Apriori Algorithm and Hybrid Apriori Algorithm in the Data Mining: A Comprehensive Review

Yahya Zakur¹ and Laith Flaih²

¹University of Mazandaran, Faculty of Mathematical Sciences, 4741613534 Babolsar, Iran
²Cihan University, Department of Computer Sciences, 44001 Erbil, Iraq

Abstract. Data mining has the potential to empower healthcare organizations by allowing them to analyze various aspects of patient information and discover connections between seemingly unrelated data. By harnessing advanced data analysis techniques, healthcare providers can identify trends in patients' medical conditions and behaviours. The Apriori algorithm is used for mining frequent item sets and devising association rules from a transactional database. The parameters “support” and “confidence” are used. Support refers to items’ frequency of occurrence; confidence is a conditional probability, while Apriori-Hybrid. Apriori-Hybrid is the combination of algorithms Apriori and Apriori-TID, which can classify large itemsets and can improve the accuracy of classification and it can also shed light on the basic mechanism. In this research, a comparison was made between the two algorithms in terms of capabilities, strengths, areas of use, and suggestions about the nature of using each algorithm.

1 Introduction

Data mining [1] is the most significant duty that is required by various companies, organizations, and government agencies among other places. Data mining is useful to meet the challenges of big data in various sectors of real life such as the challenges of big data in health and other sectors [2]. This indicates that data mining is the most active study area where new methods for extracting data from datasets are being developed. This is where the necessary data is extracted or categorized from the enormous datasets [3]. The data mining system functions, like a warehouse that stores and safeguards vast amounts of data or used datasets. When we extract or manipulate information from a source we call it data extraction or knowledge discovery [4].

2 Literature Review

The Apriori algorithm is widely recognized in the field of data mining. It is commonly employed to extract insights and patterns, from datasets making it particularly useful for comprehensively studying the connection, between multiple traits and decisions regarding basketball penalties. By utilizing a version of the Apriori algorithm we can analyze basketball

* Corresponding author: yahyazakur92@gmail.com
game data to gain insights into the tactical aspects of the sport. The Apriori algorithm is employed to extract association rules by examining itemsets, among basketball actions. The enhanced algorithm allows for more effective mining of technical moves that exhibit stronger connections in the basketball game data. As a result, the analysis results derived from this improved approach are highly instructive for coaches, players, and analysts alike. The technical and tactically directed analysis is divided into two primary sections: the technical and tactically directed action analysis and the technical and tactically directed cooperation analysis. The key action analysis employs a Markov process-based data mining algorithm to examine basketball game data, with a particular emphasis on identifying key score transfer steps and key score loss transfer steps. The algorithm can effectively identify the key actions that lead to scoring as well as the key actions that result in conceding points during the basketball game process. The analysis results derived from this algorithm hold significant practical value. The collated game data is utilized as the independent variable, and the number of games won and lost serves as the dependent variable in logistic regression analysis. By applying logistic regression, researchers aim to identify the characteristics that significantly influence the likelihood of winning games. The decision tree algorithm is utilized to identify the significant factors that influence victory and to predict team performance. This algorithm aids in comprehending the relationship between various factors and their impact on basketball game outcomes. The association rule algorithm is employed following the selection of the technical statistics of the top players from the previous three seasons. This algorithm analyzes the data to identify patterns and correlations between player performances and the outcome of the game [5].

The study aimed to explore the network association patterns between different diseases that occurred in the same individual. It specifically focused on using the association rule mining technique to identify short-term or long-term connections between these disorders. The methods used in this study involved extracting raw data from a large-scale electronic medical record database affiliated with Xuzhou Medical University's hospital. Additionally, the study collected a substantial amount of data, consisting of 1,551,732 pieces of diagnosis information obtained from 144,207 patients. This data spanned a period of five years, from 2015 to 2020. The clinic diagnoses were categorized based on the "International Classification of Diseases, 10th revision". The Apriori algorithm was employed to investigate the association patterns among the collected diagnoses. The analysis resulted in the generation of 12,889 rules initially. However, through the application of threshold filtering and manual examination, the researchers narrowed down the results to 110 disease combinations that exhibited strong association strength. To be included in the final set of associations, these disease combinations had to meet specific criteria: a minimum support of 0.001, a minimum confidence of 60%, and a lift value greater than 1. These criteria were used to ensure that only significant associations were considered, providing valuable insights into the relationships between different diagnoses within the patient population studied. Among the results obtained from the Apriori algorithm, a significant portion of the association rules were related to the circulatory system and metabolic diseases. The study's findings revealed network associations between disorders originating from various body systems within the same individuals. Moreover, the study demonstrated the significance and effectiveness of utilizing the Apriori algorithm in comorbidity or multimorbidity investigations. Additionally, the mined combinations can contribute to reducing mortality rates, as early detection and proactive interventions can lead to timely and appropriate treatments for patients with multiple co-occurring conditions [6].

