Research of Large-Scale and Complex Agricultural Data Classification Algorithms Based on the Spatial Variability

Hang Chen[1,2], Guifen Chen[1(✉)], Lixia Cai[1], and Yuqin Yang[1]

[1] College of Information Technology,
Jilin Agricultural University, Changchun 130118, China
chenhang0811@163.com, guifchen@163.com,
419513823@qq.com, 1172126066@qq.com
[2] Institute of Scientific and Technical Information of Jilin,
Beijing 130000, China

Abstract. In the actual classification problems, as a result of lack of clear boundary information between classification objects, that could lead to loss of classification accuracy easily. Therefore, this article from the spatial patterns of the sample properties to proceed, fuzzy clustering algorithm is proposed based on the sensitivity of attribute weights, through using the attribute weights to improve the classification capability between confusing samples, that is for researching and analysing soil nutrient spatial data with consecutive years to collect in Nongan town. Then through the analysis of the visualization technology to realize the visualization of the algorithm. Experimental results show that introducing weights portray attribute information could reduce the objective function value, and effectively alleviate the phenomenon of boundary data that cannot distinguish. Ultimately to improve the classification accuracy. Meanwhile, use of MATLAB to form visualization of three-dimensional image. The results provide a basis for to improve the accuracy of data classification and clustering analysis of large and complex agricultural data.

Keywords: Large-scale and complex data · Spatial variation law · Fuzzy clustering · Soil nutrients · Sensitive attribute weights

1 Introduction

The arrival of the era of precision agriculture [1, 2], makes a variety of complex link relationship between agricultural data features with apparent spatial variability [3] and the correlation. The consequent massive, diverse and dynamic changes, incomplete, uncertain and a series of features, so that each attribute internal link close, but contact between attributes relatively sparse [4]. However, data mining can effectively for data analysis, Wherein the cluster analysis can be used as an independent tool to obtain data distribution situation, so that can observe characteristics of each class, analysis some specific class to move forward a single step, Final extract useful information. But with the rising importance of data structure information and the data on the exponential growth. This shows traditional data mining algorithms have been unable to meet these needs. How to

D. Li and Z. Li (Eds.): CCTA 2015, Part I, IFIP AICT 478, pp. 45–52, 2016.
DOI: 10.1007/978-3-319-48357-3_5

introduce spatial patterns of in large-scale agriculture data [5]. And to strengthen the links between attributes for regional management in order to improve the parallel and distributed implementation strategy of clustering algorithm [6, 7]. All of these are gradually attracted researchers' attention [8]. So, on the basis of K-means algorithm, according to the interdependence of spatial unit location. Li [9], who put forward a new Spatial Contiguous K-Means Cluster algorithm, who removed a lot of debris and isolated cell and taken into account the continuity of the management partition. The actual show that the method is suitable for variable precision agriculture field management operations. Fleming et al. noted that define management zones based on oil properties, terrain and farmers' production experience. Then Appeared feature selection methods which is proposed for large-scale data sets. The purpose is to improve the data processing efficiency and rationality of the decision-making program [10]. While Cui proposed Quick association rules mining algorithm based on a large dense database of vertical data step [11].

Studied the basis of the existing methods, consider the algorithm when dealing with large data sets required scalability and efficiency. Analysis the influence of spatial variability and structural information on the temporal and spatial data, the paper proposed Fuzzy C-Means algorithm that based on attribute weights. In the case of verify its reliability, analysis the algorithm through MATLAB toolbox graphical to further improve the quality of clustering results. The results showed that the introduction of spatial variability and structure information can effectively reduce losses, due to the imbalance caused by the boundary. So, in visual processing, MATLAB played an immediate role.

2 Sensitive Attribute Weights Fuzzy C-Means Analysis

The traditional clustering algorithms in the classification process vulnerable to the sample spatial variability and structure information effect, the existence of boundary data processing hard to demarcation issues, which lead to low accuracy [12]. So, fuzzy c-means cluster algorithm was introduced to analysis of data space structure information [12]. Then constructed fuzzy similar matrix directly after standardization of data. Therefore, this study and master the premise of in soil temporal and spatial variation characteristics and laws, the combination of the spatial variability of attribute weights applied to FCM algorithm. Ultimately improving algorithm's classification capability, reducing the loss of classification precision caused by boundary spatio-temporal data.

