The Association Rules Algorithm Based on Clustering in Mining Research in Corn Yield

Bo Liu, Guifen Chen

The association rules algorithm based on clustering in mining research in corn yield

Liu Bo ,Chen Guifen(✉)

JiLin Agricultural University College of Iinformation Technology, Jilin Changchun 130118,China

Abstract: With the popularization of agricultural information technology, the use of data mining techniques to analyze the impact of different types of soil nutrient content and yield of corn has become a hot topic in the field of agriculture. Association rule mining is an important part of the field in Data mining , association rules can be found associated with agricultural data attributes. This article will use cluster analysis and association rule to analysis correlation between corn yield and soil nutrient. Firstly compare different clustering algorithm to chooses the optimal algorithm, make data collected in scientific classification, and based on expert knowledge of the collected data into different levels; then determine the type and content of different soil by association rules corn yield and soil nutrient; final inspection algorithm is correct. The results showed that: comparing K-means, hierarchical clustering analysis, and PAM, K-means algorithm to determine the optimal clustering; K value can be determined at selected intervals.K is equal to3, 4 or 6,clustering effect is good according to Sil value when K from 3 to 10 . based on the principle of association rules, clustering algorithm to select a K value associated with the combination of rule 6; After clustering algorithm of association rules, support and credibility and improve degree of accuracy is better than not clustering; by mining association rules after clustering, a great influence on the different levels of soil nutrients in corn yield. The results for the corn yield provides intelligent decision support data.

Key words: corn yield; cluster analysis; association rules; frequent itemsets; Apriori Algorithm

0 Introduction

Corn is the world's most productive acreage and grain crops, it has fast growth rate and high yield characteristics [1]. Moreover, with the development of animal husbandry and corn deep processing industry, corn has become the world's most important food crops, forage crops and cash crops [2]. 863 in support of national technology plan, the Shanghai, Beijing, Heilongjiang, Xinjiang, Jilin and other places to carry out tests to explore intelligent agriculture, the establishment of a number of precision agriculture experiment and demonstration area, and achieved gratifying results. China is a large agricultural country, China's grain output of corn yield very significant impact, where soil nutrients is one of the main factors affecting the yield of corn, so dig out the relationship between the different types and content of soil nutrients and corn yield is particularly important. But our data mining research later, there is no overall strength, compared

with developed countries there is still a big gap, has seriously hampered the development of intelligent agriculture in China.

With the emergence and speed of data collection significantly improved, we urgently need new technologies and tools to the vast amounts of data into information and knowledge available to us. Data mining [3] is a large number of known data search and analysis to discover hidden potential in the data relationship, in order to predict the future. In the last decade, data mining as decision support has been rapid development [4]. Data mining has an important direction include cluster analysis [5] and association rule mining [6], the paper first by comparing the different clustering methods to select the optimal clustering algorithm, combined with association rules algorithm, dig out different relationship between soil nutrients and corn yield type and content. Data mining is used in agriculture to promote agricultural production and direction sustained, high-yield and effective means of security is important to protect the interests of farmers and national food security.

1 Clustering and association rules

Clustering is a way to simplify data through data modeling, which uses similarity to different data into different classes. So in the same class has a great similarity, in the different classes has a larger dissimilarity . Association rules are used to find links between things. Firstly, the concept and characteristics of each clustering algorithm is analyzed, based on the value of Sil select the optimal clustering algorithm; and then by the clustering algorithm is divided into different clusters to compare and choose the most appropriate number of clusters. Finally through the association rules to determine different kinds and content of soil nutrients and corn production relations.

1.1 Select Clustering Algorithm

Different algorithms have different characteristics and adaptability. For example, K-means algorithm [7] When more dense clusters and cluster is obvious difference between clustering effect is good. But the noise and isolate sensitive data point. Specific steps are as follows: 1) Assign to each instance from its nearest cluster center to give K clusters; 2) were calculated for each cluster mean all instances, each of them as brand new cluster center. Repeat 1) and 2) until the position of the K cluster centers are fixed, assigned cluster is also fixed.

Hierarchical clustering algorithm [8] feature is not required prior to a given class number, the system can display the results in the clustering tree way, more suitable for data hierarchy, But to determine the distance matrix computation. Specific steps are as follows: 1) Each object is classified as a class, received a total of N classes, each class contains only one object. Distance between classes of objects they contain is the distance between. 2) find the closest two classes combined into one category, so the total number of class one less. 3) recalculate a new class clustering and all the old classes. 4) Repeat 2) and 3), until finally merged into a class so far.

