Analysis and Evaluation of Soil Fertility Status Based on Weighted K-means Clustering Algorithm

Guifen Chen, Lixia Cai, Hang Chen, Liying Cao, Chunan Li

Analysis and evaluation of soil fertility status Based on weighted K-means clustering algorithm

Guifen Chen[1,a],Lixia Cai[1,b],Hang Chen[1,2,c],Liying Cao[1],Chunan Li[1]

[1]Jilin Agricultural University, Changchun 130118,China;[2]Jilin Academy of Agricultural Sciences, Changchun 130124,China

[a]guifchen@163.com,[b]lixcai@163.com,[c]chenhang@163.com

Abstract. Generally K-means clustering algorithm can not distinguish the imbalance between attributes, so it can only be an independent investigation situation of each attribute but can not be comprehensive analysis of the soil fertility status. To solve this problem, this paper proposes a weighted K-means clustering algorithm to evaluate the soil fertility in Nong'an County, Jilin. The algorithm uses AHP to get the weight of soil nutrient attributes. Then combined with K-means clustering algorithm. Finally through the operational efficiency and accuracy to determine the optimal classification, that can improve the clustering algorithm of intelligent. The algorithm and the traditional K-means clustering algorithm are used in the comparison, tests showed that the weighted K-means clustering algorithm has a better accuracy, operational efficiency, significantly higher than the unweighted clustering algorithm; Comprehensive evaluation of the changes in soil nutrients after precision fertilization that used algorithm. The soil fertility status has a significantly improvement after years of continuous precision fertilizing. The results show that the improved clustering algorithm is a good method to comprehensive evaluation of soil fertility.

Keywords: AHP; Weighted K-means clustering; Optimal classification; Soil fertility evaluation

1 Introduction

3S technology (GPS, GIS and RS), networking technology and expert system (ES) technology are widely used in precision agriculture with the rapid development of information technology. That all make soil fertility data appear to rich,

multidimensional, dynamic, incomplete, uncertainty and other characteristics [1].How to be more timely and accurately show the differences in temporal and spatial data, comprehensive evaluation and correct analysis of the data have an important practical significance [11]. Data mining technology [5] is the process of generating new regular, which through the massive amounts of data classification, extraction to discover the mutual contact between data.

Related data show that variable region and the traditional region of N, P, K nutrient variation coefficients in soil fertilization were compared by ZHANG. Which indicate that variable rate fertilization has a balanced effect to soil nutrient fertilization [2]. The research on weighted space fuzzy dynamic clustering algorithm by CHEN Gui-fen, proved the effectiveness of soil fertility evaluation [3]. And Li Yan.etc. [4] Who used fuzzy clustering method to classify partition and introduced two kinds of partitions to compare and evaluate, such as fuzzy clustering index and normalized classification. That can offer the decision basis for soil management. Even K-means clustering algorithm based on the classification method could differentiate soil fertility according to soil nutrient, However, it can't consider the nutrient differences between each attribute. As a result, we use the improved K-means algorithm, considering the linkages between soil nutrients of the fertility in Nong'an country [8] and give a comprehensive evaluation.

2 Analysis of k-means algorithm

K-Means algorithm is a clustering algorithm based on partitioning method [7-17], it is first suggested and one of the more classical clustering algorithms [14-15].

2.1 The process of K-means algorithm [6]

Algorithms: k-means. Divided k-means algorithm based on the average value of the objects in the cluster.

Input: the number of clusters (k) and the database contains n objects.

Output: k clusters, so that the minimum squared error criterion.

Method:

(1) Choose k objects as the initial cluster centers;

(2) Repeat;

(3)According to the average value of the objects in the cluster, each object is (re) assigned to the most similar clusters;

(4)Update the average value of cluster. That is to say, calculate the average value of each cluster in the object;

(5) Until no change.

