Research on Top-k Association Rules Mining Algorithm Based on Clustering

Lei Wu, Zhe Wang*

Departments of Information Institute, North China University of Technology, Beijing 100144, China
*a2018312120107@mail.ncut.edu.cn
980339280@qq.com

Abstract: Until now, the association rule mining algorithm is still one of the core issues in the field of big data research. At present, there are many research related to association rule mining algorithms, mainly focusing on how to find frequent item sets and how to tailor rules, and based on this direction, there have been many classic algorithms, such as Apriori and FP-growth algorithms. However, most of the above algorithms process and analyze the entire data set, so that although related algorithms can be used to obtain the results, the results are not meticulous. To solve this problem, this paper proposes an algorithm, which first uses the k-means algorithm to perform cluster analysis on the data set to generate a specific number of clusters. Then, in different clusters, the association rules in each cluster are found by the improved Top-k algorithm combined with the correlation coefficient. By integrating clustering and the improved Top-k algorithm, the data set can be analyzed directionally to improve the accuracy and efficiency of the whole algorithm. And the final experiment shows that compared with the original algorithm, the running time is shortened by 14%.

1. Introduction
Today, our society has entered the era of big data. With the continuous integration of IoT, social networks, cloud computing and other technologies into our lives and the rapid development of existing computing capabilities, storage space and network bandwidth, this has also promoted the development of data analysis algorithms. Only by analyzing data can people discover the hidden rules and laws. In the future, human decision-making also needs to rely on the results of data analysis, not just on intuition.

In unsupervised learning, the training samples have no labeled information, which is to learn the unlabeled training samples to reveal the inherent nature and laws of the data, and provide a basis of further data analysis. The most researched and widely used in unsupervised learning is "clustering". Clustering attempts to divide the samples of the data set into several disjoint subsets, and each subset is called a cluster. Through this division, each cluster may mean some potential categories.

Association rule mining is a rule-based machine learning algorithm that can find interesting relationships between large databases. Its purpose is to use some metrics to distinguish the strong rules that exist on the database. In other words, association rule mining is used for knowledge discovery, not prediction, so it belongs to unsupervised machine learning method. The traditional association rule mining algorithms include Apriori algorithm and FP-growth algorithm, etc., by finding frequent item sets in the data, and then discover the association rules. The Top-k algorithm is a new association rule mining algorithm proposed by Fournier-Viger et al. Compared with the traditional association rule...
mining algorithm, the advantage of this algorithm is that the user can specify the number of required rules \( k \) to avoid generating too many Rules, so the efficiency is higher than traditional association rule mining algorithms. This paper adds correlation coefficient on the basis of Top-k algorithm to improve the accuracy of the algorithm and further improve the efficiency of the algorithm.

In this paper, experiments are carried out by combining clustering and the improved top-k algorithm, and the experiments are verified by online public data sets. Compared with the original algorithm, the running time is shortened by 14%.

2. Material and Methods

2.1. Cluster analysis
Cluster analysis is a technique to find the internal structure of data. It is to group data objects according to the information on the data describing the objects and their relationships. The aim is to make the objects of one group similar to each other while the objects of different groups are different from each other. In clustering, these similar groups are usually called clusters, so the objects in the same cluster are very similar, while the objects of different clusters are very different. The greater the similarity in the cluster, the greater the difference between clusters, indicating the better clustering effect. From the perspective of data mining, it can be roughly divided into four categories: divided clustering, hierarchical clustering, density-based clustering, grid-based clustering.

There are various clustering algorithms, but not all of them are suitable for this experiment. Considering the problem of algorithm efficiency, we prefer the clustering algorithm with high efficiency. The clustering method based on partition has the greatest advantages of fast convergence speed, simple and efficient for large data sets, low time complexity and space complexity. The basic principle of this algorithm is: suppose you have many scattered points that need clustering. The desired clustering effect is that the points in the cluster are close enough and the points of clusters are far enough. First, you must ensure that a pile of scattered points is finally divided into several classes, then select some points as initial center points, and then iteratively reset the data points until all points in each cluster are the closest and the distance between clusters is farthest. The most classic clustering algorithm based on partition is K-means algorithm. Distance is a measure of similarity between data objects, that is, the smaller the distance between data objects, the higher their similarity, and the greater their possibility in the same cluster. There are many methods to calculate the distance between data objects. K-means algorithm usually uses Euclidean distance to calculate the distance between data objects. The algorithm flow is as follows:

