一种基于簇中心点自动选择策略的密度峰值聚类算法

马春来[1] 单 洪[1] 马 涛[2]
(电子工程学院 合肥 230037)[1] (通信信息控制和安全技术重点实验室 合肥 230037)[2]

摘 要 针对基于密度峰值的聚类算法(CFSFDP)无法自行选择簇中心点的问题,提出了CFSFDP改进算法。该算法采用簇中心点自动选择策略,根据簇中心权值的变化趋势搜索"拐点",并以"拐点"之前的一组点作为各簇中心,这一策略有效避免了通过决策图判决簇中心的方法所带来的误差。仿真实验采用5类数据集,并与DBSCAN及CFSFDP算法进行了对比,结果表明,CFSFDP改进算法具有较高的准确度及较强的鲁棒性,适用于较低维度的数据的聚类分析。

关键词 聚类,DBSCAN,密度峰值,簇中心点
中图法分类号 TN301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.046

Improved Density Peaks Based Clustering Algorithm with Strategy Choosing Cluster Center Automatically

MA Chun-lai[1] SHAN Hong[1] MA Tao[2]
(Electronic Engineering Institute, Hefei 230037, China)[1]
(Science and Technology on Communication Information Security Control Laboratory, Hefei 230037, China)[2]

Abstract A new density peaks based clustering method (CFSFDP) was introduced in the paper. For the problem that it is difficult to decide the cluster number with CFSFDP, an improved algorithm was presented. With a cluster center automatic choosing strategy, the algorithm search for the "turning points" with the trends of cluster center weight's changing. Then we could regard a set of points whose weight is bigger than "turning points" as the cluster center. The error brought by ruling in the decision graph could be avoided with the strategy. Experiment was done to compare to DBSCAN and CFSFDP with 5 kinds of datasets. The results show that the improved algorithm has better performance in accuracy and robustness, and can be applied in clustering analysis for low dimension data.

Keywords Clustering, DBSCAN, Density peak, Cluster center

聚类是一种根据在数据中发现的描述对象及其关系将对象分组,从而达到分析数据的潜在结构并判断数据的自然簇属性以及压缩数据的存储容量等目的的分析方法[1-3]。由于聚类通过度量簇内对象的相似性及簇间对象的相异性来对无类别标签的数据进行分类,因此,从本质上讲,聚类可被视作一个无监督的学习过程[4,5]。

自经典的K-means算法被提出60年以来,聚类分析研究经历了巨大发展。然而,随着传感器技术和数据存储技术的成熟,大容量、高维度、非线性、无标签的数据激增,传统的聚类由于自身技术的缺陷,难以适用于对该类数据的处理,因此聚类作为一种基础性的数据分析手段,依然是数据挖掘、机器学习及模式识别领域中的研究热点。目前,聚类的种类主要分为基于划分的聚类、基于层次的聚类、基于密度的聚类、基于网格的聚类及基于模型的聚类等[1]。其中,基于密度的聚类方法以高密度区域作为判断依据,这种非参数方法与传统方法相比,不仅适用于处理任何形状的数据集,而且无需预先设定簇的数量[6]。

其中,DBSCAN算法采用空间索引技术来搜索对象的邻域,可发现任意形状的簇,能够有效排除噪声点和离群点,是基于密度的聚类方法中的经典代表[7],在时空数据处理[8]、图像处理[9,10]等方面都有广泛应用。但该算法依然存在着如下缺点[11-13]:1)对输入参数敏感,致使参数选择困难;2)计算量大,难以用于高维数据环境;3)难以处理密度分布不均匀的数据集。尽管该算法提出后,针对以上缺点的改进算法数不胜数,如OPTICS、DENCLUE、GDBSCAN等,但是这些算法难以同时顾及所有缺点。

2014年,Alex Rodriguez在《Science》上提出了一种新型、简洁、高效的聚类算法(CFSFDP)[14]。该算法只需计算一次距离,且不需 Eps 和 $minPts$ 参数,无需迭代,可针对各种类型的点集进行聚类。该算法相比DBSCAN及其改进算法具有优越性,但其需要通过决策图(Decision Graph)来完成对聚类中心的选择,增加了算法的冗余性。鉴于此,本文拟通过引入聚类中心选择策略来实现聚类中心的自动选择,以期在保证算法准确度及鲁棒性的基础上减少经验输入及人工参与,提高聚类效率。

