基于改进的DBSCAN算法的土壤肥力变化的分析研究

郭万春[1] 蔡丽霞[2] 陈 航[2] 陈桂芬[2]
(长春金融高等专科学校 长春 130022)[1] (吉林农业大学 长春 130118)[2]

摘 要 通常基于密度的DBSCAN算法可以有效地处理任意形状的簇,但由于时空数据具有明显的差异性,该算法不能综合分析土壤肥力状况。针对这一问题,提出了一种基于改进的DBSCAN算法来对农安镇土壤肥力状况进行分析研究。首先利用层次分析法得到土壤养分各属性的权值,以平衡数据间的差异性;其次,利用改进的DBSCAN算法对农安镇的土壤肥力数据进行分析,并将实验结果与传统的DBSCAN算法进行比较。实验结果表明,改进的DBSCAN算法对于选取Eps和minPts两个参数更加快速、有效,聚类结果更好。

关键词 层次分析,DBSCAN算法,土壤肥力,权重值
中图法分类号 TP399,S158.2 **文献标识码** A

Analysis and Research of the Soil Fertility Status Based on Improved DBSCAN Algorithm

GUO Wan-chun[1] CAI Li-xia[1] CHEN Hang[2] CHEN Gui-fen[2]
(Changchun Finance College, Changchun 130022, China)[1]
(Jilin Agricaltural University, Changchun 130118, China)[2]

Abstract Typically DBSCAN algorithm based on density can effectively deal with clusters of arbitrary shape, but the algorithm is not a comprehensive analysis of soil fertility, Since temporal data have obvious differences. To solve this problem, this paper proposes a method to analyze the situation of soil fertility in nong'an town based on improved DBSCAN algorithm. First, using AHP to get the weight of each property in order to balance the differences between the data; Secondly, the improved DBSCAN algorithm apply to analyze the data of soil fertility, and the experimental results with traditional DBSCAN algorithm were compared. Experimental results show that the improved DBSCAN algorithm more quickly and efficiently for selecting the two parameters Eps and minPts, it has better clustering results.

Keywords AHP, DBSCAN algorithm, Soil fertility, Weights

农业是人类赖以生存的基本生活资料的来源。随着信息化的推进,使得土壤肥力数据呈现丰富、多维、动态、不确定、不完整等特性[4]。如何更及时、更精确地展现这种时空数据的差异性并对数据进行综合评价具有重要的现实意义[11]。

目前,数据挖掘技术在时空数据分类方面已经得到了日益广泛的应用。作为数据挖掘主要方法之一的聚类算法,也越来越受到人们的关注[12]。聚类算法主要有划分方法、层次方法、局部方法和模型方法[1]等几种类型。DBSCAN算法属于局部方法,它可以发现任意形状的聚类[2],具有较强的聚类能力。李良厚等提出了聚类分析在立地分类与土壤肥力评价中的应用[3]。张书慧等对变量区与传统区土壤中N、P、K养分变异系数进行了比较,说明变量施肥具有均衡土壤养分的作用[5]。陈桂芬等提出了加权空间模糊动态聚类算法,并证明了该方法在土壤肥力评价上的有效性[6]。周水庚等[7]提出了一种基于数据分区的DBSCAN算法,根据时空数据的空间分布特征,将整个数据空间划分为多个较小的分区,然后分别对这些局部分区进行聚类,最后将各局部聚类进行合并。通常DBSCAN算法以超球状区域内数据对象的数量来衡量区域密度的高低,但对于高密度复杂数据,需要耗费大量的内存,且聚类效果不佳。本文首先利用层析分析法对土壤肥力数据进行加权以平衡肥力数据,之后将加权DBSCAN算法与未加权DBSCAN算法进行对比分析,结果表明,加权DBSCAN算法对于选取Eps和minPts两个参数更加有效,聚类结果更加稳定。

1 算法分析

1.1 DBSCAN算法[10]

假设有一个数据对象集合U,对于给定的minPts和Eps,DBSCAN算法描述如下:

