DOI: 10.21917/ijsc.2023.0410

IMPROVED ASSOCIATION RULE MODELLING USING VARIOUS MACHINE LEARNING MODULES FOR LARGE DATASETS

S. Nandagopal
Department of Computer Science and Engineering, Nandha College of Technology, India

Abstract

There are four modules namely Modified Apriori Algorithm (MAA), Crumb Based Association Rule Mining (CBARM), Inter-transaction Association Rule (IAR) miner and Categorized and Bounded Inter-Transaction (CBIT) proposed in this research work. The methodology of data mining is a relatively new field of study that has grown over the course of several decades of research and practise, drawing on the findings made in a wide variety of other fields of study. The reality that data mining studies and implementations are exceedingly difficult cannot be avoided in any manner. The development of data mining follows a process that is analogous to the development of any other new technology. This process begins with the presentation of an idea and is then followed by stages in which the concept is accepted, major research and exploration is conducted, incremental application is performed, and finally mass deployment occurs. The great majority of researchers working in the academic world are of the opinion that the process of data mining is still in its infancy in terms of both research and investigation.

Keywords:

Association Rule, Machine Learning, Rule Mining

1. INTRODUCTION

The Data Mining is held by most researchers working in the academic world. On the other hand, the idea of mining data for useful information has been met with a variety of responses. The range and depth of these exciting theoretical problems are attracting an increasing number of academics. At the same time, the number of fascinating theoretical questions that are being explored is growing.

Since the 1980s, when the idea of data mining was first presented, its economic relevance has emerged, and many commercial firms have lobbied for it, thereby creating a preliminary market for it. Since then, the term data mining has also been used interchangeably with data mining and data mining. Because of the efforts put forward by these commercial companies, a market of this kind already exists.

It reveals linkages between things that can't be exposed using more typical AI and statistical techniques, the association rule has a high value for research. This is because it has a high research value. The more traditional forms of artificial intelligence and statistical methods are unable to uncover these linkages. In the same breath, it satisfies the pressing demand for knowledge that may be gathered from extremely big datasets.

Countless research findings have been generated by the research institutes of the most prominent universities in the world as well as the research departments of the most successful IT firms. These discoveries have been published in a variety of academic journals. A significant number of the most recent and innovative algorithms for mining are made available to users. Users can uncover a wide variety of different sorts of knowledge, such as sequential patterns, classification, and so on, without needing to have comprehensive understanding of statistics or training in the field.

The database management solutions that are industry-standard, such as SQL Server and Oracle, interact very well with the system, which is platform-agnostic and does not require special software to run. Additionally, the technology is compatible with multiple operating systems. In addition, online analysis and mining technology have been added into the system [1] to research the benefits of data warehouses.

In this area of study, data mining refers to the process of using computers to extract instances from large data sets, and it is one of the topics covered. We concluded that the data in the dataset needed to be streamlined to provide it with a structure that was more internally consistent. Data mining is a method that involves extracting information from databases using various exploratory techniques [2]. [3] The ability to make decisions that are knowledge-driven and proactive, as well as predict future trends, grants a competitive advantage to organisations who implement these technologies. When discussing data mining, the phrase knowledge discovery in databases (often abbreviated as KDD) is frequently used [4]. KDD makes use of data mining to take care of word alternatives [5].

In the information technology, the amount of focus that has been placed on data mining and association rules over the course of the past few years has increased. A wide range of academic institutions have been conducting research and examinations into the various data mining technologies available.

2. RELATED WORKS

Recent research efforts have been focused on mining multilevel association rules. This activity has taken up a large amount of time. Among the available schools, the apriori school [6] is the one that should be evaluated first. These techniques either exhaustively discover all frequent things in every idea level, as in the ML-T2L1 algorithm [10] and the Level-Crossing algorithm [11], or they add all the ancestors of frequent items in the relevant transaction database, as in cumulate [9]. For example, the ML-T2L1 algorithm [10] and the Level-Crossing algorithm [11] both exhaustively discover all frequent things in every idea level. Examples of algorithms that do this include the ML-T2L1 algorithm [10] and the Level-Crossing algorithm [11].

Both algorithms uncover everything frequent at every thought level. An additional subgroup includes methods that are based on FP-growth [7]. FP-tree-based methods perform better than Apriori-based methods because they inherit the benefits of the FP-tree methodology, which scans the dataset in a quick and efficient manner to find highly common items. Apriori-based methods, on the other hand, assume that highly common items are likely to be present. In addition, FP-tree approaches perform a dataset scan in

a manner that is more natural. However, the ever-increasing expenses of their processing and memory create a significant bottleneck when they are utilised for the analysis of enormous amounts of data.

