IBIMA Publishing
The MENA Journal of Business Case Studies
https://ibimapublishing.com/articles/MENA/2020/467550/
Vol. 2020 (2020), Article ID 467550, 9 pages, ISSN: 2326-7038
DOI: 10.5171/2020.467550

Research Article

Data Science for Prediction of Grades in a Mathematics Course based on Performance in its Prerequisites

Muhammad Abaidullah Anwar and Rashmi Rani

AlGhurair University, Dubai, UAE

Correspondence should be addressed to: Muhammad Abaidullah Anwar; anwar@agu.ac.ae

Received date: 26 February 2020; Accepted date: 7 August 2020; Published date: 13 October 2020

Academic Editor: Azzam Jamil Falah AlRifaee

Copyright © 2020. Muhammad Abaidullah Anwar and Rashmi Rani. Distributed under Creative Commons Attribution 4.0 International CC-BY 4.0

Abstract

Data mining is one of the important techniques in data science and has been effectively and efficiently used to extract useable, previously unknown, comprehensible, useful, and actionable knowledge from a large database whether structured or unstructured. Data mining is important for supporting crucial business decisions, identifying disease diagnostics, and predicting what may happen in future academics as well. This paper presents the application of Apriori algorithm of data mining to mine associate rules among the marks scored by students in prerequisite and successor mathematics courses in an engineering degree program. The analysis of rules reveals that students who scored better marks in the prerequisite courses will 100% maintain their same performance in the successor courses. The mined rules also revealed that association rule mining could be used effectively for predicting grades and also adapting the teaching methodologies to teach mathematics courses.

Keywords: Educational Data Mining, Association Rules, Confidence, Support.

Introduction

The most important targets of any educational system are to provide students with the required knowledge and skills to implement them in successful professions within a specified period. How global educational systems effectively meet this goal is a major determinant of both economic and social progress.Baker, R. S. (2014) mentioned that in recent years, the increasing attention towards Artificial Intelligence (AI) inspired the development of data mining and analytics in the pedagogical domain. Data mining is the method to obtain additional characteristics and models from a large data set. It uses the methods of machine learning, statistics and database systems. Fayyad, U. et al. (1996) mentioned that data mining is a field of knowledge discovery in databases (KDD), which is the area of detecting individual and hypothetically

Cite this Article as: Muhammad Abaidullah Anwar and Rashmi Rani (2020), " Data Science for Prediction of Grades in a Mathematics Course based on Performance in its Prerequisites" The MENA Journal of Business Case Studies, Vol. 2020 (2020), Article ID 467550, DOI: 10.5171/2020.467550

advantageous information from a large amount of data set. The data mining that concentrates on the educational area is called Educational Data Mining (EDM). EDM refers to the techniques, tools, and research designs utilized to obtain information from educational records, typically online logs, and examination results, and then analyse this information to formulate conclusions. Berland, M. et al. (2014) mentioned that EDM is theory-oriented and focuses on the connection to the pedagogical theory. Papamitsiou, Z. and Economides, A. A. (2014) mentioned that presently, little empirical evidence exists to support a theoretical framework that is able to gain wide acceptance in the scientific community. Given that in the real world there is a great diversity of different learning contexts, they determine the analytical approaches utilized by EDM. Therefore, how EDM can be beneficial in real educational practices, as demonstrated in the research, could be crucial.

The website International Educational Data Mining Society (2011) mentioned that Educational Data Mining, one of the important techniques in data science, is an emerging field, concerned with the rising procedures for discovering exclusive and progressively large-scale data attained from educational surroundings, and uses those methods to better understand students and the settings in which they learn. Koedinger, K. et al. (2008) state that EDM permits users to extract knowledge from students' data. This experience can be used in different ways such as to substantiate and assess an educational system, enhance the quality of T & L processes, and lay the groundwork for a more effective learning process.

EDM can provide universities with a clear picture of specific hindrances to student learning. For example, students can fail in advanced subjects because they did not learn the basic information from the prerequisite subjects. Using data mining (DM) techniques to analyze students' information can help identify possible reasons for students' failures. Data mining delivers many procedures for data analysis. The enormous quantity of information presently in students' databases surpasses the human capability to evaluate and obtain the most useful information without help from computerized analysis procedures. Knowledge discovery (KD) is the process of nontrivial extraction of implicit, unknown and potentially useful information from a large database. Data mining has been used in KDD to discover patterns with respect to a user's needs. The pattern definition is an expression in language that describes a subset of data. An example of a KDD pattern definition was use by Agrawal, R. (1993).

Section 2 presents association rule mining. Section 3 presents the data and methods, and a discussion of the results is presented in section 5, followed by a conclusion.