1. First, we determine a \( k \) value, that is, we want to cluster the data set to obtain \( k \) sets.
2. Randomly select \( k \) data points from the data set as centroids.
3. For each point in the data set, calculate the Euclidean distance between each centroid and each point. According to the criterion of the closest distance, they are divided into the corresponding classes that are closest to the cluster center (most similar);
4. After putting all the data together, there are \( k \) sets. Then recalculate the centroid of each set.
5. If the distance between the newly calculated centroid and the original centroid is less than the set threshold (indicating that the position of the recalculated centroid changes little and tends to stabilize or converge), it can be considered that the clustering has reached the expected result, and the algorithm termination.
6. If the distance between the new centroid and the original centroid changes greatly, you need to repeat 3-5 steps.

2.2. Top-k Association rule mining algorithm
Top-k association rule mining algorithm is an algorithm proposed by Fournier-Viger et al. In the process of association rules mining, through practice, it is found that the minimum support is more difficult to set than the minimum confidence, because the minimum support depends on the database characteristics that most users do not know, while the minimum confidence represents the expected
confidence that users want in the rule and is usually easy to determine. Therefore, the goal of the algorithm is to mine the top-k rule of the highest support degree on the premise of satisfying the expected confidence level. The algorithm flow is as follows:

1. Scan the database: scan each single item set; then generate all item sets of 1 * 1 scale, and calculate their support and confidence, for each effective rule \{i\} -> \{j\} or \{j\} -> \{i\} calls the procedure SAVA to store the rule in L. In addition, each valid 1 * 1 rule is stored in R in order to consider the subsequent expansion, and the flag expandLR is marked as true. Then recursively select the rule r with the highest support in R such that sup (r) >= minsup and expand it. When there are no more rules in R that support higher than minsup, the loop terminates. For each rule, the flag EXPAND L R indicates whether the rule is expanded by calling the procedures EXPAND-L and EXPAND-R or by simply calling EXPAND-L.

2. SAVA Process: Its function is to raise minsup and update List L when a new valid rule R is found. The first step is to add the rule R to L. Then, if L contains more than K rules and the support is higher than minsup, the rule whose support is exactly equal to minsup in L can be deleted until only K rules are retained. Finally, the one with the smallest support for rule R in L is set to minsup.

3. Expand process: It extends rule I -> J as an extension, scans r, and finds out the one with the highest support for rule r in r as an extension object. This algorithm scans every itemset tid from tids(I∩J). In this scanning process, for each item c appearing in each item set, if all c are greater than all items in j in lexical order, tid is added to tids(I→J ∪ \{c\}). After scanning is completed, if | tids(I→J ∪ \{c\})|/| T | ≥ minsup, rule I→J ∪ \{c\} is considered frequent and added to set r for later consideration of expansion. Finally, the confidence level of each frequent rule I → j ∪ c is calculated, and the validity of the rule is determined by dividing | tids(I → j ∪ c) | by | tids (I), if the confidence level of I → j ∪ c is not less than minconf, the rule is valid, and the procedure SAVE is called to add the rule to the list l of current top-k rules.

2.3. experiment procedure
This experiment uses clustering and association rule mining algorithm to analyze the data set, which uses the data recorded by students from a website. Firstly, the data set is clustered by k-means algorithm, and the data set is divided into reasonable clusters. Then, each cluster is analyzed by Top-k algorithm. Through this method, directional data analysis can be carried out according to the dimensions of clusters, that is, a grouping analysis is carried out on the data set, so as to obtain more detailed analysis results.

2.3.1. k-means clustering analysis of data sets
The k-means algorithm needs to specify the number of clusters. This experiment uses the difference minimum method to determine the number of clusters k, which improves the shortcomings of the k-means algorithm randomly selecting the number of clusters. We use the elbow method to iterate the algorithm repeatedly 5-15 times, and record the evaluation value for each iteration. This experiment uses the SSE (sum of the squared errors) standard Equation (1) to determine the evaluation value.

\[ SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \]  

Among them, Where Ci is the ith cluster, p is the sample point in Ci, and mi is the center of mass of Ci (the mean of all samples in Ci), and SSE is the clustering error of all samples, which represents the quality of the clustering effect. And draw the results of a relationship diagram. See Fig 1, Fig 2.
From Fig 1, Fig 2 and the results after multiple iterations, we can find that when \( k = 8 \) and \( k = 9 \), the downward trend of the curve begins to slow down, and when \( k \) is greater than 8 and 9, the downward trend becomes more and more gentle. Therefore, it is more reasonable to select \( k = 8 \) and cluster the data set into 8 clusters.

