Mining Multi-dimensional Patterns for Student Modelling

Andreia Silva
Instituto Superior Técnico - Universidade de Lisboa
Av. Rovisco Pais 1, Lisbon, Portugal
andreia.silva@tecnico.ulisboa.pt

Cláudia Antunes
Instituto Superior Técnico - Universidade de Lisboa
Av. Rovisco Pais 1, Lisbon, Portugal
claudia.antunes@tecnico.ulisboa.pt

ABSTRACT

A careful analysis of educational data reveals their multi-dimensional nature, with several orthogonal dimensions from students to teachers, courses, evaluation items, topics, etc. In addition, their historical nature translates into large data warehouses, which are modeled through inter-connected huge tables that encompass data from several distinct perspectives. Despite the recent advances in big data research for this educational domain, the ability to consider these very large multi-dimensional datasets remains unexplored. In this paper, we explore a multi-dimensional algorithm in order to find multi-dimensional patterns in education, which in turn will be used to model student behaviors. Experimental results in a real case study show a significant improvement on the prediction of student results, when compared with the same classifiers trained without those patterns.

1. INTRODUCTION

The long history of education as an institution lead to huge amounts of data, requiring automatic means for exploring them. Educational data mining (EDM) [1] gives a first opportunity for exploring these data, providing the adequate tools to predict students performance and dropouts, but also for understanding student behaviors [4].

Despite the encouraging results, few approaches were dedicated to explore the multi-dimensionality of data. Definitely, the educational process encompasses a set of different entities, characterized by distinct sets of attributes. Each kind of entity is usually known as a *dimension* (e.g. students, teachers, courses). In the intersection of these dimensions occurs the educational process, with the materialization of its *events* (e.g. the marks obtained by students). Multi-dimensional models, such as *star schemas*, are recognized as the most usual schemas to model these kinds of data. They consist in a central table containing the occurring events, and a set of surrounding tables, comprising the specific data about each dimension. Figure 1 shows an example in the educational domain with 2 star schemas: one modeling student enrollments and another teachers quality assurance surveys.

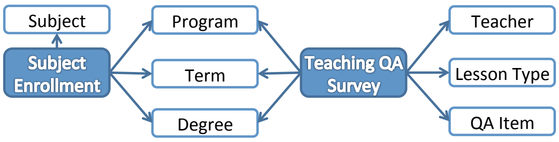

Figure 1: Example of an educational data schema.

In this work we propose a multi-dimensional methodology for analyzing educational data and improve prediction.

2. MULTI-DIMENSIONAL DATA MINING FOR EDUCATION

The prediction of future outcomes is a task mostly addressed by classification. However, results are far from being satisfactory, and one of the reasons may be the fact that the multi-dimensional relations between attributes are not being considered. Thus, we propose to use a multi-relational data mining (MRDM) algorithm to find patterns that are able to characterize different entities and their behaviors, and use the discovered information to enrich the data used for classification training, similar to what was proposed in [2].

MRDM [3] is an area that aims for the discovery of frequent relations that involve multiple tables, without joining all the tables before mining. Pattern mining, in particular, aims for enumerating all frequent patterns that conceptually represent relations among entities. These patterns can be *intra-dimensional* or *inter-dimensional*, if they contain items from the same or more than one dimension, respectively; or *aggregated*, if they result from the aggregation of events of the central table. The works on MRDM has increased, but they do not often scale with the number of facts. To overcome this, the algorithm *StarFP-Stream* was proposed [5], combining MRDM with data streaming techniques, and it is able to mine both large and growing star schemas.

The methodology proposed has four main steps: *multi-dimensional pattern mining*, *pattern filtering*, *data enrichment* and *classification*. The first consists on running an algorithm for multi-dimensional pattern mining over each star schema. After finding all the patterns, the next step is to filter the inter-dimensional and aggregate ones and choose the N best. We define a set of filters that try to capture the interestingness of a pattern: (1) *support* – The higher the support, the more events share the same characteristics represented in this pattern. However, the smaller the relations modeled; (2) *size* – The largest patterns model more relations than smaller ones. However, they tend to have the smallest supports; (3) *closed* patterns – A pattern is closed if none of its immediate supersets have the same support. Thus, a set of closed patterns (non-redundant) is more likely to be more interesting. (4) *rough independence* – If two events are independent, the occurrences of one do not influence the probability of the other, and therefore they are not interesting. Thus, $RInd(\{A_{1..n}\}) = \frac{P(A_1 \cap A_2 \cap ... \cap A_n)}{P(A_1)P(A_2)...P(A_n)}$. (5) *rough chi-square* – Chi^2 evaluates the correlation between variables. And the more correlated, the more interesting are the relations: $RChi^2(\{A_{1..n}\}) = \frac{(support(A_1 \cap ... \cap A_n) - P(A_1)...P(A_n))^2}{P(A_1)...P(A_n)}$.

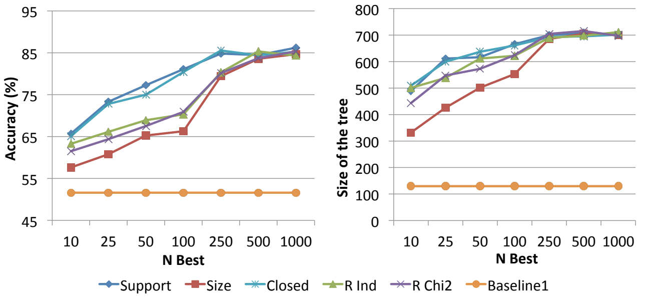

Figure 2: Accuracy and size of the model for B1.

