Data Mining Methods in Educational Process Management

ANNA HOVAKIMYAN, SIRANUSH SARGSYAN
Department of Programming and Information Technologies,
Yerevan State University,
1, Alek Manukyan st., Yerevan, 0025,
ARMENIA

Abstract: - The paper addresses the challenge of effectively managing the educational process by leveraging intelligent data analysis of student performance during learning activities. It introduces an approach centered around data clustering, specifically applied to the study of programming disciplines and languages. By utilizing clustering techniques, the paper aims to identify the most challenging topics within a given academic subject, track students' learning paths, evaluate and enhance teaching methodologies, and create personalized learning plans tailored to individual students' needs. This approach enables educators to better understand and address the diverse learning requirements of students, ultimately enhancing the overall educational experience.

Key-Words: - educational process, student's educational path, programming languages teaching, knowledge testing environments, data mining, data clustering, adaptive teaching scenarios.

Received: March 3, 2024. Revised: September 3, 2024. Accepted: October 5, 2024. Published: November 6, 2024.

1 Introduction

At present, educational institutions and training centers gather and retain a vast amount of data about the educational procedure. This information includes records of student enrollment and attendance to courses, outcomes of ongoing and final exams, charts displaying the degree to which students have attained the educational objectives stated in educational programs, and so on. Analyzing this data provides valuable insights that aid educational institution staff in efficiently arranging the educational process, [1], [2], [3], [4].

When dealing with large sets of data, data mining methods can be very useful. Note that traditional data mining algorithms may not be suitable for solving problems interested. Therefore the development of new algorithms that offer specific functionality and integrity is necessary, [1].

The process of data mining aims to extract knowledge from large amounts of raw data and has become a crucial aspect of decision-making systems. Data mining techniques are introduced into different research fields, such as statistics, databases, machine learning, artificial intelligence, data visualization, etc. [1]. It has gained significant attention and has become an important component of the activities in various organizations. For example, in the service sector, data mining helps analyze customer behavior and improve a company's performance. In the healthcare sector, it serves as an additional tool for doctors to diagnose diseases. In the marketing sector, it helps to study the market and its needs and to make informed decisions, [1].

Data mining is a continuous cyclic process that does not stop after finding a solution. The results generated from data mining lead to new business goals, which can be used to create more focused models.

As we know, the process of data mining includes the following stages: problem identification, data collection and preparation, model building, evaluation, and model deployment, [1].

In the first stage, the subject area as well as the problem are analyzed, project goals and requirements are identified, the data mining task is formulated, and a preliminary implementation plan is developed, [1], [5].

The second stage involves gathering and researching data. This step helps to determine how effectively the data gathered solves the problem. One can remove unnecessary data or add more data to improve the accuracy of the results. During this stage, the data is also subjected to statistical analysis. Proper preparation of the data can significantly enhance the quality of information that must be extracted using data mining techniques. In addition, this stage involves visualizing and interpreting the data for various stakeholders, [6].

During the model construction and evaluation stage, different modeling methods are selected and applied, and model parameters as well as hyperparameters are adjusted to optimal values. At this point, it is crucial to evaluate how well the model satisfies the initially stated business goal, [6].

The final stage is the deployment of the model, where data mining is used in the target environment. In this stage, the trained model can be used to make decisions based on real-time data, [7].

In data mining, various approaches and algorithms are used to find connections between different data characteristics. These methods include classification, regression, cluster analysis, factor analysis, social network analysis, association rules searching, sequential pattern analysis, etc. The results are used to predict future events and the forecast results are presented via various visual representations such as pie charts, histograms, Gantt charts, tables, etc. This helps people to better understand the patterns and forecast results, which can then be used for decision-making, [6].

2 Data Mining in Education

Educational Data Mining (EDM) is a specific research field in Data Mining that focuses on using various studies, techniques, and tools to extract useful information from vast amounts of data related to educational processes, [8], [9]. This data can include various details such as the performance of applicants in entrance exams, the results of students' session exams, how the independent and research work is carried out, the academic subjects that students frequently access during online learning, as well as the students’ preferred format of educational materials (text, multimedia, etc.), [10].

