Predicting Student Performance Using Enrollment Figures And Background Information

literature Review

Harmony Mncube
1272371

April 4, 2019

Supervised by Dr. Ritesh Ajoodha and Dr. Ashwini Jadhav

Contents

1 Introduction **3**
1.1 Introduction 3

2 Background **4**
2.1 Introduction 4
2.2 Domain Complexity 5
2.3 The Student Model 5
2.3.1 Levels of error analysis 5
2.3.2 The Student Model Construction 6
2.4 Logic Based Algorithms 6

3 Conclusion **8**
3.1 Summary 8

1 Introduction

1.1 Introduction

Increasing student retention or persistence is a long term goal in all academic institutions. [Kovacic, 2010] highlights the importance of student attrition for students, academic and administrative staff. After all different student apply different strategies in getting through their academic program. From [Kovacic, 2010] implies that traditionally, logistic regression and discriminant analysis are frequently used in retention studies of students academic success or failures.However, logistic regression is bias in large data and is unreliable [Kovacic, 2010].

This has led to rather highly uncertainty to decide the outcome for a students[Sison and Shimura, 1998].Therefore, in order to constructing of a qualitative representation that accounts for student behavior in terms of existing background knowledge about a domain and about students learning the domain ,it requires a model that collaboratively adapts to specific aspects of student behavior. Looking back in the last 15 years educational data mining[Kotsiantis, 2007; Al-Radaideh et al., 2006] emerged as a new application area for data mining[Kovacic, 2010], becoming well established with its own journal [Kotsiantis, 2007] which is the most significant in this research.

Good data results in good results otherwise bad data results in inconsiderable decision making thus using supervised machine learning in labelled data as suggested by [Kotsiantis, 2007] with the use of reinforcement learning and the measure of how well the system operates to help with the validation of students performance progress. Some aspects for data mining applications is in education,enrollment management, graduation, academic performance,gifted education, web-based education, retention and other areas[Kovacic, 2010]. We may rely on student participation on traceable systems [Minaei et al., 2003] to allow grouping of students based on similar trends but supervised machine learning techniques are not that easy in constructing intricate student models. In this case, features student model, the student behavior, and the background knowledge were considered but however, we can not rely only on background characteristics alone[Kovacic, 2010] since there are factors that may influence student progress.

Research on learning is actually made up of diverse sub-fields[Quinlan, 1986]. At one extreme there are adaptive systems that monitor their own perfor-

mance and attempt to improve it by adjusting internal parameters[Quinlan, 1986]. This approach, characterize a large proportion of the early learning work that produced self-improving programs for solving problems [Quinlan, 1986] and many other domains. A quite different approach sees learning as the acquisition of structured knowledge in the form of concepts or production rules. While the typical rate of knowledge elucidation by this method is a few rules per man [Quinlan, 1986] [Sison and Shimura, 1998], and an expert system for a complex task may require hundreds or even thousands of such rules. [Sison and Shimura, 1998] used machine learning to automatically extend the background and induce the student model,base on manual, tedious, time consuming and error-prone analysis of student protocols

[Kovacic, 2010] used ASSISTANT to further generalizes on the integer-valued attributes of ACLS by permitting attributes with continuous (real) values. To avoid classes being disjoint ,[Kovacic, 2010] used ASSISTANT to allow them to form a hierarchy, so that one class may be a finer division of another. This simplifies out dataset to only focus on the important features rather than spending more time on irrelevant or unnecessary data. ASSISTANT does not form a decision tree attractively in the manner referred by [Ogunde and D.A., 2014], but does include algorithms for choosing a 'good' training set from the objects available including ASSISTANT that has been used in several medical domains with promising results including TDIDT that will be discussed later.

2 Background

2.1 Introduction

This paper explores the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course program and course block), that may influence persistence or failure of students but our main focus will be on the enrolment figures . Training examples of the form $f(x_i, y_i)$ for x $\in \mathbb{R}^{\times}$ for some unknown function y=f(x).whose components are discrete- or real-valued such as age ,race , live and so on will be included. The y values are typically drawn from a discrete set of classes y =1,..., K where $y_i \in \mathbb{R}$ represents the outcome for a student(0=PCD,1=RET,etc)in the case of classification. Using x_{ij} to represent the j-th feature of x_i.If f is drawn from H according to P (h), then the Bayesian voting scheme is optimal [Dietterich, 2000]. PIXIE and ASSERT validates the output's accuracy from Bayesian voting .The model consist of

numerical value for easy use of data set,string are converted to numerical values and stored on a separate dataset(the important background information relating to the student is usually in string format ,e.g. Religion). The error of the model is monitored closely to see if we not drifting away from the expected results. Then the constructed student model classifies and predict outcome of students based on both enrollment figures,past test and or other results including background computed information . From [Dietterich, 2000], ID3 and ASSISTANT including ACLS are then used in computing valid outputs based on the data structures per attributes.Lastly the decision tress are used for final output through methods like TDIDT, CLS based on NP-Complete.

2.2 Domain Complexity

Complexities of domain and domain tasks are usually alluded because the concept of learning or classification task in the sense of attacking problems requires the classification ability. Complex linear student models are easily solved by PIXIE compared to ASSERT which models students' classification represented as vectors. Therefore, viewing these as a spectrum or possible multidimensional[Sison and Shimura, 1998] e.g. with mathematics and programming can allow easy manipulation.

