BEST PRACTICES FOR ADAPTATION OF DATA MINING TECHNIQUES IN EDUCATION SECTOR

Jikitsha Sheth, Dr. Bankim Patel

ABSTRACT

Best practices help to make the business processes smooth. As best practices are always recommended and not forced, the authors have recommended few best practices to be followed in Educational institutes so that the activities related to educational data mining becomes easy to implement. The best practices suggested are with the objective to gain and maintain data quality; as quality data leads to correct analysis.

Keywords: Educational Data mining, best practices, data quality

1. INTRODUCTION

Before a decade, student's data was collected to provide summarized academic progress report of a student. With the advent in technology and concepts, nowadays, student's data is stored to understand the student as whole which not only includes student's progress in terms of marks or grades but also their interest level in subject, behaviour during learning phase, measuring of his/her traits related to learning process, etc. Making assessment a continuous process helps to predict student's learning outcome near to accurate [7]. To extract relevant data and predict the result of student, data mining can be used. Data mining is an important member in the Product family of Business Intelligence (BI), Online Analytical Processing (OLAP), enterprise reporting and ETL [15]. Data mining has played key roles in sectors like Customer Relationship Management, Weather Forecasting, etc. One of its applications is Web usage mining techniques to discover usage patterns from Web data, mainly to better understand the needs of Web-based applications [14]. This paper tries to suggest some best practices for using Educational Data Mining techniques at educational institutes.

2. EDUCATIONAL DATA MINING AND ITS ROLE

Educational data mining (EDM) represents a process of applying data mining techniques to data related to education aspects for addressing the research issues in education domain. It addresses problems like developing predictive model for student engagement [10], adaptive feedback mechanism for learners during their online communication for collaborative learning [1], understanding behaviour of students while learning [2, 9,11]. Collaborative Recommendation systems for students can be developed using data mining that recommends options to the

students for taking enrollment decision [12]. Educational data mining helps to identify learning variables and is also used to examine such variables so as to understand the Web-based learning process [3].

Educational Data mining analysis helps in decision making process of educational bodies. To improve decision making process, data quality plays a vital role. Decisions can be on varied issues, some of them can be as follows:

- With what level of interestingness does the student learn the subject and which are the indicators playing vital role in it?
- How to enhance student motivation in learning process?
- Which alumnis' are likely to provide service in institute's progress?
- Which students are likely to fail in external exam?
- How to increase retention of staff and students in the institute?

3. DATA QUALITY ISSUES

To make sound decisions, quality information is must. It adds value to the services provided by the institute to its stakeholders namely students, staff, parents & management body in particular and to the whole society in general. Data reconciliation becomes time consuming due to poor quality data. The highest benefits of high quality data is "single version of truth" and "greater confidence in analytical systems" [6]. As EDM involves high analysis part, quality data is its high demand.

In this section we focus on data collection phase from where the data quality dilution occurs. If the data is cleaned at source system, the defects cannot spread [5].The dataset used for data mining is generated from the data of sources like flat files, tables, databases or data warehouses. Such sources may have incomplete, noisy and inconsistent data [8].

- Data may be incomplete due to reasons:
  - Data unavailability: During data collection phase, the details were not known, which resulted to null values in flat files and tables. Missing data or not application data also results to data unavailability.
  - Unidentified feature: A feature or attribute can be independent or dependent in nature. The structure of flat file or table is designed to cope up with day to day transactions. It might happen that the feature required for data mining, was not identified or was thought to be irrelevant during table design or flat file creation, so there is no data in the dataset with respect to that required feature. If it is dependent in nature, then only it can be derived and further used in analysis.

- Data may be noisy due to reasons:
  - Inputted data was erroneous due to lack of attention of the person entering data, or may be typographical error.
  - Also, presence of outliers results to noisy data.
- Data may be inconsistent due to one of the major reason - redundancy. A data which is stored at several places results to duplication of data. If some updation in required in such redundant data, it may happen that all copies of data is not updated and results to inconsistent data. In addition to this, when same data is stored with different terms and terminology at different sources, then chances of inconsistency rises.

