Hybrid CHAID a key for MUSTAS Framework in Educational Data Mining

G.Paul Suthan [1] and Lt.Dr. Santhosh Baboo [2]

**[1] Head, Department of Computer Science, Bishop Appasamy College
Race Course, Coimbatore, Tamil Nadu 641018, India**

**[2] Reader, PG and Research Department of Computer Application, DG Vishnav College
Arumbakkam, Chennai 600106, Tamil Nadu, India**

Abstract

Currently there is an increased interest in Educational Data Mining due to the compelling need for quality in higher education and the need to know student behavioural pattern to cater individual needs. The performance prediction of student kind model is quite familiar and mostly it is associated with academic performance. Our proposed framework Multi Dimensional Student Assessment (MUSTAS) has unique feature to measure the student's performance through multidimensional attributes. Each dimension and its associated factors are carefully designed to predict the student's behaviour. We propose the Hybrid CHAID algorithm , a combination of CHAID and Latent Class Modeling (LCM) as the best matched technique for our MUSTAS framework in educational data mining.

Keywords: *Data Mining, Educational Data Mining, CHAID Prediction Model, Latent Class Model.*

1. Introduction

Educational system as of now, especially in India is going through a radical transformation due to the efforts taken by UGC and HRD ministry. The reason behind this is that, quality of education is not met in higher educational Institutions. Due to this many Institutes want to be centre of excellence by going through accreditation process, such as ISO etc., to enhance their quality of education.

Many affiliated institutions want to become autonomous and in due course to become unitary university; thereby enabling them to have more freedom in syllabus and course selection. This also gives them flexibility to have tie ups with foreign Universities. The private institutions are on an increase now to cater to the growing population of youth in countries like India where population growth is high. The parents on the other hand are now looking for quality education so as to enable their child to be placed in good Multi-National Companies.

The private institutions are now having no choice of selection of students in the entry level due to enormous new institutions coming up every year in higher education. This contributes to the low calibre of students in the entry level and making the faculty to take enormous efforts to cater to these students. The faculty with station seniority is also on a decline in private institutions due to change over for higher salary or migrating to Government or other lucrative jobs. On the present scenario the institution has no choice but to have quality in education for attracting students. Therefore to meet the quality needs of the institution, the staffs have to know the behaviour pattern of the students in shorter time, so as to give coaching in accordance to their specific need.

The student needs were earlier known by knowing the students personally or through some response mechanism like feedback. Later statistical methods were used to analyse these behaviour. Now with the advent of data mining techniques and tools, this process of finding patterns in the behaviour of students can be taken to a next higher level. The tools and algorithms used exclusively for educational purpose in data mining is categorised as Educational Data Mining (EDM).

2. EDM for Higher Education

Data mining is finding hidden patterns in a large collection of data. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [2]. Mining in educational environment is called Educational Data Mining. Han and Kamber [15] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the

mining process. New methods can be used to discover knowledge from educational databases. Student data can be used to analyze trends and behaviors toward their education [2]. Lack of deep and adequate knowledge in higher educational system may prevent management to achieve quality objectives, data mining methodology can help bridging this knowledge gaps in higher education system.

Traditional classroom environments are being widely used. Here face to face contact is established between the teacher and the student. Johnson.S, Arago Shaik and Palma-Rivas [18] says that educations are of different types as public, private, elementary, primary, adult, higher, tertiary and academic education. Most of these types uses passive learning and ignore individual differences. They also sometimes do not cater to the need of students. Here the teachers monitor the student learning process by analyzing the paper records and on observation.

3. Related Studies in EDM

Educational data mining has emerged as an independent research area in recent years, culminating in 2008 with the establishment of the annual International Conference on Educational Data Mining, and the Journal of Educational Data Mining. Romero and Ventura [30] provides a comprehensive study of EDM from 1995 to 2005. It describes the need for analyzing the student data which can be used by students, educators and administrators.