Association Rule Mining

The association rule mining was introduced by Agrawal, R. (1993) in 1993 in an international conference on Management of Data, and since then it has received a notable attention of the researcher in every domain of life. The association rule mining is one of the mostly used machine learning algorithm for hidden patterns discovery in a large database. Association rule mining, in brief, employs the use of machine learning models to analyze data for hidden patterns or co-occurrences in a database. It identifies frequently occurring *if-then* associations that are called association rules. An association rule is composed of two parts; an antecedent (if) and a consequent (then), and are written as "*antecedent* → *consequent*". The antecedent is an itemset in the database and the consequent is an itemset in the database that occurs in combination with the antecedent.

To understand the association rule mining, the terms from Lustgarten, J. L. et al. (2008) are worth mentioning, as follows:

Itemset – a set of one or more items i.e. {Butter, Bread, Milk} in a transaction database.

K-itemset – an itemset which contains K items. For example, {Butter, Bread, Milk} is a 3-itemset.
Support (denoted by 's') – is the frequency or occurrence of an itemset X. It is one of the measures of interestingness that tells about the usefulness and certainty of a rule. A 5% support means total 5% of transactions in a database following the rule.
Frequent itemset –an itemset X is frequent if the support of X is no less than a minimum support threshold.

Confidence (denoted by 'c') – a conditional probability that a transaction having X also contains Y.
Strong rule – an association rule having support greater than or equal to the user-specified minimum support threshold and confidence greater than or equal to a user-specified minimum confidence threshold.

An association rule *Laptop → Wireless Mouse* is presented in Figure 1.

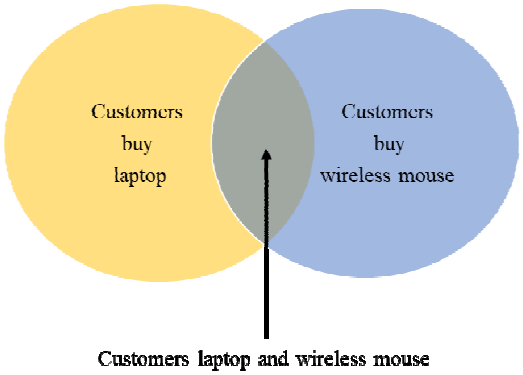

Fig. 1: *Laptop → Wireless Mouse* association rule

This direct applicability to business problems together with their inherent understandability (even for non-data mining experts) made association rules a popular mining method. Moreover, it became clear that association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems.

To predict grades in a course on the basis of performance in the prerequisite course(s), the *Apriori* algorithm described in Markus, H. (2005) was used, which is a quite favorite algorithm for frequent itemset mining and for the identification of association rules over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger itemsets as long as those itemsets appear sufficiently often in the given database. The frequent itemsets determined by *Apriori* algorithm can be used to determine the association rules which highlight general trends in the dataset. An example of few transactions in the database is presented in Table 1.

Table 1: Example transactions database

Transaction ID	Items bought
T100	{A, B, D, K}
T200	{A, B, C, D, E}
T300	{A, B, C, E}
T400	{A, B, D}

Given that the minimum support is 60% and the minimum confidence is 80%.

Step 1: Scan the database for 1-itemsets candidates and frequent 1-itemsets.

1-itemsets candidates (C-1)

Itemset	Support (%)
{A}	100
{B}	100
~~{C}~~	~~50~~
{D}	75
~~{E}~~	~~50~~
~~{K}~~	~~25~~

Frequent 1-itemsets (L-1)

Itemset	Support (%)
{A}	100
{B}	100
{D}	75

Step 2: Scan the database for 2-itemsets candidates and frequent 2-itemsets.

2-itemsets candidates (C-2)

Itemset	Support (%)
{A, B}	100
{A, D}	75
{B, D}	75

Frequent 2-itemsets (L-2)

Itemset	Support (%)
{A, B}	100
{A, D}	75
{B, D}	75

Step 3: Scan the database for 3-itemsets candidates and frequent 3-itemsets.

3-itemsets candidates (C-3)

Itemset	Support (%)
{A, B, D}	75

Frequent 3-itemsets (L-3)

Itemset	Support (%)
{A, B, D}	75

Step 4: Extract all strong association rules from the database.

Frequent Itemsets	Support (%)
{A}	100
{B}	100
{D}	75
{A, B}	100
{A, D}	75
{B, D}	75
{A, B, D}	75

Candidate Rules	Confidence (%)
A → B	100
B →A	100
~~A →D~~	~~75~~
D → A	100
~~B →D~~	~~75~~
D →B	100
~~A, B →D~~	~~75~~
A, D → B	100
B, D → A	100
~~A →B, D~~	~~75~~
~~B →A, D~~	~~75~~
D → A, B	100

The strong association rules that satisfy the minimum support (60%) and confidence (80%) thresholds are given in Table 2.