Data mining is widely used in education to analyze a large amount of data produced by information systems in schools, universities, and educational centers. Experts from various fields such as computer and communication technologies, pedagogy, psychology, and statistics work together to improve the educational process by combining traditional and innovative teaching and learning methods. The ultimate goal is to enhance the quality of education through advanced data analysis techniques (Figure 1), [11], [12].

Automated data analysis through data mining in the field of education provides detailed results that are difficult for a person to obtain through manual analysis. These results can be used by the learning process management system to establish a correlation between a student's educational path, final grades, achievements, and educational goals [9], [10], [13].

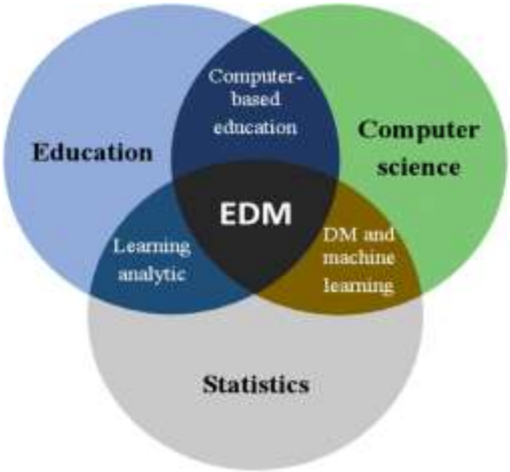

Fig. 1: Data mining infrastructure in education

Data mining can be a valuable tool for educators and researchers in the field of education. It can assist in analyzing curriculum to ensure high-quality learning for students, altering the list and content of academic disciplines and workshops according to student and employer requirements, studying the educational paths of students and their connection with educational goals, identifying patterns and anomalies in student work, suggesting the most effective criteria for selecting courses in asynchronous learning, and much more, [3], [4], [9], [12].

To comprehensively evaluate the entire educational process and identify patterns between students' academic achievements and different educational approaches, it is necessary to analyze and visualize information. Research has shown that pre-processing algorithms should be applied to learning process data before any specific data mining methods can be applied, [1], [3], [7].

Data mining methods are utilized in various forms of learning, such as offline, online, and hybrid formats. They are applied to analyze the learning outcomes of individual subjects and the entire educational program, [3], [11], [13]. In a single subject, it is possible to examine the results of mastery of individual topics and patterns that reflect the pace of mastery of a specific topic, depending on the degree of mastery of other topics. Ultimately, this assessment can provide an evaluation of the students' acquisition of knowledge, competencies, and skills in the entire academic discipline, [9], [11], [13].

It is possible to establish a correlation between accessibility and the degree of mastery of specific topics within an academic discipline. The same method can be used to evaluate the entire educational program and determine the extent to which the educational goals have been met, [14].

Automated testing environments can generate data that can be used to build machine learning models. With a target variable in mind, regression models can be constructed to predict expected output values. These predictive models can be used to develop warning and recommendation systems, to prevent undesirable academic outcomes, [11], [15].

3 Evaluating the Effectiveness of Implementing the Educational Program

One of the key factors in evaluating the effectiveness of an educational program is to determine the extent to which students have achieved the goals outlined in the program. This is done by assessing the educational outcomes produced by various components of the curriculum including final and current exams, tests, independent and research work, practice, projects, final papers, and so on. The points obtained by the students for these components are then used for certification purposes. Every unit of the syllabus is linked with indicators that show whether it covers educational objectives in the form of knowledge, abilities, competencies, and skills that students acquire. Gathering, processing, and analyzing the data will enable students to be grouped based on the extent to which they have achieved the stated educational objectives.

We represent the knowledge, skills, competencies, and abilities acquired by a student with an ID as a vector $F^{ID} =< F_1^{ID}, F_2^{ID}, \ldots, F_{k_i}^{ID} >$, $(3 \leq k_i)$, where F_j^{ID} is the j-th knowledge/ability/ competence/skill acquired by the student identified by this ID.