### 2.3.2. Top-k algorithm for data set analysis

After clustering the data set using the k-means algorithm, the data set can be divided into several clusters, and then the Top-k algorithm is executed for each cluster separately. First, we need to specify the number \( k \) of rules to be derived, and the expected minimum support and minimum confidence. The support is defined as follow:

\[
support = \frac{p(X \cup Y)}{I}
\]  

(2)

Where \( X \) is the previous data set, \( Y \) is the last data set, and \( I \) is the entire data set. The confidence is defined as follow:

\[
confidence = \frac{p(Y \mid X)}{p(X)}
\]  

(3)
Because the Top-k algorithm only uses the two parameters of support and confidence to restrict the rule, and does not take into account the promotion of the previous item to the latter item, it cannot effectively distinguish whether the rule is valuable, so add New metric correlation coefficient. The correlation coefficient is defined as follows:

\[
\text{Lift}(X \Rightarrow Y) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X) \times \text{Sup}(Y)}
\]

\[
\text{Lift}(X \Rightarrow Y) > 1, \text{ Positive correlation}
\]

\[
\text{Lift}(X \Rightarrow Y) = 1, \text{ Standalone}
\]

\[
\text{Lift}(X \Rightarrow Y) < 1, \text{ Negative correlation}
\]

The correlation coefficient Lift reflects how much the occurrence probability of item set Y changes from the appearance of item set X. In practical application, both positive and negative correlations need to be paid attention to. When lift is greater than 1, it means that the previous item has a positive promotion effect on the latter item, and the larger the value, the greater the promotion effect. On the contrary, it is the inhibitory effect, and when lift is 1, it can be seen from the formula that the occurrence of X has no effect on the occurrence of Y, so we need not consider the case of lift 1. Therefore, on the basis of support and confidence, we can identify valuable rules by adding correlation coefficients. First, after clustering the data set, using support, confidence, and correlation coefficients equal to 0.1, 0.3, and 10, respectively, the experiment is performed. Fig 3 is the algorithm flow chart.

![Algorithm flowchart](image_url)

Fig. 3 Algorithm flowchart
3. Result

The following algorithm is used in combination with clustering and the improved Top-k algorithm to perform algorithm verification and result analysis on the user course selection records of a website in 2019, which is limited to space. In this paper, we select some clustering results based on the clustering results Several rules of high confidence.

As follows: Table 1 is the association rules related to the front-end technology-related courses. From the results, we can see that the front-end framework of Spring Boot can launch html5, docker and other courses. Spring Boot is an open source framework of java, which is designed based on Spring4.0 and is one of the most popular open source frameworks of java at present. While html5 is a popular front-end scripting language, it can be seen from the results. most users like to choose html5 courses and Spring Boot courses at the same time, the confidence is 0.66, and angular is another front-end script Language is also closely related to html5. According to the results, users that choose angular courses also tend to choose html5 courses, and the confidence level of the rule is 0.8. It can be seen that the results of this algorithm are more accurate.

| Association rules     | Confidence |
|----------------------|------------|
| Spring Boot->html5   | 0.66       |
| Spring Boot->docker  | 0.66       |
| angular->html5       | 0.8        |

Table 2 is the result of mining association rules for image processing related technologies. It can be seen that cocos2d-x was once a popular game development engine, and react.js and cocos2d-x also have a great dependency. Unity-3d, which has been popular in recent years, is also inseparable from courses related to react.js and animation, with a rule confidence of 0.834.

| Association rules       | Confidence |
|-------------------------|------------|
| react.js->cocos2d-x     | 0.83       |
| Animation->unity-3d     | 0.834      |
| react.js->unity-3d      | 0.834      |