To minimize the ill-effects of incomplete, noisy and inconsistent data, data preprocessing is done. It includes data cleaning, integration, transformation and reduction. Data preprocessing itself is a time consuming and computationally extensive process [8].

If best practices are followed during data collection and data manipulation, the cost of data preprocessing can be highly minimized. Also, it will help to maintain data quality which would lead to accurate data analysis and quality decision making. This paper suggests some of the best practices to be followed in academics so that data mining process and other relevant techniques can be adapted which can further provide accurate results.

4. BEST PRACTICES

Following best practices should be followed at the time of data collection and dataset designing:

1. Use a methodology like CRISP-DM [4] before initializing a data mining project. This will help to understand the project and data requirements clearly. This will further help to identify the features required during data mining in advance, so data completeness can be achieved.
   CRISP-DM is methodology containing following six phases [4].
   a. Business Understanding: During this phase the requirements are understood. Also, the objectives of projects are realized. This in itself is a data mining problem definition. It helps to identify the relevant attributes in advance so that unavailability of required data is not observed.
   b. Data Understanding: In this phase initial data collection. Here a clear measure on data quality can be provided so that ultimate analysis is accurate.

c. Data Preparation: From the collected data, attribute selection is done as per the problem definition. The data is transformed and cleaned in case required.
d. Modeling: Identification of technique to model the problem and reach towards solution is done at this phase.
e. Evaluation: Continuous evaluation regarding objectives achieved is done at this phase.
f. Deployment: The result model is deployed.

All these phases ultimately ensure encompassment of relevant attributes and minimization of noise.

2. Human inspections at the time of data collection and data entry will reduce noisy and inconsistent data.
   For reduction of noisy data, the source of data should be checked and corrected. It is not so that once the correction done, no future error will occur. One has to frequently keep inspection on the source, before, during and after the data collection. Such inspections if not automated, should have human factor involved. If the data collection process is computerized, it should have simple and good user interface. Data validation feature should be strongly imbibed in it.
3. Consolidation of data from various sources will reduce the data preprocessing phase. If the data source is a database, record linkage concept can be useful. Data mining itself is one of the ways to identify record linkage using algorithms like Naïve Bayes algorithm and Rule-based induction techniques.
4. While designing the dataset features, one should identify composite attributes and break them till they become pure simple attributes. Having such atomic data provides fine-grained features at the time of analysis. For example if we are storing address, it is better to have separate attributes to store HOUSE, STREET, AREA, CITY, STATE, COUNTRY details instead of a composite attribute ADDRESS. This will be helpful while result analysis is required on few of the former attributes [13].
5. While designing the dataset, one should identify hierarchy concept in the attribute and break data to the lowest possible level. So mining will be possible at multiple levels of abstraction. For example, in education domain, to store USERCATEGORY the COURSE, YEAR, SEMESTER, STUDENT hierarchy will be helpful. During analysis, aggregation operation can be easily applied on various levels as and when required.
6. Data documentation and process documentation should be emphasized.

Once the data is collected and stored in the data source, following practices should be observed during data manipulation:

1. If record requires to be deleted, do not delete it. Instead of deleting the records, archive it. Because in analysis of data, historical data is required to analyze past patterns. If such past data gets deleted, analysis may not be accurate.
2. If archiving of record is required, it is better to set such record as blocked instead of shifting the records from one source to another. When such archived data is required for future data mining, one has to go for ETL process which is costly and may result to data loss. By using the mechanism of record status, just by setting its state from blocked to unblocked, the record is now ready for mining.
3. If data is required to be modified, do not replace the old data with the new one. Instead, a well planned mechanism should be adopted to store the old data and relate it with new data. Thus, the history of values stored in that feature should be managed.
4. The process of modification suggested in above point 3 should also be time stamped.