PAM algorithm [9] is less sensitive to noise and is not affected by the order of the input data, which is insufficient to determine the high computational cost of clustering centers required for large data clustering process is slow. Specific steps are as follows: 1) choose K objects as the initial cluster centers. 2) In addition to the open cluster center point of the sample to calculate the

distance to each cluster center will classify the sample from the sample to the center of the nearest sample point. 3) and then calculated for each category, in addition to other classes outside the center of sample points and minimum distance to all other points, the minimum point as a new cluster centers. 4) Repeat 3) until the position of the two cluster centers unchanged.

Many data by understanding the structure and background information, you can know which algorithms are relatively good, but in many cases, our understanding of the data is not much, so we have to choose an objective criteria to evaluate [10]. Between the use of class compactness evaluated within the class separation and clustering is a common approach, one of the most classic is Silhouette indicators. Silhouette indicators both for the number of clusters optimal estimation can also be applied to evaluate the quality of clustering [11]. Therefore, this paper Silhouette indicators as the number of selection and clustering clustering algorithm to determine the objective evaluation criteria.

Having provided a sample data set is divided into clusters, the clusters in the sample and the average of all other samples of dissimilarity or distance for the sample to an average of all samples of another class or dissimilarity distance, wherein, and. Thus, the index is calculated Silhouette samples are as follows:

$$\mathrm{Si}l(t)=\frac{[b(t)-a(t)]}{\max\{a(t),b(t)\}} \tag{1}$$

All the samples in a cluster compactness of the average value of sil said tightness and separability; Sil average value of all samples of a data set may reflect the quality of clustering results , the greater the Sil value represents the better the quality of clustering.

The main steps of Silhouette clustering algorithm selection method based on the validity of indicators designed as follows: 1) setting a given data set, candidate cluster algorithm; 2) to specify the output number of classes K, respectively for each candidate cluster algorithm for data collection clustering; 3) calculating an average value of clustering results Sil each candidate clustering algorithm; 4) comparing the average value Sil, Sil value corresponding to the maximum average selected candidate clustering algorithm for the optimal algorithm.

In this paper, the national "863" plan "Maize Precise Operation System and Application" project demonstration base - Jilin Nong'an some experimental data from 2005 to 2010, and use MATLAB R2014a to cluster analysis. When K is equal to 3, the average Sil K-means algorithm is the highest; when K is equal to 4, the average value of K-means algorithm Sil also the highest. K-means algorithm proved optimal clustering algorithm, the results shown in Figure 1.