2.2 Advantages and disadvantages of k-means algorithm

Using k-means algorithm [12] to clustering, the effect is good. While the result is a dense cluster, the differences between clusters are obvious. When we deal with large data sets, this algorithm is relatively scalable and efficient, because of its complexity is O (nkt), where, n is the number of all objects, k is the number of clusters, t is the number of iterations. Typically k«n and t«n. This algorithm often ends with a local optimum. However, k-means method [13-18] is only used in the case of the average value of clusters were defined. This attribute data is not applicable for processing symbol attribute data, it also requires user to give the value of k (the number of clusters to be generated) in advance. In addition, for the "noise" and outlier data is sensitive, a small amount of such data can have a significant impact on the average value.

3 Analysis of weighted k-means algorithm Based on AHP

3.1 Using AHP to determine the weight coefficients

The algorithm is as follows:

Step 1: Construct paired comparison matrix;

Step 2: Take any n-dimensional normalized initial vector $\mathbf{w}^{(0)}$;

Step 3: Calculation $\tilde{\mathbf{w}}^{(k+1)} = \mathbf{A}\mathbf{w}^{(k)}, k = 1,2,\cdots$;

Step 4: Normalization $\tilde{\mathbf{w}}^{(k+1)}$;

Step 5: For the pre-specified precision ε, when $\left| w_i^{(k+1)} - w_i^{(k)} \right| < \varepsilon, \quad i = 1,2,\cdots,n$

Established, $\tilde{\mathbf{w}}^{(k+1)}$ shall be eigenvector; otherwise return Step 2;

Step 6: Calculate the maximum eigenvalue $\lambda = \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{w}_i^{(k+1)}}{w_i^{(k)}}$;

Step 7: Calculate consistency index $CI = \frac{\lambda - n}{n-1}$;

Step 8: Calculate consistency ratio $CR = \frac{CI}{RI}$;

Step 9: If $CR < 0.1$ is established, through consistency test; otherwise reconstruct paired comparison matrix;

Step10: If all the layers are calculated. And we can obtain the weight vector of total target ,

$A = (a_1, a_2, \cdots, a_m)$; Otherwise, return back to Step 1.

3.2 The establishment of the weighted k-means model

In this paper, we used the weighted fuzzy dynamic clustering approach to process spatial data, which is proposed by CHEN [3].

(1) Data's standardization

Since in practical problems, different data generally have different dimensions, in order to have the amount of different dimensions can be compared, we need to standardized data, which data are compressed to the [0, 1] interval. Now we use the range transformation,

$$x'_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{xi_j\} - \min\{x_{ij}\}} \tag{1}$$

(2) Weighted calculation

$$Y = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1m} \\ x'_{21} & x'_{22} & \cdots & x'_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x'_{n1} & x'_{n2} & \cdots & x'_{nm} \end{bmatrix} \cdot \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & a_m \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix} \tag{2}$$

(3) Fuzzy similar matrix

Calculating Closeness r_{ij} of fuzzy sets i and fuzzy sets j ,

$$r_{ij} = \frac{\sum_{k=1}^{m} y_{ik} \cdot y_{jk}}{\sqrt{\sum_{k=1}^{m} y_{ik}^2} \cdot \sqrt{\sum_{k=1}^{m} y_{jk}^2}} \quad (3)$$

Resulting in fuzzy similar matrix $R = (r_{ij})_{n \times n}$.

4 The application of weighted k-means algorithm

4.1 Data Sources

Application and research of soil fertility data after precision fertilization for many years [9-10], which is based on the "863" plan --"research and application of corn precise operating system" project demonstration base in Nong'an County, Jilin. And we select the representative soil nutrient data to integrated analysis, such as, alkaline hydrolysis nitrogen, available potassium and available phosphorus. From Nong'an County during 2007 to 2011 years.