2.1 Construction of Sensitive Attribute Weights

Firstly, we analyzed spatial patterns of soil nutrient which according to experience of experts in the field and soil fertility characteristics of test area [13]. The results showed that available p variation coefficient was 31.12 %, and the number of rapidly available potassium was 21.51 %, and available nitrogen was 11.69 % [14]. Secondly, the space coefficient of variation was introduced to the algorithm and AHP could solve the weight coefficients. Now, specific steps are as follows:

(1) Construct pair wise comparative matrix;
(2) Selected any n-dimension normalization original vector $\mathbf{w}^{(0)}$;
(3) Calculate $\tilde{\mathbf{w}}^{(j+1)} = \mathbf{A}\mathbf{w}^{(j)}, j = 1, 2, \cdots$;
(4) Normalize the $\tilde{\mathbf{w}}^{(j+1)}$;
(5) For a given precision ε in advance, when $\left|w_i^{(k+1)} - w_i^{(k)}\right| < \varepsilon, j = 1, 2, \cdots, n$, Then $\tilde{\mathbf{w}}^{(j+1)}$ is the requirement feature vector; otherwise return (2);
(6) Compute $\lambda = \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{w}_i^{(j+1)}}{w_i^{(j)}}$;
(7) Calculation $CI = (\lambda - n)/(n\text{-}1)$;
(8) Calculated $CR = CI/RI$;
(9) If the $CR<0.1$, then through the consistency check; Otherwise, reconstructing paired comparative matrix;
(10) when all layers are calculated out, to obtain the total target weight vector $A = (a_1, a_2, \ldots, a_m)$; if not, return back to (1).

2.2 Construction of Attribute Weights Fuzzy C-Means Algorithm

When deal with clustering problems with fuzzy concepts [15, 16], each sample is not divided into one class strictly, but belongs to a category at a certain membership. Define:

$J(U, V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m d_{ik}^2 (U = (u_{ik})_{c\times n}, d_{ik} = \|x_k - v_i\|)$, which based on the membership degree matrix [17]. Objective function refers to the sum of weighted square distance between the samples and the center of the cluster. The optimization class refers to make the objective function to take the minimum class. If all points of a class are closed to the center of the class, the value of the goal function is very small.

2.3 MATLAB Modeling Tool

MATLAB is an interactive programming language based on matrix manipulation, the main functions of MATLAB including data analysis, numerical calculation and engineering drawing and so on [18], it can also marked and print the graphics. This paper adopts MATLAB to process and analyze the data, compares this kind of algorithm with the traditional fuzzy c-means, compares the accuracy of the algorithm through the objective function value, and realize the 3d visualization process of the data.

3 Experimental Results and Discussion

3.1 Data Sources

The use of 3S (these are GPS, GIS, RS) and sensor technology acquisition Nongan town soil nutrient information, then based on the geolocation of arable land, then positioning collecting position of the point. Select the main factor affecting soil fertility

[19] (nitrogen, phosphorus, potassium) as the sample data for research, the sampling distribution is as follows:

The plum blossom in Fig. 1 sampling methods is the five sampling method, that is the grid on the four horns and on the center of the grid as the soil samples mixed grid soil sample. Collecting the data content to farmers in 2010 in the town of partial data, for example, as be showed in Table 1.

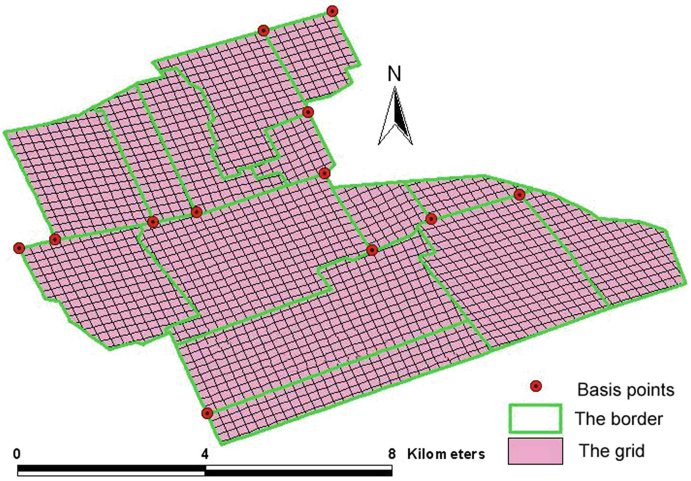

Fig. 1. Soil Sampling Point in Nongan town

Table 1. The sample data

Town name	Alkaline hydrolysis nitrogen (mg/kg)	Available phosphorus (mg/kg)	Available potassium (mg/kg)	Latitude	Longitude
Nongan town	118	30.8	208	44.51535	125.23695
Nongan town	118	36.3	198	44.50502	125.2354
Nongan town	118	42.8	227	44.50268	125.23857
Nongan town	103	32.1	198	44.51077	125.25425
Nongan town	132	15.8	227	44.50702	125.2542
Nongan town	165	10.8	237	44.51245	125.25438
Nongan town	147	30.8	217	44.51385	125.2544
Nongan town	143	21.3	198	44.5057	125.25578

