we present a BN model with different search algorithms which is different from what is already presented in the literature. We use, socio-economic, personal and other native domain knowledge attributes to determine the level of performance by employing Bayesian network modeling technique.

3. Data Source

In this work, the main source of data for the study is complied by collecting student's key demographic details, family details, socio-economic details, and previous academic performance at secondary level from different schools.

The above details were collected from the students who appeared for their final examinations in April 2009, in person by a questionnaire. The marks obtained by these students at their final examinations were collected from the concerned Chief Education Officer (CEO). A total of 6000 data sets were collected from three educational districts of Tamilnadu based on the average literacy rate of Tamilnadu. After cleaning, we had 5650 data sets and they have been used for model construction. An essential step in prediction model construction is selecting the features used for classification. The features, as shown in Table-1, are considered for BN model construction. The dependent variable is HScGrade - percentage of marks obtained in higher secondary examination is taken as class variable.

Table 1Features used in performance prediction

Variable Name	Description
(1) SEX	student's sex
(2) BMI	student's body mass index
(3) VAcuity	student's eye visual acuity
(4) COMM	student's community
(5) PHD	physically handicapped or not
(6) FHBT	student's food habit
(7) FAM-SIZE	student's family size
(8) LArea	student's living area
(9) No-EB	number of elder brothers
(10) No-ES	number of elder sisters
(11) No-YB	number of younger brothers
(12) No-YS	number of younger sisters
(13) JIFamily	student's family status
(14) TranSchool	mode of transportation to school
(15) Veh-Home	own vehicle at home
(16) PSEdu	student had primary education
(17) EEdu	type of institution at elementary level
(18) StMe	type of secondary syllabus
(19) XMARK-P	percentage of marks obtained at secondary level
(20) MED	medium of Instruction
(21) PTuition	private tuition- number of subjects
(22) GROUP	group of study
(23) TYP-SCH	type of school
(24) LOC-SCH	location of school
(25) SPerson	sports/athletic
(26) SpIndoor	type of indoor game
(27) SpOutdoor	type of outdoor game
(28) CStudy	care of study at home
(29) FEDU	father's education
(30) FOCC	father's occupation
(31) FSAL	father's monthly income
(32) MEDU	mother's education
(33) MOCC	mother's occupation
(34) MSAL	mother's monthly income
(Response Variable) (35) HScGrade	marks/grade obtained at HSc Level

Classification and Prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict data trends. Since HScGrade is a class variable, the feature selection as well as student performance prediction models are extensively studied by varying the number of cases of class variable – ***HScGrade****viz.* two-case (***pass, fail***), three-case (***very-good, good, poor***), five-case (***excellent, very-good, good, fair, poor***) and seven-case values(***O***, ***A***, ***B***, ***C***, ***D***, ***E***, ***F).*** The labeling of class values have

been fixed based on the marks obtained at their higher secondary examinations as follows:

i) for two cases problem , "*pass*" for all passed students with 40% and above, and "*fail*" for all failure students below 40% of marks;

ii) for three case problem, "*very good*" representing marks from 80% to 99%", "*good*" representing marks from 60% to 79%,"poor" representing marks from 40% to 59%, and failed students;

iii) for five case problem, "excellent" indicating marks 90% and above, "*very good*" representing marks from 75% to 89%", "*good*" representing marks from 60% to 74%, "fair" representing marks from 50% to 59%, and "poor" indicating marks below 50% and failed students;

iv) for seven case problem, "***O"*** representing *marks 90% to 100%,* "***A"*** means marks from80% - 89%, ***B*** representing marks from 70% - 79%, ***C*** representing marks from 60% - 69%, ***D*** indicating marks from 50% - 59%, ***E*** indicating marks from 40% - 49%, and ***F*** representing marks below 40%.

4. Experimental Setup and Results

The belief network modeling software employed for the purpose of the experiment is the Bayesian Network in WEKA [10]. Building a selective Naive Bayesian classifier involves singling out the feature variables that best serve to separate the different classes under study. The selection of appropriate feature variables generally is based on data.

We have used all three types feature selection techniques and arrived the conclusion that hybrid based feature selection method outperformed over other two feature selection approaches[8]. The study also reveals that the rank based information gain (ING) method performed well on four different class values of the class variable - HScGrade. We have used three rank based feature selection methods viz., Consistency subset Evaluation (CFS), Chi-Square based attribute evaluation (CHI) and Information Gain Attribute Evaluation(ING). Based on the Peak value of ROC and F1-Measure values we have chosen best subsets for a given cardinality.