Cluster analysis of data should be conducted on both individual components and groups of components while excluding others, [16]. Typically, the number of clusters is set to three, corresponding to the categories of "bad," "satisfactory," and "good." Visualizing the analysis results can aid in performing a SWOT analysis on the implementation of an educational program by highlighting the program's strengths, weaknesses, and possible risks. The outcomes of the analysis can be used to take corrective measures to eliminate disadvantages and enhance educational processes.

As a part of this research, some methods have been created to evaluate the outcomes of students' education in programming subjects, including programming fundamentals and programming languages. To accomplish this, the e-judge interactive environment was used, which enables automated verification of code written in various programming languages, [17].

As is known programming is a discipline that involves creating algorithms to solve problems effectively. A good algorithm should be optimal in terms of complexity. For the student, it is essential to have the ability to create correct algorithms and evaluate their complexity. Programming training involves mastering the syntax of a programming language, as well as acquiring the skills to build effective and accurate programs in that language. This means that the program should work correctly for all proposed test cases.

Automated systems designed to support programming disciplines such as e-judge, Code Signal, etc. are equipped with tools that help check the syntax of programs in the desired programming language. They support the launch of the programs on specific input data and provide results on test cases. These results are available in a convenient format for both manual and automated analysis, [17].

This paper examines the use of data intellectual analysis methods to improve education by analyzing data from knowledge-testing environments. The purpose of this study is to create a toolset for exploring students' educational paths, monitoring their progress, identifying challenging topics, and implementing effective teaching methods. To solve the problem, data clustering is proposed, [16].

Our research is conducted to evaluate students' achievements in programming disciplines. The students are given tasks to programming in different programming languages such as C, C++, and Assembler. The students must implement the tasks in the e-judge environment, [17]. The tasks are provided with test cases that the e-judge system uses to test the programs.

The e-judge system presents program execution results on test cases in different forms. If there are compilation errors in the program that point gaps in learning by students the syntax of the programming language, the system returns a "Compilation Error" message. If not all tests have been passed, an "Execution Error" message is returned. If the program is ineffective and exceeds the maximum operating time, the system returns a "Maximum Operating Time has been Exceeded" message. Finally, if the program executes successfully on all the test cases, the system returns an "Awaits for Confirmation" message, [17].

Data clustering can be used to identify how students learn the syntax of programming languages, whether they create complete programs with all the necessary functionality, how effective their programs are, and which tasks are the most challenging for them.

The data that needs to be analyzed consists of vectors that represent students. Each vector reflects the progress made by the respective student in a training module section and provides a summary of his task completion. The student vector contains various components such as the student's ID, number of tasks completed, task numbers, number of completed tasks that ended in a particular status, and the dates of the first and last submission of completed tasks. These vectors will be used to cluster the data.

Clustering is a process of dividing a set of objects into groups called clusters. The main goal is to place similar objects in one cluster while keeping significantly different objects in separate clusters, [16]. A cluster can be defined as a group of objects that share common properties. Data clustering helps us to understand the key issues related to learning and to identify weaknesses in the educational process. The results of data analysis can be used to create an individualized learning approach for each student by developing an adaptive learning scenario, [11], [18].

The process of cluster data analysis is carried out in several stages. First, features are selected based on which clustering will be performed. Next, a measure of distance between objects is chosen. Then, a clustering method is selected. Finally, the reliability of the clusters is interpreted and assessed, [16].

In this work, two cluster analysis methods were used: the K-means method and the DBSCAN method, [16].

The K-means method involves selecting K (K>=2) centroids randomly among the data. The distance from each data point to each centroid is then calculated, and each data point is assigned to the cluster with the nearest centroid. After this, the centroids are recalculated by taking the average of all points in a given cluster. The algorithm ends when the centroids of the newly formed clusters do not change, or the data remains in one cluster without moving from one cluster to another, or if the maximum number of assigned iterations has been completed, [7], [15], [16].

