Unsupervised multi-scale fuzzy clustering algorithm application in the evaluation of soil fertility

Liying Cao , Helong Yu, Li Ma Guifen Chen*
School of Information Technology Jilin Agricultural University Changchun China
E-mail: caoliying99@163.com

***Abstract*—this paper adopts statistical learning theory and optimization theory to the analysis of the algorithm theory, probe into its theoretical foundation. The existing theoretical analysis on the basis of the establishment of clustering model algorithm design, code realization and finally a lot of different data set of test, choose soil data as a test database, will be in the database on a large number of data mining experiment to verify the performance of the proposed algorithm. The test result feedback back will further deepen the theoretical research or correct theory already mistakes, new theory and will continue to guide experiments, both mutual promoting common development.**

Keywords- Data mining, clustering algorithm, unsupervised, multi-scale

I. INTRODUCTION

With accurate job expansion in agricultural production, agricultural scientists and managers to obtain and accumulated a great deal of precision agriculture production process is closely related to all aspects of attribute data and spatial data, these reflect the true nature of agricultural production job process, the specific the data is a valuable asset guidance area precise job production [1]. The agriculture data has rich, multidimensional, dynamic, incomplete, uncertain characteristics [2], how to be more timely, more accurately show the difference this time, space is also before us an important research content. Agricultural production is affected by many factors of complexity, how to adopt one or several ways to dig out from the large amounts of data law hidden within the data, and then according to the law to determine reasonable soil nutrient management unit, the development of farmland soil nutrients the partition management method [3], the rational development and utilization of the soil, increase productivity is important [4].

More than a decade, we in the process of implementation of precision agriculture in the national "863" project, due to the use of computers, networks, 3S technology has accumulated a large number of spatial data, these spatial data space, timeliness, multi-dimensional, the characteristics of the mass, complexity and uncertainty [5]. In order to efficiently use existing resources in agriculture to achieve the production target optimization run, you must find out the land within the soil properties and productivity of spatial variability of circumstances, and to conduct a comprehensive analysis of these results and the correct evaluation of [6,7].

II. RESEARCH METHODS

Unsupervised multi-scale fuzzy clustering algorithm is a new kind of mean shift clustering algorithm。Mean shift clustering algorithm is an unsupervised statistical iterative clustering algorithm, it is derived from the pattern recognition of non-parametric kernel function probability density estimates.

For a given data set X={x1 ， x2 ， … ， xn} ， Unsupervised multi-scale fuzzy clustering algorithm to determine the final clustering divided the main steps is as follows:

Step 1:

Determine the scale factor of η upper bound $\overline{\eta_{\max}}$ and lower bound $\overline{\eta_{\min}}$

Step 2:

Discrimination η ， Generate a sequence of discrete points $\{\eta\}$ =l，2，…，t。

Step 3:

For $\eta=\eta$ k (k=1，2，…，t) ，to run partitioning algorithm, and record the number of clusters of different $\eta=\eta$ k (k=1，2，…，t).

Step 4:

Calculated unsupervised multi-scale fuzzy clustering algorithm clustering validity index V (n q)(q=2，3，…，ss-1)。

Step 5:

Make q =arg max {V (n (q))} ，Calculation of the cluster number (n (q)) into the optimal clustering the division is unsupervised and multi-scale fuzzy clustering algorithm is finally cluster classification results.

The value of V((n(q))) is judge data set X whether there is clustering structure of an index .Users can customize set threshold θ ($\theta\geq1$)(In this paper, the experimental always set θ=1.5)， if V(n(q))< θ ， It can judge X do not have we expect data structure. If the user wants to data multi-scale analysis or want to structure information of data for further study of the following step 6 can help get more about data clustering structure information.

Step 6:

Setting the threshold value θ ($\theta \geq 1$) ， for each meet V((n(q)))$\geq \theta$ of the q, calculation n(q) the corresponding optimal clustering division.

Clustering divided from above to get a different number of clusters, the user can conduct multi-scale analysis of the clustering structure of the data set, which is conducive to the user to grasp the relationship between different clustering results at different scales, deepen deep data structure levels of understanding.