(1)选择数据库中任意一个不属于任何聚类且满足条件的对象p,创建一个新的聚类;

(2)根据该聚类中的对象,循环收集密度可达的对象加入该聚类,直到没有新的对象加入为止;

(3)若不存在不在任何聚类内的对象则结束,否则执行(1)。

1.1.1 利用层次分析法确定权重系数[6]

其算法如下:

Step 1 构造成对比较矩阵;

本文受吉林省世行项目(2012Z04)资助。

郭万春(1972—),女,讲师,主要研究方向为人工智能与数据挖掘、计算机应用,E-mail:Guowanchun1234@163.com。

Step 2　任取 n 维归一化初始向量 $w^{(0)}$；

Step 3　计算 $\widetilde{w}^{(k+1)}=Aw^{(k)}$，$k=1,2,\cdots$；

Step 4　归一化 $\widetilde{w}^{(k+1)}$；

Step 5　对于预先给定的精度 ε，当 $|w_i^{(k+1)}-w_i^{(k)}|<\varepsilon$，$i=1,2,\cdots,n$ 成立时，$\widetilde{w}^{(k+1)}$ 即为所求特征向量；否则返回 Step 2；

Step 6　计算最大特征值 $\lambda=\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{w}_i^{(k+1)}}{w_i^{(k)}}$；

Step 7　计算一致性指标 $CI=\frac{\lambda-n}{n-1}$；

Step 8　计算一致性比率 $CR=\frac{CI}{RI}$；

Step 9　若 $CR<0.1$ 成立，通过一致性检验；否则，重新构造成对比较矩阵；

Step 10　若所有层都计算完成，获得总目标的权重向量 $A=(a_1,a_2,\cdots,a_m)$；否则，返回到 Step 1。

1.2　建立加权 DBSCAN 模型

本文运用陈桂芬等[6]提出的加权空间模糊动态聚类方法对数据进行预处理。

①数据标准化

由于在实际问题中，不同的数据一般有不同的量纲，为了使有不同量纲的量也能进行比较，需要进行数据标准化，即将数据压缩到[0,1]区间上。对原始数据进行极差变换：

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

②加权运算

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

③建立模糊相似矩阵

计算模糊集 i 与模糊集 j 的贴近度 r_{ij}：

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

从而得到模糊相似矩阵 $R=(r_{ij})_{n\times n}$。

2　实验与分析

2.1　实验数据集

实验数据来自国家“863”计划项目示范基地——吉林省农安县连续多年进行了精准施肥后的土壤肥力[9,10]数据。本文从该地区中选择了具有代表性的 2007－2011 年农安县农安镇的土壤碱解氮、土壤有效磷和土壤速效钾养分数据进行综合分析，部分数据如表 1 所列。

表 1　部分采样数据

Town name	Available nitrogen (mg/kg)	Available phosphorus (mg/kg)	Available potassium (mg/kg)	Latitude	Longitude
Nongan town	110.0	10.8	180.0	44.50040	125.24821
Nongan town	138.0	12.7	170.0	44.50205	125.2497
Nongan town	152.0	46.4	100.0	44.50305	125.2346
Nongan town	138.0	9.10	130.0	44.50648	125.2486
Nongan town	141.0	20.6	210.0	44.25682	125.0319
Wanshun town	138.0	16.4	173.0	44.52530	124.9508

2.2　算法应用

首先，依据式(2)对农安镇土壤养分数据进行评价，构造成对比矩阵 B。其次，用层次分析法求得土壤中碱解氮、有效磷和速效钾 3 种养分权重分别为 0.3782、0.2032 和 0.4185。最后，利用加权和未加权聚类算法对试验区土壤养分数据进行聚类分析。其中，邻域半径 R 为 0.09，P 的领域中包含的对象($P_{\min}$)不少于 4 个。实验结果如表 2 所列。