GIS-based sampling points shown in Figure 1:

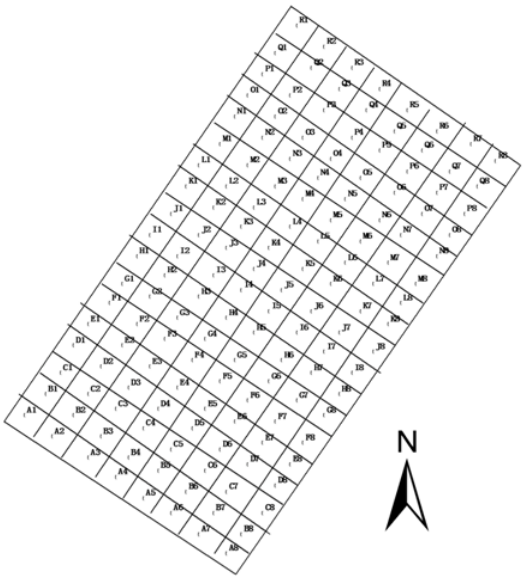


Fig. 1. GIS-based sampling points

The sampling method from figure1 is “five plum blossom sampling”, namely taking soils from the four corners and the center, then blending these soils together as the sample. The samples have been taken and tested in Tab.1, from 2007 to 2011; through tests we have acquired the data of soil, such as spatial coordinates, organic matter, alkaline hydrolysis nitrogen, available potassium, available phosphorus, soil humidity and PH value. Select the main factors which affecting fertility as the sample data, part of the data shown in Table 1.

Table 1. Part of the sample data

Town name	Alkaline hydrolysis nitrogen (mg/kg)	Available phosphorus (mg/kg)	Available potassium (mg/kg)	latitude	longitude
Nong’an town	154.0	28.0	208.0	44.58417	125.2898
Nong’an town	136.0	31.3	217.0	44.49895	125.2512
Nong’an town	132.0	16.3	198.0	44.49926	125.2507
Nong’an town	125.0	36.8	140.0	44.51392	125.2540

4.2 Application of algorithm

First, Standardization of soil nutrient data (the data come from Nong'an town during 2007 to 2011). Then we can analysis spatial patterns of soil nutrients data according to the test area of N, P and K. The results show that the test area available phosphorus in the soil spatial variability of the maximum. It is shown in Figure 2:

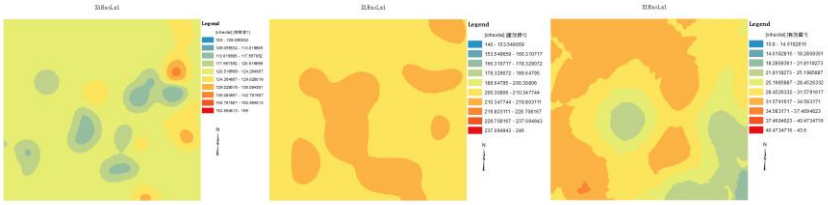


Fig. 2. Soil nutrient (N, P, K) spatial variation in Xi haolai, Nong'an town

After that, Evaluation of the results based on the status of spatial variability and local soil characteristics, then construct pair wise comparison matrix B (Equation 2).

Second, the author use AHP to get the nutrient weights of soil alkaline hydrolysis nitrogen, available potassium, available phosphorus, three nutrient weights are 0.3782, 0.2032 and 0.4185. CI = 0.026420513 << 0.1, which is closer to the complete consistency. Then, compared with the classical k-mean clustering algorithm, the weighted K-means clustering algorithm has a significantly higher accuracy, and operational efficiency. We can see the results have shown in Table 2.