	K-means	Hierarchical	PAM
K=3	0.7605	0.7271	0.6027
K=4	0.729	0.6687	0.5772

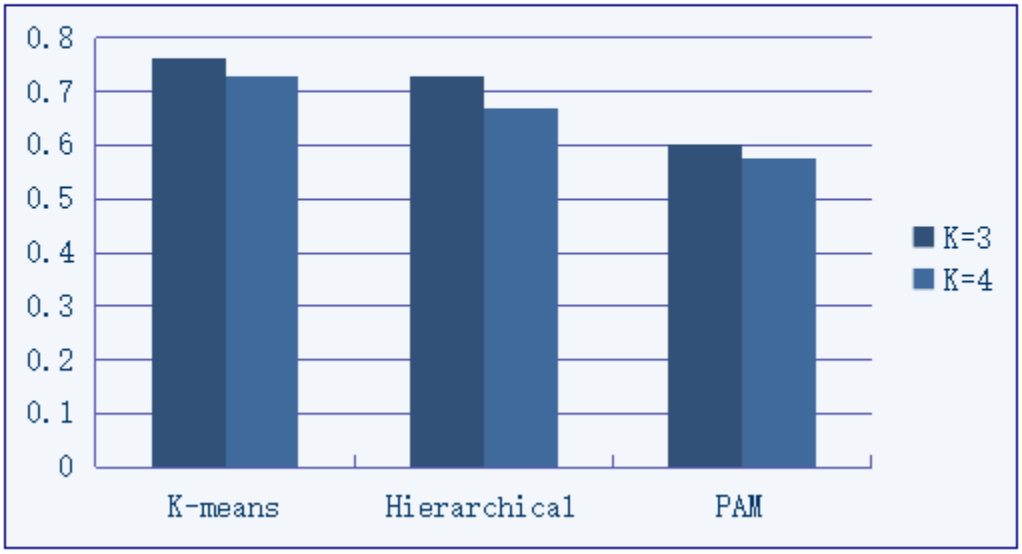


Figure 1 each algorithm of Sil value comparison

1.2 Clustering algorithm K value selection

Different values of K greater impact on clustering results, so traverse K 3-10 is the case, the corresponding value of Sil. as shown in picture 2. when K is equal to 3,4,6 clustering is better. According to the association rules algorithm is the core elements of frequent itemsets, therefore this article selects the K value is 6 k-means algorithm and its combined association rules algorithm.

K	3	4	5	6	7	8	9	10
Si1	0.7605	0.729	0.1123	0.7407	0.5569	0.2224	-0.3486	0.1073

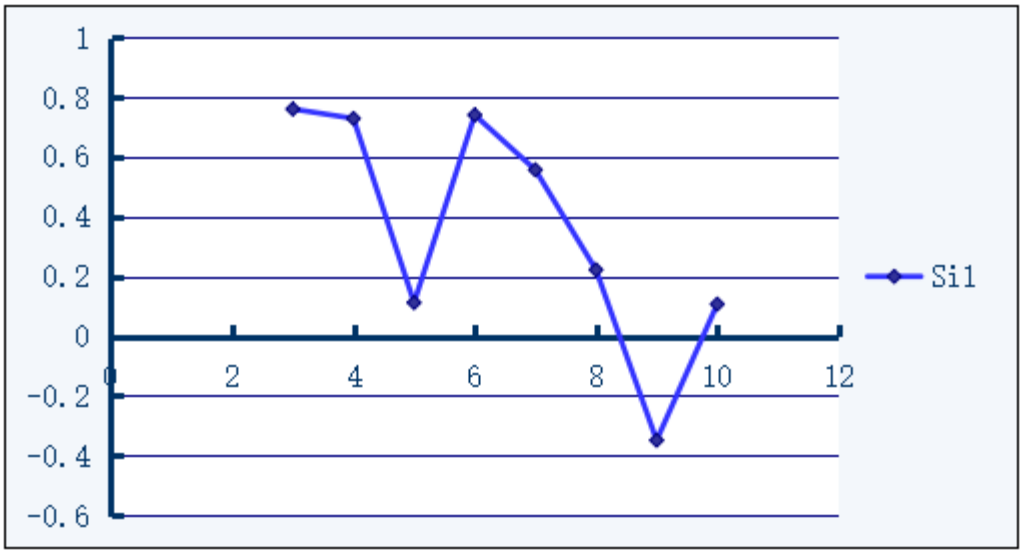


Figure 2 different K value corresponding to the Sil value comparison

1.3 Association Rules Algorithm

1993, R.Agrawal, who designed a program called Apriori algorithm [12], the algorithm is the most influential algorithm, it laid the foundation for the Association Rules algorithm. This algorithm has two steps: The first step is mining frequent item sets. That support is greater than those specified by the user support selected projects, as frequent item set; the second step is based on frequent item sets to generate strong association rules. That association rules support and confidence are greater than or equal to the user specified support and confidence [13].

Support abbreviated as sup, Refers to a rule before or after a corresponding number of support as a percentage of total number of records. Formula is as follows:

$$\sup(A => B) = \sup(A \cup B) = \frac{\left|\left\{TR \middle| TR \supseteq A \cup B\right\}\right|}{|n|} \quad (2)$$

Confidence abbreviated as conf, Said the DB contains the percentage of things at the same time also contains B. Formula is as follows:

$$\mathrm{conf}(\mathrm{A} => B) = \frac{\sup(A \cup B)}{\sup(A)} \quad (3)$$

Lift, it is the credibility of the ratio of the target density of the overall goal. So when the lift is greater than 1, the rules are better able to predict the results, rather than based on how often the data items to guess whether the result will be. Formula is as follows:

$$lift(\mathrm{A} => B) = \frac{conf(A => B)}{\sup(B)} = \frac{p(B|A)}{p(B)} \quad (4)$$

2 Case analysis

Based on the national "863" plan "corn precision operating systems research and application" project demonstration base - county agriculture of jilin province from 2005 to 2010, part of experimental data, experiment with clustering analysis and association rules method, validation and analysis of data mining method to the correlation between corn yield and soil nutrient.