1 CFSFDP算法

CFSFDP(Clustering by Fast Search and Find of Density Peaks)算法[14]的假设是类簇的中心由一些局部密度较低的点围绕,且这些点距离其他有局部高密度的点的距离都比较

到稿日期:2015-04-02 返修日期:2015-09-08
马春来(1989-),男,博士生,主要研究方向为机器学习、大数据情报挖掘,E-mail:eviiive@163.com;单 洪(1965-),男,教授,博士生导师,主要研究方向为战场无线网络、大数据情报挖掘;马 涛(1979-),男,博士,讲师,主要研究方向为战场无线网络、机器学习。

远。通过计算最近邻距离，得到聚类中心，并依据密度大小排序，将剩余点划分至所属类别。

对于数据集 $D=\{p_1,p_2,\cdots,p_n\}$ 中的每一个点 p_i，计算局部密度 ρ_i 及相邻密度点距离 δ_i。定义局部密度 ρ_i 为以 d_c 为半径的圆内的数据点个数：

$$\rho_i=\sum_j \chi(d_j-d_c) \tag{1}$$

其中，$\begin{cases}\chi(x)=1, x<0\\ \chi(x)=0, x\geqslant 0\end{cases}$，$d_j$ 为 p_i 到其它点的距离，d_c 为距离阈值。

定义相邻密度点距离 δ_i 为在比点 p_i 密度高的数据点中，与 p_i 最相邻的点到 p_i 距离：

$$\delta_i=\min_{j:\rho_j>\rho_i}(d_j) \tag{2}$$

算法的流程如表 1 所列。

表 1　CFSFDP 算法流程

CFSFDP	
Step1	根据簇点集数目确定 d_c 并计算每一点的局部密度 ρ_i。
Step2	将密度点按由高到低排序。
Step3	令 $\delta_1=\max_j(d_j)$，并按式(2)求得 δ_i 并存储与之对应的标号。
Step4	根据 δ_i 及 ρ_i 的关系决策图，选取簇的中心点。
Step5	根据簇中心点、数据对象标号及密度边界阈值，将剩余点分到各簇或边界域。

通过观察密度及相邻距离的关系来确定簇的中心点，算法的决策图如图 1 所示。利用决策图选取簇中心点的过程：根据 ρ 及 δ 所确定的点的分布位置，以“最上最右”的点为起点，按照“向下向左”的原则，用矩形框选择与剩余点差异最大的一组点。即可表示为在数据集 D 中寻一组点集 C，使其满足差异度 $diff(C,C^C)$ 最大，其中 $C=\{p_i\mid\rho_i>\rho_{min}$ 且 $\delta_i>\delta_{min}\}$，$(\rho_{min},\delta_{min})$ 为矩形左下点。

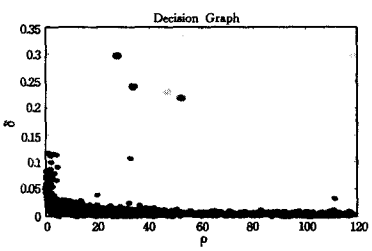

图 1　CFSFDP 算法决策图

由于这一过程需人工选择，算法每运行一次需要对簇中心点重新进行选择，增加了算法的冗余性。同时，人工辅助选择带有较强的主观性，不利于实现算法的批量自动化应用。

2　CFSFDP 改进

2.1　基本思路

相比于 DBSCAN 及其改进算法，CFSFDP 能够有效发现不同形状、不同密度的簇，且无需 *Eps* 和 *minPts* 等参数。然而，该算法在选取聚类中心时没有研究 δ_i 及 ρ_i 对聚类中心选取的影响。由于在选取聚类中心时需要人工辅助决策，这不仅增加了聚类成本，还使聚类过程带有一定的随意性和主观性，不利于算法的实现与应用。

鉴于此，本文通过分析 δ_i 及 ρ_i 的关系，提出一种基于 δ_i 及 ρ_i 这两个参数的统计特性的聚类中心选取策略，对基于密度峰值的聚类算法进行改进，从而实现聚类中心选取的自动化。该策略的基本思想：根据聚类中心的选取原则，以相邻距离 δ_i 及密度 ρ_i 的归一化乘积评测聚类点的差异度，根据差异度的统计特性及变化规律选取差异度最大的一组点作为聚类中心。在确定聚类中心之后，根据相邻距离标号划分至各簇，完成聚类。

