Online at: https://researchtrendsjournal.com **ISSN No:** 2584-282X

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 2; Issue 5; 2024; Page No. 109-113

Received: 08-06-2024
Accepted: 18-07-2024

To study multidimensional dataset acquisition from various data sources, with a particular focus on data banks of recognised school education departments

[1]Gopinath Puppala and [2]Dr. Ajay Kumar Chaurasia

[1]Research Scholar, Department of Computer Application, Maharaja Agrasen Himalayan Garhwal University, Uttarakhand, India
[2]Assistant Professor, Department of Computer Application, Maharaja Agrasen Himalayan Garhwal University, Uttarakhand, India

Corresponding Author: Gopinath Puppala

Abstract

These days, association rule mining is a crucial area of study. Both the hybrid dimensional association rule mining and the general survey of multidimensional association rules are included in this study. The various methods for mining multidimensional and hybrid dimensional association rules are demonstrated in this paper. The conditional hybrid dimensional association rule is also explained in this work, which also determines the optimal method for mining the multidimensional and conditional hybrid dimensional association rule. Educational data mining has the potential to use a vast quantity of research to address a variety of learning, cognitive, and assessment-related educational issues. Although student enrollment has significantly increased in the national setting of the school education system, student performance remains inadequate. The Indian government's Ministry of Human Resource Development has released comprehensive data regarding the school system's dropout rates. Before creating a program for pupils to improve their performance, the school education system requires that students' learning behaviors be analyzed. Additionally, early student performance prediction helps management take corrective action to improve student accomplishment. Data mining offers a wide range of methods for analyzing and forecasting student performance. Given the state of the educational system today, a decision support system that relies solely on mining techniques is unable to manage the massive information and provide complicated answers. The goal of this study is to offer a strategy for the educational system in schools to address the intricate educational problems and increase student performance. Data mining technologies have been taken into consideration in this research project in order to address educational questions.

Keywords: Hybrid, data mining, performance, educational, student

1. Introduction

The platform that allows people and society to contribute to the nation's progress is education. Knowledge and talents are the weapons that every person possesses. Giving young people the proper direction and skills has a big impact on the country's overall development and economic expansion. Over the past ten years, as the number of students enrolled has increased, so too has the amount of data pertaining to students in the education sector, making it challenging to assess the students' academic achievement. Thus, it is now crucial to create strong tools for evaluating student academic data in order to extract knowledge that is useful. Students' performance will be improved with the use of this new information. Numerous factors, including socioeconomic status, examination results, and demographics, influence pupils' academic performance. Numerous factors that correspond to distinct entities-students, faculty, infrastructure, and the learning environment-are involved in the education sector. These aspects are multifaceted. Each student's basic, personal, academic, and exam information is described by the student entity. In a similar vein, the facilities and instructor profile offered in a particular course are represented by the faculty and infrastructure entities. There may be direct or indirect connections between different student characteristics that should be noted. Furthermore, the need to create new technologies that make intelligent use of knowledge and information arises from the excessive proliferation of databases. Therefore, in order to analyze student performance, an automated decision support system that finds instructional patterns is required.

1.1 Data Mining

The process of obtaining knowledge from diverse data sources is known as data mining. It is the technique that looks for recurring patterns and methodical connections between variables by analyzing vast amounts of data. After determining the fundamental relationships between various entities, it entails validating the patterns found to new data subsets. As seen in Figure, data mining is an iterative process that begins with problem definition in phase 1, moves on to data collection through sampling and data preparation through transformation in phase 2, builds a model and evaluates it in phase 3, and concludes with knowledge deployment.

Data mining is the process of semi-automatically or automatically analyzing vast amounts of data to find previously unknown, valuable patterns, data record clusters, rare records, and dependencies. Many fields have chosen to use data mining talents in order to obtain important facts and retrieve data more quickly. A few data mining trends [2] that highlight the pursuit of information include the structure of the coupled and communicative data mining atmosphere: -

- Domain-specific applications, including data mining for biomedicine (DNA), banking, retail, and telecommunications;
- Visual data mining, which includes interactive visual mining, data visualization, and visualization of the mining process and outcome.
- One type of data mining technique that is helpful for finding trends on the internet is web mining.