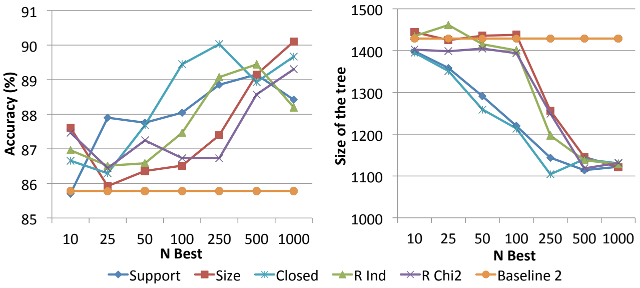

Figure 3: Accuracy and size of the model for B2.

Once we have the best patterns, we can use them as features for classification training by extending individual records with the multi-dimensional patterns, represented as boolean attributes (true or false) whenever an entity satisfies (or not) the particular pattern. We can then finally run classification algorithms on these enriched data and observe the results.

3. AN EDUCATIONAL CASE STUDY

In this case study we used the data from the *Information Systems and Computer Engineering* program, offered in *Instituto Superior Técnico – Universidade de Lisboa*, in Portugal. From the data warehouse, we have chosen the 2 stars in Figure 1, modeling student performances in their enrollments and teacher evaluation for their lectures. Our main goal is to test our multi-dimensional methodology for predicting student results on the 10 most representative courses of more advanced years (3rd-5th), based on the frequent behaviors found in the first 2 years. There were more than 650 students enrolled in some of those courses and 36 teachers lecturing them. There were 1830 enrollments to predict. We tested our enriched data with two baselines (without patterns). The first (B1) consists in the joining of the student, course and teacher dimensions, plus the student average grade, and the second (B2) contains also the specific grades on the most representative courses of the 1st and 2nd years (23 courses). Student grades were categorized and classification results are the average of several 10-cross fold validations, given by *C4.5* (available in *Weka*).

For finding student behaviors, there were more than 17 thousand enrollments that were used for pattern mining. We used an implementation of *StarFP-Stream* [5] made in Java (JVM version 1.6.0 37), and data in the fact table were aggregated per each pair student–term, so that we could find frequent sets of courses attended per term. We found, e.g. that it is frequent to succeed to both SIBD, PLD, AM3 and AN in the same term, and to fail to AN course in the 2nd season. For finding teacher behaviors, we used the surveys of the courses we were predicting, in previous years (1088 survey questions). Data in this star schema was aggregated per survey id, in order to find frequent sets of evaluations given by students to their teachers.

Figures 2 and 3 (left) show the accuracy of the classification step, over the B1 and B2, and corresponding datasets enriched with patterns from student behaviors (i.e. patterns of *Enrollments Star*). As expected, since B2 has more information about the background of the student, it achieves better accuracy than B1 (a 35% improvement). It is interesting to see that we can predict 50% of the grades of students based solely on their characteristics and average grade from years 1 and 2 (B1). When we add the patterns, we can see that the accuracy improves in both cases. In B1, the improvement is huge, of about 35%, because we are adding the behavior information about students, that was not present before. In B2, it allows classification to achieve an accuracy of 90%. Although only 4%, this improvement indicates that patterns are chosen instead of specific courses, and this may result in models with less over fitting, and therefore more accurate when predicting new instances. Also, results show that the more *N* best patterns are chosen, the better the accuracy, in general. When analyzing the different filters, both the *size* and *closed* filters achieved better results.

Figures 2 and 3 (right) analyze the size of the trees created by the classifier (i.e. the size of the model). We can see that for B2, also as expected, the trees resulting from classifying the enriched datasets are smaller than the base tree. In the B1 case, the models of the enriched datasets are larger because the baseline does not have much information, and when we add patterns, they are chosen for building the tree. Nevertheless, for similar values of accuracy (85%), the tree for B1 is much smaller than for B2.

4. CONCLUSIONS

In this paper we proposed a multi-dimensional methodology for mining educational data. It is general, and may be applied to different domains and with different algorithms. Experiments on a real case study show that we can take into account the multi-dimensionality of the educational data to discover frequent behaviors, and also to improve prediction. This work is partially supported by FCT – Fundação para a Ciência e a Tecnologia, under project educare (PTDC/EIA-EIA/110058/2009) and PhD grant SFRH/BD/ 64108/2009.

5. REFERENCES

[1] R. Baker, T. Barnes, and B. J. Educational data mining 2008. In *EDM 2008: Proc. of 1st Intern. Conf. on Educational Data Mining*, page 2, 2008.

[2] J. Barracosa and C. Antunes. Anticipating teachers performance. In *Proc. of Int. W. on Knowl. Discovery on Educational Data (KDDinED@KDD)*. ACM, 2011.

[3] S. Džeroski. Multi-relational data mining: an introduction. *SIGKDD Explor. Newsl.*, 5(1):1–16, 2003.

[4] C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE Trans. on Systems, Man, and Cybernetics, Part C*, 6(40):601–618, 2010.

[5] A. Silva and C. Antunes. Finding patterns in large star schemas at the right aggregation level. In *Proc. of the 9th Intern. Conf. on Modeling Decisions for AI*, pages 329–340. Springer, 2012.