2.2　聚类中心选择策略

由图 1 可看出，在 δ_i 及 ρ_i 归一化的条件下，一般遵循选取沿坐标轴正 45°方向偏离原点最远的一组数据点集为聚类中心的原则。但是，仅凭该原则难以从量化的角度确定簇中心点，尤其是在多个点极为接近的情况下，要找到沿正方向且偏离原点最远的一组点集十分困难。因此，为量化数据集点偏离原点的程度，在 δ 及 ρ 的归一化之后，根据其正比例关系，引入簇中心权值的概念。

定义 1　簇中心权值

$$\gamma_i=\delta_i\cdot\rho_i \tag{3}$$

为了找到偏离度最大的一组点集，将簇中心权值按降序排列并取前 $m(m=30)$ 个点。由图 2 可以看出，簇中心权值总体呈下降趋势，但下降程度不同。因此，偏离原点的程度差异最大的点可认为是簇中心权值“总体”下降趋势由急变缓的“拐点”。

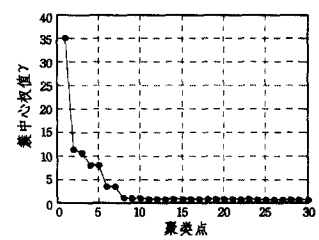

图 2　聚类点偏离度

用两点线段的斜率表征簇中心权值下降趋势，即

$$k_i^m=\frac{\delta_{i+m}-\delta_i}{m} \tag{4}$$

k_i^m 表示在区间 $[i,i+m]$ 簇中心权值的平均变化率，该参数描述了某一区间内 γ 的总体变化趋势。

定义 2　拐点

$$x=\arg(\max(\frac{k_i^1}{k_1^{i-1}})) \tag{5}$$

由式(5)可以看出，k_1^{i-1} 为第 1 个点到第 i 个点的斜率，表示点集 $\{1,2,\cdots,i\}$ 的平均变化率。k_i^1 为第 i 个点到第 $i+1$ 个点的斜率。因此，拐点的意义可表示为偏离度的“总体”趋势变化最快的临界点 x。图 3 所示为偏离度变化趋势，最大值点即为拐点。

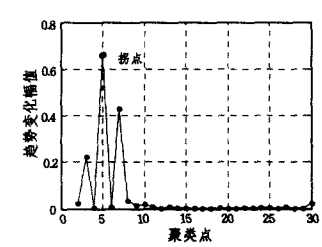

图 3　拐点判别图

表 2 给出了簇中心选取策略的具体步骤，在 Step3 的计算中，由于簇类别数远小于总数目，可取前 m 个点作为候选点。

表 2　聚类中心选择策略

簇中心选取策略	
Step1	求取每个点的差异度 γ。
Step2	将簇中心权值按降序排列。
Step3	求取 k_i^1 及 k_1^{i-1}，并求取 $\frac{k_i^1}{k_1^{i-1}}$ 的最大值，确定拐点 i=x。
Step4	以拐点之前的点{1,2,…,x}作为簇中心点
Step5	根据簇中心点、数据对象标号及密度边界阈值，将剩余点分到各簇或边界域。

3　仿真实验

3.1　仿真环境与实验参数

为验证算法的有效性，本文使用 5 类数据集，分别采用 DBSCAN、CFSFDP 及基于聚类中心选择策略的 CFSFDP 3 种算法，从准确度、鲁棒性及计算量 3 个方面对算法的性能进行了评估及分析。实验环境及参数设置如表 3 所列。

表 3　实验环境及参数设置

	参数	参考值
硬件环境	CPU	Intel(R) Core(TM) i5-3320M
	主频	2.6GHz
	内存	6GB
软件环境	OS	Windows 7 Ultimate
	软件	MATLAB 7.0
仿真参数	d_c	0.033
	minPts	4
	Eps	0.3

所采用数据集的属性如表 4 所列。

表 4　数据集属性

数据集名称	简记	样本数	类别数	维度	来源
Gaussian Density Peaks	GDP	1000～4000	5	2	文献[14]
Aggregation	AG	788	6	2	文献[15]
Iris	IR	150	4	3	UCI
Pima Indians Diabetes	PID	768	2	8	UCI
Waveform	WF	5000	3	21	UCI

3.2　准确度分析

为评估算法的准确度，本文从聚类中心选取准确度及聚类效果准确度两个方面进行了分析。

(1)聚类中心选取准确度

聚类中心选取准确度分别选用无标签 GDP 数据集和有标签数据集进行实验。首先，以 2000 点的 GDP 数据集为例，分别采用 CFSFDP 算法及引入簇中心选取策略的改进算法对聚类误差平方和进行对比。图 5－图 7 表示聚类中心数分别为 4、5、6 时的聚类效果。