The study of educational datasets using a variety of computational techniques is known as educational data mining. Educational Data Mining is the outcome of applying data mining techniques to the field of education.

In order to make the system scalable, distributed data mining divides the job among multiple locations. The intrusion detection system, a technique to safeguard network systems employed in real-time, is a component of real-time data mining.

Multi-database mining is a distributed data mining technique where data is dispersed over several databases. A single dataset that can be mined using conventional methods can be produced by combining these databases using a variety of methods.

Semantic query optimization and database knowledge querying are aided by intelligent query responding. Database systems can use data mining technologies to intelligently respond to queries.

Data mining can help achieve information security and privacy protection.

In addition to these, the fields of business analysis, production control, risk management, science exploration, and customer retention can all benefit from the use of data mining tools. Educational Data Mining (EDM) is the term of the data mining field that focuses on education. It covers the methods, resources, and research strategies used to gather data from academic records, such as internet logs and test results, and then analyzes this data to make judgments.

2. Motivation

Although the size of India's school system has grown over the past 10 years, the system's quality remains inadequate in terms of student achievement and retention rates. With reference to the MHRD education census conducted between 2011 and 2014, certain facts about the student dropout rate are revealed. Statistics on student dropout rates from 2011 to 2014 are shown in Figure. According to the 2011 education survey, the dropout rate was 28.9% for I through V classes, 42.4% for I through VIII, and 52.8 for I through X standard. In 2012, the dropout rate was 27.0% for I to V classes, 40.6% for I to VIII, and 49.3% for I to X. Likewise, in 2013, the dropout rate was 27.0% for I to V classes, 40.6% for I to VIII, and 49.3% for I to X. Furthermore, the situation remained unchanged until V class in 2014. Dropout rates were 19.8% for classes up to VIII, 36.3% for classes up to X, and 47.4% for classes up to X.

3. Objectives of the study

1. Subject-oriented multidimensional dataset acquisition from various data sources, with a particular focus on data banks of recognised school education departments.
2. Integrate and preprocess the dataset to remove anomalies before converting it to a consistent format for optimal feature extraction and analysis.

4. Research Methodology

The new field of study known as "educational data mining" (EDM) focuses on creating methods for examining the unique data that comes from educational settings. The suggested architecture of the educational data mining system that corresponds to the analytics used to examine student data for educational inquiries is the main topic of this chapter. Exam data from the Punjab School Education Board in India over a ten-year period is used to validate the suggested architecture. The gathered dataset is organized as a central warehouse and includes a variety of student variables.

An association rule classification method that identifies the recurring pattern from multidimensional educational data is proposed in order to find the relationships among student features. The administration will have a better understanding of the reasons behind the high student failure and dropout rates thanks to the patterns that have been found. Developing new policies to improve pupils' academic achievement would also be beneficial. The high-level and detailed designs of the research technique, as stated in the following sections, are described in the proposed system architecture.

Each module of the suggested system architecture has a more thorough label thanks to the study methodology's precise design. The "Multidimensional Data Extraction," "Association Rule Classification System," "Data Visualization," and "Validation & Testing" modules make up the thorough design. Multi-dimensional data extraction is the first step in this process, which is further broken down into several stages: gathering pertinent data from federated educational sources, eliminating anomalies, dealing with missing values, configuring working attributes, and finally converting the data into the required format. To prepare it for the following module, the pre-processed data is loaded into the warehouse. The pre-processed dataset is mined by the association rule classification system in the following phase. The purpose of the hybridization of association and classification rule mining is to uncover the hidden dependencies between the dataset's attributes for pupils. As part of the mining process, the domain expert might also include the educational pattern in the form of association rules. To find the workable ruleset, the significant restrictions that correspond to the included rules have also been entered. The ruleset repository contains the resulting pattern that was found under this module. All of the result sets found using the association rule classification method are sent into the data visualization stage, which shows the result set graphically. Validating the result set both objectively and subjectively is the final step in the suggested process. The subjective validation provides an empirical description of the testing process's conforming condition. In a similar vein, objective validation takes into account the threshold and constraint values that need to be established at the beginning.

