International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 04 - Issue 10 || October 2018 || PP. 35-45

Correlation analysis of user power consumption behavior based on improved Apriori algorithm

Liu Haiqing[1], Qu Jinmeng[1],Li Yuancheng[1]
[1]*(School of control and Computer Engineering, North China Electric Power University, China)*

Abstract: With the increasing number of smart meters, The client generates a large data and there are the potential information of users'electricity consumption behavior behind these data. By mining the similarities of different users' electricity consumption behaviors, the paper finds out the relationship between various factors and different electricity consumption behavior patterns, which can be used to customize and analyze service applications for power companies, users and governments. In the process of association analysis, an improved algorithm based on prefix itemsets and vertical Boolean matrix is proposed. By using the frequent itemsets storage method of prefix itemsets, the efficiency of the algorithm in the join step and pruning step is greatly improved. At the same time, the vertical Boolean matrix is used to store the transaction database and the "and"operation is used to calculate the support degree, which further optimizes the efficiency, reduces the occupied memory space and improves the efficiency of the algorithm. Compared with other algorithms, the effectiveness and feasibility of the proposed method are verified.

Keywords: Apriori algorithm; power consumption; association analysis; prefix item set; vertical Boolean matrix

I. INTRODUCTION

This paper comprehensively and effectively analyzes the relationship between the various types of electricity data and the various economic, conscious, demographic and other factors related to users' electricity consumption behavior , discovers different user electricity consumption patterns, digs out the correlations between various factors and different electricity consumption behavior patterns, and makes a thorough analysis of different user electricity consumption behavior patterns. It helps to provide customized personalized services to government, residents and business users. In this regard, scientific research on user electricity consumption behavior has been gradually carried out at home and abroad.

The classification of residential users' electricity consumption behavior is indispensable to power companies, in order to understand the personalized needs of users and provide targeted services K-means clustering algorithm is one of the most popular methods in previous studies. However, in the traditional K-means algorithm, the initial cluster center is randomly selected, which makes it easy to be subjected to local optimum and difficult to converge to the global minimum. In order to overcome this disadvantage, an improved K-means algorithm based on simulated annealing algorithm is proposed in literature [1]. By using simulated annealing algorithm, the optimal clustering center can be obtained. Finally, experiments show that the proposed method can extract typical power consumption patterns and has better performance than the traditional K-means algorithm. With the development of intelligent power market and the change of load characteristics, in order to build global energy interconnection and improve the effect of demand response projects, literature [2] mainly discusses the clustering analysis of user behavior and the target user selection of demand response of smart grid. Firstly, the user electricity classification system is built, and the decomposition model is constructed by fuzzy membership function according to the user information obtained by load detection. Then choose the most valuable sensitive and the lowest cost of demand response as the optimization parameters to construct the user selection model. Analyzing the characteristics of users' electricity consumption helps to improve the energy efficiency of DSM. The paper[3] presents a method for analyzing user's power consumption behavior in big data environment. Firstly, the clustering algorithm is used to cluster the daily load curve of the power grid in the past year. Under different date clusters, the electricity consumption behavior of residents and non-residents is analyzed. Finally, the experimental results show that this method provides effective support for orderly and intelligent power consumption of residents and non-residents, which is conducive to peak load transfer and stable operation of power grid.

Document [4] proposes that the accuracy of load forecasting can be improved by analyzing the correlation between different electricity consumption behaviors. Various machine learning techniques are used to analyze problems from family level and district level. Firstly, the algorithm for solving problems under different scales is determined , and then the main time segments of these time series are found to help improve the accuracy of short-term load forecasting. Data mining technology is widely used in power load analysis, which is conducive to decision-making analysis of the power industry. The paper [5] introduces an incremental tree induction algorithm by using cascading and sharing methods, and the minied important classification rules are used to analyze the user's electricity consumption behavior. These users generally involve users in the industrial sector, education industry users and other regular industry users.

In a word, the research on user behavior analysis is still in its infancy under the large data environment, and there is not much research on this aspect at home and abroad. In previous studies, the similarity of users' power consumption is clustered or simple association analysis were used , but the influence factors of specific users' electricity consumption behavior were not excavated. By analyzing the influence of large data on user behavior analysis, the appropriate correlation model between power consumption patterns and influencing factors is established and the problem of user behavior analysis under large data environment can be better solved.

II. APRIORI ASSOCIATION RULE ALGORITHM

2.1 Apriori algorithm thought and process

Apriori algorithm is a basic algorithm in association rules, and is one of the most widely used association rules data mining algorithms. The core idea is to generate frequent itemsets by generating candidate sets and selecting pruning.

