How to Determine Minimum Support in Association Rule

Erna Hikmawati
School of Electrical and Information Engineering
Bandung Institute of Technology
Ged. Achmad Bakrie, Lt. 2
Jl. Ganesha No.10, Bandung 40132
+62-22-2502260
Bandung, Indonesia
erna@pasim.ac.id

Kridanto Surendro
School of Electrical and Information Engineering
Bandung Institute of Technology
Ged. Achmad Bakrie, Lt. 2
Jl. Ganesha No.10, Bandung 40131
+62-22-2502260
Bandung, Indonesia
endro@informatika.org

ABSTRACT

The growth of increasingly complex data now raises new challenges in the world of technology. Large volumes of data store a lot of knowledge that can help in the decision-making process. One way to find knowledge in big data is by the Association Rule. The association rule is a technique in data mining that can produce rules based on the frequency of items appearing from a transactional database. One thing that is critical in the association rule is the determination of the minimum support value used to determine which items will be included in the formation of rules. If the minimum support value that is set is too small, it causes too many items to be involved in establishing the rules. Conversely, if the minimum support is too large, the number of items involved in forming the rule is too small. The problem in determining the minimum support value greatly affects the accuracy of the resulting rule. In this paper, various methods will be discussed to determine the minimum value of support through study literature based on related research. In addition, it explains research opportunities that can be done in the future about the minimum value of support determination in the association rule.

CCS Concepts

• Information systems→ Information systems applications→Data mining → Association rules.

Keywords

Data Mining; Association Rule; Minimum Support

1. INTRODUCTION

The development of technology has a major influence on all fields of life. Existing companies, agencies and government organizations must apply information technology to support their competitive advantage. It triggers the emergence of software development in various fields. The emergence of software in the various fields is in line with the growth of existing data growth. At this time, data can be collected anywhere, anytime, from various devices and in any way [13]. However, the amount of data owned will not be worth anything if its contents are not analyzed. In the world of Information Technology, it is known as Data Mining. The data mining is one of the stages in the Knowledge Discovery Database (KDD) [21] which contains the process of extracting data in order to find useful information or knowledge. With data mining, the researcher can find useful knowledge or information on big data [7,13,15,18]. The information generated later can be used as a basis for decision making or basis in the preparation of plans and strategies. In addition, Data mining is one of the most popular technologies that discover potential customer knowledge from business databases to assist a policy decision [19].

One of the tasks of Data Mining is the Association Rule. The association rule is a way to find association relationships or correlations between a set of items [21]. Association rule which is also called Frequent Itemset Mining is a basic concept in the data mining and is the most common way to find association relationships based on the frequency of occurrence of itemset [4,12,13,18]. The first concept of Association Rule Mining or Frequent Itemset Mining suggests that to get frequent itemset mining, the initial step must be conducted to determine the threshold called minimum support. The minimum support here serves to determine which items will be included in the process of forming the rules. The minimum value of support is determined by the user, so it is very intuitive. Moreover, the process of forming a rule can be done repeatedly with different minimum support values to get the appropriate rule.

The process of determination of the minimum support value is very influential on the rules formed and is not easy for the user. The determination of a minimum support value that is too low causes too many items to be involved in the rule formation process. Conversely, the determination of a minimum support value that is too high causes the items involved to be small and there may be a lot of non-reputable information. In addition, in terms of time and memory usage, a minimum value of support that is too low requires more time and memory when compared to a higher minimum value of support [4,15].

There are several studies that focus on the determination of the minimum support value with different methods. This paper will discuss the methods of determining the minimum support value in various cases through the study literature of related research. In addition, it will also explain the research opportunities that can be done in the future about the minimum support value determination in the association rule.

This paper is organized as follows: related research will be presented in part II. Discussions, opportunities and research directions are explained in section III. The conclusion and future work of this paper is presented in section IV which describes the closing statement.

ICSCA 2020, February 18–21, 2020, Langkawi, Malaysia
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7665-5/20/02…$15.00
https://doi.org/10.1145/3384544.3384563

2. RELATED WORK

In this section, we will review methods in the previous studies for determining the minimum value of support and basic concepts association rule.

2.1 Basic Concepts Association Rule

Association rule mining can be defined formally as follows: Let I = {i 1, i 2, . . ., i m} be a set of literals, called items. Let DB be a set of transactions, where each transaction T is a set of items such that T ⊆ I. Note that the quantities of the items bought in a transaction are not considered, meaning that each item is a binary variable indicating whether an item was bought or not. Each transaction is associated with an identifier called a transaction identifier or TID [9,21].