By employing both traditional analytics tools like statistics and graphs and those connected to artificial intelligence like rule induction, it was emphasized the significance of
data mining. It is a special process or attitude for analyzing knowledge. As UUM's postgraduate database contains a wealth of information, the university may profit from the knowledge obtained from that data. Additionally, the Apriori algorithm's ability to mine student data has not been examined. Therefore, our investigation gives evidence that Apriori has data mining ability. Discovering it influences students' decisions to enrol in postgraduate programs is also important. Discovering these elements could be significant for UUM since it would enable management to direct advertising to different groups of new students. The KDD methodology was used in this investigation. The results then confirmed that the Apriori algorithm can identify affect variables. The most inevitable outcome was understanding the influencing variables that can encourage students to join, which can increase the number of students at that university. The most well-known association rule algorithm is the apriori algorithm, which is efficient at recognizing patterns in a wide range of data. The created rules were clear. There used to be a limitation on the type of data needed for training. It took a lot of actual transactions to get any useful data. Based on this finding, it could be best to train the same data using a different software program and a sampling approach that employs association rule algorithms. This study has included comparisons of variables other than the algorithm. Data mining may seem like a highly interesting concept, but it must only be utilized following strict guidelines and without any degrees of prejudice or common assumptions. This effort can assist "UUM" in making the best promotions possible [7].

In this study, the expansion of the food and beverage industry was credited for the culinary industry's quick expansion. Therefore, this forced owners of food and beverage businesses to make the right choices to survive in a highly competitive market. They must also be willing to constantly innovate and adapt to meet consumer needs with products that can draw customers in and decide on strategies for sales-driving promotions. Data from stored transactions provided information, such as consumer spending trends, that could be obtained using data mining techniques. For calculating and identifying association rules, the Apriori methodology is considered very helpful. A database containing 150 transaction datasets was used to derive the outcome of association rules using apriori methods. Both calculation and implementation were made using Rapid-miner. This time, the creator of the test aimed to accurately determine the laws of association and establish a significant correlation between the elements. In this study, the researchers came to the result that the domain of data mining may also provide new information patterns that were valuable for predicting sales trends. Fish restaurants might use the information on the item association rules in customer purchase patterns provided by the deployment of the apriori algorithm to create a promotional strategy to improve sales depending on the items bought at the same time. Compared to when marketing strategies were previously used, this method produced item association rules in consumers purchasing patterns that were more accurate [8].

The current period has seen enormous progress in many facets of animal health care. Numerous institutes are examining the use of information technology to convey knowledge and skills in the animal health care sector using expert systems. The distribution of veterinary and scientific data to livestock owners is a barrier to any state's progression in animal health care. It might be challenging to evaluate an expert system's performance since knowledge cannot be quantified. Expert systems, however, cannot be given explicit proof of their accuracy. Approaches that guarantee the accuracy of the findings for their users can be added to expert systems. The research created a unique association algorithm based on support and confidence that can repeatedly scan datasets to find item groupings that are comparable in terms of length and satisfy a certain increment. In this study, associative rules were found based on clinical diagnostic criteria, and a data mining approach was created for storing and
analyzing data for various illnesses in nursing animals. The clinical data storage dataset being developed for this project should be capable of rendering data in acceptable formats and providing metadata that accurately expresses data semantics. When used to collect clinical data, the Apriori Algorithm can verify expert system findings on the co-occurrence of diseases and symptoms. It was demonstrated that the Apriori algorithm was a good confidence association rule [9].

Communication of this study was to investigate the feasibility of utilizing an intelligent medical communication system based on the Apriori algorithm, along with a cloud follow-up platform, to deliver continuous nursing care to breast cancer patients who were undergoing treatment outside of a hospital setting. To expedite the data mining process using the Apriori algorithm, Graphics Processing Units (GPUs) are employed in conjunction with Amazon Web Services (AWS). Simultaneously, a cloud-based intelligent mobile medical communication system is being developed, encompassing various modules to enhance healthcare services and patient management. The system comprises the following components an appointment management module, log-in, my workstation, patient records, follow-up centre, satisfaction management, propaganda centre, and education centre is also being constructed at the same time. The study participants are divided into two groups: the control group, which consists of 163 individuals receiving regular telephone follow-up, and the intervention group, which comprises 216 participants receiving continuous nursing intervention through various nursing approaches. The compliance of the study participants, as well as their quality of life before and after nursing, and the degree of function restriction in the wounded limb, are all evaluated using a cloud-based intelligent medical communication system. The Apriori algorithm's execution duration is inversely proportional to the number of cluster nodes and inversely proportional to the amount of data. The proportions of complete compliance data, poor compliance data, and overall compliance in the intervention group show statistical significance compared to the control group (P < 0.05). The ratio of complete compliance data, the proportion of poor compliance data, and the overall compliance percentage in the intervention group significantly differ from the control group (P < 0.05). The proportion of patients in the intervention group experiencing limited and severely reduced functional activity of the affected limb is significantly lower than that in the control group (P < 0.05). The postoperative nursing satisfaction rate of the intervention group is significantly higher than that of the control group (P < 0.01). Additionally, the proportion of fundamentally satisfied and dissatisfied patients in the control group is larger than that in the intervention group (P < 0.05) [10].