Galit [11] developed a system to warn weak students. Han and Kamber [15] discovered relationship among data. Henrik [16] found hidden relationships. Walters and Soyibo [21] discovered relationship between academic performance and nature of their schools. Z.N. Khan [36] found Girls with high socio-economic status were relatively higher achievers in science stream and boys with low socio-economic status were relatively higher achievers in general. Hijazi and Naqvi, [33] using regression found factors like mother's education and student's family income were highly correlated with the student academic performance. A.L Kristjansson, Sigfusdottir and Allegrante [1] found that Body Mass Index (BMI) affects higher academic achievement. Moriana et al. [17] used Analysis of variance (ANOVA) and it was observed that group involved in activities outside the school yielded better academic performance. Al-Radaideh, et al. [29] prescribed Decision Tree model had better prediction than other models. Cortez and Silva [25] found Decision Tree and Neural Networks in some areas give same accuracy. Gong, Rai, Beck and Heffernan [12] found Impact of self decipline on learning co-related with higher incoming knowledge and fewer mistakes but the actual impact of learning was only marginal. Perera et al. [28] got the Big 5 theory for teamwork as a driving theory to search for successful patterns of interaction within student teams. Madhyastha and Tanimoto [21] investigated the relationship between consistency and student performance with the aim to provide guidelines for scaffolding instruction. Beck and Mostow [6]; Pechenizkiy et al. [27] discovered which types of pedagogical support are most effective, either overall or for different groups of students or in different situations. McQuiggan et al. [24], found whether students are experiencing poor self-efficiency. Baker [3] identified students who are off-task. D'Mello et al. [8] studied on students who are bored or frustrated. Dekker et al. [7] Romero et al. [31]; Superby et al. [34] found factors that predict student failure or non-retention in college courses. Barnes [5] developed algorithms which automatically discover a QMatrix from data. Desmarais & Pu [9] and Pavlik et al [26] developed algorithms for finding partial order knowledge structure (POKS) models that explain the interrelationships of knowledge in a domain. Walters and Soyibo [35] conducted a study to determine Jamaican high school students and found positive significant relationship between academic performance of the student and the nature of the school. Ryan S.J.D. et al. [32] explore that prediction and discovery model are increasing while relationship mining are not used much.

4. CHAID Prediction Model

Chi-squared Automatic Interaction Detection (CHAID) analysis which was first proposed by Kass, 1980[10] is one of post hoc predictive segmentation methods. The CHAID, using of decision tree algorithms, is an exploratory method for segmenting a population into two or more exclusive and exhaustive subgroups by maximizing the significance of the chi-square, based on categories of the best predictor of the dependent variable. Segments obtained from CHAID analysis are different from cluster type models because the CHAID method, which is derived to be predictive of a criterion variable, is defined by combinations of predictor variables as described by Magidson, [22]. CHAID technique depends on interactions among the independent variables, finding those that explain the greatest differences within the dependent variable. Thus, a CHAID decision tree demonstrates how the predictors are differently formed and predicts a dependent variable that shows nominal and continuous scaling. Educators can identify the key influencers or significant drivers in certain students using CHAID analysis, which results in a tree like diagram commonly called a decision tree. Decision trees have several advantages as explained by Bakken [4]. The type of representation makes the resulting classification model easy to use. Moreover, decision trees are suited for

exploratory knowledge discovery because they are non-parametric and make no assumptions about the underlying probability distribution. Decision trees are also efficient to higher-order interactions. They are relatively quickly constructed for large datasets compared to other classification models as presented by Magidson and Vermunt, [23]

5. Latent Class Modeling- LCM

Latent class (LC) modeling was initially introduced by Lazarsfeld and Henry [19] as a way of formulating latent attitudinal variables from dichotomous survey items. In contrast to factor analysis, which posts continuous latent variables, LC models assume that the latent variable is categorical, and areas of application are more wide-ranging. The methodology was formalized and extended to nominal variables by Goodman [13,14], who also developed the maximum likelihood (ML) algorithm that serves as the basis for many of today's LC software programs. In recent years, LC models have been extended to include observable variables of mixed scale type (nominal, ordinal, continuous and counts), covariates, and to deal with sparse data, boundary solutions, and other problem areas.