Table 2. The comparison result

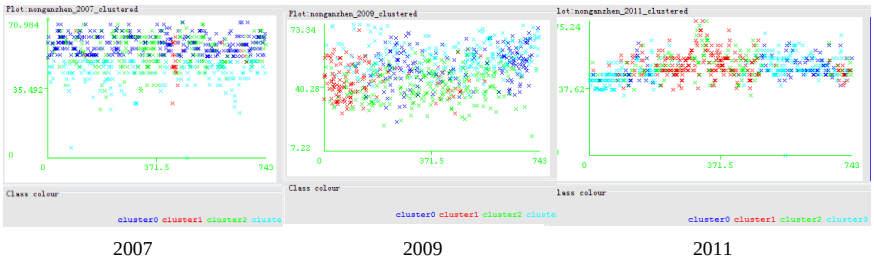
Algorithm	Average accuracy (%)	Average running time(s)
K-means	95.03	0.08
K-Wmeans	96.91	0.06

From table 2, the weighted and unweighted k-means both are better classification methods (the accuracy are 96.91%, 95.03% respectively, and the running times are 0.06 s, 0.08 s respectively). When we use unweighted clustering, different nutrients in the soil will offset the gap between the highest and lowest and delimit in the same class, while weighting the gap will be assigned to different classes, this method can reflect the real situation of soil nutrient. Weighted k-means for the "noise" and outlier data is not very sensitive; a small amount of this kind of data does not have great influence on the average value.

Finally, using the weighted k-means algorithm, weighted clustering the soil nutrient data of Nong'an town for five consecutive years from 2007 to 2011. Experimental results as shown in Table 3 and Figure 3.

Table 3. The result of soil nutrient data weighted clustering

	2007		2008		2009		2010		2011	
	Clustered Data	Percentage (%)	Clustered Data	Percentage (%)	Clusterd Data	Percentage (%)	Clustered Data	Percentage (%)	Clustered Data	Percentage (%)
Cluster 0	310	42	247	33	161	22	198	27	99	13
Cluster 1	26	3	257	35	126	17	254	34	235	32
Cluster 2	225	30	194	26	234	31	151	20	134	18
Cluster 3	183	25	46	6	223	30	141	19	276	37



2007 2009 2011

Fig. 3. Clustering results

The Table 3 and Figure 3 show that, in the same category case, the degree of similarity between the data is gathered and the differences between clusters are decreasing year by year after a continuous precise fertilization. The value of Cluster 0 decreased from 42% in 2007 to 13% in 2011, cluster 1 from 3% in 2007 years rose to 35% in 2010.All above shows that the weighted k-means algorithm is an effective method for soil fertility evaluation. After a continuous precise fertilization, alkaline hydrolysis nitrogen,available phosphorus and available potassium, the comprehensive similarity of three kinds of nutrient data have been improved year by year. The results

tally with the actual situation, so weighted k-means algorithm is an effective method of fertility evaluation.

5 Results and Discussion

Analysis and evaluation of soil nutrient data by using weighted k-means algorithm. The data of Nong'an town for five consecutive years from 2007 to 2011. We can see that significant changes in soil fertility occur after five consecutive years of precision fertilization. Experimental results show that the weighted k-means algorithm is an effective method of fertility evaluation.

(1)AHP is used to determine the initial weight values; the weighted original decentralized data can avoid the shortcomings that unweighted k-means algorithm does not distinguish between data imbalance between attributes as well as sensitive to "noise" and isolated points data .

(2)The use of weighted and unweighted k-means algorithm for comparative analysis soil nutrient data about soil alkaline hydrolysis N, available P and available K from Nong'an town in 2011, and the results showed that weighted K-means clustering algorithm has better effect than unweighted k-means algorithm, in the terms of accuracy which is increase1.88% and operating efficiency which is increase 25%.

(3)From the experimental results of the algorithm, after five consecutive years of precision fertilization on soil nutrient data in Nong'an town, Comprehensive similarity in increase year by year.This evaluation results and the actual situation is consistent, provides a new reference for analysis of soil fertility status in future.

The initialization of weighted and unweighted k-means clustering algorithm should depend on iterative method to determine the number of clusters that is relatively close to the true value and the initial center. However, this article only analysis verification of soil nutrient data that after five consecutive years of precision fertilization on soil nutrient data in Nong'an town. And the improved algorithm has not tested the application of large data sets or fully confirmed that the validity of new algorithm for massive data sets. How to simplify the clustering algorithm and combine with soil nutrient data of many years, more townships (towns) and soil types. They are all problems. So these parts still need further research.