FUMFA specific details are as follows:

Data re-expression

Initialize m=LCM ={x1}

FOR =2 to N

Find ck ： d (xi ， ck) =minl≤j≤d(xi,cj)

If d (xi,ck)>⊙ AND(m<q)then

m=m+1

cm = {xi}

Else

Ck=Ck ∪ {xi}

If necessary, the profile vector to express

End {if}

End {For}

II.UMFA clustering

Step 1 make j=1， Setting threshold ε >0。

Step 2 make v (0) =cj, Use update formula:

$$v^{(l+1)} = \frac{\sum_{k=1}^{n} n_k c_k \overline{d}(v^{(l)}, c_k)}{\sum_{k=1}^{n} n_k c_k \overline{d}(v^{(l)}, c_k)}$$

Calculation of cj convergence point，Recorded as Pj。If j < $\overline{n}$ ， make j =j +1，Repeat step 2.

Step 3 for $\forall_{pa}$, $\forall_{pb}$ (1≤a ， b≤ $\overline{n}$ ， a≠b) ， if $\|\text{Pa - Pb}\| \leq \varepsilon$ ， Sa and Sb in the data points are divided into a class; otherwise divided into different classes.

Define the point to set mapping：

$F : H^c \mapsto P(M_{fc})$,

$H^c = \underbrace{H \times H \cdots, H}_{c\uparrow}$, $F(W) = \{U \in M_{fc} \mid (U,W)$

Meet iteration constraint style;

Defined functions

$G : M_{fc} \mapsto H^c$, $W = G(U)$

Defined by the iterative formula.

Define mappings

$T_m = A_2 \circ A_1$

$A_1 : M_{fc} \times H^c \mapsto H^c$, $A_1(U,W) = G(U)$;

$A_2 : H^c \mapsto P(M_{fc} \times H^c)$, $A_2(W) = \{(U,W) \mid U \in F(W)\}$.

T_m subscript m denotes the fuzzy coefficient ,To sum up

$T_m(U,W) = \{(\hat{U},\hat{W}) \mid \hat{W} = G(U), \hat{U} \in F(\hat{W})\}$.

The KFCM is convergence theorem. Located contains at least one non-zero elements of fuzzy coefficient, initial iteration point

$U^{(0)} \in M_{fc}$,

$U^{(k)} \in \{U \mid W = G(U^{(k-1)}), U \in F(W)\}$ $(k = 1,2,\cdots)$ The KFCM algorithm iterative sequence is defined

$\{U^{(k)}\}_{k=1,2,\cdots}$.

Iterative sequences or converges to a point in the Ω or the presence of at least one promoter sequence converges to a point in the Ω .

Given KDF convergence of the algorithm, and the results were eventually given the convergence theorem unified form of promotion.

Z will convergence theorem:

Set V is a distance space. Point $z^{(1)} \in V$, $A : V \mapsto P(V)$ is V on the point to set mapping,

By A defined algorithm with $z^{(1)}$ for initial point produce sequence

$\{z^{(k)}\}_{k=1,2,\cdots}$ .make $\Omega \subset V$ said solution set.

I. All the point $z^{(k)}$ belong to the tight subset of V ;

II. There exists a continuous function $J : V \mapsto R$ make

(a) if $z \notin \Omega$,then for any $y \in A(z)$, $J(y) < J(z)$,

(b) if $z \in \Omega$, or algorithm terminated, or for any $y \in A(z)$, $J(y) \leq J(z)$;

III. If $z \notin \Omega$, Mapping A in z point is closed;

Then, the algorithm to stop in a solution or any a convergent subsequence limit is a solution.