Mining multilevel associations is an amazing strategy, but there are a range of alternative methods that can make it even more effective. Mining multilevel associations is an excellent strategy. A method that extracted multilevel membership functions by utilising the methodology that is used in Ant Colony Systems. Before being put into action, this strategy did not call for a predetermined minimum degree of support as a prerequisite. Vejdani method by fixing the functions for each item and then computing minimum supports and optimised multilevel association rule mining by taking advantage of OLAP and data mining technologies in multilevel association rule mining [16], which brought efficiency and flexibility.

In addition, research has been done on genetic algorithm (GA) [17]-based techniques for mining association rules. These techniques are part of the mining of association rules. Methods that are founded on GA are able to search through a very large number of potentially suitable candidates for an association rule in a very short amount of time. Techniques that are based on GA have been shown in previous research [18] to be capable of identifying generalizable prediction rules.

The GA-based approaches, as opposed to the greedy rule induction techniques, carry out a global search on association rules and are better able to deal with data that contains attribute interactions. This demonstrates that solutions based on GA are superior to other approaches when it comes to coping with complicated data. In the past, a great deal of attention has been paid to mining single-objective rules [19] and mining multi-objective rules [20] as two examples of single-level association rule mining with GA. The disciplines of data mining are represented by both examples.

In the context of big data research, strong association rules are frequently presented in numerous forms. Mining multilayer association rules in large data sets hence calls for more efficient methodologies. Researchers are faced with a dilemma because of this, and they must discover solutions. The purpose of this research is to offer a GA-based multilevel association rule mining approach with the intention of identifying multilevel association rules in enormous volumes of data in a more effective and efficient manner.

3. MODIFIED APRIORI ALGORITHM (MAA)

An explanation of the difficulty involved in mining multilevel association rules is as follows: The catalogue tree provides a condensed definition of the multi-level categorization relationships among the items, and domain knowledge is made up of a set of objects that are denoted by $I = i_1, i_2,...,i_n$.

We refer to a node in one graph as the ancestor of a node in another graph if there is a path between the two graphs that a node in the first graph can take to get to the node in the second graph. If there isn't such a path, then we don't consider the node in the first graph to be the ancestor of the node in the second graph. The database will only show the leaf nodes in its representation. The term transaction T refers to a collection of objects that are saved in the database D and are represented as the form T_I. The phrase transaction T also refers to the form T_I. A one-of-a-kind Transaction Identification Number (TID) will be generated for each one of your individual transactions. It is common practise to assume that words beginning with T refer to branches.

It is important to keep in mind that the transaction in question is able to manage the item in question if it already contains the item x_I in question or if the item x is the parent of certain things that also exist in the transaction in question. A transaction X_I is said to be compatible with a different deal T in a more general sense if the first contract, deal T, is compatible with everything that is included in the second deal, X. Inferences of the form X_Y, where X_I, Y_I, and X_Y are some of the distinguishing features of multilevel association rules. It is possible to provide evidence that Y ancestors = 0, which shows that no element in Y can be traced back to any element in X as a direct ancestor of that element. This can be done by demonstrating that Y ancestors are equal to 0. This is since a rule of the form x ancestor (x) is regarded as being superfluous since it is always accurate. The following are some of the reasons for this: Both X and Y allow for items to be brought in from any level to be used in the game.

It is true that XY holds true in the set D of all possible transactions if and only if s_0, where s is the proportion of transactions in D that conform with the rule. If s_0 is not true, then it is not true that XY holds true in the set D of all possible transactions. It is possible to express the probability that X and Y are both correct using the equation $XY = P$. (X Y). In the data set D, the confidence level of the rule XY is c, where c is the fraction of the data points that support X that also support Y. This confidence level is derived from the fact that X is supported by more data points than Y. The degree of assurance that can be derived from following the rule is dependent on this fraction being employed. The expression $P(Y|X)$ is a formula that represents the conditional probability that this will take place. Therefore, confidence (XY) equals $P(Y|X)$ in the event when X is accurate while Y is inaccurate, and support (XY) equals $P(XY)$ in all other cases.

$$\text{Support}(Y \Rightarrow X) = P(Y \cup X),$$

$$\text{Confidence } (X \Rightarrow Y) = P(Y|X).$$

The following are some recommendations that we have regarding the process of analysing the rules that were discovered through the application of multilayer abstraction.