Let X be a set of items. A transaction T is said to contain X if and only if X ⊆ T. An association rule implies the form X ⇒ Y, where X ⊆ I, Y ⊆ I, and X ∩ Y = Ø. The rule X ⇒ Y holds in the transaction set DB with confidence c if c % of the transactions in DB that contain X also contain Y. The rule X ⇒ Y has support s in the transaction set DB if s % of the transactions in DB contain X ∪ Y. Confidence denotes the strength of implication and support indicates the frequency of the patterns occurring in the rule [9,21].

The problem of mining association rules may be decomposed into two phases:

1. Discover the large itemset, that is, the sets of items that have transaction support s above a predetermined minimum threshold.
2. Use the large itemset to generate the association rules for the database that have confidence c above a predetermined minimum threshold [9,21].

2.2 Traditional Association Rule Mining

The Association Rule Mining was first proposed by Agrawal [2], then it was developed by many other researchers. The Traditional Association Rule Mining can be categorized into two categories: level wise and pattern growth [12]. In the level wise approach, a candidate item is generated for each level, so it is considered inefficient because it takes a long time to scan the database multiple times. This method is better known as the Apriori Algorithm [1].

The pattern growth method is an improvement of the apriori algorithm. In the pattern growth, there is no need to generate candidates for each level. The next step after the itemset elimination process with minimum support is forming a Frequent Pattern-Tree (FP-Tree). With the FP-Tree, a database scan is only once conducted so it is more efficient in terms of time and memory usage [16]. In determining the minimum support, both use the minimum support value specified by the user.

2.3 Multiple Minimum Support

Most of the association rules set a minimum support threshold for all the items, but in fact, the different items may have different criteria for assessing their importance. The support must vary for different items [11]. Therefore, many studies have revealed the method for association rule by applying different minimum support for each item called multiple minimum support [7,8]. However, the implementation of multiple minimum support increases the user's responsibility to determine the minimum support for each item.

2.4 High Utility Itemset

High-Utility Itemset (HUI) mining is one of the Data Mining Task that has gained popularity in recent years because it is used in various fields. HUI Mining aims to find itemset that has high utility (such as profit) in the transactional database [14]. The Association Rule Mining only represents the item sets based on the frequency whose items appear in the transactions. Other factors such as weight, profit or interestingness in the item are not considered. These factors affect the decision making. It is not enough if the decision making is only seen from the frequency of occurrence of the item, for example, the sale frequency of diamonds will be smaller than the sale of the clothes, but the diamonds produce greater profits [12]. These studies utilize the minimum support specified by the user. There are even some studies which combine it with multiple minimum support.

2.5 Without Minimum Support

Another view in determining the minimum support value in the association rule is to eliminate the minimum support in the rule formation process. There are two methods of association rule without considering the minimum support value, namely: Top-k and skyline. In the top-k association rule method, a minimum support value is not needed, but a k-value is needed which indicates the number of rules to be produced [12,15]. In this way, it is easier for the user to determine the value of k because it explicitly knows the number of rule results to be obtained. Determining the minimum support value is more difficult, because the user does not know how many final results that will be produced by the rule. The top-k association rule method was first proposed by Fournier and Cheng Wei Wu [6,20].

Skyline is defined as points that are not dominated by other points [3]. The skyline itemset algorithm is described by points that are scattered in a 2-dimensional plane. Being compared with other points, the skyline is located on the lowest line.

Skyline is used for the users who may need more than one aspect of decision making, for example, in choosing a hotel, two factors will be considered: the distance of the hotel to the downtown and its price. Hotels which are far from the downtown have cheaper prices than the hotels in the countryside. However, as a visitor, the user will look for the cheapest hotel and close to the downtown. Furthermore, the SKYMINE algorithm was designed to find the skyline frequent-utility patterns (SFUPs) and developed by Jerry Chun-Wei Lin [12] and Jeng-Shyang Pan [15].

2.6 Research Review on the Association Rule

As time goes by, the research on the association rules grows by using the methods that have been proposed in the previous studies. This is triggered using Association rules in various fields such as retail, warehouse or distribution center, electronic sales, banking, insurance and health [10]. The determination of the minimum support value is the first step in the Association Rule process. The different method in the determination of the minimum support value will affect the performance of the association rule method. In addition, the purpose of the establishment of rules also affects the determination of what threshold must be applied before the rule is formed. Table 1 presents the results of the papers review on the association rule by comparing the methods, advantages and disadvantages of each method used.