According to the basic idea of the Apriori algorithm described above, the Apriori algorithm [6] can be described by the following implementation steps.

Step 1: Scanning transaction database D, calculating the C1 support of all 1-candidate itemsets and comparing with min_sup, selecting candidate sets whose support is greater than min_sup to generate 1-frequent itemset set L1.

Step 2: Let the 1-frequent itemset set L1 join itself （L1∞L1）, prune it and generate the 2-candidate itemset set C2, then select the candidate set whose support degree is greater than the minimum support degree min_sup to generate the 2-frequent itemset set L2.

Step 3: By using the 2-term frequent itemset set L2, the 3-term frequent itemset set L3 is generated. The operation is repeated and iterated until the final results satisfy the final condition k- frequent item sets or k-candidate itemsets.

The flow chart of the Apriori algorithm is shown in Figure 1,

2.2 Apriori algorithm analysis

With the gradual increase of data,the shortcomings of the algorithm slowly emerge through the continuous study of Apriori algorithm. The current Apriori algorithm has two serious performance defects[7]:

(1) Due to frequent scanning of database, I/O load is very serious. Through the analysis of the implementation process of the algorithm, we can conclude that the number of elements in the candidate itemset Ck determines the number of times that the algorithm scans the database. For each candidate element, you need to scan the database once. This creates a very large I/O load.

(2) candidate itemsets Ck are generated by Lk-1 self join of frequent itemsets, resulting a large number of redundant candidate sets. That is to say, a frequent itemset containing 104 1- items can generate about 107 2-candidate sets. Compared with the number of frequent itemsets, the number of candidate itemsets increases exponentially, which requires high spatial and temporal efficiency of the algorithm.

Because of this, many scholars have improved the Apriori algorithm. The existing improved algorithms are as follows [8]:

(1)Hash based optimization technology: this technique can be applied to compress K candidate set Ck (k > 1). When Lk-1 is generated by Ck-1, we can get all the k-items corresponding to any transaction. Then we scatter them into different hash buckets. For each hash bucket we add a bucket count and reduce candidate items that need attention by comparing the value of each bucket count with the set support threshold. This method is very effective when implemented in k=2.

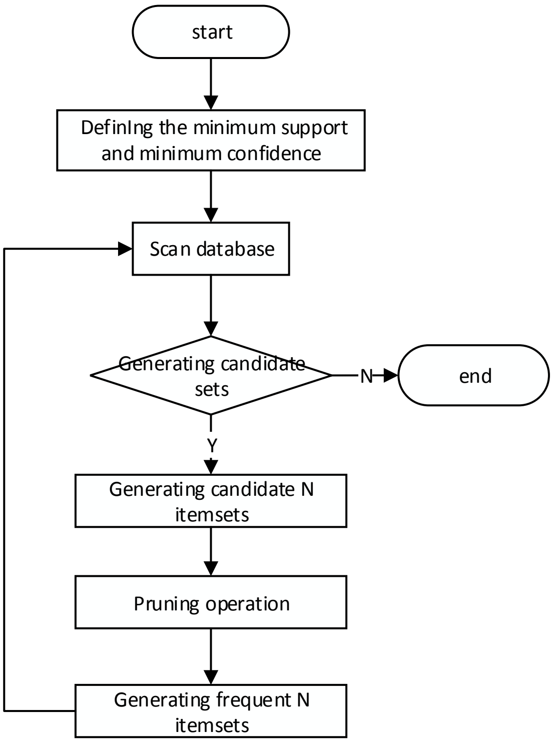

start
DefinIng the minimum support and minimum confidence
Scan database
Generating candidate sets
N
end
Y
Generating candidate N itemsets
Pruning operation
Generating frequent N itemsets

Fig.1.The flow chart of the Apriori algorithm

(2) Transaction compression optimization technology: if there is no frequent (k-1) - itemsets in a transaction, then there can not be any frequent k-itemsets in this transaction. We can realize the compression of large-scale database by deleting such transactions.

(3)partitioning technology: partitioning technology is to divide database D into N non-overlapping parts, the size of each part must be able to put into memory. We set the minimum support for transactions to min_sup in transaction database D, and the value of the minimum support count for a partition is the product of min_sup and the number of transactions in that partition. The candidate itemsets of database D are composed of each partitioned local frequent itemsets, and determining whether the corresponding candidate itemsets are added to the global frequent item set when the database D is scanned twice[10].

(4) Sampling technique: Randomly select the sample S that can be put into memory in database D, and then search for frequent itemsets in memory S. This is a way to exchange effectiveness with sacrifice accuracy. We can use a threshold lower than the minimum support to find frequent itemsets Ls localized to S, thereby reducing the omission of frequent itemsets in database D [11].