Looking to behavioral aspect, an organization can carry on a task in smooth manner only if the organization culture supports it. So before adapting techniques, an environment of transparency and trust is required among all the stake holders of the education body i.e. student, staff, managing body, etc.

So, apart of all the above practices,

- Transparency in data collection and usage should be there, which implies that students should be made aware about nature of the data collected and should also be informed about the mechanism by which it is collected (when the means/mechanism is not direct like that in Web based tutoring systems, online examination systems, etc.).
- Students should be assured confidentiality of their information.
- Students should be allowed to manage the integrity of their personal data.
- Data quality should be considered as continuous process instead of restricting it to data collection, data cleaning or ETL phases.
- Commitment of high level managers namely Head of the Department, Dean and management bodies in monitoring and evaluation of data quality should be high.

5. CONCLUSION

The use of Data mining has widely increased to improve teaching learning process. The Education Data mining approach is bringing paradigm shift in usage of Information Technology in Education Sector. To effectively adapt the data mining techniques in educational institutes, authors have suggested few best practices in this paper. Implementing such best practices helps to manage the 'quality factor' in data which further leads to accurate result analysis.

REFERENCES:

1. Anjewierden, A., Kolloffel B., and Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes, In Proceeding of International Workshop on Applying Data Mining in e-Learning (ADML'07), pp.23-32.
2. Baker, R. and Carvalho, D. (2008). Labeling Student Behavior Faster and More Precisely with Text Replays, In Proceedings of the 1st International Conference on Educational Data Mining, pp.38-47.
3. Ben-Zadoki, G., et. al. (2009). Examining online learning processes based on log files analysis: A case study, In Research, Reflections and Innovations in Integrating ICT in Education (Ed. A. Méndez-Vilas,et.al.), FORMATEX, pp.55-59.
4. Chapman, P., et. al. CRISP-DM 1.0, Step-by-step data mining guide, SPSS, CRISP-DM Consortium, http://www.crisp-dm.org/download.htm, accessed on Oct 2009.
5. Eckerson, W., Data Quality and the Bottom Line, TDWI Report Series, 2002.
6. Eckerson, W., Excerpt from TDWI's Research Report - Data Quality and the Bottom Line, Business Intelligence Journal, Dec 2001, http://www.tdwi.org/research/display.aspx?ID=6589 accessed on Dec 2009.
7. Examination Reforms and Continuous and Comprehensive Evaluation (CCE) in CBSE, http://www.cbse.nic.in/cce/index.html, accessed on Dec 2009.
8. Han, J. and Kamber M. (2001). Data Mining: Concepts and Techniques, San Francisco, Morgan Kaufmann.
9. Jeong, H., and Biswas, G.(2008). Mining Student Behavior Models in Learning by-Teaching Environments, In Proceedings of the 1st International Conference on Educational Data Mining, pp.127-136.
10. Lloyd, N., Heffernan, N. and Ruiz C. (2007). Predicting student engagement in intelligent tutoring systems using teacher expert knowledge, Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education.pp.40-49.
11. Mavrikis, M. (2008). Data-driven modelling of students' interactions in an ILE, In Proceedings of the 1st International Conference on Educational Data Mining, pp.87-96.
12. Sacin, C., Agapito J., et.al.(2009). Recommendation in Higher Education Using Data Mining Techniques, Proceedings of 2nd International Conference on Educational Data Mining, Spain.

13. Sheth, J., Patel B., and Bhatti, D. (2010). Improper Internet Usage: Controlling through Policy Model and Identifying through Data Mining, National Journal of Computer Science & Technology, Vol. 02(1), pp.16-21
14. Srivastava, J., Cooleyz, R., et. al. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD, Vol. 01(2), pp.12-22.
15. Tang, Z., Maclennan, J. (2005). Data mining with SQL Server 2005, Wiley Publications.