Table 1. Paper Review on the Association Rule

Title	Method/Algorithm	Strengths	Weaknesses	Use of Minimum Support
1. **Efficient mining of high utility itemsets with multiple minimum utility thresholds** [10]	a. MHUI (Mine High Utility Itemsets) b. Multiple minimum utility c. Suffix minimum utility	a. The MHUI algorithm can produce high utility itemsets in only one phase without the process of determining candidates. b. The MHUI algorithm can speed up and requires less memory than the previous algorithms.	a. Need to provide a minimum utility for each item b. Not yet applicable with Top-k HUI and data stream mining.	In this study, the minimum support value is formulated to be the minimum utility threshold. The minimum utility threshold value is different for each item and must be determined by the user at the beginning.
2. **Mining of skyline patterns by considering both frequent and utility constraints** [12]	a. Skyline Concept b. SKYMINE Algorithm c. Depth-First Search (SKYFUP-D) Algorithm d. Breath First Search (SKYFUP-B) Algorithm	a. You do not need minimum support and minimum utility, but the sets are based on the frequency and utility. b. It produces 2 more efficient algorithms for higher frequency and utility patterns and returns a set point that does not dominate as a decision-making solution c. Can reduce the search space in determining SFUP d. Has better performance in terms of runtime, memory usage, search space size and scalability	Not yet applicable to data streams, uncertain data and dynamic databases	In this study, it does not use minimum support, but it uses a maximum utility (ultimax) obtained from each iteration performed by the utility list structure.
3. **An efficient algorithm for mining the top-k high utility itemsets, using novel threshold raising and pruning strategies** [4]	a. Top-k high utility itemset mining b. kHMC algorithm c. strategy RIU, CUD, and COV d. EUCPT technique	a. Proposed more efficient algorithm for top-k high utility itemsets b. The proposed algorithm uses a utility-list structure to improve performance when finding a high utility itemset c. The proposed algorithm uses several strategies to find an internal minimum utility called RIU, COV and CUD.	Only applicable in top-k high utility itemset mining	Not use minimum support but it uses minimal utility. The minimum utility used is not specified by the user. At the beginning, it is set to 0 which will be increased during the process.
4. **Mining high-utility itemsets in dynamic profit databases** [14]	a. High utility Itemset b. MEFIM (iMEFIM) algorithm	a. Can find HUI in a database that has a dynamic profit b. The performance of the iMEFIM algorithm is better in terms of runtime, memory usage and scalability	Only applicable to High Utility Itemset	Not using minimum support but using the minimum utility threshold specified by the user.
5. **Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France** [23]	a. Multicriteria decision methods (Electre II) b. Apriori Algorithm	Can produce the most relevant and interesting rules using the ELECTREE II method	a. Using apriori algorithm b. The minimum support value is determined by the user	Using the minimum support specified by the user.
6. **Information fusion from multiple databases using meta-association rules** [17]	a. Meta-association rules b. Crisp meta-rules c. Fuzzy meta-rules	a. Can produce rules from several databases b. The resulting rule is more accurate because it combines the strength or validity of previous information c. Generates the rules that are more easily managed for inspection by humans	The database used as an input to obtain the rule must have the same items.	Using the minimum support specified by the user.

From the six papers reviewed, there are 2 (two) papers using minimum support, 3 (three) papers using minimum utility, and 1 (one) paper using maximum utility threshold. In terms of terms, minimum support, minimum utility and maximum utility have different terms but in terms of function, they all have the same function.

The minimum utility is the minimum support that has been formulated with each item's utility. In the maximum utility, what is stored is the biggest utility for every iteration process and its value will change every time it finds a larger utility, so it is called the maximum utility.

The function of the three is the same, that is to reduce the items involved in the rule formation process. In addition, the existence of this threshold can cut the time and memory needed in the rule formation process. Therefore, whatever the term is, this threshold is needed in the association rule process. The association rule process which does not use a threshold will involve all existing items, requires more time and memory, and the resulting rule may not be what you want. Producing too much or too little rules will make the decision maker confused.

Because the minimum support is needed by the association rule, there should be a special method to determine the minimum support value that is suitable with the characteristics and the amount of data even though it uses various methods. With this special method, it can cut the time and memory used in the process of forming the rules. This minimum support value should be determined before the rule formation process begins in order to eliminate which items have no effect and will not be involved in the rule formation process.