Table 3 shows the mining results of association rules for testing related technologies. From the table, we know that various testing technologies are closely connected. The confidence level of security testing and automated testing in mining results is 0.75. The confidence level of the functional test introduced by automated testing is 0.66. The functional test has a confidence level of 0.66 for the performance test.

| Association rules               | Confidence |
|---------------------------------|------------|
| Safety testing->automated testing| 0.75       |
| Automated testing->functional testing| 0.66      |
| Functional test->performance test| 0.66      |

Based on the above analysis results, the clustering results of the algorithm are reasonable. The clustering analysis results divide the data set into 8 categories, and the records of each category have
certain similarities. The results after mining through the improved Top-k association rule mining algorithm are also more accurate, which is basically in line with the public perception in real life. Large, HTML5 and other front-end scripting languages are also chosen by most people at the same time. In the "test-related technology" category, the results also show that the relevance of various test technologies is also very large, which proves the effectiveness of the algorithm.

4. Discussion
The improved algorithm in this experiment firstly uses k-means algorithm to cluster data sets to generate different clusters, and the data in these clusters have certain similarity, so that subsequent algorithms can process each cluster separately without analyzing the entire data set. Therefore, the efficiency of the algorithm can be effectively improved. The implementation process of Tok-k algorithm first prunes the support threshold rules to filter out the non-conforming rules and generate candidate sets, then filters the extended rules with support and confidence, and finally gets the result set through such iteration. Because the two criteria of support and confidence are not enough to judge valuable rules, correlation coefficient is added to judge, and the rules are pruned again to reduce the number of rules in the candidate set. Therefore, the calculation time required by the algorithm will be reduced, so the running efficiency of the new algorithm will be better than that of the original algorithm.

In order to verify the improved effect of the algorithm, the author compares the execution time of top-k algorithm and the improved algorithm through experimental simulation. The experimental environment is: win10 64-bit operating system, 16G memory, and intel i7-9700F 3.00GHz CPU. Run the algorithm 10 times and record the running time. From Fig 4, we can see that with the increase of support, the running time of the new algorithm and the old algorithm shows an upward trend in the overall trend, but under the same specified support, the running time of the new algorithm is obviously less than that of the original algorithm.

At the same time, in order to study the influence on the efficiency of the algorithm when the value of K set by the user increases, we also run the algorithm 10 times, increasing the value of K by 50 each time, gradually increasing the value of K from 50 to 500, and recording the change trend of the algorithm's memory consumption. As shown in Fig 5, with the constant increase of k value, the
memory consumption of the algorithm basically does not change too much, which illustrates the stability of the algorithm. And we can see that when the value of K is less than 400, the running time of the algorithm increases slightly, while when the value of K is greater than 400, the running time of the algorithm increases greatly. This may be because the number of generation rules specified by the user increases to a certain extent, and the processing time of the algorithm for data increases suddenly. Therefore, using this algorithm, it is recommended not to generating too many rules, in order to avoid the algorithm running for too long.

5. Conclusions
Based on the clustering and the improved Top-k algorithm pair, this paper analyzes and analyzes the data generated after preprocessing the data on the students' course selection to find out the association rules between the courses. Firstly, the k-means algorithm is used to cluster and analyze the data set. The data set is subdivided by clustering and divided into several clusters, in which the data in each cluster has a certain degree of similarity. Then, each cluster is analyzed by the improved Top-k algorithm, and finally the association rules are obtained. In this way, the algorithm can be prevented from directly analyzing the entire data set, effectively solving the problem that the algorithm analysis is not detailed, and the integration of the improved Top-k algorithm improves the overall operation efficiency. Through the running results of the algorithm and the performance analysis results, it can be verified that the new algorithm can effectively improve the defect of imprecise analysis caused by direct analysis of the entire data set, and has higher efficiency and more accurate results than the original algorithm.

Acknowledgements
This thesis was completed under the guidance of Associate Professor Wu Lei and the School of Information, North China University of Technology. Deep professional knowledge, rigorous learning attitude, work style of excellence, tireless teaching of noble virtues, self-improvement, noble manners of tolerance, and simple and charming personality have a profound impact on me. I couldn’t help setting a great academic goal, but also mastering basic research methods, and also made me understands a lot of truth about being human and treating others. From the topic selection to the completion of the thesis, every step is completed under the guidance of the tutor, and the tutor has put
a lot of effort into this. Here, I would like to express my high respect and heartfelt thanks to my teacher. Thank you very much!