- ***Support Confidence***: The rules for which there is more evidence are given preference since this implies that the rules can be applied to a larger range of scenarios. Additionally, the rules for which there is more evidence that the rules have been reliably detected in the domain statistics are given preference. This is since the accumulation of supporting data for a rule suggests that the rule can be utilised in a greater variety of contexts.
- ***Interest***: The rule that should be followed is the one that is located at the level of the catalogue tree that contains the most relevant information. Exploring simple correlations could lead to the discovery of laws that are so commonplace that they are irrelevant to the question at hand.

It has been said that the amount of computational complexity required for multilevel association mining is almost exhaustive. This is because it contains data that changes in the degree of abstraction that it holds, which is the reason for this result.

In the context of big data where the number of catalogue items and transactions is expanding at an exponential rate, the conventional technique compute and memory consumptions will grow at a rate that is comparable to that increase. This is because the number of catalogue items and transactions is growing at an exponential rate. Even though the FP-tree algorithm makes mining association rules more efficient, it has difficulty mining multilevel association rules, particularly cross-level association rules. Cross-level mining is particularly difficult for it. Because taking careful note of the strategy is so critical, it is important to keep this point in mind. A fresh heuristic approach is required in order to successfully mine multilevel association rules within the context of research involving massive amounts of data.

3.1 CRUMB BASED ASSOCIATION RULE MINING (CBARM)

The architecture that is supplied for mining association rules is one that is built on a crust, and the method that is offered for it makes use of that architecture. The proposed method moves on to the subsequent phase of extracting rules of varying durations after first converting the input data items into a format known as Kavosh. After the if conditions have been evaluated, the overall length of the rule in this document is computed with the help of the number of fields that have been evaluated. The recommendation that was made requires the data to first be transformed into the Kavosh format before it can be considered an effective strategy. Because of this format, nodes can independently carry out their respective responsibilities. This standardised format is utilised to convert the several kinds of data that are sent in for processing. Because it simplifies the process of generating key-value pairs, the crumb-based method is one that can make use of the uniform format that has been suggested. This is because the crumb format makes it possible.

3.2 INTER-TRANSACTION ASSOCIATION RULE (IAR) MINER

Our methodology is founded, initially and fundamentally, on the inter-transactional association rules architecture, which is a very effective instrument for the development of predictive rules. This architecture is a tool that allows us to analyse the relationships between different types of transactions. When it comes to making these kinds of projections, our standard operating procedure always considers how important it is to provide timely information concerning unusual occurrences.

The assumption that the forecast will be of the most value in the period right prior to the event of interest, X_t (the prediction period or monitoring window). This window pops in the moment the user makes the decision to start caring about having a prediction produced, and it closes the moment it becomes clear that the prediction was made too late to be of any use. It starts when the user makes the decision to start worrying about having a prediction prepared and ends when it is too late to do anything about it (warning time).

We describe a method for efficiently mining predictive patterns within the time span of the prediction, with the warning time set w time points before the event X that is being forecasted. This method is applicable during the time of the prediction.

Only the transactions that are pertinent to those times are collected and stored using a sliding window during a single trip through the database. This is done on each individual journey through the database. If we are looking for an uncommon event, like the one we are looking for, then the number of these time periods (windows) that are saved will be high.

The first stage in the process of extracting the vital information is recording the monitoring window that corresponds to each instance of the target event. This should be done whenever the target event occurs. The database goes through a transformation that maps the relative temporal information of everything that is contained within the window as part of the process of window capture. This takes place when the window is being captured. The framework that is utilised to define associations between transactions makes the shift that was outlined earlier much simpler to carry out.

3.3 CATEGORIZED AND BOUNDED INTER-TRANSACTION (CBIT)

The ITP tree oversees maintaining a record of any inter-transaction item sets that occur often and can be accessed several times. This is done so that in the future it will be simpler to derive closed pattern sequences. Either the prefix pruning or the hash pruning method that we will discuss here can be utilised to get rid of open patterns that are found in an ITP tree when it is being traversed in a DFS-style fashion. The prefix pruning method is what eliminates many open patterns, and the closed patterns are obtained by hashing the remaining patterns for subpattern check into the same bucket as the prefix pruning method.