(5)Dynamic itemset counting: The Apriori algorithm determines the new candidate after scanning the database D. Dynamic itemset counting is different from it. It divides database D into different blocks, each block has a start flag, and can add new candidate itemsets at any start point. We can add new candidate itemsets from any start sign. This technique can dynamically analyze the support of all counted itemsets. If we can already determine that any subset of an itemset belongs to a frequent itemset, we can add it to the candidate and make it a new candidate. If all subset of an itemset has been determined to be frequent, it is added to the new candidate set [12].

2.3 Apriori algorithm optimization

(1)transformation of original database based on vertical Boolean matrix.

For any original database D that takes item I as row and transaction T as column is converted into a block of vertical Boolean matrix Rmn.

The element in the matrix :$r_{I_i T_j}=\begin{cases}1, & I_i \in T_j \\ 0, & I_i \notin T_j\end{cases}$ i=（1,2,3⋯.m）， j=(1,2,3⋯.n)。

Where m is the number of items in the original database, n is the number of transactions in the original database. represents the first I item, represents the first j transaction.

In addition, a row of weights is added in front of the matrix. The weight value represents the number of transactions with the same content, and the original value is 1. Add a row defined as sum to the back of the matrix to record the number of items contained in each transaction.

For example, there is a transaction database, as shown in Table 1, and its corresponding vertical Boolean matrix is shown in Table 2.

Table 1 Transaction database

Tidset	Itemset
T1	I1、I2、I5
T2	I2、I4
T3	I2、I3
T4	I1、I2、I4
T5	I1、I3
T6	I2、I3
T7	I1、I3
T8	I1、I2、I3、I5
T9	I1、I2、I3

Table 2 Vertical Boolean matrix

Tid	T1	T2	T3	T4	T5	T6	T7	T8	T9
weight	1	1	1	1	1	1	1	1	1
I1	1	0	0	1	1	0	1	1	1
I2	1	1	1	1	0	1	0	1	1
I3	0	0	1	0	1	1	1	1	1
I4	0	1	0	1	0	0	0	0	0
I5	1	0	0	0	0	0	0	1	0
sum	3	2	2	3	2	2	2	4	3

(2)support count based on vertical Boolean matrix

This paper enumeration the support degree by combining matrix and "and" operations.Carry out "and" operations on the corresponding rows of the item set in the matrix and the result is then summed up with the corresponding weight product. The support count for Ii (Ij) in association rules is defined as follows:

$$\mathrm{Support\_count}(\mathrm{I_i}，\mathrm{I_j}) = (\mathrm{Tidset}(\mathrm{I_i}) \wedge \mathrm{Tidset}(\mathrm{I_j})) * \omega^{\mathrm{T}}$$

Where "" is the “and” operator , is the corresponding weight in the matrix.

For example, for table 2, we have the following expressions for the support degree of I1 and I2.

$$\mathrm{Support\_Count}_{(\mathrm{I_1},\ \mathrm{I_2})} = \left(\mathrm{Tidset}(\mathrm{I_1}) \wedge \mathrm{Tidset}(\mathrm{I_2})\right) * \omega^{\mathrm{T}}$$

= ((1 0 0 1 1 0 1 1 1) $\wedge$(1 1 1 1 0 1 0 1 1))*(1 1 1 1 1 1 1 1 1)T

If the total support count of candidate frequent itemsets obtained by sum operations of matrices is greater than or equal to the set minimum support count, the itemsets are frequent itemsets. Usually, frequent k-itemsets are recorded as Lk.

(3) storage structure of frequent itemsets based on prefix itemsets

In the improvement of this algorithm, we adopt a <key , value> data storage method for frequent itemsets. For example, for each itemset in a frequent k-itemset, the key represents the first (k-1) item of the k-itemset, and the value represents the set of k-itemsets with the same key value.

For example, with a minimum support of 2, the data storage structure of the prefix itemsets corresponding to the transaction database shown in Table 1 is shown in Table 3

Table 3 storage structure of frequent itemsets based on prefix itemsets

	key	value
Frequent 1- itemsets	NULL	I1、I2、I3、I4 、I5
Frequent 2- itemsets	I1	I2、I3、I5
	I2	I3、I4 、I5
Frequent 3- itemsets	I1、I2	I3、I5

In a frequent 1-itemset, because there is only one item in each frequent itemset, there is no key value, whose key is NULL.