3. DISCUSSIONS, OPPORTUNITIES AND RESEARCH DIRECTIONS

Based on the objects and methods of the Association Rule, the development of Association Rule research can be divided into at least five categories: High Utility Itemset, Top-k Association Rule, Skyline, Multi Criteria and Meta Association Rule. Each category in the association rule research has its own development. Future studies mostly aim to improve the performance of the algorithm or method used in the previous studies. The improved performance can be seen from the time, memory, search space and scalability used. All association rule research developments originate from the Association Rule Mining (ARM)/Frequent Itemset Mining (FIM) technique proposed by Aggrawal in 1993. Some studies combine existing methods with ARM/FIM techniques to produce new methods such as Skymine Algorithm which is the result of the merging of High Utility Itemset and The Skyline Operators, Meta Association Rule which is a merger between ARM/FIM and Fuzzy Association Rule, and Multi Criteria Decision Analysis (MCDA) which is a combination of ARM/FIM techniques with Multi Criteria Analysis.

Most of the research that produced this new method still uses the traditional ARM / FIM technique (Apriori Algorithm), even though at this time many researchers have pointed out the weaknesses in this apriori algorithm, one of which is the process of forming rules that are not efficient because they generate candidates in each process. In addition, the user in the apriori algorithm must determine the minimum support and minimum confidence values at the beginning.

By reviewing the previous research, the Association Rule is divided into two categories. There are Association Rule with minimum support and Association Rule without minimum support. The research that does not use minimum support does not mean that it does not use thresholds for the items involved. The research only formulated minimum support to other forms and methods. Table 2 illustrates the methods in the association rule research and their relationship with the determination of the minimum support value by the user.

Table 2. Summary of using Minimum Support

No	Method	Determination of the minimum support value by the user?
1	High Utility Itemset	Yes
2	Skymine Algorithm	No
3	Top-k Association Rule	No
4	Multicriteria Decision Analysis	Yes
5	Meta Association Rule	Yes

In addition, to determine the minimum support method, another thing to consider in the association rule is to evaluate the suitability of the resulting rule. Suitability of the rules produced in the association rule process is one of the important things and is used as a measurement of the success of the association rule method. There are many rules produced which may not be used entirely in the decision-making process. The method required is a method that can choose the best rule according to the user needs based on the assessment of certain criteria. The feedback is needed for the rules that have been formed so we can learn from the previous rule formation. Furthermore, another problem arises when the source of the dataset used for the association rule comes from several datasets. The required methods is those which can combine several datasets or databases in the association rule process.

Based on the results of a review of several scientific papers, here are some research opportunities that can be done:

1. Creating a method or technique to determine the minimum support value that is in accordance with the conditions of the dataset.
2. Creating methods to measure the accuracy of the rules produced in accordance with certain criteria.
3. Combining the concept of multicriteria decision analysis with the latest association rule method to get the best rule.
4. Creating an association rule method for data streams and dynamic databases.
5. Combining the concept of multicriteria decision analysis with the meta association rule to produce the best rules from multiple datasets.
6. Creating an adaptive association rule method by determining the parameters included in the calculation in accordance with the wishes of the user to produce a rule that is useful for various parties.

4. CONCLUSION AND FUTURE WORK

By studying previous studies, in association rule need a method of determining the appropriate minimum support value and is not specified by the user. Several new methods / algorithms proposed in the previous studies can be developed or combined with other methods, so an appropriate method is produced to determine the minimum value of support in the various cases to produce an appropriate rule.

In addition, based on several research opportunities that have been discussed previously, for the future work, the research will be conducted based on the results of the paper analysis and the latest development needs that requires methods that can produce adaptive rules. The adaptive rule here means a rule that can adjust automatically from various sides, including Automatic determination of the minimum support value, selection of the best rule according to certain criteria, involving various data sources, adapting current issues and summarizing feedback from the users.

5. REFERENCES

[1] Rakesh Agrawal. 1994. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, San Francisco, CA, USA, 487–499.

[2] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, Harry Road, and San Jose. 1993. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, DC, USA, 207–261.

[3] S. Borzsony, D. Kossmann, and K. Stocker. 2001. The Skyline operator. In *Proceedings 17th International Conference on Data Engineering*, IEEE Comput. Soc, Heidelberg, Germany, 421–430. DOI=https://doi.org/10.1109/ICDE.2001.914855

[4] Quang-Huy Duong, Bo Liao, Philippe Fournier-Viger, and Thu-Lan Dam. 2016. An efficient algorithm for mining the top- k high utility itemsets, using novel threshold raising and pruning strategies. *Knowl.-Based Syst.* 104, (July 2016), 106–122. DOI=https://doi.org/10.1016/j.knosys.2016.04.016

[5] Fatima Zahra El Mazouri, Mohammed Chaouki Abounaima, and Khalid Zenkouar. 2019. Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France. *J. Big Data* 6, 1 (December 2019). DOI=https://doi.org/10.1186/s40537-018-0165-0