The prefix pruning method is responsible for getting rid of the vast majority of open patterns. During the process of designing the prefix pruning strategy, both the ITP-tree definition and the DFS methodology were utilised. The l prefix of any length-l pattern pl in the ITP-tree is equivalent to the same thing: pattern cp_{l+1}, which is the length-l prefix of the pattern parent. If a pattern p_l is closed, then there is no super pattern of p whose support is equal to the support of p_l. This is because the meaning of the word closed states that this is impossible. This is due to the absence of the pattern, which leads to this conclusion.

Any child pattern that is one more than the c_{pl}. on the ITP tree, the pattern of pl is regarded to be a superpattern of p_l. This is because the child pattern of pl is the same as the prefix pattern of p_l. Therefore, p_l is not a closed pattern, and DFS will not yield it even if the extra criteria of closed are satisfied, specifically the condition that $sup(pl) = sup(cpl+1)$. When it comes to filtering open patterns inside the same ITP-tree branch, the prefix pruning method has a specific strength that sets it apart from other methods. It is referred to as hash pruning, and it is the process of further filtering nonclosed patterns across separate branches.

If $p_{l.list}$ were $a_1, a_2,..., a_n$, then each pattern pl that was not eliminated by the prefix trimming operation would be hashed based on three attributes of *pl*. The first identifying characteristic is a $sup(pl)$, which is then followed by the initial time stamp in the *pl.list* file. The distinction between the two letters a, which are represented by the numerals a_2 and a_1, is the third distinguishing characteristic.

Each pattern on its own allows one to evaluate whether a certain pattern in a specific bucket is, in fact, a subpattern of the

other patterns that are contained within the same bucket. The designs in the bucket are examined, beginning with the ones that are the shortest in length and working their way up. Because of these actions, the patterns that are contained within the buckets after they have been emptied are those that are considered to be closed.

4. EXPERIMENTAL RESULTS

In this part of the article, we are going to put together several different tests so that we can evaluate how useful our design is. In addition to comparing the GA-based approach that we employ with the traditional association rule mining algorithm, we also utilise the latter as a point of reference for the work that we undertake. This is because the former method is based on genetic algorithms. Every single one of the examinations is carried out on a personal computer that is outfitted with a Core i5 CPU and a total of 4 gigabytes of random-access memory.

In the first experiment, it is explored whether valid association rules can be mined in a predetermined length of time, given a fixed number of generations with which to begin. This is done to determine whether or not this is possible. The number of individuals in the first generation rises from forty in Dataset 1 to one thousand in Dataset 2, and from fifty in Dataset 2 to one hundred and sixty in Dataset 3. In Dataset 1, the number of individuals in the first generation was forty.

We can draw the conclusion that the performance of the algorithm that is based on the GA will be comparable to that of the method that is based on randomization if the population is too small. Even if we can quickly establish adequate association rules if the population is large enough, the amount of complexity involved in computing continues to increase at a rapid pace. However, the two datasets make it abundantly clear that a solution that is feasible has been found: most appropriate association rules have been mined in a reasonable amount of time. Because utilising this method with these specified populations led to favourable outcomes, we have made the decision to make it our standard practise for Datasets 1 and 2.

By utilising examples taken from both serial and parallel data mining algorithms, we illustrate how to acquire insights from data mining and present them in a clear and concise manner. It is safe to say that the data parallel programme can be trusted if the results it produces are the same as those produced by the data serial programme.

The Table.1 and Table.2 contain the results of the data mining methods that were carried out in serial and parallel for the 150 M file and the 250 M file, respectively. The frequency of occurrence matrix, or FIM, that is included in the table provides an illustration of the item categories that are encountered the most frequently.

Table.1. Data mining of training and testing

Algorithm	Testing	Training
FIM1	21.483	26.46
FIM2	68.541	84.42
FIM3	46.035	56.7
FIM4	28.644	35.28

Table.2. Computational Time and Memory

Algorithm	Time	Memory
FIM1	29.26	30.78
FIM2	103.18	108.54
FIM3	69.3	72.9
FIM4	43.12	45.36

Although it needs to be able to deal with a varied range of objects as a direct result of parallel algorithms, the integrity of the serial technique has been maintained throughout. When compared to one another, there is no noticeable difference between the results of a single set and those of four when it comes to the overall outcome. In this aspect, the parallel technique is better to other approaches since it accurately excavates the sets of often occurring objects and satisfies the requirement for only the barest minimum of assistance.