2.3 The Student Model

The primarily qualitative representation of student knowledge about a particular domain by [Sison and Shimura, 1998] somehow accounts for student behaviour.By being qualitative, we simple mean that is is either numerical(information in quantities; that is, information that can be measured and written down with numbers). This model can only account for computational utility rather than in cognitive fidelity[Sison and Shimura, 1998].

2.3.1 Levels of error analysis

Relationship between the actual and desired behaviours are determined , these discrepancies behavior or behavioral-level errors are named after incorrect behaviour based on their importance when dealing with simple integer behavior. It then becomes non-trivial when we trade our interests to less valuable task[Sison and Shimura, 1998](more complex behaviours like programs) resulting in the significance of this knowledge level meaning that

behavioral error can also be due to inconsistency or and insufficient knowledge[Quinlan, 1986],[Sison and Shimura, 1998].Therefore ,all information is important[Quinlan, 1986] because it gives a statical picture .

Usually student models are concerned with the error at learning level[Sison and Shimura, 1998],this allows the model to know what it is dealing with for better prediction accuracy. Since our model is to predict students' outcome, therefore, having a strong model can bring about accuracy in model predictions.

2.3.2 The Student Model Construction

There are two approach that seem rather considerable compared to other methods, [Sison and Shimura, 1998] looked at background knowledge to transform the student behaviour so that he can quantify the relationship between the student behaviour and the problem given if they have some common relationship to classify the student progress so as to create a strong model. The second approach [Sison and Shimura, 1998] used was to synthesis elements from the background knowledge or input data, taking into account the construction of the student model from a single behaviour(e.g. Sakai) or analytic approach, while taking a close look at system that construct their student models from multiple behaviour. These two methods reduces data (removing redundant and irrelevant features) to allow data mining algorithms to function and work effectively because the training examples may be corrupted by some random noise.

The background knowledge feature is a set of discrete, mutually exclusive values. These values are calculated based on information gathered about a specific student such as religion , interests , participation etc using the ASSISTANT approach.

2.4 Logic Based Algorithms

[Kovacic, 2010] highlighted the used key demographic variables and assignment marks in the supervised machine learning algorithms (decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines) to predict student's performance. His decision trees classified instances by sorting them based on fea-

tures with high information gain at root node.NP-complete was then used as their appropriate way of constructing binary decision trees optimally. [Kotsiantis, 2007] emphasis starting from the root node ,with the feature that best describes the data using both information gain and gini index ,with the aid of myopic [Quinlan, 1986] to measure estimates of each attribute independently proceeding down down to the leaves. The same procedure is adopted by [Kotsiantis, 2007] for each partition of the divided data,creating sub-trees until the training data is divided into subsets of the same class .

Both TDIDT and CLS systems attempts to minimize the cost of classifying an object. This cost has components of two types: the measurement cost of determining the value of property A exhibited by the object, and the misclassification cost of deciding that the student belongs to class RET when its real class is PCD. CLS uses a look-ahead strategy similar to minimax. At each stage, CLS explores the space of possible decision trees to a fixed depth, chooses an action to minimize cost in this limited space, then moves one level down in the tree. Depending on the depth of look-ahead chosen, CLS can require a substantial amount of computation, but has been able to unearth subtle patterns in the objects shown to it.

ID3 is then used to embeds a tree-building method in an iterative outer shell[Kovacic, 2010], and abandons the cost-driven look ahead of CLS with an information-driven evaluation function to evaluate the output from the above two methods.from [Kovacic, 2010; Mason et al., 2018] then An arbitrary object will be determined to belong to class P with probability p/(p + n) and to class N with probability n/(p + n) then using bayesian probability to validate if a test feature belongings to that class .However, when other variables beside demographic were included, the naive Bayes classifier is found to be the most accurate algorithm for predicting students' performance[Mason et al., 2018].[Mason et al., 2018] used decision trees, neural networks and linear discriminant analysis for the early identification of three categories of students: low, medium and high-risk students.

The expected information required for the sub-tree C_i is I(p_i, n_i).Then the expected information required for the tree with A as root is then obtained as the weighted average. Bayesian formalism determine the probability that the object has value A_i of A by examining the distribution of values of A in C as a function of their class.

3 Conclusion

3.1 Summary

In precis, there's combined evidence on whether the contribution of historical records to the early prediction of student achievement is widespread or not. It relies upon on the listing of variables included, students population and type techniques used. Even when the historical records became appreciably associated with the instructional performance, the prediction accuracy was pretty low with a standard accuracy.

References

Al-Radaideh, Q., Al-Shawakfa, E., and I. Al-Najjar, M. (2006). Mining student data using decision trees. *The International Arab Journal of Information Technology - IAJIT*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15, London, UK, UK. Springer-Verlag.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Kovacic, Z. J. (2010). Early prediction of student success : Mining students enrolment data.

Mason, C., Twomey, J., Wright, D., and Whitman, L. (2018). Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression. *Research in Higher Education*, 59(3):382–400.

Minaei, B., Kashy, D., Kortemeyer, G., and Punch, W. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. volume 1, pages T2A– 13.

Ogunde, A. and D.A., A. (2014). A data mining system for predicting university students' graduation grades using id3 decision tree algorithm. *Computer Science and Information Technology*, 2:1–26.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Sison, R. and Shimura, M. (1998). Student Modeling and Machine Learning. *International Journal of Artificial Intelligence in Education (IJAIED)*, 9:128–158.