Enriching Information Security in Educational Data Mining Using Big Data

[1]**Gopal R. Chandangole,** [2]**Aparna V. Mote,** [3]Prashant M. Mane

Assistant Professor
Computer Engineering
Zeal College of Engineering and Research, Pune, Maharashtra

***Abstract*— Presently, educational institutions compile and store huge volumes of data, such as student enrolment and attendance records, as well as their examination results. Mining such data yields stimulating information that serves its handlers well. Rapid growth in educational data points to the fact that distilling massive amounts of data requires a more sophisticated set of algorithms. This issue led to the emergence of the field of educational data mining (EDM). Traditional data mining algorithms cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a preprocessing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. One such preprocessing algorithm in EDM is clustering. Many studies on EDM have focused on the application of various data mining algorithms to educational attributes. In this paper, to view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each type of user, we discuss his privacy concerns and the methods that can be adopted to protect sensitive information. We briefly introduce the basics of related research topics, review state-of-the-art approaches, and present some preliminary thoughts on future research directions. Besides exploring the privacy-preserving approaches for each type of user.**

***IndexTerms*— Data mining, clustering methods, educational technology, systematic review.**

I. INTRODUCTION

EDM as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students its aim is to develop models to improve learning experience and institutional effectiveness. While DM, also referred to as Knowledge Discovery in Databases (KDDs), It refers to collecting similar objects together to form a group or cluster. Each cluster contains objects that are similar to each other but dissimilar to the objects of other groups. An educational institution environment broadly involves three types of actors namely teacher, student and the environment. Interaction between these three actors generates voluminous data that can systematically be clustered to mine invaluable information.

Data clustering enables academicians to predict student performance, associate learning styles of different learner types and their behaviors and collectively improve upon institutional performance. Various methods have been proposed, applied and tested in the field of EDM. It is argued that these generic methods or algorithms are not suitable to be applied to this emerging discipline. It is proposed that EDM methods must be different from the standard DM methods due to the hierarchical and non-independent nature of educational data [5]. Educational institutions are increasingly being held accountable for the academic success of their students [4]. Notable research in student retention and attrition rates has been conducted by Luan [1]. For instance, Lin [9] applied predictive modeling technique to enhance student retention efforts.

The e-commerce websites use recommender systems to collect user browsing data to recommend similar products. There have been efforts to apply the same strategy in the educational information system. One such successful system is the degree compass.

Data mining has attracted more and more attention in recent years, probably because of the popularity of the ``big data'' concept. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [6]. As a highly application-driven discipline, data mining has been successfully applied to many domains, such as business intelligence, Web search, scientific discovery, digital libraries, etc.

The term ``data mining'' is often treated as a synonym for another term ``knowledge discovery from data'' (KDD) which highlights the goal of the mining process.[10]

II. EDUCATIONAL DATA AND CLUSTERING METHOD

EDC is based on data mining techniques and algorithms and is aimed at exploring educational data to find predictions and patterns in data that characterize learners' behavior. The system has clustered various research works that have been conducted exclusively within educational attributes related to clustering algorithms. The system will now provide a detailed analysis on various aspects of educational attribute collated with the application of clustering algorithms to help improve the education system.

III. EDUCATIONAL DATA MINING USING CLUSTERING

As we know that clustering algorithms can broadly be divided into hierarchical and non-hierarchical types. So, it would be easier if the research conducted could equally be partitioned according to the clustering algorithm used. Wook, *et al*. [4] have evaluated undergraduate student's academic performance on end of semester exam. They applied a combination of data mining methods such as Artificial Neural Network (ANN), Farthest-First method based on *K*-means clustering and Decision Tree as a classification approach. The data set comes from the faculty of science and defense technology. DM techniques to student academic data have been provided. In this study, Apriori algorithm was applied to academic records of students to obtain the best association rules which help in student profiling. *K- m*eans clustering was used to group students categorically. The data is obtained from student academic record file.

In Fig. 1, we show the educational data clustering process. The first stage is the data pre-processing stage in which the researcher must first understand the domain and complexity of the educational dataset collected thereafter should be able to identify the attributes that have garbage or missing values. By garbage values we refer to values that are not marked to be present for the attribute.