表 2　两种算法的对比结果

算法	聚类数	孤立点数	运行时间(s)
加权 DBSCAN	2	24	0.27
DBSCAN	5	76	0.25

由表 2 可知，在运行效率基本一致的情况下，实验结果的分类数和孤立点数有明显差异(2、5；24、76)，加权 DBSCAN 算法的聚类结果比未加权的 DBSCAN 算法更加接近实际情况。未加权聚类时会将土壤中不同养分的高低差距抵消而划入同一类中，而加权之后会分配到不同的类中，能更好地反映土壤养分的真实情况。

结合上述情况，利用加权 DBSCAN 和未加权的 DBSCAN 算法对农安县农安镇 2011 年的土壤肥力数据进行反复测试及比较。部分实验结果如表 3 所列。

表 3　部分对比结果

	加权 DBSCAN					未加权的 DBSCAN				
Eps	0.07	0.075	0.08	0.09	0.095	0.07	0.075	0.08	0.09	0.095
minPts	4	4	3	3	4	4	4	3	3	4
聚类数	3	2	3	3	2	11	7	10	9	4
孤立点数	63	53	37	16	22	162	144	89	53	60

从表 3 可以看出，Eps、minPts 两个参数的设置对实验结果的影响很大，加权 DBSCAN 比未加权的 DBSCAN 算法的聚类结果更加稳定，易于选取 Eps 和 minPts 两个参数的范围(加权 DBSCAN 从 0.08 开始聚类结果就比较稳定，未加权 DBSCAN 变化程度一直较大，难以缩小范围)。经过反复测试，我们发现选取 Eps 为 0.09、minPts 为 3 的聚类结果最佳，这一结果也与农安镇实际情况相符。因此，加权 DBSCAN 算法更容易对土壤肥力数据进行分类。

结束语　从以上分析与比较可以看出，本文提出的加权 DBSCAN 算法比未加权的 DBSCAN 算法更能有效地分析土壤肥力，对土壤肥力演变规律的研究具有一定的参考意义。

(1)该算法利用层次分析法对原始分散数据进行加权处理，避免了未加权 DBSCAN 算法没有区分数据各属性之间的不平衡性的缺点。

(2)利用加权与未加权 DBSCAN 算法对农安县农安镇 2011 年土壤中碱解 N、有效 P 和速效 K 3 种土壤养分数据进行聚类的对比分析,结果表明加权 DBSCAN 算法比未加权的算法易于选取 Eps 和 minPts 两个参数,改进后的 DBSCAN 算法聚类效果更好。

虽然 DBSCAN 算法能够发现任意形状的簇,但土壤肥力数据具有空间性、不确定性、复杂性、地域性等特点,当数据量较大时,所要求的内存支持大;当数据分布不均匀时,由于使用统一的全局变量,使得聚类的效果差。另外,DBSCAN 算法需要人为确定 Eps 和 minPts 2 个参数,导致聚类过程需人工干预才能进行。本文提出的加权 DBSCAN 算法未能综合考虑上述问题,证实该算法对海量数据集聚类的有效性,如何简化聚类算法并结合多年多个乡(镇)以及土壤类型的数据进行验证,还需要进一步的深入研究。

参 考 文 献

[1] 卜东波. 聚类/分类理论研究及其在文本挖掘中的应用[D]. 北京:中国科学院技术研究所,2000

[2] Ester, Martin, Kriegel H P, et al. A Density Based Algorithmfor Discovering Clusters in Large Spatial Data bases withNoise [C]// Proceedings of the 2nd International Conference onKnowledge-Discovery and Data Mining (KDD-96). Ortland, Oregon, 1996

[3] Li Liang-hou, Li Ji-yue. Application of Clustering Analysis in classifying Site Type and Evaluating Soil Fertility[C]// 2010 Third International Conference on Education Technology and Traning (ETT 2010). 2010:468-471

[4] Turner B L, Meyer W B. Land use and land cover in global environmental change: considerations for study [J]. Int. SoiSci., J., 1991, 130: 669-680