The ING method attains its maximum ROC value with 9 top ranked attributes for two class variable HScGrade, whereas ING method with 13 top ranked attributes yields highest F1-Measure for three class variable HScGrade. The ING method also exhibits better performances on both five class and as well as seven class variableHScGrade with top 19 and 23 ranked attributes respectively.

We use a simple **Select Estimator** algorithm for finding the conditional probability tables of the Bayes Network. This estimator along with various search algorithms like Hill Climbing (HC), K2 (K2), LAGD Hill Climbing (LC), Repeated Hill Climbing (RC), TabuSearch (TS) and Network Augmented with Tree (TN) are usedfor constructing Bayesian Network structure..

We have trained the Bayesian Network with 2, 3, 5 and 7 categories of class value with 10-fold cross validation method. Cross-validation is a standard statistical technique, in which, certain amount of data is reserved for testing and uses the remainder for training. It is customary, to predicting the accuracy of the classifier models on given a single, fixed sample of data is to use stratified 10-fold cross-validation as suggested by Frank et al in [11]. Table-2illustrates the predictive accuracy of BN models with different search algorithms against different class values of class variable – HscGrade.

Acomparison of the percentage of predictive accuracy of all six algorithms , as shown in Table-2, reveals that, BN models with TAN search algorithm achieves better performance over others types of search algorithms for all types of class values as given in figure-1.

Table 2 BN Models prediction accuracy in percentage

Search Algorithm	Predictive Accuracy			
	2-class (IG-9)	3-class (IG-13)	5-class (IG-19)	7-class (IG-23)
HillClimber (HC)	82.2575	59.1647	40.214	31.4981
K2	82.2575	59.1647	40.214	31.0321
LAGDHill Climber(LC)	84.5713	59.7169	40.7318	33.8799
RepeatedHill Climber(RC)	82.2575	59.1647	40.7318	31.4981
TabuSearch (TS)	84.8809	60.5627	52.1401	31.4636
TAN(TN)	84.9154	63.3069	52.1401	42.3196

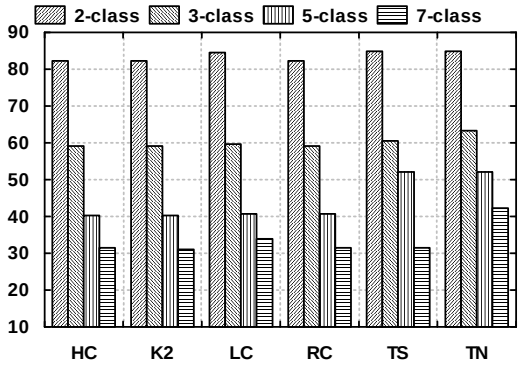

Fig.1. Prediction Accuracy of BN Models

The Bayesian Network structure with higher predictive accuracy obtained through TAN search algorithm for 2-class, 3-class, 5-class and 7-class are depicted in figure 2(a) – 2(d). These network structures show casual relationships between the class variable HScGrade with demographic, previous

academic performance and socio-economic factors. Also it reveals that the class variable HScGrade has strong dependency on i) marks obtained at secondary level (XMARK-P), Type of transportation to school (TranSchool), medium of instruction (MED), sibling structure (number of brothers and sisters) and economic status of the family.