Let us take an example, consider a nominal attribute `student response' with allowed values like `yes' or `no'. Now, if this attribute is coded with a value like `NA' then it should be treated as a garbage value and must be removed.

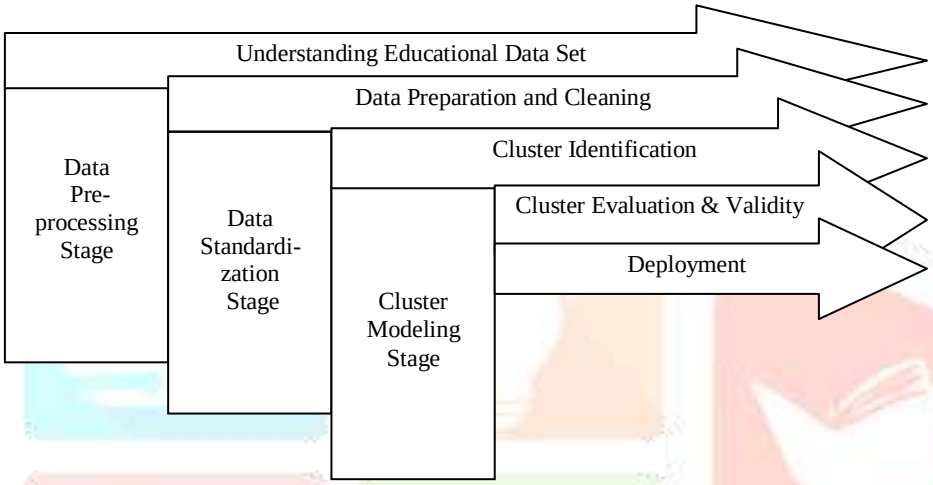

Figure 1: Educational data clustering process.

There should also be defined a standard to fill missing values. So, for example, referring to the above attribute `Student response' the missing values could be filled as `?'. This activity is termed as data standardization. Once the data is cleaned, it should then be analyzed. Perhaps the easiest way is to determine relationships between various attributes that constitute the dataset. For example, Weka uses various machine learning algorithms (like Correlation attribute evaluator, One R attribute evaluator, gain ratio attribute evaluator, Principle component analysis attribute evaluator) that can easily determine the most significant attributes within the dataset. Once such significant attributes are found they can then be used to train and cluster the whole dataset to create data models. Post which new data having same or similar attributes can be applied to these data models to reveal interesting insights.

Learning portfolios are records that are created during the learning process. Note taking, assignments, test paper reports, test papers etc. are examples of learning portfolio. In their analytical paper Chen, *et al*. [5] applied *K*-means, Farthest First and EM clustering algorithms and statistical *t*-test to the student portfolios of an e-learning system. Using clustering methods in this study they were able to cluster students' e-learning performance. Using t-test they were able to evaluate mid-term and final term exam performance of the clusters with high & low online learning frequency. 162 subjects used in this study were junior students of the department of computer engineering at Chung Yuan Christian University. This data was taken from i-learning [2] eLearning software being used in Taiwan. Their tests found that there was a positive correlation between students with high online eLearning frequency and higher scores. It was also found that the student portfolio of click times and duration of the study of learning materials at the beginning of the semester does not show any correlation with midterm and final term exam results. They also found that student participation in online discussion forums showed significant effect on their exam results.

In a similar analytical work conducted by Perera, *et al*. [3], *K*-means & EM clustering algorithms from WEKA was used to find group similarities. In this study their experiments revealed the same result for *k* D 3 for *K*-means. Hierarchical agglomerative clustering with Euclidean distance was used for this purpose. The student teams were required to use TRAC [1] for online collaboration. TRAC is an open source, professional soft-ware development tracking system. The researchers collected the data over three semesters, for student cohorts in 2005 and 2006. The data size was 1.6 Mbytes in my SQL format and it contained approximately 15,000 events. The key contribution of this research is improved understandings of how to use data mining to build mirroring tools that can help small long-term teams improve their group work skills