[5] 张书慧,马成林,李伟,等. 变量施肥对玉米产量及土壤养分影响的试验[J]. 农业工程学报,2006,22(8):64-67

[6] 陈桂芬,曹丽英,王国伟. 加权空间模糊动态聚类算法在土壤肥力评价中的应用[J]. 中国农业科学,2009,42(10):3559-3563

[7] 周水庚,等. 基于数据分区的 DBSCAN 算法[J]. 计算机研究与发展,2000,37(10)

[8] 盛骤,谢式千,潘承. 毅概率论与数理统计(第 3 版)[M]. 北京:高等教育出版社,2004

[9] Umeda M, Kaho T, Iida M, et al. Effect of variable rate fertilizing for paddy field[C]// 2001 ASAE annual international meeting, 2001, 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] 任兴平,何忠龙,孟增辉. 改进 DBSCAN 算法中参数 Eps 值的确定[J]. 现代化技术,2007,(11):120-121

[12] 刘志勇,耿新青. 基于模糊聚类的文本挖掘算法[J]. 计算机工程,2009,35(5):44-45

(上接第 411 页)

步分析发现,GPU 架构下的 EMD 算法对道线数更敏感,而 CPU 架构下的 EMD 算法则似乎对采样点数更敏感些。

图 7 给出了原始地震信号剖面图与 EMD 分解后利用 Hilbert 变换[12]提取的瞬时频率 IMF 第 1 和第 2 分量。从 IMF 提取的分量图上看,高频分量包含了较多的噪音,而低频分量所刻画的能量强弱表现较高频分量要更清晰、更丰富。

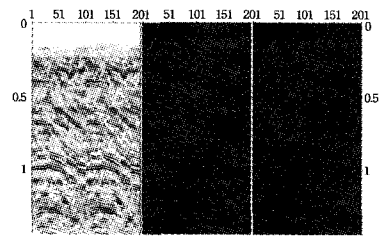

图 7 原始信号与瞬时频率第 1 和第 2 分量

参 考 文 献

[1] 皮红梅,刘财,王典. 利用 Hilbert-Huang 变换提取地震信号瞬时参数[J]. 石油地球物理勘探,2007,42(4)

[2] Xie K, Wu P, Yang S. GPU and CPU cooperation parallel visualisation for large seismic data[J]. Electronics Letters, 2010(17)

[3] Waskito P, Miwa S, Mitsukura Y, et al. Parallelizing Hilbert-Huang transform on a GPU[J]. Networking and Computing (ICNC), 2010

[4] Yu Wen-mao, Xie Kai, Yu Huo-quan, et al. Hilbert-Huang Transformation of Large Seismic Data Based on GPU[C]// Intelligence Science and Information Engineering (ISIE). 2011

[5] NVIDIA 官方网站[OL]. http://www.nvidia.cn/object/product_tesla_C2050_C2070_cn.html

[6] 韩俊刚,刘有耀,张晓. 图形处理器的历史现状和发展趋势[J]. 西安邮电学院学报,2011,16(3)

[7] Huang N E. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis[J]. J. Proc. R. Soc. Lond. A, 1998, 454: 903-995

[8] 万剑怡,孙永强,薛锦云,一种基于设计模式的三阶段并行程序设计方法[J]. 计算机研究与发展,2002,39(3)

[9] Huang N E, Wu Z. A review on Hilbert-Huang transform: method and its applications to geophysical studies[J]. Reviews of Geophysics, 2008, 46(2)

[10] Flandrin P, Rilling G, Goncalvés P. Empirical mode decomposition as a filterbank[J]. IEEE SIGNAL Proc Lett., 2004, 11: 112-114

[11] Nvidia Corporation. NVIDIA CUDA C Programming Guide Version 4.2[Z]. 2012

[12] 刘慧婷,程家兴,张旻. 利用 Hilbert 变换提取信号瞬时特征的算法实现[J]. 微机发展,2003,13(6)