III. EXPERIMENT STUYING

Experimental place of the national "863" demonstration zones the Yushu City gongpengzi town on the 13th Village F, 3, located in the northwest of Yushu City, 26 km away from the city, the longitude between 126.315738-126.317017 latitude in 44.999859-45.002761 between . Temperate semi-humid temperate climate zone, the annual average temperature of 4 ° C, the frost-free period is about 135 days, with an average annual rainfall of 500 to 700 mm. Demonstration base of corn precise operating system development and application of the national "863" project, a total area of approximately 375 acres in the experimental field. We use GPS, GIS, RS and sensor technology to obtain the experimental ground soil nutrients and corn growth information to obtain the spatial information of the sampling points, first by GPS, then the use of GIS technology land divided into 40m×40m grid unit, A1 ~ L10 sampling point, the soil position grid in Figure 1-year shows.

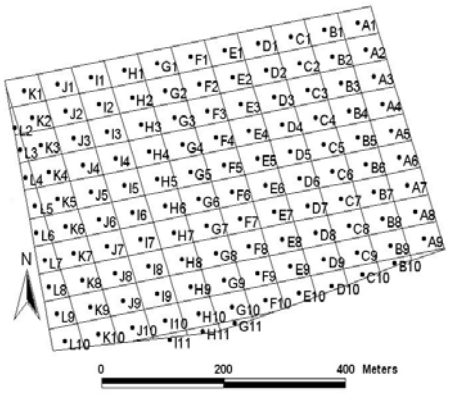

Figure 1. soil position grid chart

Within this grid cell sampling, sampling deep degrees 25 cm sampling method for the five-point the Plum sampling method, the upcoming four grid as the grid angle and grid center of the soil samples mixed soil samples. Collected soil samples for testing for seven consecutive years in 2003-2010, and the spatial coordinates of the soil organic matter, available nitrogen, available phosphorus, available potassium, soil moisture and PH property values after clustering process, select one from the 63 sample points as sample data, some data are shown in Table 1 below.

TABLE I. YUSHU CITY GONGPENGZI TOWN ON THE 13TH VILLAGE TESTING GROUND FOR MANY YEARS PART OF THE BASIC DATA

No.	Longitude	Latitude	2003 Available Phosphorus (ppm)	2008 Available Phosphorus (ppm)	2010 Available Phosphorus (ppm)
3-A1	126.3157	45.00276	30.24	16.33	16.84
3-A2	126.3158	45.00241	27.22	16.57	15.32
3-A3	126.3159	45.00206	37.91	15.11	16.65
3-A4	126.3161	45.00171	31.26	31.92	15.79
3-A5	126.3162	45.00137	23.79	18.52	13.98
3-A6	126.3163	45.00102	20.97	19.98	16.27
3-A7	126.3164	45.00067	29.64	20.22	23.15
3-A8	126.3165	45.00032	27.02	16.81	21.05
3-A9	126.3167	44.99998	20.57	18.52	13.60
3-B1	126.3152	45.00267	26.42	18.52	11.31
3-B2	126.3153	45.00232	13.51	20.22	18.75
3-B3	126.3154	45.00197	14.72	23.88	9.40
3-B4	126.3156	45.00163	19.76	20.47	13.03
3-B5	126.3157	45.00128	12.70	16.57	13.60
3-B6	126.3158	45.00093	26.62	16.33	16.46
3-B7	126.3159	45.00058	25.00	16.08	11.31
3-B8	126.3161	45.00024	11.70	14.13	7.10
3-C1	126.3147	45.00258	36.50	18.52	9.78
3-C2	126.3148	45.00223	36.10	33.63	11.12
3-C3	126.3150	45.00189	33.47	20.71	13.79
3-C4	126.3151	45.00154	25.41	25.59	17.80

Data mining, fuzzy clustering algorithm based on the division of the objective function class needs artificially assumed initialization Center of (or membership matrix) to run clustering algorithm, the cluster results depends on the initial parameter settings, and how to set up these the parameters still no reasonable theoretical basis of reliable options. The project will start to study this type of algorithm optimization and statistical theory, research and improvement of such clustering model, proposed the use of the fixed points of the domain of attraction for clustering thinking, based on the establishment of a new unsupervised multi-scale fuzzy clustering model, and explore its theoretical basis, from a theoretical point of view, the clustering framework established based on this method. A reasonable basis for the theoretically given parameters selected, the original algorithm needs to avoid the application on human lack of initialization of the cluster centers (or membership matrix).