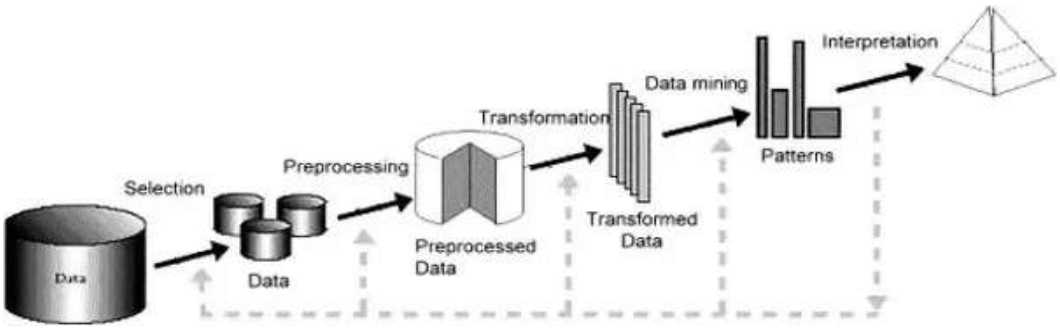

FIGURE 2. An overview of the KDD process.

IV. Privacy Concern and PPDM

The information discovered by data mining can be very valuable to many applications people have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining [2]. Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc. To deal with the privacy issues in data mining, a subfield of data mining, referred to as *privacy preserving data mining* (PPDM) has gained a great development in recent years. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone number, should not be directly used for mining. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded. After the pioneering work of Agrawal et al. [7], [6} numerous studies on PPDM have been conducted.

V. Approaches to Privacy Protection

A) Limit the Access

A data provider provides his data to the collector in an active way or a passive way. By active we mean that the data provider voluntarily opts in a survey initiated by the data collector, or fill in some registration forms to create an account in a website. By passive we mean that the data, which are generated by the provider's routine activities, are recorded by the data collector, while the data provider may even have no awareness of the disclosure of his data. When the data provider provides his data actively he can simply ignore the collector's demand for the information that he deems very sensitive. If his data are passively provided to the data collector, the data provider can take some measures to limit the collector's access to his sensitive data.

B) Trade Privacy for Benefit

In some cases, the data provider needs to make a tradeoff between the loss of privacy and the benefits brought by participating in data mining. For example, by analyzing a user's demographic information and browsing history, a shopping website can offer personalized product recommendations to the user. The user's sensitive preference may be disclosed but he can enjoy a better shopping experience. Driven by some benefits, e.g. a personalized service or monetary incentives, the data provider may be willing to provide his sensitive data to a trustworthy data collector, who promises the provider's sensitive information will not be revealed to an unauthorized third-party. If the provider is able to predict how much benefit he can get, he can rationally decide what kind of and how many sensitive data to provide.[8]

C) Privacy False Data

As discussed above, a data provider can take some measures to prevent data collector from accessing his sensitive data. However, a disappointed fact that we have to admit is that no matter how hard they try, Internet users cannot completely stop the unwanted access to their personal information. So instead of trying to limit the access, the data provider can provide false information to those untrustworthy data collectors the following three methods can help an Internet user to falsify his data:

1) Using ``sock puppets'' to hide one's true activities.[7] A sockpuppet12 is a false online identity though which a member of an Internet community speaks while pretending to be another person, like a puppeteer manipulating a hand puppet. By using multiple sock puppets, the data produced by one individual's activities will be deemed as data belonging to different individuals, assuming that the data collector does not have enough knowledge to relate different sock puppets to one specific individual. As a result, the user's true activities are unknown to others and his sensitive information (e.g. political preference) cannot be easily discovered.

2) Using a fake identity to create phony information. In 2012, Apple Inc. was assigned a patient called ``Techniques to pollute electronic profiling'' which can help to protect user's privacy. [6]This patent discloses a method for polluting the information gathered by ``network eavesdroppers'' by making a false online identity of a principal agent, e.g. a service subscriber.

The clone identity automatically carries out numerous online actions which are quite different from a user's true activities. When a network eavesdropper collects the data of a user who is utilizing this method, the eavesdropper will be interfered by the massive data created by the clone identity. Real information about of the user is buried under the manufactured phony information.