2.1 Cluster analysis

In order to select frequent item sets, the first use of K-means algorithm on data collected scientific classification, data mining software Weka 3.7 data divided into six categories. And select the item appears set frequency greater than or equal to 15% based on expert recommendations [14]. The third term due to the frequency appears set to 14%, the frequency of the sixth item appears set to 14%, it is discarded. The other four groups composed of frequent item sets. As shown in Figure 3.

	one	two	three	four	five	six
Quantity and ratio	18(16%)	28(25%)	16(14%)	17(15%)	17(15%)	16(14%)
N	134.29	153.2439	164.2363	140.9376	155.9112	130.6275
P	11.0711	25.6339	11.8875	11.08	12.1053	11.8338
K	93.8889	94.3571	95.625	99.4118	113.5294	115.375
Yield	8955.5389	10331.5393	9408.8187	7611.8882	7721.8588	9834.2437

	one	two	three	four
Quantity and ratio	18(22.5%)	28(35%)	17(21.25%)	17(21.25%)
N	134.29	153.2439	140.9376	155.9112
P	11.0711	25.6339	11.08	12.1053
K	93.8889	94.3571	99.4118	113.5294
Yield	8955.5389	10331.5393	7611.8882	7721.8588

Figure 3 clustering results and screening

2.2 Data into Level

Based on the experience of experts of different data into different levels [15]. The N, P and K from A to F is divided into six different levels, the yield from A to C is divided into three different levels. And based on the knowledge of experts just four frequent itemsets obtained by clustering into different levels, as shown in FIG.

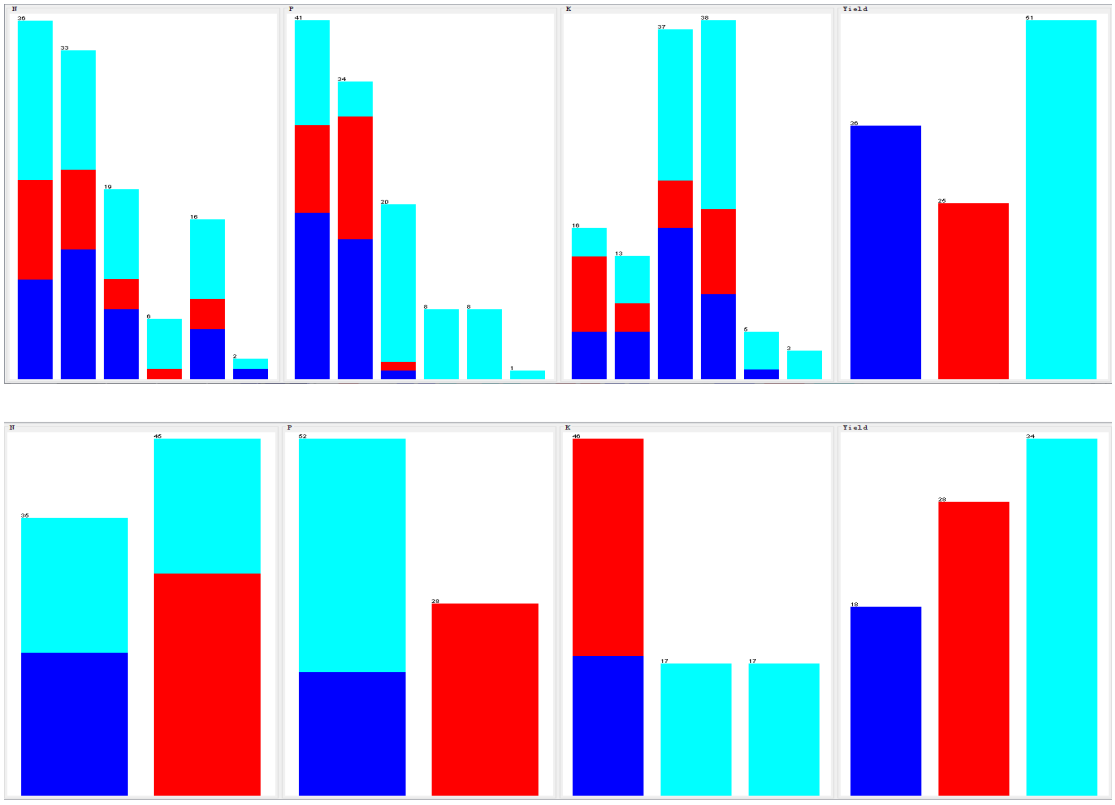
	N	P	K	Yield
A	175-190	35-40	125-130	9526-10987
B	160-175	29-35	116-125	8066-9526
C	145-160	23-29	107-116	6606-8066
D	130-145	17-23	98-107	
E	115-130	11-17	89-98	
F	100-115	5-11	80-89	