IV. CONCLUTION

UMFA algorithm in order to test the proposed clustering performance in three real data sets with the most commonly used clustering algorithm carried out comparative experiments, and the UMFA applied to the MRI image segmentation, as shown in Figure2.

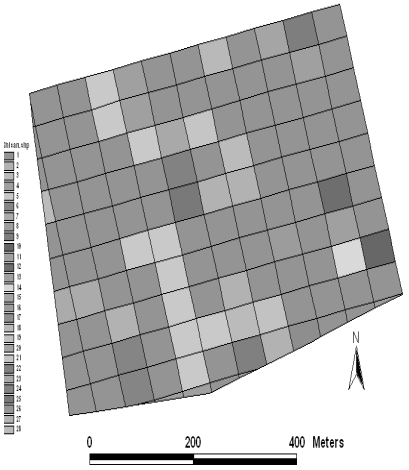

Figure 2. Results of UMFA on the image of MR

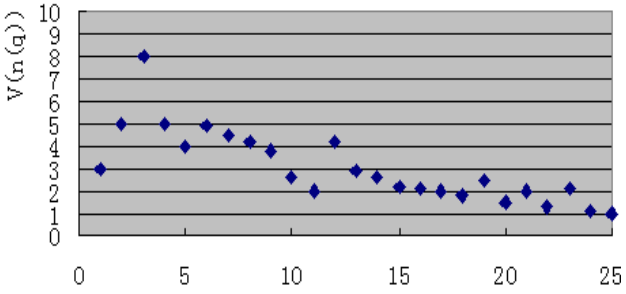

Figure 3. Effectiveness index value

Reasonable basis for data mining based on the division of the target function class fuzzy clustering need artificially assumed to initialize the center to run the clustering algorithm clustering results depend on the initial parameter settings, the project will theoretically give parameter selection.

ACKNOWLEDGMENT

This work was funded by the Youth Foundation of Jilin Agricultural University under Grant No. 201136, the National High-Tech Research and Development Plan of China under Grants Nos.2006AA10A309, Changchun Technology Correspondent Project (2009245), Jilin provincial "125" science and technology research subject(No. 201247 and 201248), the research and application of facilities for the safety of vegetables production technology based on Internet of Things(2011-Z20), Jilin province science and technology development program (201101114).

REFERENCES

[1] Bach MP, Simicevic V, Leskovic d. data mining in telecommunications: case study of cluster analysis[M]//katalinic b. Annals of Daaam for 2009 & Proceedings of the 20th International Daaam Symposium. 2009: 491-492.

[2] Bazzan ALC. Agents and Data Mining in Bioinformatics: Joining Data Gathering and Automatic Annotation with Classification and Distributed Clustering[M]// Agents and Data Mining Interaction. 2009: 3-20.

[3] Chen G-F, Wang G-W, Ma L, et al. Research on Spatially Weighted Fuzzy Dynamic clustering algorithm and spatial data mining Visualization[M]. 2009: 60-66.

[4] Daruru S, Marin N, Walker M, et al. Pervasive Parallelism in Data Mining: Dataflow solution to Co-clustering Large and Sparse Netflix Data[M]. 2009: 1115-1123.

[5] Dong J. Data Mining of Time Series Based on Wave Cluster[M]. 2009: 697-699.

[6] Karmaker A, Rahman SM, Society IC. Outlier Detection in Spatial Databases Using Clustering Data Mining[M]. 2009: 1657-1658.

[7] Marroquin ID, Brault J-J, Hart BS. A visual data-mining methodology for seismic-facies analysis: Part 1-Testing and comparison with other unsupervised clustering methods[J]. Geophysics, 2009,74(1): P1-P11.

[8] Qu Z, Wang X. Application of Grey Relational Clustering and Data Mining In Information Extraction[M]. 2009: 3-6.

[9] Qu Z. Application of Information Technology in Enterprise E-Commerce Based on Grey Relational Clustering and Data Mining[M]. 2009: 7-10.