3.2 Data Processing

Processing soil nutrient data in Nongan town from 2008 to 2012, Take on [0,1] evenly distributed random number to determine the initial membership degree matrix. Which

determined cluster center by iteratively. Among them, step l iteration of the cluster centers is:

$$v^{(l)} = \frac{\sum_{k=1}^{n} (u_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^{n} (u_{ik}^{(l-1)})^m}, \ i = 1, \ 2, \ldots c$$

where c is the number of classes, m > 1.

3.3 Application and Analysis of Algorithms

Preprocessing the data after combined with the analysis of the soil nutrient spatial variation. Algorithm through continuous iterative to adjust the size of the objective function value in order to achieve the classification of soil fertility. To objects in the cluster based on the continuity of time and space, after processing the sample data, we use sensitive attribute weights fuzzy C-means algorithm to analysis the data from 2008 to 2012. Experiments show that when taking membership degree exponent 8, clustering result is obvious. In the case of the same power exponent value, compared with the traditional fuzzy C-means clustering algorithm, after repeated experiments and found that the accuracy and operational efficiency of the improved algorithm are both higher traditional algorithm. Wherein the results of 2011 as shown in Table 2.

Table 2. The result

Algorithm	Mean objective function	Average running time (s)
Fuzzy C-Means	15.118689	0.217
Attribute weights Fuzzy C-Means	11.989009	0.180

From Table 2 we can see that under the same conditions, the objective function value is smaller, and the accuracy of the relative increase 21.7 %, also it has a higher operating efficiency. That because the sample edge has no clear demarcation point, and Fuzzy C-Means could improve this problem when dealing with data. On this basis, introducing of spatial variation regularity. Without prejudice to the classification results, the better management area is divided. Combined with the results of the above analysis, using MATLAB visualization toolbox for data processing. The results obtained in Fig. 2 and Table 3.

The Table 3 and Fig. 2 show that after years of continuous precise fertilization, the similar degree of data is improved in gradually, the discrepancy between categories gradually become smaller, the soil fertility difference is leveling off. All above shows that after precise fertilization, the plot of soil in Alkeline-N, Olsen-P and Olsen-K three nutrients data integrated similarity increased year by year; On the other hand also proved that the attribute weights are C clustering algorithm is suitable for evaluation of soil fertility.

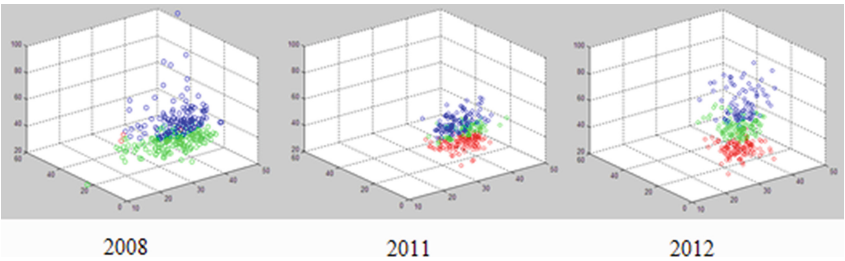

Fig. 2. The three dimensional clustering figure

Table 3. The clustering results

Clustered data	2008	2009	2010	2011	2012
Cluster 0	3	93	44	74	82
Cluster 1	123	99	114	108	108
Cluster 2	174	108	142	118	110

4 Conclusion

The attribute weights fuzzy c-means algorithm is used to analysis and evaluation for Nongan county Nongan town soil nutrient data for five consecutive years (2008–2012). The test results show that after five consecutive years of precise fertilization, soil fertility condition had the obvious change. The attribute weights c-means clustering algorithm is an effective methods of research and evaluation of the soil fertility, in line with the farmers, the change trend of soil fertility.

Firstly, the algorithm consider the spatial variability of soil fertility, combined with AHP to determine the sensitive attribute weights. the original scattered data not only retain the traditional algorithms consider the problem of difficult to deal with the boundary points by using the concept of fuzzy sets, and to overcome the imbalance between the various properties and is sensitive to “noise” and outlier data shortcomings.

Secondly, the paper used the sensitive attribute weights fuzzy c-means algorithm to do the clustering analysis for Nongan soil data in 2011 which included the soil alkaline hydrolysis nitrogen, available phosphorus and available potassium three nutrients data. The results show that the algorithm was 21.7 % higher than that of traditional algorithm of relative accuracy and efficiency increased by 17 %, the improved algorithm clustering effect is better.