(3) connection steps based on prefix itemsets

According to Theorem 3, it can be concluded that the join steps only need to be performed in frequent itemsets with the same key value. For example, if frequent k-itemsets generate candidate (k+1) -itemsets, it is only necessary to merge any two values from the value under the same prefix key, and then merge the prefix key value into a candidate (k+1) -itemset. For example, the frequent 2-itemsets in Table 3 are self-joined to produce candidate 3-itemsets. The value sets with the prefix key = I1 are joined to obtain {(I2, I3), (I2, I4), (I3, I4)} and the corresponding candidate sets are {(I1, I2, I3), (I4),(I1, I2, I4)}.The value sets with the prefix key = I2 are joined to obtain{（I3、I4）、（I3、I5）、(I4 、I5)} and the corresponding candidate sets are {（I2、I3、I4）、（I2、I3、I5）、(I2、I4 、I5)}.

(4) pruning steps based on vertical Boolean matrices and prefix itemsets

There are two main steps in pruning operation based on vertical Boolean matrix.

The first step: when the transaction database is transformed into a vertical Boolean matrix, the transactions covering the same item are merged and the corresponding weights are accumulated. For example the set of transactions T3 and T6 in the transaction database in Table 1 is the same, they are all (I2, I3). The same sets of transactions T5 and T7 are all (I2, I3). The vertical Boolean matrix corresponding to this pruned transaction database is shown in Table 4.

Table 4 vertical Boolean matrices after merging the same transaction

Tid	T1	T2	T3	T4	T5	T8	T9
weight	1	1	2	1	2	1	1
I1	1	0	0	1	1	1	1
I2	1	1	1	1	0	1	1
I3	0	0	1	0	1	1	1
I4	0	1	0	1	0	0	0
I5	1	0	0	0	0	1	0
sum	3	2	2	3	2	4	3

The second step:In the process of finding frequent k-itemsets, the columns whose sum is less than K in a vertical Boolean matrix can be deleted according to theorem 2.For example, when calculating Table 4 frequent 3- itemsets for corresponding databases, the value of sum must be greater than or equal to 3. Table 4 shows that the sum values of transactions T3, T4 and T5 do not meet the requirements, and the corresponding columns need to be deleted. The Boolean matrix after pruning is shown in Table 5.

Table 5 Boolean matrices for pruning undesirable transactions

weight	T1	T4	T8	T9
weight	1	1	1	1
I1	1	1	1	1
I2	1	1	1	1
I3	0	0	1	1
I4	0	1	0	0
I5	1	0	1	0

sum	3	3	4	3

2.4 algorithm description

The complete algorithm can be described as:

(1) set minimum support min-sup.

(2) generate the corresponding vertical Boolean matrix R based on transaction database D. Add a row of weights before the matrix to represent the number of same transactions. Add a row of item count sum to the matrix to record the number of items contained in each transaction. Then the same transaction is merged and the weights are accumulated. Use "and" operations to calculate the support degree of each item, and find out the frequent 1- itemsets L1. The frequent 1- itemsets L1 is stored in a prefix itemset.

(3) find frequent (k+1) - itemsets L (k+1) from the frequent k- itemsets Lk that have been obtained.

A.The vertical Boolean matrix is pruned according to the condition sum > min-sup, and the columns that do not meet the condition are deleted. The vertical Boolean matrix R 'after pruning is generated.

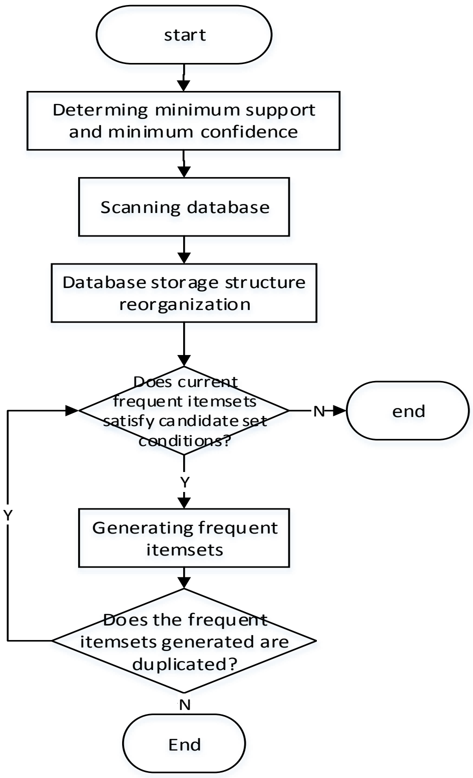
B. generate candidate (k+1) - itemsets based on frequent k- itemsets. Based on the pruning operation of prefix itemsets, the candidate itemsets are pruned, and the candidate itemsets which do not meet the condition of frequent itemsets are deleted.