According to the data, the parallel technique did not improve the efficiency of mining, which may be due to the overhead associated in job scheduling. Since the parallel approach is beginning to emerge gradually and only requires a portion of the time to mine, it is superior to the serial method, which has been in use until recently. When applying the proposed algorithm to a database with a restricted storage capacity, the information presented in Table.3 and Table.4 provides specifics on the amount of time needed, the acceleration ratio, and the speed needed to complete the process.

Table.3. Mining Time

Parameters	Association Mining
Time (s)	62.12
Acceleration	24.27
Speed	7.67

The method that has been discussed yields parameters that are both incredibly exact and straightforward to manipulate. When the accuracy of their system is evaluated in terms of timing, velocity, and acceleration, they find that it has a very high level of precision.

The proposed method is evaluated in terms of its accuracy and its recall rate in comparison to methods that are state-of-the-art. A graphical and a numerical representation of the data that was acquired is provided in the following table and figure, respectively.

Table.4. Recall Rate

Recall (%)	CBARM	IAR	CBIT
10	82.5561	82.9653	84.2952
20	63.6306	66.8019	69.7686
30	51.5592	53.8098	56.5719
40	33.3498	42.4545	45.7281
50	28.4394	37.0326	39.7947

The findings of the comparison make it abundantly evident that the method that was provided possesses a high precision ratio in contrast to other current approaches that have the same recall rate. A successful regulation of the noisy data is possible with the

aid of the method that was provided, and this is made possible thanks to the regulation of mining time, speed, and acceleration. In addition, a high rate of precision is maintained, which is evidence that the method is as exact as it is possible to be given the constraints of the method.

5. CONCLUSION

Most of the research that is described in this paper is predicated on a study of previously published algorithms for the mining of association rules. This paper is offered as part of a larger body of work. A brand-new approach that we refer to as association rule mining is shown here. It combines the characteristics that make DIC and FP trees the most beneficial. When compared to methods that have been utilised in the past, the objective of this strategy is to deliver improved performance when working with big datasets. This goal was established with the intention of achieving this objective. This project objective is to design and create new algorithms that make use of these overlaps so that the existing algorithms can be made more efficient.

REFERENCES

[1] I.H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science*, Vol. 2, No. 3, pp. 160-178, 2021.

[2] I.H. Sarker and A. Ng, "Cybersecurity Data Science: An Overview from Machine Learning Perspective", *Journal of Big data*, Vol. 7, pp. 1-29, 2020.

[3] X. Zhou and Q. Jin, "Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things", *IEEE Internet of Things Journal*, Vol. 7, No. 7, pp. 6429-6438, 2020.

[4] T. Alam and Z. Abbas, "A Model for Early Prediction of Diabetes", *Informatics in Medicine Unlocked*, Vol. 16, pp. 100204-100209, 2019.

[5] Z. Wu, J. Cao and Y. Ge, On Scalability of Association-Rule-based Recommendation: A Unified Distributed-Computing Framework", *ACM Transactions on the Web (TWEB)*, Vol. 14, No. 3, pp. 1-21, 2020.

[6] J. Surendiran, S. Theetchenya and M. Dhipa, "Segmentation of Optic Disc and Cup Using Modified Recurrent Neural Network", *BioMed Research International*, Vol. 2022, pp. 1-13, 2022.

[7] I. Ullah and S.W. Kim, "A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", *IEEE Access*, Vol. 7, pp. 60134-60149, 2019.

[8] A. Telikani and A. Shahbahrami, "A Survey of Evolutionary Computation for Association Rule Mining", *Information Sciences*, Vol. 524, pp. 318-352, 2020.

[9] P. Ghavami, "Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing", Walter de Gruyter, 2019.

[10] I. Lee and Y.J. Shin, "Machine Learning for Enterprises: Applications, Algorithm Selection, and Challenges", *Business Horizons*, Vol. 63, No. 2, pp. 157-170, 2020.

[11] Y. Mourdi and W. Berrada Fathi, "A Machine Learning-based Methodology to Predict Learners' Dropout, Success or Failure in MOOCs", *International Journal of Web Information Systems*, Vol. 15, No. 5, pp. 489-509, 2019.

[12] S. Neelakandan and D. Paulraj, "An Automated Exploring and Learning Model for Data Prediction using balanced CA-SVM", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, pp. 4979-4990, 2021.