	one	two	three	four
Quantity and ratio	18(22.5%)	28(35%)	17(21.25%)	17(21.25%)
N	D	C	D	C
P	E	C	E	E
K	E	E	D	C
Yield	B	A	C	C

Figure 4 data into level

2.3 Relationship between the various soil nutrients and Yield

Respectively to observe the alkali solution N, P and K impact on production [16]. The results in FIG. 5 shown in FIG.



	Not clustering	After clustering
N and yield relations	When N grade D, production with 39% yield for B level	When N grade D, production with 51% yield for B level
	When N grade C, production with 44% yield for A level	When N grade C, production with 62% yield for A level
P and yield relations	When P grade D, production with 90% yield for A level	When P grade C, production with 100% yield for A level
	When P grade D, production with 41% yield for C level	When P grade E, production with 62% yield for C level
K and yield relations	When K grade E, production with 42% yield for B level	When K grade E, production with 39% yield for B level
	When K grade D, production with 53% yield for A level	When K grade E, production with 61% yield for A level

Figure 5 is not clustering and clustering after the relationship between the soil nutrient and yield

2.4 Data Association

To ignore an abnormal event and a small probability event. Selection criteria are: support between 15-100 percent, and lift the value of more than 1.0 and lift the top five values and yield-related association rules, as shown in Figure 6.

Not clustering	P=D 20 ==> Yield=A 18 conf:(0.9) < lift:(1.98)> lev:(0.08) [8] conv:(3.63)
	P=E 34 ==> Yield=C 14 conf:(0.41) < lift:(1.84)> lev:(0.06) [6] conv:(1.26)
	K=E 37 ==> Yield=B 16 conf:(0.43) < lift:(1.35)> lev:(0.04) [4] conv:(1.14)
	N=D 33 ==> Yield=B 13 conf:(0.39) < lift:(1.23)> lev:(0.02) [2] conv:(1.07)
	K=D 38 ==> Yield=A 20 conf:(0.53) < lift:(1.16)> lev:(0.02) [2] conv:(1.09)
After clustering	N=D P=E K=E 18 ==> Yield=B 18 conf:(1) < lift:(4.44)> lev:(0.17) [13] conv:(13.95)
	N=C P=C K=E 28 ==> Yield=A 28 conf:(1) < lift:(2.86)> lev:(0.23) [18] conv:(18.2)
	N=D P=E K=D 17 ==> Yield=C 17 conf:(1) < lift:(2.35)> lev:(0.12) [9] conv:(9.78)
	N=C P=E K=C 17 ==> Yield=C 17 conf:(1) < lift:(2.35)> lev:(0.12) [9] conv:(9.78)
	N=D P=E 35 ==> Yield=B 18 conf:(0.51) < lift:(2.29)> lev:(0.13) [10] conv:(1.51)

Not clustering	After clustering
When P grade D, production with 90% yield for A level	When N grade D,P grade E and K grade E,production with 100% yield for B level
When P grade D, production with 41% yield for C level	When N grade C,P grade C and K grade E,production with 100% yield for A level
When K grade E, production with 42% yield for B level	When N grade D,P grade E and K grade D,production with 100% yield for C level
When N grade D, production with 39% yield for B level	When N grade C,P grade E and K grade C,production with 100% yield for C level
When K grade D, production with 53% yield for A level	When N grade D,P grade E,production with 51% yield for B level

Figure 6 after clustering and clustering of relative contrast

2.5 Examine the validity of

To test the accuracy of the algorithm, based on J48 algorithm [17] 10-fold cross-validation [18]. The first data set is divided, which will in turn nine as training data, one as the test data, test. Each test will draw the appropriate accuracy. Then the average of 10 times as a result of the correct rate of accuracy of the method of estimation. As shown in Figure 7.