C.Combining the vertical Boolean matrix R', the support degree of each candidate set after pruning is calculated by using “and” computing. Comparing the support of each candidate set with the minimum support, the frequent (k + 1) - itemsets are obtained according to the conditional support_countmin-sup of the frequent itemsets.

(4) cyclic steps (3) until the frequent itemsets can not be regenerated.

After optimization, the flow chart of the Apriori algorithm is shown in Figure 2.

Fig. 2 flow chart of Apriori algorithm after optimization


start
Determing minimum support and minimum confidence
Scanning database
Database storage structure reorganization
Does current frequent itemsets satisfy candidate set conditions?
N
end
Y
Generating frequent itemsets
Y
Does the frequent itemsets generated are duplicated?
N
End

2.5 algorithm performance analysis after optimization

Compared with other improved Apriori algorithm, the improved algorithm only scans the main database once, which will greatly reduce the I / O load and time. In the subsequent steps, only a specific set of dynamic arrays need to be scanned to avoid unnecessary transactions and itemsets. In addition, the number of transactions decreases with the increase of frequent itemsets. Therefore, this algorithm can greatly reduce the running time and storage space.

In the third step, the candidate k-itemset Ck is generated by frequent k-1 itemset Lk and frequent 1 itemset L1 join, which not only avoids repeated comparisons between frequent K-2 itemsets, but also avoids duplicate redundant candidate sets. In addition, it does not need to consider whether all the combined k-1 items of CK belong to Lk-1, but only needs to be connected. Frequent itemsets are pruned once for the first time, so the running time of the algorithm will be greatly reduced and the computational efficiency of the algorithm will be improved.

In the process of calculating the degree of support, the reconstructed database only needs the intersection of Tid lists corresponding to subsets Ik-1 and I1. Tid transaction numbering lists are sorted in ascending order, so there is no need to repeat loop matching scan when calculating the intersection of two list sets, only one traversal can get the intersection of two sets, thus avoiding frequent scanning of the database, thus greatly improving the time efficiency of the algorithm.

In order to compare the efficiency before and after optimization, the optimized Apriori algorithm is tested with mushhroom data set. The data and a total of 8124 transactions, on average, each transaction has 119 data items. Under different support, the optimized Apriori algorithm and the MC_Apriori algorithm based on compression matrix are compared. The test results are shown in Figure 3.

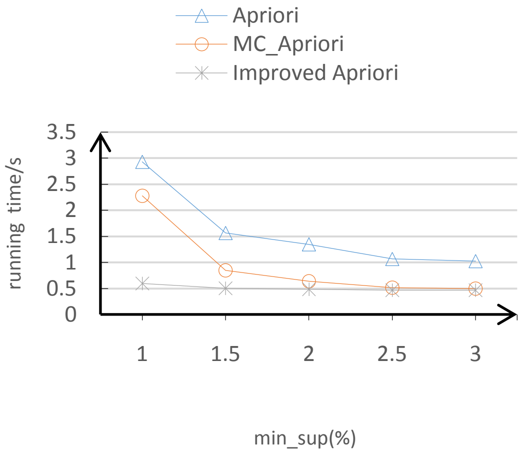


Fig.3 run time diagram with different supports

It can be concluded from Figure 3 that compared with Apriori algorithm, the optimized Apriori algorithm takes less time at runtime than Apriori algorithm, and also has a certain reduction compared with MC_Apriori algorithm. Moreover, when the support count is small, we will see a more prominent effect

III. CORRELATION ANALYSIS OF ELECTRICITY CONSUMPTION BEHAVIOR OF POWER USERS BASED ON IMPROVED APRIORI

3.1 user factor analysis

Each type of user's electricity consumption behavior is not only related to the current electricity price, weather and other uncertain factors, but also objectively related to the type of housing, age composition, the number of major electrical appliances, the economic conditions of users, the number of members of the user composition and other factors. Subjectively, household energy consumption will also be. Members of the family are influenced by their attitudes toward electricity use, and the educational level of each member of the family is also related to this subjective attitude. Generally speaking, there are many factors affecting the consumer's electricity consumption behavior, and these factors have some relevance, which can be summarized as the following four aspects [8]:

(1) Economic factors: the most important indicators of economic factors are the number of family income sources and the income of each family member. Secondly, housing type, housing size and the number of major household appliances and other factors can also reflect the economic situation of a family.

(2) Population factors: As far as a family is concerned, population factors mainly refer to gender composition, age structure and the size of family members and other factors.