Table 2: Strong association rules

Candidate rule	Support (%)	Confidence (%)
A → B	100	100
B →A	100	100
D → A	75	100
D →B	75	100
A, D → B	75	100
B, D → A	75	100
D → A, B	75	100

Data and Methods

Data Collection

The data consists of the total marks (0 – 100) scored by the forty-four students who passed the second course in Calculus (SC) and Differential Equations (DE) course of an engineering degree program taught by the same instructor in two consecutive semesters of fifteen weeks. The course Preliminary Calculus is a prerequisite of the Differential Equations course. The sample marks data is presented in Table 3.

Table 3: Sample dataset

S. No.	Marks scored by students	
	2nd Course in Calculus	Differential Equations
1	63.7	76.3
2	68.5	70.5
3	76.2	81.3
4	62.8	81.5
5	63.0	62.8
6	66.3	76.5
7	76.8	91.3
8	81.9	76.5
9	76.1	70.3
10	54.2	63.0

Data Cleaning

There is a number of data cleaning techniques in the literature such as filling missing values, binning, regression, and clustering – some are mentioned in Kiron, D. et al. (2012). The authors did not use any algorithm or technique to handle the missing values in the marks dataset. The data of the students who passed a Calculus course in the preceding semester but did not pass the DE course in the successor semester was not included in this study. The students who did not register to a course taught by the same instructors in the two consecutive semesters were also not included in finding the association rules.

In educational systems, assessment is a predictable component as it has a wide impact on learning. The student's performance in a course is also based on the teaching methodology, the faculty's behavior towards the student, and the faculty's tone and communication style in the classroom. These factors influence the participation of the student in a class. If a different faculty will teach the course, the student's association in class will be affected, and hence it affects the student's grades. The value of the data in this study is that one faculty taught the prerequisite as well as the successor courses to the same students. This ruled out the above mentioned side effects of two courses taught by two different faculties.

Data Discretization

The *Apriori* algorithm works on nominal attributes that require the discretization of any numerical data in the datasets. Discretization is a typically preprocessing step for machine learning algorithms that transformed a continuous-valued feature to a discrete one as mentioned in Lustgarten, J. L. et al. (2008). Cios, K. J. et al. (2007) state that the goal of discretization is to reduce the number of possible values a continuous attribute takes by partitioning them into a number of intervals. The marks scored by students in two courses in this study are numerical, and the authors have converted them to a nominal data consisting of ten categories using MS Excel. The categories and their corresponding range are presented in Table 4.

Table 4: Numerical to nominal – range and category

Sr. No.	Marks Range	Category
1	90 < Marks <= 100	C-10
2	80 < Marks <= 90	C-9
3	70 < Marks <= 80	C-8
4	60 < Marks <= 70	C-7
5	50 < Marks <= 60	C-6
6	40 < Marks <= 50	C-5
7	30 < Marks <= 40	C-4
8	20 < Marks <= 31	C-3
9	10 < Marks <= 20	C-2
10	0 < Marks <= 10	C-1

A sample of the scored marks of the students in two courses transformed into categorical data is presented in Table 5. Now, the scored marks have been made to be eligible for applying mining association rules, since association rule mining supports nominal data only in Weka.

Table 5: Scored marks transformed to Nominal Data

Marks Numerical Data		**Marks as Nominal Data**	
2nd Course in Calculus	**Differential Equations**	**2nd Course in Calculus**	**Differential Equations**
63.7	76.3	C-7	C-8
68.5	70.5	C-7	C-8
76.2	81.3	C-8	C-9
62.8	81.5	C-7	C-9
63.0	62.8	C-7	C-7
66.3	76.5	C-7	C-8
76.8	91.3	C-8	C-10
81.9	76.5	C-9	C-8
76.1	70.3	C-8	C-8
54.2	63.0	C-6	C-7

Results and Discussion

The Apriori algorithm was applied on the dataset using a publically available free data mining tool developed at Waikato University and available at Weka (2019). The Weka tool allows users to apply most of the data mining algorithms on a given dataset. The summary of the performance categories of students in two courses produced by Weka is presented in Figure 2 (a) and Figure 2 (b).

