

# Data-Mining using Nearest Neighbour Method

Baljeet Kaur<sup>1</sup>, Kailash Bahl<sup>2</sup>

## ABSTRACT:

As an interdisciplinary discipline, data mining (DM) is popular in education area especially when examining students' learning performances. It focuses on analyzing educational related data to develop models for improving learners' learning experiences and enhancing institutional effectiveness. Therefore, DM does help education institutions provide high-quality education for its learners. Applying data mining in education also known as educational data mining (EDM), which enables to better understand how students learn and identify how improve educational outcomes. Present paper is designed to justify the capabilities of data mining approaches in the filed of education. The latest trends on EDM research are introduced in this review. Several specific algorithms, methods, applications and gaps in the current literature and future insights are discussed here.

**Keywords:** *Data Mining (DM), Association rule, Nearest Neighbour Method*

## INTRODUCTION

One of the biggest challenges that educational institutions facing today is the exponential growth of educational data and how to apply this data to improve the quality of managerial decisions [1]. Education Institutions would like to know, for instance, which students will enroll in particular course programs, and which students will need assistance for graduation. Through the analysis and presentation of data they collected, or data

mining, the challenges of these student or learners are able be effectively addressed. Data mining enables organizations to uncover and understand hidden patterns in vast databases by using their current reporting capabilities. And these patterns are then built into data mining models and applied to predict individual behavior and performance with high accuracy. In this way, resources and staff can be allocated by institutions more effectively. Data mining may also, for example, efficiently allocate resources with an accurate estimate of how many students will take action before he or she drops out.

## BACKGROUND OF DATA MINING

Data mining is an interdisciplinary subfield of computer science [3-5]. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD [6]. Data mining techniques have their roots in machine learning, artificial intelligence, computer science, and statistics etc. [7]. And data mining is an exploratory process, but it can be used for confirmatory investigations [8]. It is different from other searching and analysis techniques because data mining is highly exploratory, where other analyses are typically problem-driven and confirmatory. Through the combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge, hidden trends and patterns are able to be uncovered. These trends and patterns form the predictive models that enable to assist organizations with uncovering useful information then guide decision-making [9].

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a cycle process for development and analysis of data mining models [10]. As the demand for data mining increases and more algorithms are created, CRISP-DM ensures practices that everyone can follow, and it gives specific tips and techniques on how to understand business data by deploying a data-mining model. CRISP-DM has six phases including business understanding, data understanding, data preparation, modeling, evaluation, and deployment [10].

## ALGORITHMS OF DATA MINING

Data mining relies on disciplines like classification, categorization, estimation, and visualization. Classification assists with identifying associations and clusters, and separates subjects under study. E.g., education institutions can use classification comprehensively to analyze student's characteristics. Categorization applies rule induction algorithms to handle categorical outcomes. Estimation includes predictive functions or likelihood deals with continuous outcome variables. Estimation and classification use unsupervised or supervised modeling techniques. Visualization uses interactive graphs to demonstrate mathematically induced data and scores, and is much more sophisticated than traditional bar charts or pie charts.

## ASSOCIATION RULE

Association rules are used to find relations between different items [15]. Back to 1995, the analysis method of association rule was frequently utilized in most studies on educational data mining because of its less extensive expertise while comparing with other methods [20, 21]. However, after the year of 2005, as researchers frequently adopting clustering and classification methods, the trend changed.

## NEAREST NEIGHBOUR METHOD

Nearest neighbour method, also called the k-nearest neighbour technique sometimes, classifies each record in the dataset based on a combination of different classes of the k record(s), which is similar to that in a historical dataset (where  $k \geq 1$ ).

To choose the appropriate algorithms, researchers need design the data and align it with the desired output. As for small-scale, they can opt for clustering approach since it does not require necessary splitting data in classification.

## CONCLUSION AND FUTURE INSIGHTS

Data mining is a powerful analytical tool to enhance decision making and analyzing new patterns and relationships for organizations. And EDM contains techniques including data mining, statistics, machine learning. DM need to analyze data coming from teaching and learning, tests learning theories, and policy decision-making etc. There are a number of opportunities exist in EDM, from an analysis at organizational level to the analysis at individual level. What's more, EDM is widely used and applied by learners, researchers and teachers, even institutions.

Recently, there are several studies focus on applying EDM into admissions and enrollment, but we don't know exactly how institutions using data mining to enhance student learning or improving related educational processes. And results from EDM research are typically obtained from the narrow context of specific educational settings. Therefore, the need for studies to examine in the broader context is necessary. For the overall EDM work to be completed, the urgent need of examining how to widespread the adoption of educational data mining is necessary. Further more, research

indicates the area of educational data mining is concentrated in western cultures and subsequently, other countries like Asians may not be represented in the related researches and studies. Therefore, applications across multiple contexts should be considered in development of future models [39].

## **REFERENCES**

- [1] Koedinger K, Cunningham K, Skogsholm A, Leber B. (2008) An open repository and analysis tools for fine-grained, longitudinal learner data. In: *First International Conference on Educational Data Mining*. Montreal, Canada; 2008, 157-166.
- [2] Baker, R., & Yacef, K.(2009). The State of Educational Data mining in 2009: A Review Future Visions. *Journal of Educational Data Mining*, 1(1).  
  
*Data Mining Curriculum*. ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.
- [3] Clifton, Christopher (2010). *Encyclopedia Britannica: Definition of Data Mining*. Retrieved 2010-12-09.
- [4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Retrieved 2012-08-07.
- [5] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). *From Data Mining to Knowledge Discovery in Databases*. Retrieved 17 December 2008.
- [6] Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Pearson Education.
- [7] Berson, A., Smith, S., & Thearling, K. (2011). An Overview of Data Mining Techniques Retrieved November 28, 2011,
- [8] Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2012). Analytics: The Widening Divide. *MIT Sloan Management Review*, 53(2), 1-22.
- [9] Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice*, 12(2), 137-153, doi:10.1057/dddmp.2010.35.
- [10] Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *SIGKDD Explor. Newsl.*, 13(2), 3-6. doi: 10.1145/2207243.2207245
- [11] Corbett, A. T. (2001). *Cognitive Computer Tutors: Solving the Two-Sigma Problem*. Paper presented at the Proceedings of the 8th International Conference on User Modeling 2001.
- [12] Baker RSJd (2010). Data mining for education. In McGaw B, Peterson P, Baker E, eds. *International Encyclopedia of Education*. 3rd ed. Vol. 7. Oxford, UK: Elsevier; 2010, 112118.
- [13] Pechenizkiy, M., Calders, T., Vasilyeva, E., & De Bra, P. (2008). Mining the student assessment data: Lessons drawn from a small scale case study. *Educational Data Mining 2008*, 187.
- [14] He, W. (2012). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*.
- [15] Shih, Y. C., Huang, P. R., Hsu, Y. C., & Chen, S. Y. (2012). A complete understanding of disorientation problems in web-based learning. *The Turkish Online Journal of Educational Technology*, 2012, 11(3).
- [16] Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop* at the ECAI 2004.
- [17] Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *Knowledge and Data Engineering*, IEEE *Educational Data Mining*. Boca Raton, FL: CRC Press; 2011, 123-142.



Tane J, Schmitz C, Stumme G. (2004). Semantic resource management for the web: an elearning application. In: *International Conference of the WWW*. New York; 2004, 1-10.