6. MUSTAS Framework

The Multidimensional Students Assessment (MUSTAS) framework is a novel model, which consist of demographic factors, academic performance of the student and dimensional factors. The dimensional factors has further sub divided into three dimensions respectively self assessment, institutional assessment and external assessment. The main objective of this framework is to identify the contribution of selected dimensions over academic performance of the student, which helps to teachers, parents and management about the student's pattern. Understanding of the pattern may facilitate to redefine the education method, additional care on weakness, and promoting their abilities. The academic performance shows the present ability of the student and the demographic factors shows his personal lifestyle. Construction of this framework strongly believes each aspect considered for this framework is closely related to one another.

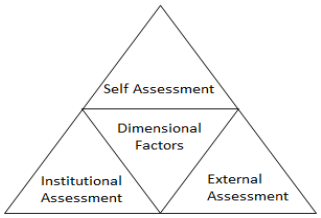

Fig.1 Dimensions of MUSTAS

The dimensional factor helps to measure the student's attitude. Self assessment is measured through five questions, which express their personal interest towards studies. Institutional assessment is specially designed for lecturers/faculty and institution's support towards studies. The third dimension is external assessment, which is designed to measure an external attribute contribution towards their studies.

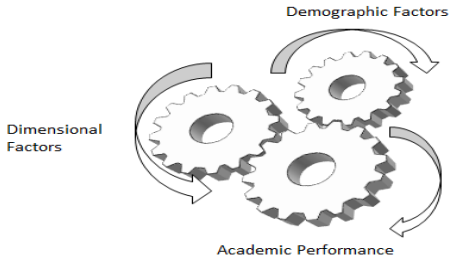

Fig.2 MUSTAS Framework

7. Proposed Model

CHAID based Performance Prediction model in EDM was analysed by Ramaswamy, [20] and the results have proved to be accurate when compared to some other models in terms of accuracy in prediction. One limitation of CHAID is that segments are defined based on a single criterion variable. Given situations where multiple criteria exist, it is not clear how one should go about obtaining a single common segmentation. Using one dependent variable as the criterion may result in one set of segments, while use of an alternative dependent variable will likely yield a different set of segments. Moreover, the categories of a predictor may merge in different ways depending upon which dependent variable is used, again leading to different segments.

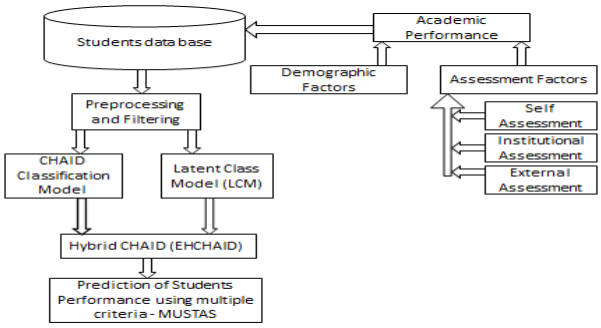

Fig.3 Hybrid CHAID in Educational Data Mining

In addition, when multiple dependent variables do exist, they may be of different scale types (nominal,

ordinal, continuous, count, etc.). Using a 3-category response variable as an example Magidson [22] showed that CHAID segments resulting from treating the dependent variable as ordinal (using profitability scores for the categories) differed substantially from segments derived from the nominal algorithm which ignored the scores. The hybrid approach resolves the need to choose between different segmentations because indicators with differing scale types can be used in extended LCMs, yielding a single LC solution. An important advantage of this hybrid approach over approaches based on specific measures for node homogeneity rather than a model is that the LC model used here can handle dependent variables of different scale types.

The evaluation of student attitude is important to predict the academic performance on 3-dimensions (Self Assessment, Institutional Assessment and External Assessment). A LCM was fit to these data, using academic performance as an active covariate and the eight demographic factors as inactive covariates. This model may be viewed as a kind of unsupervised regression with 12 dependent variables, plus the 11 attribute ratings. This LCM yielded 3 segments. The segments are Good, Average and Poor with respect to the attribute ratings and in their feedback as percentage. These percentages are displayed in the root node of the hybrid CHAID tree. The hybrid CHAID used the 3-category latent variable (segments) as the dependent variable and again utilized the 8 demographics as the predictors.