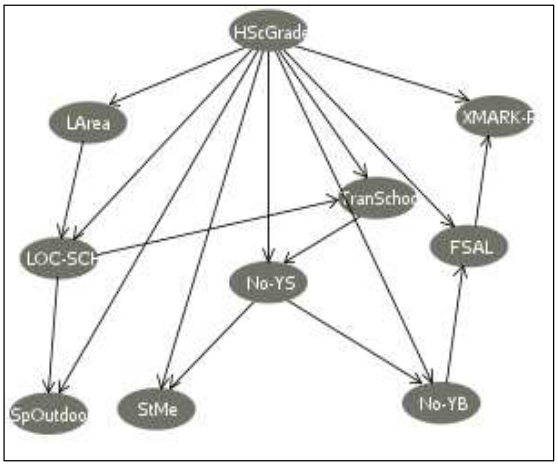

Fig.2(a). Bayesian Network model for 2-class

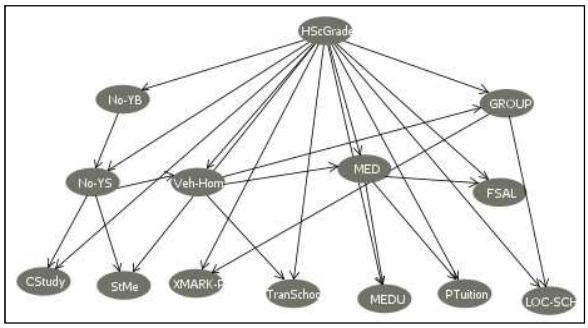

Fig. 2(b). Bayesian Network structure 3-class

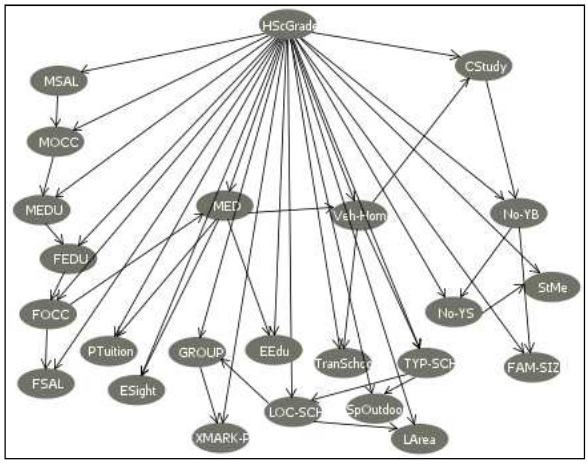

Figure 2(c). Bayesian Network structure for 5-class

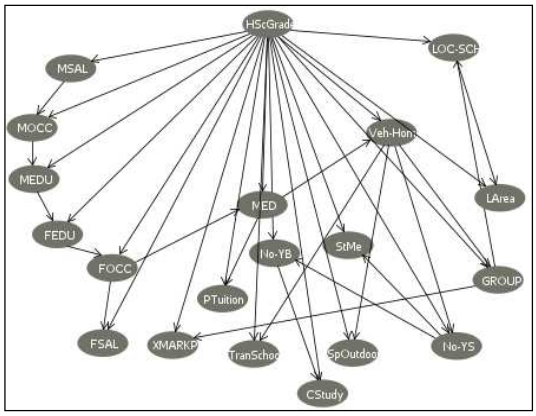

Figure 2(d). Bayesian network structure for 7-class

5. Conclusion

The main objective is to construct BN model to analyse the casual relationship between values of socio-economic and other academic factors with students academic performance at higher secondary level. Then enhance the prediction of academic performance by exploring the dependence between them. Better Bayesian Classifier models with different categories of class values are achieved for predicting student performance. The predictions of student‘s academic performance can be useful in many contexts. For admissions, it is important to be able to identify excellent students for allocating scholarships and fellowships, as well as getting the desired groups at the higher secondary level. The BN model provides a robust mechanism to predict the academic performance of higher secondary students with better predictive accuracy.

References

[1] R. Bekele, and W. Menzel, “A Bayesian Approach to Predict Performance of a Student (BAPPS): A Case with Ethiopian Students”, in Proceedings of International Conference on Artificial Intelligence and Applications.453 (191), 2005.

[2] N.Friedman, D. Geiger, and M. Goldszmidt, Bayesian Network Classifiers, Kulwer Academic Publishers, Boston.

[3] D. E. Goldberg, Genetic algorithms in search, optimization and machine learning, An imprint of Pearson Education, 2000.

[4] D. Heckerman. “A Bayesian approach for learning causal networks”, in Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QU, pages 285-295. Morgarti Kaufmann.1995.

[5] F. Jensen ., “Bayesian Networks and Decision Graphs”, Springer-Verlag, 2002.

[6] N. T Nghe, P. Janecek, and P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance". Paper presented at 37th ASEE/IEEE Frontiers in Education Conference, October 10 – 13, 2007.

[7] Z. A Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan, "Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks", Workshop in Educational Data Mining held at the Eight International Conference on Intelligent tutoring Systems. Taiwan. 2006.

[8] M. Ramaswami. and R. Bhaskaran., "A study on feature selection techniques in educational data mining," Journal of Computing vol. no. 11, pp. 7-11, 2009.

[9] M. K. Xenos, "Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks", Computers & Education, Vol. no. 43(4), 345- 359, 2004.

[10] Weka 3.5.6. An open source data mining software tool developed at university of Waikato, New Zealand. Retrieved from: http://www.cs.waikato.ac.nz/ml/weka/

[11] I. H. Witten,., and E. Frank,. Data mining – Practical Machine Learning tools and Techniques (2nd Ed.). San Francisco,CA: Morgan Kaufmann Publisher., An imprint of Elsevier. 2005