Thirdly, using the algorithm to analysis soil nutrient data which precision fertilization consecutive for five years, the results show that the whole plot soil in alkali solution nitrogen, available phosphorus and available potassium in three kinds of nutrient data integrated similarity increased with each passing year. The results of the experiment are consistent with the actual situation of soil fertility, which provides a new reference for the analysis of the status of soil fertility in the future.

Fuzzy clustering is a rather ambiguous concept, the two clustering algorithms should be repeated iterations based on the exponent value of the objective function membership degree, so as to determine the relatively close to the true clustering value. We know MATLAB can handle large-scale data, and the formation of the visual clustering results. Covered in this article the agricultural data mostly a single soil nutrient data. Face of the growing complexity of massive agricultural data, the original matrix processing mode is not enough. In the future, attention should be application testing large data sets, in order to confirm the validity of the algorithm of massive data clustering.

Acknowledgment. The paper was supported by the national "863" project (2006AA10A309), the National spark plan project (2015GA660004), National Spark Plan (2008GA661003) and Shi Hang projects of Jilin province (2011-Z20).

References

1. Zhang, S.: Research of precision agriculture automatic variable fertilization theory and technology based on GPS, GIS, pp. 1–3. Jilin University, Jilin (2003)
2. Chen, L.: Theoretical and experimental studies on variable-rate fertilization in precision farming, pp. 98–104. China Agricultural University, Beijing (2003)
3. Xiang, J., Fu, Q., Wang, Z.: Applied research progress of spatial variability theoretical in soil properties analysis. Soil Water Conserv. Stud. **15**(1), 250–253 (2008)
4. Enlai, Z., Wenning, H., Hang, L., Rong, Yu., Zhu, L.: Compression algorithm of Beidou position redundant data based on time series clustering. Comput. Eng. **2**, 40–42 (2012)
5. Chen, G.: Research and application of spatial data mining technology for precision agriculture. Jilin University, ChangChun (2009)
6. Yun, G.X.F., Xingjie, F.: Incremental - K medoids clustering algorithm. Comput. Eng. **7**(3l), 181–183 (2005)
7. Leung, K.W.T., Ng, W., Lee, D.L.: Personalized concept-based clustering of search engine queries. IEEE Trans. Knowl. Data Eng. **20**(11), 1505–1518 (2008)
8. Deng, M., Liu, Q.L., Wang, J.Q., Shi, Y.: A general method of spatio-temporal clustering analysis. Sci. China Press **42**(1), 111–124 (2012)
9. Li, X., Pan, Y.C., Zhao, C.: Precision agriculture management zones based on spatial continuity of the clustering algorithm. Agric. Eng. J. **21**(8), 78–82 (2005)
10. Li, Z., He, C.: A large data set suitable for feature selection method. Comput. Sci. **33**(4), 184–186 (2007)
11. Cui, J., Li, Q.: Quick association rules mining algorithm based on a large dense database of vertical data step. Comput. Sci. **39**(1), 134–137, 151 (2012)
12. Jing, H., Li, D., Duan, Q., Han, Y., Chen, G.: A fuzzy c-means clustering based algorithm to automatically segment fish disease visual symptoms. Sens. Lett. **10**, 1–8 (2012)
13. Zhao, Y., Han, H., Cao, L., Chen, G.: Study on soil nutrients spatial variability in YuShu City. Comput. Comput. Technol. Agric. **2**, 1–7 (2012)
14. Chen, G.F., Tsao, L.I., Wang, G.: Application of weighted spatial fuzzy clustering algorithm in soil fertility evaluation. Chin. Agric. Sci. **42**(10), 3559–3563 (2009)
15. Backer, E., Jain, A.K.: A clustering performance measure based on fuzzy set decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **3**(1), 66277 (1981)

16. Li, Y., Shi, Z., Cifang, W., Li, F., Cheng, J.: Definition of management zones based on fuzzy clustering analysis in coastal saline land. Sci. Agric. Sin. **40**(1), 114–122 (2007)
17. Helong, Yu., Dayou, L., Guifen, C.: Determination of the soil nutrient management zones based on weighted fuzzy clustering. Trans. Chin. Soc. Agric. Mach. **40**, 177–182 (2009)
18. Bai, X., Xiong, S.: The implementation and application of mathematical experiment system based on MATLAB. Nanchang University (2012)
19. Umeda, M., Kaho, T., Iida, M., Lee, C.K.: Effect of variable rate fertilizing for paddy field. In: 2001 ASAE Annual International Meeting, 2001, Paper Number 01(Part. II)