Acknowledgment

This work was supported by the national "863" project (2006AA10A309), National Spark Plan (2008GA661003) and Shi Hang of Jilin province projects (2011-Z20).

References

1.Turner BL,Meyer WB.Land use and land cover in global environmental change:considerations for study [J]. Int.SoiSci, J., 1991, 130:669-680.

2.Zhao Jiewen, Hu Huaiping, Zhou Xiaobo. Application of Support Vector Machine to apple classification with near—infrared spectroscopy[J]. Transactions of the CSAE, 2007, 23(4):149-152.

3.CHEN Gui-fen, CAO Li-ying,, WANG Guo-wei.Application of Weighted Spatially Fuzzy Dynamic Clustering Algorithm in Evaluation of Soil Fertility[J]. Scientia Agricultura Sinica, 2009, 42(10): 3559-3563.

4.Li Y, Shi Z, Wu C F, Li F, Cheng J L. Definition of management zones based on fuzzy clustering analysis in coastal saline land[J].Scientia Agricultura Sinica,2007,40(1):114-122.

5.Amirmahdi,Neda Rajabpour,Ali Naserasadi.A Survey on Data Mining Approaches[J]. International Journal of Computer Applications (0975–8887).2011, 36(6):14-18.

6.Sun Shibao Qin Keyun. Improved k - average clustering algorithm research [J]. Computer engineering, 2007, 33(13):200-201,209.

7.SUN Ji-Gui,LIU Jie,ZHAO Lian-Yu.Clustering Algorithms Research[J]. Journal of Software, Vol.19, No.1, January 2008:48 − 61.

8.Yang C, Everitt J H, Bradford J M. Comparisons of uniform and variable rate nitrogen and phosphorus fertilizer applications for grain sorghum[J]. Transactions of the American Society of Agricultural Engineers, 2001, 44(2): 201-209.

9.Umeda M, Kaho T, Michihisa IIDA, Choung Keun LEE. Effect of variable rate fertilizing for paddy field. 2001 ASAE annual international meeting, 2001, Paper Number. 01(Part. II).

10.Wittry D J, Mallarino A P. Comparison of uniform-and variable-rate phosphorus fertilization for corn-soybean rotations [J]. Agronomy Journal, 2004, 96(1): 26-33.

11.Lianghou Li,Jiyue Li.Application of Clustering Analysis in Classifying Site Type and Evaluating Soil Fertility.2010 Third International Conference on Education Technology and Training (ETT 2010) ,2010:468-471.

12.Chawan,Saurabh R Bhonde, Shirish Patil.Improvement of K-Means clustering Algorithm. International Journal of Engineering Research and Applications (IJERA) ISSN. 2012,2(2) :1378-1382.

13.Lai Yuxia, sd, Yang Guoxing. K-means clustering analysis based on genetic algorithm [J]. Computer engineering, 2008, 34(20):200-202.

14.BACKER E, JA IN A K. A clustering performance measure based on fuzzy set decomposition [J]. IEEE Trans on Pattern Analysis and Ma chine Intelligence, 1981, 3 (1): 66277.

15.ZAHN C T. Graph2theoreticalmethods for detecting and describing gestalt clusters [J]. IEEE Trans on Computers, 1971, 20 (1): 68286.

16.TAN Yong, RONG Qiusheng. Implementation of A Clustering Algorithm Based on High Density [J].Computer Engineering, 2004, 30(13):119-121.

17.Leung K W T,Ng W,Lee D L. Personalized Concept-based Clustering of Search Engine Queries[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(11): 1505-1518.

18.Gao Xiaomei Feng Yun, Feng Xingjie. Incremental - K Medoids clustering algorithm [J]. Computer engineering, 2005, 7 (3l):181-183.