[6] Philippe Fournier-Viger, Cheng-Wei Wu, and Vincent S. Tseng. 2012. Mining Top-K Association Rules. In *Advances in Artificial Intelligence*, Leila Kosseim and Diana Inkpen (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 61–73. DOI=https://doi.org/10.1007/978-3-642-30353-1_6

[7] Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Justin Zhan. 2017. Mining of frequent patterns with multiple minimum supports. *Eng. Appl. Artif. Intell.* 60, (April 2017), 83–96. DOI=https://doi.org/10.1016/j.engappai.2017.01.009

[8] Ya-Han Hu and Yen-Liang Chen. 2006. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decis. Support Syst.* 42, 1 (October 2006), 1–24. DOI=https://doi.org/10.1016/j.dss.2004.09.007

[9] Mehmed Kantardzic. 2011. *DATA MINING Concepts, Models, Methods, and Algorithms* (Second Edition ed.). John Wiley & Sons, Inc., Hoboken, New Jersey.

[10] Srikumar Krishnamoorthy. 2018. Efficient mining of high utility itemsets with multiple minimum utility thresholds. *Eng. Appl. Artif. Intell.* 69, (March 2018), 112–126. DOI=https://doi.org/10.1016/j.engappai.2017.12.012

[11] Yeong-Chyi Lee, Tzung-Pei Hong, and Wen-Yang Lin. 2005. Mining association rules with multiple minimum supports using maximum constraints. *Int. J. Approx. Reason.* 40, 1–2 (July 2005), 44–54. DOI=https://doi.org/10.1016/j.ijar.2004.11.006

[12] Jerry Chun-Wei Lin, Lu Yang, Philippe Fournier-Viger, and Tzung-Pei Hong. 2019. Mining of skyline patterns by considering both frequent and utility constraints. *Eng. Appl. Artif. Intell.* 77, (January 2019), 229–238. DOI=https://doi.org/10.1016/j.engappai.2018.10.010

[13] José María Luna, Philippe Fournier-Viger, and Sebastián Ventura. 2019. Frequent itemset mining: A 25 years review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* (July 2019). DOI=https://doi.org/10.1002/widm.1329

[14] Loan T.T. Nguyen, Phuc Nguyen, Trinh D.D. Nguyen, Bay Vo, Philippe Fournier-Viger, and Vincent S. Tseng. 2019. Mining high-utility itemsets in dynamic profit databases. *Knowl.-Based Syst.* 175, (July 2019), 130–144. DOI=https://doi.org/10.1016/j.knosys.2019.03.022

[15] Jeng-Shyang Pan, Jerry Chun-Wei Lin, Lu Yang, Philippe Fournier-Viger, and Tzung-Pei Hong. 2017. Efficiently mining of skyline frequent-utility patterns. *Intell. Data Anal.* 21, 6 (November 2017), 1407–1423. DOI=https://doi.org/10.3233/IDA-163180

[16] Jiawei Han Jian Pei. 2000. Mining Frequent Patterns without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, Dallas, Texas, USA, 1–12. DOI=https://doi.org/10.1145/342009.335372

[17] M.D. Ruiz, J. Gómez-Romero, M. Molina-Solana, M. Ros, and M.J. Martin-Bautista. 2017. Information fusion from multiple databases using meta-association rules. *Int. J. Approx. Reason.* 80, (January 2017), 185–198. DOI=https://doi.org/10.1016/j.ijar.2016.09.006

[18] Heungmo Ryang and Unil Yun. 2015. Top-k high utility pattern mining with effective threshold raising strategies. *Knowl.-Based Syst.* 76, (March 2015), 109–126. DOI=https://doi.org/10.1016/j.knosys.2014.12.010

[19] Chieh-Yuan Tsai and Sheng-Hsiang Huang. 2015. Integrating Product Association Rules and Customer Moving Sequential Patterns for Product-to-Shelf Optimization. *Int. J. Mach. Learn. Comput.* 5, 5 (October 2015), 344–352. DOI=https://doi.org/10.7763/IJMLC.2015.V5.532

[20] Cheng Wei Wu, Bai-En Shie, Vincent S. Tseng, and Philip S. Yu. 2012. Mining top-K high utility itemsets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, ACM Press, Beijing, China, 78. DOI=https://doi.org/10.1145/2339530.2339546

[21] Chengqi Zhang and Shichao Zhang. 2002. *Association rule mining: models and algorithms*. Springer, Berlin ; New York.