Accordingly, data mining can empower healthcare organizations to predict trends in a patient's medical condition and behavior by analyzing various aspects and uncovering connections between seemingly unrelated information. Raw data from healthcare organizations are often vast and diverse. For the data to be utilized, it must be collected, organized, and integrated, allowing for seamless integration of medical information systems. Using conventional research methods, health data mining affords unbounded opportunities to examine a large number of obscure or concealed data patterns. Association rule mining (ARM) is a powerful technique for discovering relationships within data and the most used and influential algorithm in ARM is the Apriori algorithm. However, the Apriori algorithm generates many rules and does not guarantee the efficiency and value of the knowledge it produces. To address this concern, a solution is proposed in the form of an enhanced Apriori algorithm (EAA) called EAA-SMO, which leverages a context ontology to improve efficiency. This methodology incorporates sequential minimal optimization (SMO) techniques. The fundamental concept of ontology is to establish hierarchical structures of conceptual clusters related to specific subjects. These clusters consist of "similar" concepts,
representing precise categories of knowledge within a given domain. Each cluster exhibits intriguing rules derived from the correlation between its items. Furthermore, the rules developed within these clusters can be classified as prediction models, specifically designed for anomaly detection using SMO regression. The experimental analysis shows that the proposed method has enhanced accuracy by 2% and reduced the execution time by 25% when compared to using a semantic ontology [11].

On the one hand, the Internet's rapid expansion has facilitated numerous conveniences and amenities in daily life. However, it has also introduced a significant threat in the form of intrusion, posing dangers to the security and privacy of internet users. Detecting and preventing intrusions is crucial for securing and optimizing network communication. In the literature, numerous intrusion detection techniques have been proposed. However, when applied to big data applications, many of these techniques are either excessively complex or fail to deliver satisfactory performance. This paper therefore presents a novel Hybrid Mayfly Apriori-Intrusion. The detection mechanism is designed specifically to detect data application intrusions. We employ Mayfly optimization and the Apriori algorithm for intrusion detection in this proposed method. In contrast to classification-based methods, we process network data to construct an Apriori rule based on item sets. The suggested approach detects itemsets or transactions, as intrusions. To assess its effectiveness it is compared to established algorithms, like Artificial Neural Networks, Random Forests K Nearest Neighbors and Support Vector Machines. This comparative analysis aims to evaluate how well the proposed intrusion detection mechanism performs and its overall efficacy. In the end the suggested approach has proven its worth through achievements, such, as attaining a 97% accuracy rate, 99% precision rate and 97% recall rate. These outcomes indicate that the suggested mechanism is well-suited for intrusion detection in big data [12].

Software quality has been a major concern of the IT industry and software companies for several decades. Defect prediction is a valuable technique that gives evaluators insight into the software product's functionality. The primary objective of software defect prediction techniques is to identify and report defect-prone areas. During the Software Development Life Cycle (SDLC), early detection of software defects can have positive effects on the development process and the finished software product. Including a reduction in development expenses. Additionally, the time spent on additional testing and revision during the post-production and maintenance periods is reduced, resulting in software that is more reliable. Indeed, software metrics play an indispensable role in constructing defect prediction models. These metrics quantify a multitude of software attributes. Publicly accessible datasets extracted from software programs, such as those made available by the National Aeronautics and Space Administration (NASA), have been heavily utilized in software engineering-related research. The datasets containing information on associated Software Metrics at the module level offer valuable insights into the characteristics of software modules and their potential relationships with defects. The proposed concept of a novel hybrid data mining technique that combines Clustering and a Modified Apriori Algorithm has the potential to enhance the precision and effectiveness of Software Defect Prediction. Combining Clustering and a Modified Apriori Algorithm, the proposed hybrid data mining technique aims to reduce the number of association rules generated. This is accomplished by employing an intriguing metric known as "spread" [13].