5. Results and Data Interpretation

In recent years, choosing the appropriate value set has also emerged as one of the applications of numerous study fields. Datasets with many attributes are becoming common in many study fields due to the quick development of database technologies. The gathered data set includes a number of student attributes that match their domain values. Certain attributes in the gathered data set have direct values, whereas other attributes include metadata about other properties. Therefore, it becomes vital to set up working qualities that contribute to the best possible outcomes of the learner's performance in order to pick attributes for future analysis. We draw the conclusion from the experimental investigation that dealing with non-contiguous (just nominal values) characteristics yields more valuable outcomes than contiguous (containing both nominal and numerical values) attributes or a composite value set. As a result, there is less computing overhead, which enhances mining process performance. As a result, working attributes have been built up based on contiguous values after the data has been pre-processed into the necessary form. As previously stated, just 23 of the approximately 115 attributes in the gathered dataset have been configured for processing. The attributes that have been put up for analysis in this study and correlate to their domain value are listed in Table 1.

Table 1: List of set-up attributes corresponding to their domain values collected from PSEB.

Sr. No.	Attribute Name	Description	Domain Values
1	Sex	Students' Gender	{Male, Female}
2	Area	Students' Learning Area	{Rural, Urban}
3	Punjabi Result	Result of Punjabi Subject	{Pass, Fail}
4	Punjabi Grade	Grade of Punjabi Subject	{A+, A, B, C, D, E}
5	English Result	Result of English Subject	{Pass, Fail}
6	English Grade	Grade of English Subject	{A+, A, B, C, D, E}
7	Hindi Result	Result of Hindi Subject	{Pass, Fail}
8	Hindi Grade	Grade of Hindi Subject	{A+, A, B, C, D, E}
9	Math Result	Result of Mathematics Subject	{Pass, Fail}
10	Math Grade	Grade of Mathematics Subject	{A+, A, B, C, D, E}
11	Science Result	Result of Science Subject	{Pass, Fail}
12	Science Grade	Grade of Science subject	{A+, A, B, C, D, E}
13	SS Result	Result of Social Study Subject	{Pass, Fail}
14	SS Grade	Grade of Social Study Subject	{A+, A, B, C, D, E}
15	Physical Result	Result of Physical Education Subject	{Pass, Fail}
16	Physical Grade	Grade of Physical Education Subject	{A+, A, B, C, D, E}
17	Computer Result	Result of Computer Science Subject	{Pass, Fail}
18	Computer Grade	Grade of Computer Science Subject	{A+, A, B, C, D, E}
19	Agriculture Result	Result of Agriculture Subject	{Pass, Fail}
20	Agriculture Grade	Grade of Agriculture Subject	{A+, A, B, C, D, E}
21	Final Result	Final Result of Matriculation Course	{Pass, Fail, Reappear}
22	Final Grade	Final Grade of Matriculation Course	{A+, A, B, C, D, E}
23	District	All districts name of Punjab state	District name subject to year of establishment

The PSEB, India student data set is used to extract these working attributes and the domain values that go with them. Different information about the students is described by each attribute. The student's gender is defined by the "Sex" attribute, which has the domain values Male or Female. The student's learning area is further defined by the "Area" element, which has the domain values "Rural" or "Urban." The following attributes-Punjabi Result, English Result, Hindi Result, Math Result, Science Result, SS Result, Physical Result, Computer Result, and Agriculture Result-define whether students passed or failed in their respective subjects. Similar to this, the following attributes-Punjabi, English, Hindi, Math, Science, SS, Physical, Computer, and Agriculture-define students' grades in a variety of disciplines. The possible domain values are A+, A, B, C, D, and E. Pass, Fail, and Reappear are the potential domain values for the ultimate Result attribute, which specifies the student's ultimate outcome in the relevant course. In a similar vein, Final Grade indicates the student's entire course grade, with possible domain values of A+, A, B, C, D, and E. To create a single consistent data set for additional processing, the chosen attribute is combined into a single data-set.

Table 2: Sample of a dataset having area, Gender, and subject-Wise result with domain values PASS (P), Fail (F), and Reappear (R).