(3) Work factors: mainly including the number of household work and type of work, as well as the daily working hours and weekly vacation time will have a certain impact on residents'electricity consumption behavior.

(4) consciousness factors: the environmental awareness, attitude and attitude of intelligent power consumption. This is also related to the educational level of family members. In addition, in the daily use of electricity, electricity efficiency is a strong sense of environmental awareness of residents. Moreover, these residents are relatively concerned about the energy consumption policy and energy conservation and emission reduction. The electricity consumption of residents will also be affected by the attitudes of residents to monthly electricity bills.

3.2 data feature extraction and discretization processing

In this paper, the main source of household electricity information is a residential survey of an Irish power company. These household survey data are respectively residential demographic factors, economic factors, awareness factors, work factors and other aspects of a detailed survey. However, because the survey information is too complicated and the data format does not conform to the situation, it is necessary to extract the survey data and format conversion.

In order to study the important factors affecting residential electricity consumption behavior in real life, this paper firstly studies the survey data of residential buildings, retains useful information for analyzing the influencing factors of residential electricity consumption, deletes useless fields, and discretizes the continuous data which can not be processed. The final result is shown in the table below. According to the above clustering results of residential users, 300 users are labeled and the type attributes of users'electricity consumption behavior are set.

Then ranking attribute values, transforming relational data into transaction data, transforming it into an instance of one-dimensional association rules for association analysis, using the optimized Apriori algorithm to find frequent k-item sets in transaction database under the given minimum support.

Table 6 the main influencing factors and their variable values are extracted.

influence factor	attributes	Variable value	Meaning of variable value
Demographic factors	Permanent population	1-10	Housing population
	Number of children	1-10	Population below 15 years of age
	Number of elderly people	1-10	Population over 65 years of age
	Average age	1-100	Average age of housing
	Average education level	1-5	1. there is no formal education, 2. primary 3. middle schools, 5. masters and above.
	per capita income	1-5	1. No more than 15,000 Euros 2.15,000 to 30,000 Euros 33,000 to 50,000 Euros 45,000 to 75,000 Euros 5.75,000 or more Euros
economic factors	Heating mode	1-6	1. Electricity (Central Electric Heat Storage) 2. Electricity (Heater Plug-in) 3. Gas 4. Oil 5. Solid Fuel 6. Renewable
	Cooking method	1-4	1. rice cooker 2. gas cooker 3. fuel stove 4. solid fuel cooker
	Number of main electrical appliances	1-10	Number of electrical appliances commonly used in residence
	Do you use electric shower?	0-1	yes= 1, no =0
	Do you use air conditioning?	0-1	yes= 1, no =0
	Do you use refrigerators?	0-1	yes= 1, no =0
	House size	1-3	1.<60 square meter 2.60-130 square meter 3.>130 square meter
	Residential type	1-5	1. apartments 2. houses, 3. detached houses, 4. balconies, 5. bungalows.
	Will electricity bills affect the use of electricity?	0-1	yes= 1, no =0
Consciousness factor	Does the replacement of electrical appliances consider energy rating?	0-1	yes= 1, no =0
	Are they satisfied with the current electricity price?	1-5	1. very satisfied 5. very dissatisfied.
	Will you pay attention to the use of electrical appliances	0-1	yes= 1, no =0

	everyday?		
Working factors	Do you go to work on Saturday?	0-1	yes= 1, no =0
	Number of family workers	1-5	Number of workers
	Daily working hours	1-4	1.8 hours 2.8-10 hours 3.10-12 hours more than 4.12 hours.

3.3 experimental results and analysis

This paper selects 300 residential electricity consumption data from the smart meter data. The original data is 336 dimensions. Through adding missing data and eliminating erroneous data, the data is reduced to 48 dimensions by autoencoder algorithm [14]. Finally, the original data is embedded by US-ELM and K-m is used. Eans algorithm [1,15] clustering, eventually gathered into 5 categories. The 5 category is described as follows:

Mode 1: The electricity consumption is high, and has obvious peak and valley characteristics. The electricity consumption in the noon time of a day is obviously higher than that in the morning and evening.

Mode 2: high power consumption throughout the day, and the average power consumption in each period is relatively small.

Mode 3: Similar to Mode 1, it also has obvious peak and valley characteristics, but the daily electricity consumption is normal, belonging to the middle level.

Mode 4: electricity consumption is at a low level throughout the day.

Mode 5: similar to mode 2, but electricity consumption is generally lower than mode 2The association rules mining for transformed data requires minimum confidence and minimum support. The higher the minimum support set, the faster the algorithm will execute and the fewer rules will be obtained. Using the optimized Apriori algorithm, the converted data can be mined directly, and the minimum support is 0.35. The potential relationship with different types of residential electricity consumption behavior can be seen. The main relations are as follows:

1. The per capita income of household users ranges from 50,000 to 75,000 euros, and 46% of the households larger than 130 square meters conform to Model 2.