The DBSCAN algorithm, short for Density-based spatial clustering of applications with noise, is a clustering technique that takes into account the distribution density of a random variable. It works by grouping points that are located close to each other and labeling the points that are situated in areas of low density as noise. The algorithm identifies neighboring points within a specific neighborhood of a chosen point. If the number of such points exceeds a given threshold, a new cluster is created. Otherwise, the point is marked as noise. Then the algorithm assigns all the points from the given neighborhood to the same cluster as the main point. These steps are repeated for unvisited points as well as for those that are marked as noise, [16].

Clustering provides an easily interpretable pattern of results by using the centroid values of each cluster. The centroid represents the most typical data or prototype in a cluster. However, it does not necessarily describe any specific instance in that cluster.

This work addresses the challenges of analyzing task complexity, tracking individual student progress, and evaluating educational paths in a course topic.

To categorize the tasks based on their level of difficulty and identify topics that may require further study, k-means cluster analysis is conducted [16]. The problems are categorized into three different levels of difficulty: easy, medium, and advanced. This is accomplished by computing the percentage of correct answers for each problem, and then conducting cluster analysis to group the tasks based on their level of difficulty for students (Figure 2). The DBSCAN method is utilized to assess students' individual work on a specific task. The data processed by the method pertains to each student and the particular task, reflecting the amount of effort the student put into solving the given task. The data vector includes information about the student's ID number, task number, number of attempts, and the date when results such as "compilation error," "working time exceeded," or "incorrect answer" were recorded.

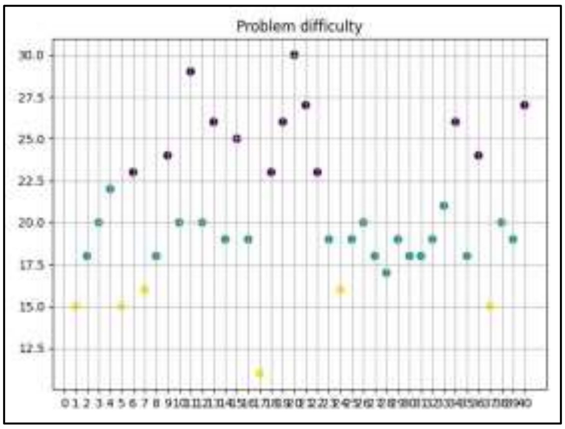

Fig. 2: Cluster analysis of tasks by complexity

The DBSCAN method clusters a set of these vectors, and the results are visualized (Figure 3).

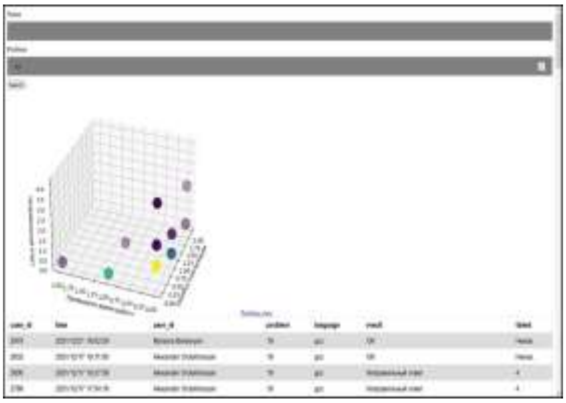

Fig. 3: DBSCAN analysis of one specific task

By using statistical analysis of data provided by the e-judge system, one can generate summary information that includes both solved and unsolved tasks. This information can be used to develop an effective learning strategy (Figure 4), [11], [18].

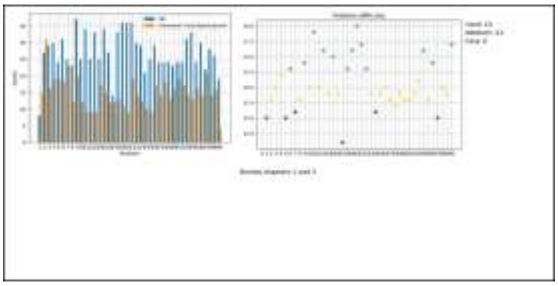

Fig. 4: Information about solutions to problems

The teacher can obtain information about the student΄s progress regarding the specific task. Filtering is performed by the student name and a task, and the student’s progress in solving the problem is visualized (Figure 5).