Frequent itemset mining is a fundamental and indispensable data mining technique for discovering valuable patterns and correlations in data. Due to the large volume and intricacy of the data, working with big data presents unique challenges for frequent itemset mining. The aforementioned frequent itemset mining algorithms, such as Apriori, FP-growth, and Eclat, are often implemented on big data processing engines like MapReduce and Spark to
handle large-scale datasets efficiently. The paper's proposed hybrid algorithm for mining frequent patterns on sparse big datasets over the Spark platform is an innovative approach to address the challenges associated with large-scale data processing. By integrating the advantages of the Apriori and the Eclat algorithms. Experiments are conducted to evaluate the efficacy of the hybrid algorithm with four phases that was proposed. In these experiments, the elapsed time of the proposed hybrid algorithm is compared with three other algorithms: parallel Fp-growth, YAFIM, and Eclat-Spark. The results indicate that the proposed hybrid algorithm outperforms YAFIM, Eclat, and Fp-growth, particularly when using a high minimum support threshold[14].

3 Apriori Algorithm

Finding patterns in huge data sets is a technique known as data mining or knowledge discovery [15]. Each technique in data mining has a particular goal, and it is a developing and difficult challenge to extract significant and previously unrecognized patterns from large datasets [16]. Association rule mining, one of the most well-known data mining techniques, identifies unanticipated data relationships [17]. The fundamental concept [18] is to determine if the presence of certain items in a database of item sets (such as purchasing baskets) suggests the occurrence of other items with a comparable probability. One of the well-liked methods for extracting association rules from data sets is the apriori algorithm [19]. The algorithm is founded on two critical measures: support and confidence. The support reflects the frequency of an itemset's occurrence in a dataset, indicating its prevalence. A higher support threshold ensures consideration of itemsets with sufficient occurrences [20]. On the other hand, confidence gauges the reliability of an association rule by measuring the proportion of transactions containing both the antecedent and consequent items [21]. A higher confidence threshold guarantees the generation of strong and significant rules.

3.1 Advantages and Disadvantages of Apriori Algorithm

Below are some advantages of the Apriori algorithm [22]:
1- This algorithm is among association rule learning algorithms, this one is the simplest and clearest.
2- The resulting rules are simple to understand and express to a user.
3- Since it is fully unsupervised, it does not require labelled data. As a result, you may use it in a variety of settings because unlabeled data is frequently easier to access.
4- Based on this implementation, a variety of adaptations have been suggested for various use situations. For instance, there are association learning algorithms that take into consideration the ordering of items, their quantity, and linked timestamps.
5- This algorithm is exhaustive, so it discovers all the rules with the required support and confidence.
6- Below are some disadvantages of the Apriori algorithm [23]:
7- The dataset for the Apriori technique includes a large number of itemsets.
8- The data set has a low minimum support level for the Apriori method.
9- The amount of time required to hold many candidates sets with numerous frequent item sets.
10- Consequently, it is ineffective when used for vast amounts of datasets.
3.2 Hybrid Apriori Algorithm

The apriori-Hybrid algorithm is built on the concept that using the same algorithm for all passes over data may not be the most efficient approach, as mentioned in [24]. The statement from reference [25] suggests that Apriori and Apriori-TID exhibit different performance characteristics during the multiple passes over data in frequent itemset mining. Based on the experimental observations, the Apriori-Hybrid technique was developed as a response to the varying performance characteristics of Apriori and Apriori-TID during multiple passes over data in frequent itemset mining [26]. Estimating the size of the candidate itemset (Ck) at the end of each pass is essential for deciding whether to switch from Apriori to Apriori-TID in the Apriori-Hybrid technique. Accurate estimation is required to guarantee that Apriori-TID can efficiently process candidate item sets without exceeding the available memory capacity [27].

The advantages [28] of the Hybrid Apriori algorithm:
1- Increased productivity.
2- More possibilities for continuous learning.
3- Better accuracy of relationships.

The disadvantages of the Hybrid Apriori algorithm [29].
1- It is a type of network expensive.
2- The design of a hybrid network is very complex.
3- There is a change in the hardware to connect one topology with another topology.
4- Usually, hybrid architectures are larger in scale, so they require a lot of cables in the installation process.
5- Hubs which are used to connect two distinct networks are very costly. And hubs are different from usual hubs as they need to be intelligent enough to work with different architectures.
6- Installation is a difficult process.

4 Discussion and Future Works

Data mining is primarily performed using association rules. The Apriori algorithm is one of the most prevalent algorithms used to derive frequent item sets from large databases. It requires numerous scans over the database, resulting in a large number of disc reads and placing an enormous load on the system, whereas the Hybrid Apriori Algorithm produces more accurate and quicker search results. Based on the results of this study, it is evident that the hybrid Apriori algorithm outperforms the classical Apriori algorithm for solving complex problems. As a result, future research should focus on a variety of applications, such as public health, education, and industry.

References


27. M. Krishnamoorthy, R. Karthikeyan, *Method For Mini Frequent Patterns From Large Data-Sets*. European Journal of Molecular & Clinical Medicine, 9(07), 2022.