Area	Gender	English Result	Math Result	Science Result	Social study Result	Punjabi Result	Hindi Result
U	M	P	P	F	P	P	P
U	F	P	P	P	P	P	P
R	F	F	P	P	F	P	P
R	F	P	P	F	P	P	P
U	M	F	P	P	F	P	P
R	M	P	P	P	P	P	P
U	M	P	P	F	P	F	P
U	F	P	P	P	P	P	P
U	F	P	P	P	F	P	F
R	F	P	F	P	F	P	P

The relational file is changed into a transactional file by using the aforementioned procedure on the example dataset. Only attribute names that meet the constraint equal to a given domain value are included in the resulting transaction file. The pass-based transaction file, shown in Figure, only includes characteristics that meet the pass values from the domain value set; the remaining attributes are absent because they lack pass values.

Table 3: The Asymptotic Significance (Pearson) Values

	Value	Degree of freedom	Asymp. Sig. (2-sided)
Pearson Chi-Square	69.260[a]	4	.001
Likelihood Ratio	37.620	4	.000
Linear-by-Linear Association	18.629	1	.000
N of Valid Cases	175		

The pupils' academic success in the senior secondary course is indicated by Question. As a result, section two outlines the inquiries to confirm whether any DSS or analytical tool that corresponds to different functions is available. In order to verify their dependencies, we have compared each of the questions in section two to question. To statistically validate question with all of the section two questions, the Chi-Square test has been used in addition to cross-tabulation. As previously mentioned, question inquires about the senior secondary pupils' performance in relation to the "Excellent," "Good," "Average," and "Poor" scaling parameters. The existence of any DSS or analytical tool to examine the academic achievement of the students corresponding to the scaling parameters "Yes" and "No," however, is the primary question in section two. In order to confirm their dependency, a statistical test has also been performed between questions. The cross-tabulation of questions, which correspond to 175 responses, is shown in Table 4.

Table 4: Statistics Describes the Cross-Tabulation of Question

Scaling Parameters		Q		Total
		Yes	No	
	Excellent	1	7	8
	Good	21	110	131
	Average	24	12	36
Total		46	129	175

Table 5: The Asymptotic Significance (Pearson) Values

	Value	Degree of freedom	Asymp. Sig. (2-sided)
Pearson Chi-Square	38.191[a]	2	.001
Likelihood Ratio	34.425	2	.000
Linear-by-Linear Association	31.761	1	.000
N of Valid Cases	175		

Likewise, the asymptotic significant values of the variables have also been calculated using the Chi-Square test. The asymptotic significance (Pearson) values of the Chi-Square test used to confirm the dependence between questions are highlighted in Table. The dependencies between variables 1.2 and 2.0 are well-described by the 0.001 asymptotic significance value. Consequently, it was statistically demonstrated that question in order to work. The 46 "Yes" responses to question 2.0, which states that they already use a mechanism to evaluate students' performance, are highlighted in Table. Consequently, in order to verify the dependability of the alternatives requested in section two, validation is required. The availability of varied functions corresponding to the scaling parameters "Yes, Definitely," "Maybe yes, but not sure," and "No, not at all" is the subject of questions. In order to verify their dependencies, all of the scaling parameters in question have been statically validated.

6. Conclusion

Giving young people the proper direction and skills has a big impact on the country's overall development and economic expansion. The platform that allows people and society to contribute to the nation's progress is education. Although the number of students enrolled in Indian schools has significantly increased in recent years, the quality of education as measured by student achievement and retention rates remains poor. With reference to the MHRD's 2011–2014 education census, several points are revealed. Numerous factors influence pupils' academic achievement. This domain has a large number of attributes that correspond to the relevant entities and describe specifics about each of them. Therefore, in order to determine the direct and indirect relationships between different features, the collection of attributes must be analyzed. This study is an effort to use data mining techniques to the field of education in order to assess student academic data and improve the quality of the school-level educational system. Getting the student dataset from PSEB was the first stage in this research project. Over two million students' records of matriculation and senior secondary courses are gathered from the dataset of previous decades. This dataset includes each student's academic, personal, and exam information. Additionally, a pilot study was carried out to examine student performance from 1994 to 2019 in relation to PSEB enrollment and pass rate. The primary goal of this pilot project was to verify that the chosen dataset was appropriate for the stated research goal.