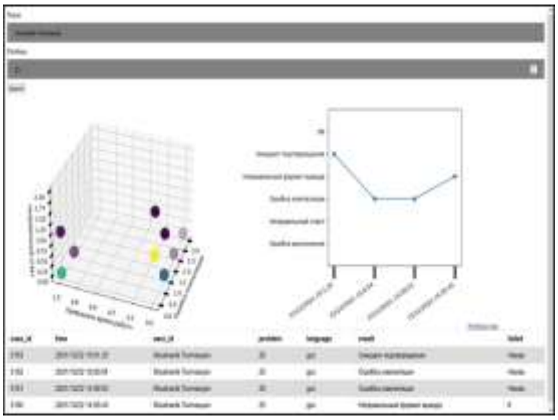

Fig. 5: Filtering by the student’s name and a specific task

4 Conclusion

Most of the studies on the use of data mining techniques in the field of education (EDM) focus on the use of e-learning technologies such as Moodle, WebCT, Blackboard, Classroom, etc., and student knowledge automated testing using tools like Google Forms, Moodle, e-judge, Code Signal, etc., [17], [19].

Currently, the main challenge is to evaluate the level of attainment of the final educational outcomes (learning outcomes), where student performance plays a crucial role. Predicting student performance based on their learning path can help both students and teachers enhance their learning and teaching methods. Frequently, the current research is limited by the statistical techniques utilized for processing data. The application of data mining methods to the educational process is an important issue that requires attention. This includes data collection, problem formulation, clarification of the methods used, determination of forecasting goals, and practical application of the results obtained.

One practical way to implement research results is to develop recommendation systems for personalized learning. These systems will provide students with a learning path that is tailored to their needs. Recent research on personalized learning systems examines some simple features such as learner preferences, interests, and learning and testing behavior.

This paper demonstrates the outcomes of a cluster analysis that was applied to data related to teaching programming and programming languages. The purpose of this analysis is to identify the shortcomings in the educational process and to develop new teaching and learning strategies that are tailored to the needs of the students.

If more data becomes available in the future, it will be possible to apply additional data analysis methods that will yield more effective results. It would also be desirable to have the ability to automatically connect to databases of automated testing systems to process real-time data, [17], [19].

The analysis via data mining methods provides insights into the importance of required courses in the syllabus. A valuable tool for teachers is an analysis tool that predicts the degree to which a student will achieve her/his educational goals.

We consider also an opportunity to use machine learning technologies to predict the expected learning outcomes for a student, as well as to create warning and recommendation systems to avoid undesirable academic outcomes, [6], [15].

References:

[1] A. A. Barseghyan, M. S. Kupriyanov, I. I. Kholod, M. D. Tess, S. I. Elizarov. *Analysis of data and processes: textbook.* 3rd ed., BHV-Petersburg, 2009 (in Russian).

[2] Kovalev E.E. A System Model and Tools for Modernization of Federal and Regional Digital Services of Statistics and Data Analytics in Education, *Lecture Notes in Networks and Systems (Volume 1).* Germany. Springer Nature, 2021.

[3] Fiofanova O. A. Data Analysis Competencies in Professional Standards: From Data-Experts to Evidence-Based Education / Advances in Natural, Human-Made, and Coupled HumanNatural Systems Research, *Lecture Notes in Networks and Systems (Volume 1).* Germany, Springer Nature, 2021.

[4] Ch. Fischer, Z. A. Pardos, R. Sh. Baker, .J.Williams, P. Smyth, R. Yu, S.Slater,R.Baker, M. Warshauer, Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*. Vol. 44, issue 1, 2020. pp. 130–160. doi: https://doi.org/10.3102/0091732X20903304.

[5] Nong Ye. *The Handbook of Data Mining*. CRC Press, 2004.