	Not clustering	After clustering
correct	63%	100%
Error	37%	0

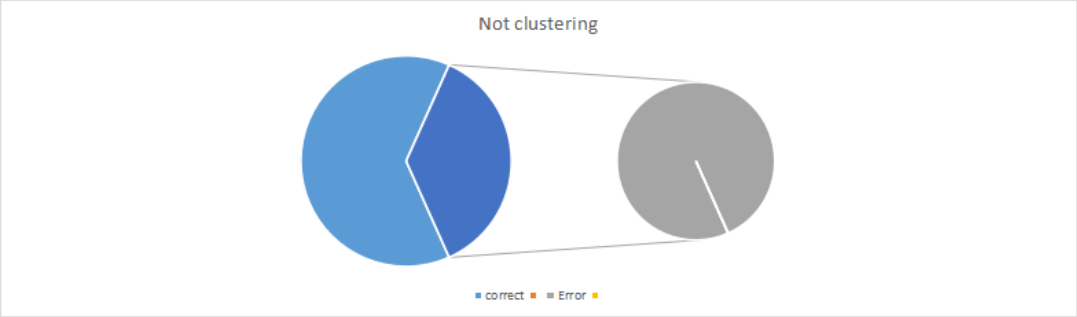


Figure 7 not clustering and clustering accuracy after contrast

2.6 when yield is equal to A, the relationship between the various soil nutrients

A full-time to the Yield. Relationship between N, P and K between the three shown in Figure 8.

Not clustering	N=D 12 ==> K=E 6 conf:(0.5) < lift:(1.59)> lev:(0.04) [2] conv:(1.18)
	K=E 16 ==> N=D 6 conf:(0.38) < lift:(1.59)> lev:(0.04) [2] conv:(1.11)
	P=C 8 ==> K=D 5 conf:(0.63) < lift:(1.59)> lev:(0.04) [1] conv:(1.22)
	K=D 20 ==> P=C 5 conf:(0.25) < lift:(1.59)> lev:(0.04) [1] conv:(1.05)
	N=C 16 ==> K=D 8 conf:(0.5) < lift:(1.27)> lev:(0.03) [1] conv:(1.08)
	K=D 20 ==> N=C 8 conf:(0.4) < lift:(1.28)> lev:(0.03) [1] conv:(1.06)
	N=C 16 ==> K=E 6 conf:(0.38) < lift:(1.2)> lev:(0.02) [0] conv:(1)
	K=E 16 ==> N=C 6 conf:(0.38) < lift:(1.2)> lev:(0.02) [0] conv:(1)
	P=D 18 ==> K=E 6 conf:(0.33) < lift:(1.06)> lev:(0.01) [0] conv:(0.95)
	K=E 16 ==> P=D 6 conf:(0.38) < lift:(1.06)> lev:(0.01) [0] conv:(0.94)
After clustering	N=C 28 ==> P=C K=E 28 conf:(1) < lift:(1)> lev:(0) [0] conv:(0)
	P=C 28 ==> N=C K=E 28 conf:(1) < lift:(1)> lev:(0) [0] conv:(0)
	K=E 28 ==> N=C P=C 28 conf:(1) < lift:(1)> lev:(0) [0] conv:(0)

Figure 8 production for A while after not clustering and clustering of relations between the soil nutrient

3 Conclusion and Analysis

By the above operation, we can see:

1. Choose a good clustering algorithm and K values of great influence on the clustering results.

According to Comparative K-means, hierarchical clustering analysis, and PAM, K-means algorithm to determine the optimal clustering algorithm shown in Figure 1.
2. the value may be determined based on Sil K value chosen in the range from 3 to 10, when K is equal to 3,4,6 there will be a good clustering results, as shown in Fig.2
3. After clustering accuracy of association rules algorithm is better than not clustering accuracy of association rules algorithm, show that using the association rules before clustering analysis of data processing is very necessary.. As shown in Figure 7. By the respective association rules diagram can be drawn: clustering support after the credibility and lift are better than not clustering, association rules show before using cluster analysis of data processing when the actual decision is significant .
4. the impact of different levels of soil nutrients for corn production is very high. When soil nutrients N content of C, P content of C, K content of E grade, grade A larger yield ratio.

4 Discussion and Outlook

In recent years, as we get the data and access to information ability increase [19], expanding our database [20]. A growing number of agricultural production, scientific research, business management to use the database, and this trend will continue to rise. In this era of information explosion, information overload is everyone has to face the problem. How can you in a sea of information, find useful, hidden knowledge, as our top priority. Only make data resources, mining the data potential and useful knowledge, to make data have the effect of support decision making and prediction. Otherwise, a large amount of data it may become a waste, even burden [21]. Therefore, data mining is a powerful guidance and direction to the development of the future, is a hot research field in the world today, its research has broad application prospects and great practical significance.