2. 43% of the households whose per capita income is less than 15,000 euros and whose house size is less than 60 square meters conform to Model 4.

3. 63% of the households whose per capita income is between 50,000 and 75,000 euros and whose houses are larger than 130 square meters and whose energy rating is taken into account by replacing appliances conform to Model 1.

4. 53% of the households with a per capita income of 50,000 to 75,000 euros and a house size of 60-130 square meters and a family workforce of four are in the same category.

5. The per capita income of household users ranges from 30,000 to 50,000 euros, the size of the house is 60-130 square meters, and the number of household workers is 4. 53% of the household users who pay attention to the daily use of electrical appliances conform to mode 3.

To some extent, the potential relationship found indirectly reflects the relationship between household electricity consumption behavior and household population, household structure, income status, residential type, energy-saving awareness and other aspects, which can make users have a clearer understanding of electricity consumption, and can also provide a basis for customized power grid personalization strategy. The parameters of the traditional household electricity load forecasting model can further improve the accuracy of power load forecasting. Different models can be developed and different strategies can be developed. Through the implementation of different strategies, it is helpful to improve the level of refined operation management and demand side management of power enterprises, so as to further expand the breadth and depth of service for grid companies, and provide data support for future power demand side policy making.

Comparing the time consumed by the classical Apriori algorithm and the improved algorithm under different support, the test results are shown in Figures 4.

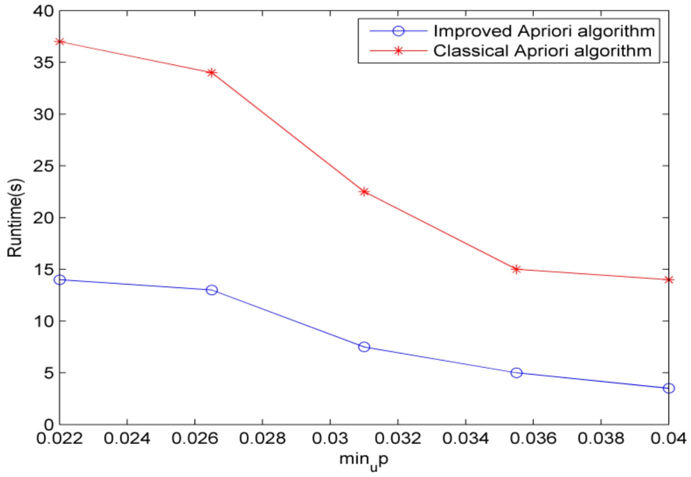


Figure 4 algorithm runtime

The above is the result of experimental operation. The abscissa is the support set by the algorithm, and the ordinate is the running time of the algorithm under a certain support. From the point of view of time consumption, the improved Apriori algorithm is better than the classical Apriori algorithm in time consumption. The time consumption of the improved Apriori algorithm changes slowly, and the change trend of the classical Apriori algorithm is relatively large. However, with the increase of support, the advantages of the improved algorithm on two types of datasets are decreasing. The main reason for this phenomenon is that the number of candidate frequent itemsets decreases with the increase of support.

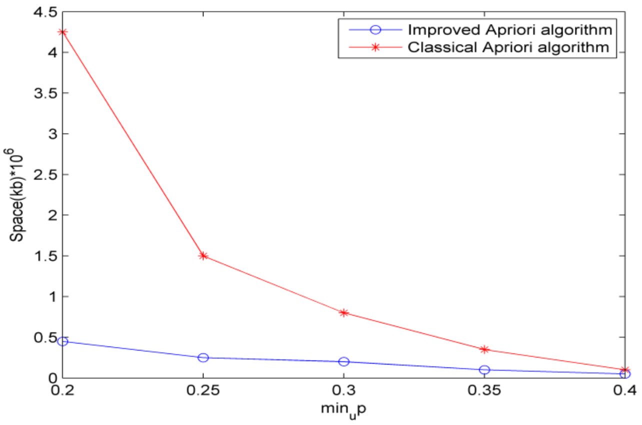


Fig 5 Memory consumption of algorithm

The abscissa represents the degree of support set by the algorithm, and the ordinate represents the memory consumption of the algorithm under a certain degree of support. The experimental results show that the improved Apriori algorithm reduces the memory consumption compared with the classical Apriori algorithm. In the case of low support, the change trend of memory consumption of the improved algorithm is gentle, while the change trend of the classical algorithm is larger. With the increase of support and the decrease of frequent itemsets, the difference of memory performance between the improved algorithm and the classical algorithm becomes smaller and smaller.