Following the gathering of the necessary dataset, the obtained dataset is subjected to multidimensional data extraction in order to get it ready for analytical mining.

Anomalies are eliminated first in this method, and then working qualities are built up. Setting up the working attributes is a crucial step in the data preparation process since choosing the appropriate domain value helps to extract useful information from the dataset of choice. We came to the conclusion during the data preparation pilot research that using non-contiguous attributes yields more valuable results than using contiguous attributes or a combined value set.

7. References

1. Tai-Chang H, An-Jin Shie, Li-Chen Chen. Course planning of extension education to meet market demand by using data mining techniques – an example of Chinkuo Technology University in Taiwan. Expert Systems with Applications. 2008;34:596-602.
2. Knowles JE. Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. JEDM: Journal of Educational Data Mining. 2015;7:18-67.
3. Chun-Hsiung L, Gwo-Guang L, Yungho L. Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning. Expert Systems with Applications. 2009;36:1675-1684.
4. Vialardi C, Bravo J, Shafti L, Ortigosa A. Recommendation in higher education using data mining techniques. International Working Group on Educational Data Mining. 2009.
5. Ning F, Jingui L. Work in progress - A decision tree approach to predicting student performance in a high-enrollment, high-impact, and core engineering course. In: 39th IEEE Frontiers in Education Conference; 2009 Oct 18-21; IEEE; c2009. p. 1-3.
6. Thai N, Janecek P, Haddawy P. A comparative analysis of techniques for predicting academic performance. In: 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports; 2007 Oct 10-13; IEEE; c2007.
7. Paulo C, Silva A. Using data mining to predict secondary school student performance. In: ECEC-FUBUTEC'2008: European Concurrent Engineering and Future Business Technology Conference; 2008 Jun 25-27; Portugal; c2008.
8. Ramaswami M, Bhaskaran R. A CHAID-based performance prediction model in educational data mining. International Journal of Computer Science Issues (IJCSI). 2010;7:10-18.
9. Ministry of Human Resource Development (MHRD). Students' dropout rate from 2011 to 2014 census. In: Education Statistics at Glance; 2015. Available from: www.mhrd.gov.in
10. Borah M, Jindal R, Gupta D, Deka C. Application of knowledge-based decision technique to predict student enrollment decision. In: International Conference on Recent Trends in Information Systems; 2011 Dec 19-21; IEEE; c2011. p. 180-184.
11. Braga D, Campi A, Klemettinen M, Lanzi P. Mining association rules from XML data. In: International Conference on Data Warehousing and Knowledge Discovery; 2002 Sep 2-5; Springer; c2002. p. 21-30.
12. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data; 1993 May 26-28; ACM; c1993. p. 207-216.
13. Merceron A, Yacef K. Interestingness measures for association rules in educational data. In: Educational Data Mining 2008; c2008.
14. Minaei-Bidgoli B, Kashy DA, Kortemeyer G, Punch WF. Predicting student performance: an application of data mining methods with an educational web-based system. In: 33rd Annual Frontiers in Education Conference, FIE 2003; 2003 Nov 5-8; IEEE; c2003. p. T2A-13.
15. Pang-Ning T, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002 Jul 23-26; ACM; c2002. p. 32-41.
16. Buldu A, Üçgün K. Data mining application on students' data. Procedia - Social and Behavioral Sciences. 2010;2:5251-5259.
17. Abdullah Z, Herawan T, Ahmad N, Deris M. Mining significant association rules from educational data using critical relative support approach. Procedia - Social and Behavioral Sciences. 2011;28:97-101.
18. Kiran RU, Krishna Re P. An improved multiple minimum support-based approach to mine rare association rules. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining; 2009 Dec 7-10; IEEE; c2009. p. 340-347.
19. Tanimoto SL. Improving the prospects for educational data mining. In: Track on Educational Data Mining, Workshop on Data Mining for User Modeling; 2007 Jul 15-19; 11th International Conference on User Modeling; c2007. p. 1-6.