8. Conclusion

We believe the academic performances of the students are not always depending on their own effort. Our investigation shows that other factors have got significant influence over student's performance. Hence, we introduce the hybrid CHAID with MUSTAS framework in education domain as multidimensional evaluation method to classify the pattern of student through classification tree. This proposal will improve the insights over existing methods.

References

[1] A. L. Kristjansson, I. G. Sigfusdottir, and J. P. Allegrante, "Health Behavior and Academic Achievement Among Adolescents: The Relative Contribution of Dietary Habits, Physical Activity, Body Mass Index, and Self-Esteem", Health Education &Behavior, (In Press).

[2] Alaa el-Halees, 2009 Mining Students Data to Analyze e-Learning Behavior: A Case Study.

[3] Baker, R.S.J.D., 2007. "Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems." In Proceedings of the ACM CHI 2007: Computer-Human Interaction conference, pp1059-1068.

[4] Bakken,S.K., 2005. "Use of chi-squared automatic interaction detector in the prediction of vocational rehabilitation outcomes among veterans with substance use disorders". Doctoral dissertation, Univeristy of Wisconsin-Madison.

[5] Barnes, T., 2005. "The q-matrix method: Mining student response data for knowledge." In Proceedings of the AAAI-2005 Workshop on Educational Data Mining.

[6] Beck, J.E. and Mostow, J., 2008. "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students." In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, pp353-362.

[7] Dekker, G., Pechenizkiy, M. and Vleeshouwers, J., 2009. "Predicting Students Drop Out: A Case Study." In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, T. Barnes, M. Desmarais, C. Romero and S. Ventura Eds.,pp41-50.

[8] D'mello, S.K., Craig, S.D., Witherspoon, A.W., McDaniel, B.T. and Graesser, A.C., 2008. "Automatic Detection of Learner's Affect from Conversational Cues." User Modeling and User-Adapted Interaction vol 18. pp45-80.

[9] Desmarais, M.C. and Pu, X., 2005." A Bayesian Student Model without Hidden Nodes and Its Comparison with Item Response Theory. "International Journal of Artificial Intelligence in Education vol 15,pp291-323.

[10] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Applied Statistic, Vol. 29, 1980, pp. 119-127.

[11]Galit.et.al.,2007."Examining online learning processes based on log files analysis" : a case study.Research, Refelection and Innovations in Integrating ICT in Education.

[12]Gong, Y., Rai, D., Beck, J. and Heffernan, N. 2009. "Does Self-Discipline Impact Students' Knowledge and Learning? "In Proceedings of the 2nd International Conference on Educational Data Mining, pp61-70.

[13]Goodman, L.A..,1974a." Exploratory latent structure analysis using both identifiable and unidentifiable models." Biometrika,vol61,pp215-231.

[14]Goodman, L.A.,1974b." The analysis of systems of qualitative variables when some of the variables are unobservable". Part I: A modified latent structure approach, American Journal of Sociology, vol79,pp1179-1259.

[15]Han,J. and Kamber, M., 2006. "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[16]Henrik (2001) Clustering as a Data Mining Method in a Web-based System for Thoracic Surgery: © 2001

[17] J. A. Moriana, F. Alos, R. Alcala, M. J. Pino, J. Herruzo, and R. Ruiz, "Extra Curricular Activities and Academic Performance in Secondary Students", Electronic Journal of Research in Educational Psychology,Vol. 4, No. 1, 2006, pp35-46.

[18]Johnson, S., Arago, S., Shaik, N., & Palma-Rivas, N. (2000). "Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments." Journal of Interactive Learning Research, 11(1), pp29–49.