3) Using security tools to mask one's identity. When a user signs up for a web service or buys something online, he is often asked to provide information such as email address, credit card number, phone number, etc. A browser extension called MaskMe,13

which was release by the online privacy company Abine, Inc. in 2013, can help the user to create and manage aliases (or *Masks*) of these personal information. Users can use these aliases just like they normally do when such information is required, while the websites cannot get the real information. In this way, user's privacy is protected

VI. PRIVACY-PRESERVING PUBLISHING OF SOCIAL NETWORK DATA

Social networks have gained great development in recent years. Aiming at discovering interesting social patterns, social network analysis becomes more and more important. To support the analysis, the company who runs a social network application sometimes needs to publish its data to a third party. [10]However, even if the truthful identifiers of individuals are removed from the published data, which is referred to as naïve anonym zed, publication of the network data may lead to exposures of sensitive information about individuals, such as one's intimate relationships with others. Therefore, the network data need to be properly anonym zed before they are published.

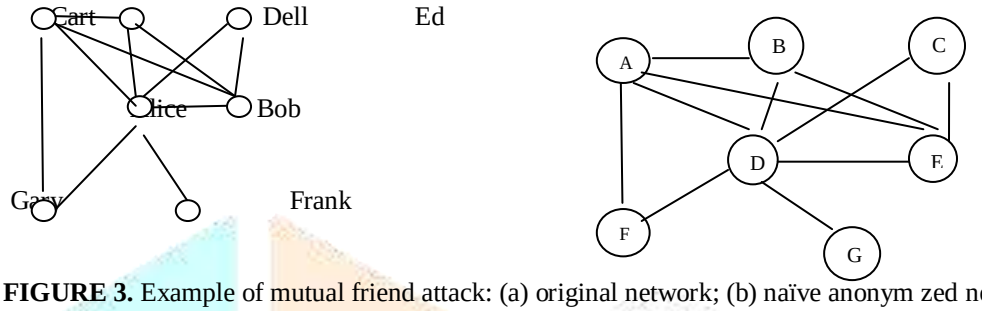

FIGURE 3. Example of mutual friend attack: (a) original network; (b) naïve anonym zed network.

VII. CONCLUSION

This paper has presented over three decade's systematic review on clustering algorithm and its applicability and usability in the context of EDM.[2] This paper has also outlined several future insights on educational data clustering based on the existing literatures reviewed, and further avenues for further research are identified. In summary, the key advantage of the application of clustering algorithm to data analysis is that it provides relatively an unambiguous schema of learning style of students given a number of variables like time spent on completing learning tasks, learning in groups, learner behavior in class, classroom decoration and student motivation towards learning. Clustering can provide pertinent insights to variables that are relevant in separating the clusters. and how to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years.[9] In this paper we review the privacy issues related to data mining by using a user-role based methodology. We differentiate four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. Each user role has its own privacy concerns

REFERENCES

[1] C. Romero and S. Ventura, ``Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601_618, Nov. 2010.
[2] J. Ranjan and K. Malik, ``Effective educational process: A data-mining approach," *Vine*, vol. 37, no. 4, pp. 502_515, 2007.
[3] V. P. Bresfelean, M. Bresfelean, N. Ghisoiu, and C.-A. Comes, ``Determining students' academic failure pro_le founded on data mining methods," presented at the ITI 30th Int. Conf. Inf. Technol. Interfaces, Jun. 2008, pp. 317_322.
[4] J.P. Vandamme, N. Meskens, and F.-J. Superby, ``Predicting academic performance by data mining methods," *Edu. Econ.*, vol. 15, no. 4, pp. 405_419, 2007.
[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.
[6] L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining," in *Proc. Austral. Inst. omput. Ethics Conf.*, 1999, pp. 89_99.
[7] R. Agrawal and R. Srikant, ``Privacy-preserving data mining," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 439_450, 2000.
[8] Y. Lindell and B. Pinkas, ``Privacy preserving data mining," in *Advances in Cryptology*. Berlin, Germany: Springer-Verlag, 2000, pp. 36_54.
[9] C. C. Aggarwal and S. Y. Philip, *A General Survey of Privacy- Preserving Data Mining Models and Algorithms*. New York, NY, USA: Springer-Verlag, 2008.
[10] S. Matwin, ``Privacy-preserving data mining techniques: Survey and challenges," in *Discrimination and Privacy in the Information Society*. Berlin, Germany: Springer-Verlag, 2013, pp. 209_221.