[6] Mohammed J.Zaki and Wagner Meira. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. 2nd ed., Cambridge University Press, 2020.

[7] Siranush Sargsyan, Anna Hovakimyan, Varditer Kerobyan. An Approach to Developing and Implementing a Recommendation System. *International Journal of Economics and Management Systems*. ISSN: 2367-8925, Vol.7, 2022, pp. 270-273, [Online]. https://www.iaras.org/home/caijems/an-approach-to-developing-and-implementing-a-recommendation-system (Accessed Date: October 20, 2024).

[8] A. H.Cairns, B.Gueni, M. Fhima, A. Cairns, S.David, N. Khelifa. Process Mining in the Education Doмain. *International juornal on Advances in Intellegent Systems*, Vol.8, no.1&2, 2015, pp.219-232, [Online]. https://www.thinkmind.org/library/IntSys/IntSys_v8_n12_2015/intsys_v8_n12_2015_18.html (Accessed Date: October 20, 2024).

[9] Ginica Mahajan, Bhavna Sahini, EducationalData Minig: a state-of-the-art survey on tools and techniques used in EDM, *International Journal of Computer Applications & Information Tecnology*, Vol.12, No.1, 2020, pp. 310-316, [Online]. https://www.researchgate.net/publication/340983783 (Accessed Date: October 20, 2024).

[10] Boyarinov D. A. Knowledge integration maps in the context of the issue of automated pedagogical information processing *Problemy sovremennogo obrazovaniya.* 2019, No. 6, pp. 232–239, [Online]. http://www.pmedu.ru/images/2019-6/22.pdf (in Russian). (Accessed Date: October 20, 2024).

[11] Khan M. A., Khojah M., Vivek V. Artificial Intelligence and Big Data: The Advent of NewPedagogy in the Adaptive E-Learning System in the Higher Educational Institutions of Saudi Arabia. *Education Research International.* Vol.2, 2022, Article ID 1263555, 10p, doi: https://doi.org/10.1155/2022/1263555.

[12] C.Romero, S.Ventura, M. Pechenizkiy and R.Baker. *Handbook of Educational Data Mining.*, Taylor&Francis, 2010.

[13] V.Sothavilay, K. Yacef and R.A.Calvo, Process mining to support student's collaborative writing, *3rd International Conference on Educational Data Mining Proceedings*, Pittsburgh,PA, June 11-13, 2010, pp.257-266, [Online]. https://educationaldatamining.org/EDM2010/uploads/proc/2010%20Proceedings%20Preface,%20TOC.pdf (Accessed Date: October 20, 2024).

[14] M. Pechenizkiy, N.Treka, E.Vasilyeva and P. De Bra. Process mining online assesment data, 2nd International Conference on *Educational Data Mining (EDM09) Proceedings*, Cordoba, Spain.July 1-3, 2009, pp.279-288, [Online]. https://www.educationaldatamining.org/EDM2009/uploads/proceedings/edm-proceedings-2009.pdf (Accessed Date: October 20, 2024).

[15] Andreas C. Müller, Sarah Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly, 2017.

[16] L. Kaufman and P. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, Wiley&Sons, Inc. 1990.

[17] eJudge, [Online]. http://ejudge-y.ispras.ru/c/, http://ejudge-y.ispras.ru/asm/ (Accessed Date: June 26, 2024).

[18] Boyarinov D. A. Pedagogical Model for Creating Individual Learning Paths Based onEducational Maps. *VI International Forum on Teacher Education,ARPHA Proceedings 3,*

Kazan, 2020, pp. 277–289. https://doi.org/10.3897/ap.2.e0277.

[19] C.Romero, S.Ventura and E. Garcia. Data mining in course management systems: moodle case study and tutorial. *Computers& Education*, Vol.51, No.1, 2008, pp.368-384, https://doi.org/10.1016/j.compedu.2007.05.016.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Anna Hovakimyan carried out the gathering of information from the e-judge system.
- Siranush Sargsyan was responsible for the implementation of clustering algorithms.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.