[19] Lazarsfeld, P.F., and Henry, N.W.,1968." Latent Structure Analysis. "Boston: Houghton Mill.
[20]M. Ramaswami and R. Bhaskaran." A CHAID Based Performance Prediction Model in Educational Data Mining." Madurai Kamaraj University, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010
[21]Madhyastha.T.and Tanimoto, S., 2009." Student Consistency and Implications for Feedback in Online Assessment Systems."In Proceedings of the 2nd International Conference on Educational Data Mining, pp81-90.
[22]Magidson, J.,1994." The CHAID approach to segmentation modeling: Chi-squared automatic interaction detection." In R. P. Bagozzi (Ed.), In advanced methods of marketing research. Cambridge, MA: Blackwell.pp118-159.
[23] Magidson, J., &Vermunt, J. K.,2005. "An extension of the CHAID tree-based segmentation algorithm to multiple dependent variables." In Weihs C, Gaul W (eds), Classification: The Ubiquitous Challenge. Springer:Heidelberg.
[24]Mcquiggan, S., Mott, B. and Lester, J. 2008." Modeling Self-Efficacy in Intelligent Tutoring Systems: An Inductive Approach. "User Modeling and User-Adapted Interaction 18, pp81-123.
[25]P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp5-12.
[26] Pavlik, P., Cen, H. and Koedinger, K.R., 2009." Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models." In Proceedings of the 2nd International Conference on Educational Data Mining, pp121-130.
[27]Pechenizkiy, M., Calders, T., Vasilyeva, E. and Debra, P., 2008. "Mining the Student Assessment Data: Lessons Drawn from a Small Scale Case Study." In Proceedings of the 1st International Conference on Educational Data Mining,pp187-191.
[28]Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaiane, O., 2009. "Clustering and sequential pattern mining to support team learning." IEEE Transactions on Knowledge and Data Engineering vol21, pp759-772.
[29]Q. A. AI-Radaideh, E. M. AI-Shawakfa, and M. I. AI-Najjar, "Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.
[30]Romera, C. and Ventura, S., 2007." Educational Data Mining: A Survey from 1995 to 2005." Expert Systems with Applications 33, 125-146.
[31]Romero, C., Ventura, S., Eapejo, P.G. and Hervas, C., 2008." Data Mining Algorithms to Classify Students." In Proceedings of the 1st International Conference on Educational Data Mining, pp8-17.
[32]Ryan S.J.D. Baker and Kalina Yacef." The State of Educational Data Mining in 2009: A Review and Future Visions"
[33]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.
[34]Superby, J.F., Vandamme, J.-P. and Meskens, N., 2006. "Determination of factors influencing the achievement of the first-year university students using data mining methods." In Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), pp37-44.
[35]Y. B. Walters, and K. Soyibo, "An Analysis of High School Students' Performance on Five Integrated Science Process Skills", Research in Science & Technical Education, Vol. 19, No. 2, 2001, pp133-145.
[36] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp84-87.

G. Paul Suthan has done his Under-Graduation and Post-Graduation at Bishop Heber College, affiliated to Bharathidasan University and Master of Philosophy at Manonmaniam Sundaranar University. He is currently pursuing his Ph.D in Computer Science in Dravidian University, Kuppam, Andhra Pradesh. Also, he is working as the Head of the Department of MCA, Bishop Appasamy College of Arts and Science, Coimbatore, affiliated to Bharathiar University. He has organized various National and State level seminars, and Technical Symposium. He has participated in various National conferences and presented papers. He has 14 years of teaching experience. His research areas include Data Mining and Artificial Intelligence.

Lt.Dr.S.Santhosh Baboo, aged forty two, has around twenty years of postgraduate teaching experience in Computer Science, which includes Six years of administrative experience. He is a member, board of studies, in several autonomous colleges, and designs the curriculum of undergraduate and postgraduate programmes. He is a consultant for starting new courses, setting up computer labs, and recruiting lecturers for many colleges. Equipped with a Masters degree in Computer Science and a Doctorate in Computer Science, he is a visiting faculty to IT companies. It is customary to see him at several national/international conferences and training programmes, both as a participant and as a resource person. He has been keenly involved in organizing training programmes for students and faculty members. His good rapport with the IT companies has been instrumental in on/off campus interviews, and has helped the post graduate students to get real time projects. He has also guided many such live projects. Lt.Dr. Santhosh Baboo has authored a commendable number of research papers in international/national Conference/journals and also guides research scholars in Computer Science. Currently he is Reader in the Postgraduate and Research department of Computer Science at Dwaraka Doss Goverdhan Doss Vaishnav College (accredited at 'A' grade by NAAC), one of the premier institutions in Chennai.