References
[1] Fournier-Viger P., Wu CW., Tseng V.S. (2012) Mining Top-K Association Rules. In: KosseimL., Inkpen D. (eds) Advances in Artificial Intelligence. Canadian AI 2012. Lecture Notes in Computer Science, vol 7310. Springer, Berlin, Heidelberg.
[2] Zhao Q, Zang Z, Huang M, Zhang X.D, Ouyang X.P. (2020) Effects of H and He on the clustering behavior of transmutation elements in tungsten[J]. Nuclear Inst. and Methods in Physics Research, B, 470.
[3] Choksi K, Jain S, Pindoriya N. (2020). Feature based clustering technique for investigation of domestic load profiles and probabilistic variation assessment: Smart meter dataset. Sustainable Energy, Grids and Networks. 22. 100346. 10.1016/j.segan.2020.100346.
[4] Wei K, Sergio R, Levi W, Elijah K, Su H. (2020) Sensitivity of sequence methods in the study of neighborhood change in the United States[J]. Nuclear Inst. and Methods in Physics Research, B, 470.
[5] Burin B, Kittisak B, Veerasak P, Manussabhorn P, Korakot N. (2020) Determination of whether morphometric analysis of vertebrae in the domestic cat (Feliscatus) is related to sex or skull shape[J]. Springer Singapore, 95(3).
[6] Evangelos T, Juri Y, Fujun H. (2020) Post-consensus analysis of group decision making processes by means of a graph theoretic and an association rules mining approach[J]. Elsevier Ltd, 94.
[7] Akbar T, Amir H. G, Asadollah S. (2020) A survey of evolutionary computation for association rule mining[J]. Elsevier Inc., 524.
[8] Abediniangerabi B, Shahandashti S.M, Makhamalbaf A. (2020) A data-driven framework for energy-conscious design of building facade systems[J]. Elsevier Ltd,29.
[9] Lakshmi K.S, Vadivu G. (2020) Network based approach for discovering novel gene-phenotypic association and disease co morbidities using ontological data[J]. Elsevier B.V., 167.
[10] Xin C.H, Guo Y.G, Lu X.B. (2019) Using reinforcement learning to improve the algorithm of association rule mining of Treap in large database [J / OL]. Computer application research: 1-6.
[11] Raymond M, Francisco C, Fabio C, Jenny C. (2020) A product-centric data mining algorithm for targeted promotions[J]. Elsevier Ltd, 54.
[12] Luo L, Ge Q, Zhou Y, Xia B. (2020) Mining equipment data association rules based on Apriori algorithm [J / OL]. Command control and simulation: 1-5.
[13] Gao Q, Bu X.Y. (2020) Mining model design of association rules based on incremental FP growth algorithm [J]. Information technology and informatization: 169-170.
[14] Jing T, Ding W, Liu C. (2020) Multi step spectrum occupancy prediction method based on association rule mining [J]. Application of electronic technology, 46 (03): 80-85.
[15] He Z. (2020) Application of association rule technology based on MapReduce in power marketing big data [J]. Shanxi electric power: 45-49.
[16] Zhang J.Y, Cui H.J, Peng Y.M, Li Y, Shen W, Zhang X, Sun C.Y. (2019) Study on the rule of traditional Chinese medicine in capecitabine related hand foot syndrome based on cluster analysis and association analysis [J]. Modern Chinese medicine clinical, 26 (06): 12-19.
[17] Wang J.B, Yang L. (2019) Research on the relationship between traffic violations and accidents based on K-means clustering [J]. Traffic construction and management, (05): 92-95.
[18] Yan I. (2016) Research and application of cluster method based on grey correlation analysis [D]. Chongqing University of Posts and Telecommunications.
[19] Jungyeol H, Reuben T, Dongjoo P. (2020) Application of association rules mining algorithm for hazardous materials transportation crashes on expressway[J]. Accident Analysis and
[20] Mu S. (2012) grey correlation and clustering model construction and application of panel data [D]. Nanjing University of Aeronautics and Astronautics.

[21] Abderrahmane E, Khalil M. (2020) Breast cancer classification with reduced feature set using association rules and support vector machine[J]. Network Modeling Analysis in Health Informatics and Bioinformatics, 9(4).