The experimental results show that the Apriori algorithm based on prefix itemsets and vertical Boolean matrix is effective and feasible. The improved algorithm has better time and space performance than classical algorithm. Moreover, as the support becomes smaller and smaller, the performance of the improved algorithm is more and more obvious.

The performance of the improved Apriori algorithm is compared with the classical Apriori algorithm in memory space under different support. The experimental results are shown in figures 3 and 4 below.

IV. CONCLUSION

In the process of association analysis, aiming at the shortcomings of existing association analysis methods, this paper presents the optimized Apriori algorithm, which reduces the memory space and improves the efficiency of the algorithm. Compared with different algorithms, the effectiveness and feasibility of the proposed method are verified. On this basis, 300 household electricity consumption data selected from the smart meter data are used as experimental data. The original data are pretreated by adding missing data, eliminating erroneous data and reducing dimension data with autoencoder algorithm. Then the original data is embedded by US-ELM and clustered by K-means algorithm. And they are clustered into 5 categories. The clustered and subdivided residential user data are analyzed by using the improved Apriori algorithm. Finally, the correlation information between different user types and 5 modes of power consumption is obtained.

REFERENCES

[1] Li Kangping, Wang Fei; Zhen Zhao, Mi Zengqiang, Sun Hongbin, Liu Chun, Wang Bo, Lu Jing. Analysis on residential electricity consumption behavior using improved K-means based on simulated annealing algorithm[A]. IEEE Power and Energy Conference at Illinois(PECI)[C] p6 pp., 2016

[2] Qian Cheng, Chen Min, Gao Ciwei, Li Huixing, Shen Tugang. Research on the analysis of user's electricity behavior and the application of demand response based on global energy interconnection[A]. China International Conference on Electricity Distribution (CICED)[C], p7 pp., 2016

[3] Jian Liu, Jiakui Zhao, Yan Chen, Hongfa Li, Qiucen Huang, Hong Ouyang, Qingli Hao, Yaozong Lu. Analysis of customers' electricity consumption behavior based on massive data[A]. 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)[C], p1433-8, 2016

[4] Humeau Samuel, Wijaya Tri Kurniawan, Vasirani Matteo, Aberer Karl. Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households[J]. Sustainable Internet and ICT for Sustainability, SustainIT 2013

[5] Minghao Piao, Meijing Li, Keun Ho Ryu. Using Significant Classification Rules to Analyze Korean Customers' Power Consumption Behavior: Incremental Tree Induction using Cascading-and-Sharing Method[A]. Proceedings of the 2010 IEEE 10th International Conference on Computer and Information Technology (CIT 2010)[C], p1649-53, 2010

[6] Bakker V, Bosman M G C, Molderink A, et al. Demand side load management Using a three step optimization methodology[C]. Smart Grid Communications (Smart Grid Comm), 2010 First IEEE International Conference on. IEEE, 2010: 431-436.

[7] 向保林. 浅析数据挖掘技术在电力市场中的运用[C]. 全国电力行业信息化年会论文集. 2004: 113-117

[8] 白晶. Apriori算法及其在智能小区用电分析中的应用研究[D].华北电力大学,2014.

[9] 林其友,陈星莺.数据挖掘技术在电价预测中的应用[J].电网技术,2006,30(23): 83-87

[10] 王佳,史雪飞.在电网中应用数据挖掘技术的探讨[J].商情,2011, (45): 68-68

[11] 刘耕砚.数据挖掘中Apriori算法改进及在电信BI上的应用[D].昆明理工大学,2008

[12] Elhillali Y, Tatkeu C, Deloof P, et al. Enhanced high data rate communication system using embedded cooperative radar for intelligent transports systems[J]. Transportation Research Part C: Emerging Technologies, 2010, 18(3): 429-439

[13] 刘玉文. 基于十字链表的Apriori算法的研究与改进[J]. 计算机应用与软件，2012,05：267-269

[14] Xuesong Wang, Yi Kong, Yuhu Cheng. Dimensionality Reduction for Hyperspectral Data Based on Sample-Dependent Repulsion Graph Regularized Auto-encoder[J]. Chinese Journal of Electronics ,2017,26(6):1233-1238

[15] Shizhi Chen, Xiaodong Yang,Yingli Tian. Discriminative Hierarchical K-Means Tree for Large-Scale Image Classification[J]. IEEE Transactions on Neural Networks and Learning Systems,2015,26(9):2200-2205