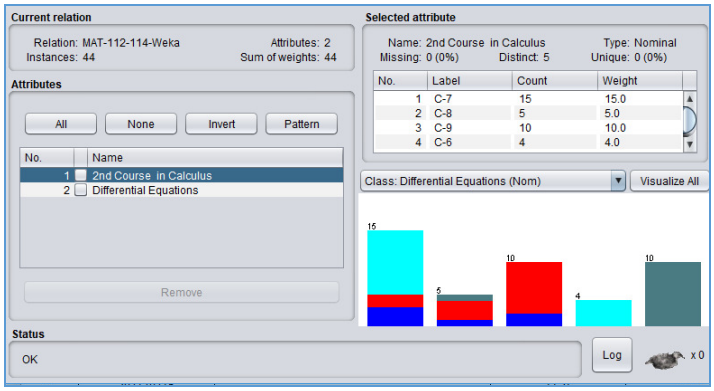

Fig. 2 (a): Summary of 2nd Course in Calculus data

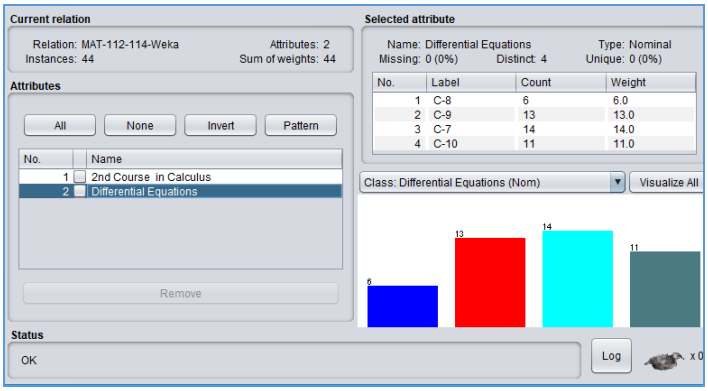

Fig. 2 (b): Summary of Differential Equations data

The Weka generated a set of association rules with the minimum confidence of 70% and minimum support of 20%, as follows:

Number of cycles performed to mine the association rules = 18

The generated sets of large itemsets are as follows:

Size of set of large 1-itemsets = 9
Size of set of large 2-itemsets = 4

The strong rules of interest selected from the association rules mined with a different combination of supports and confidences are presented in Table 6.

Table 6: Strong association rules mined

Rule No.	Antecedent – 2nd Course in Calculus	Consequent – Differential Equations	Confidence (%)
1	C-10	C-10	100
2	C-6	C-7	100
3	C-9	C-9	80

The analysis of rules reveals that students who scored marks in the range of 80 < marks <= 90 and 90 < marks <= 100 in the prerequisite course will 100% maintain their same performance in the successor course i.e. Differential Equation.

There is only one strong rule that reveals that the authors are 80% confident that students who scored marks in the range of 50 < marks <= 60 in the prerequisite course i.e. 2nd Course in Calculus, can improve their performance in the successor course. i. e. Differential Equation.

The analysis of the data disclosed that there are only two students (2/44*100 = 4.5%) whose performance was lowered in the successor course as compared to the performance in the prerequisite course, but it was not mined as a strong rule.

Conclusion

The *Apriori* algorithm proved to be successful in finding the hidden association among prerequisite and successor courses in mathematics to predict grades. The association rules generated in this study were later verified by the expert instructors involved in teaching the related course. The study can further be extended to finding association rules among the related courses in a given baccalaureate program in any discipline.

Acknowledgements

The authors wish to acknowledge the support provided by the AlGhurair University, Dubai, UAE.

References

- Baker, R.S., (2014), Educational data mining: An advance for intelligent systems in education. IEEE Intelligent systems, 29(3), pp. 78–82.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., (1996), From data mining to knowledge discovery in databases. AI magazine, 17(3), pp. 37.
- Berland, M., Ryan|Blikstein, and Paulo, (2014), Educational Data Mining and Learning Analytics: Applications to Constructionist Research. Technology, Knowledge and Learning, 19(1–2), pp.205–220.
- Papamitsiou, Z. and Economides, A. A., (2014), Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), pp. 49.
- International Educational Data Mining Society, (2011), [Online], [Retrieved December 10, 2019], http://educationaldatamining.org/.
- Koedinger K., Cunningham K., Skogsholm A., and 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.
- Agrawal, R., Imielinski, T., and Swami, A. N., (1993), Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.
- Markus Hegland (2005), The Apriori Algorithm – a Tutorial, WSPC/Lecture Notes Series.
- Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., and Haydock, M. (2012), Analytics: The Widening Divide. MIT Sloan Management Review, 53(2), 1-22.
- Lustgarten J. L., Gopalakrishnan V., Grover H., and Visweswaran S., (2008), Improving Classification Performance with Discretization on Biomedical Datasets, AMIA Annual Symposium, Proc., pp.445–449.
- Cios K. J., Pedrycz W., Swiniarski R. and Kurgan L., (2007), Data Mining A Knowledge Discovery Approach, Springer.
- Weka (2019), Weka Tool, [Online], [Retrieved September 20, 2019], http://www.cs.waikato.ac.nz/ml/weka/
- Dunham, M. (2003), Data Mining: Introductory and Advanced Topics, Upper Saddle River, NJ: Pearson Education.