Acknowledgment

Funds for this research was provided by National "863" project "corn precision operating systems research and demonstration" (2006AA10A309), National spark plan "corn" digital technology integration and demonstration(2008GA661003), Jilin provincial departments of world bank loan project(2011-Z20), Jilin province rural project in 2015(Study and demonstration of corn precision operation system based on Internet of things)

References

1. Anthony G. O'Donnell,Melanie Seasman,Andrew Macrae,Ian Waite,John T. Davies. Plants and fertilisers as drivers of change in microbial community structure and function in soils[J]. Plant and Soil . 2001 (1-2)
3. Lichen Hua, Guzhong Jun, Tang Lisong etc. Different Fertilization Models oasis farmland soil microbial community abundance and Enzyme Activity [J]. Journal of Soil. 2012 (03)
4. Long Chu now, Li Xiangjun. Summary of data mining [J]. Neijiang and Technology. 2006 (02)
5. Zhang Chengzhi present situation and the latest developments on Data Mining [J]. Nanjing Industry and Technology College. 2003 (02)

6. Huang Xiang, Cai Bi Ye, Meng Ying A clustering algorithm of PSO & PAM [J]. Journal of Computer Engineering and Applications. 2013 (04)
7. Guimei Liu,Hongjun Lu,Wenwu Lou,Yabo Xu,Jeffrey Xu Yu. Efficient Mining of Frequent Patterns Using Ascending Frequency Ordered Prefix-Tree[J]. Data Mining and Knowledge Discovery . 2004 (2)
8. An Aifen. Improved k-means initial clustering center selection method [J]. Shanxi Normal University (Natural Science). 2013 (01)
9. Xianghong Li, Wang Xiaohan, Luo Shuyun study courses relational structure [J] based on hierarchical clustering methods. China Education Innovation Herald. 2011 (26)
10. Zhang Zhao, Wang Suozhu, Zhang Yu An SOM and PAM clustering algorithm [J]. Journal of Computer Applications. 2007 (06)
11. any Harumi hole Lei. Comparative data mining clustering technology research methods [J] Analysis of Science and Technology Information (Academic Research). 2008 (24)
12. Xiejuan Ying, Ma Qing, Xie Weixin. A new algorithm for the optimal number of clusters [J] OK. Shaanxi Normal University (Natural Science) 2012 (01)
13. Cui, Li Qiang, the Kingdom division An association rule mining for large transaction database algorithm [J]. Journal of Air Force Radar Academy. 2011 (03)
14. Zhu Xiaoyu, Wang Li Dong, Wang Guangyang An improved Apriori association rule mining algorithm [J]. Computer Technology and Development. 2006 (12)
15. Liugui Xia, Cui Duo, Cao Bang and research on data mining [J]. Industrial Technology & Economy. 2000 (03)
16. Li Zhaojun, Yang Jiajia, Fan Feifei etc. under Different Fertilization film Maize dry matter accumulation and P uptake of [J]. Plant Nutrition and Fertilizer Science. 2011 (03)
17. Duga silver, Ru US, NI Wu Zhong. Save N Phosphorus stability control potassium fertilization on yield and nutrients accumulation Rice [J]. Plant Nutrition and Fertilizer Science. 2013 (03)
18. Wangshu Yan, Geng Guohua, Lee Byung Chun decision tree algorithm for medical image data mining [J]. Journal of Northwest University (Natural Science) 2005 (03)
19. Zhang Yingtang, Ma Chao, Li Zhining etc. Based on a fast cross-validation leave nuclear ELM online modeling [J]. Shanghai Jiaotong University in 2014 (05)
20. Bai Hongwei, Ma Zhiwei, Song Yaqi. Monitoring data processing based on the state of the smart grid cloud [J]. East China Electric Power. 2011 (09)
21. advanced search, Chen Hung Evaluation of Information Resources Quality [J]. Chinese Journal of Library Science. 2010 (02)
22. Guo-Jie Li, Cheng Flags Great science research data: a major strategic areas of technology and the future economic and social development - and scientific thinking [J] Status of Large Data ACADEMY OF 2012 (06).