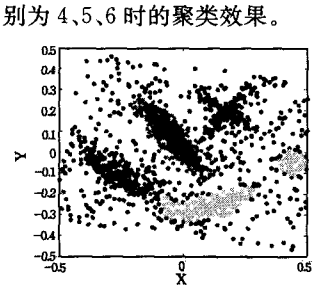

图 5　4 类簇聚类效果

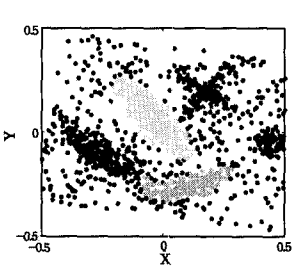

图 6　5 类簇聚类效果

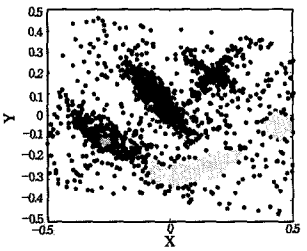

图 7　6 类簇聚类效果

为进一步量化簇中心数对聚类效果的影响，表 5 给出了在 GDP 不同点数的情况下聚类误差平方和对比。

表 5　聚类中心选取方法的误差对比

	聚类中心数目	1000 点	2000 点	4000 点
决策图选取法	4	12.75	23.14	48.60
	5	8.71	17.69	41.38
	6	10.04	21.78	42.57
策略选取法	5	8.71	17.69	41.38

由图 5－图 7 及表 5 可以看出，在聚类中心数为 5 时，聚类的误差平方和最小，这说明了聚类中心数的最优取值为 5；而采用了聚类中心选取策略的 CFSFDP 算法可自动确定最优的聚类中心数目。

采用 4 类带标签的数据集进一步验证该策略对其他数据集的有效性，统计结果如表 6 所列。

表 6　策略选取中心数目

	AG	IR	PID	WF
簇类别数	6	4	2	3
策略选取的中心	6	4	2	3

由表 6 可看出，对于不同类型的数据集，采用策略选取的 CFSFDP 算法能够有效给出最优的簇数目，这使得 CFSFDP 算法不再需要通过决策图的方法进行人工选择，从而有效减少了由决策图判断带来的主观误差。

(2)聚类准确率及 F-measure

为考察改进的 CFSFDP 与传统的 DBSCAN 算法的聚类准确率，本文采用 4 类有标签测试集，进行了 100 次蒙特卡罗实验并统计了聚类准确率及 F-measure[16]，实验结果如图 8、图 9所示。

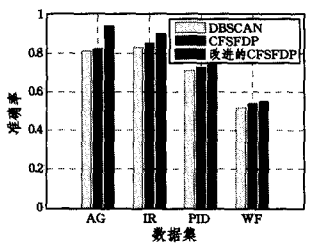

图 8　算法的准确率对比

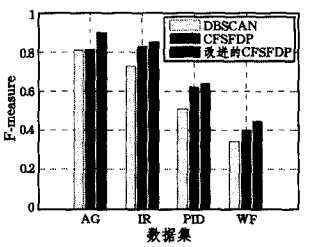

图 9　算法的 F-measure 对比

由图 8、图 9 可看出，CFSFDP 及其改进算法在准确度及 F-measure 方面都明显高于 DBSCAN 算法，说明了基于密度峰值的聚类算法的优越性：CFSFDP 算法以密度峰值作为聚类中心，仅凭相邻距离及密度即可完成簇的划分。而 DBSCAN 算法需要通过判断核心点、边界点及噪声点来划分簇，边界设置不合理会造成误判从而易致簇聚类准确度降低。标准的 CFSFDP 算法的准确度及 F-measure 与 DBSCAN 算法相当，但略低于改进后的 CFSFDP 算法的，这是由于人工辅

助决策的主观性导致并非每次聚类都能选择出合适的簇中心，从而使平均准确度变差。而改进后的算法能够自动选取合适的簇中心点，从而具有较高的准确度。

3.3 鲁棒性分析

为验证改进的 CFSFDP 算法对噪声的抗干扰能力，以 AG 实验数据集为例，分别加入聚类样本总数的 0%、5%、10%、15%人工合成噪声，采用准确率及 F-measure 两种度量方法，并与传统的 DBSCAN 及标准的 CFSFDP 进行对比。其中，噪声由以下两类等比例高斯噪声合成，其参数设置如表 7 所列。

表 7 噪声参数

	1	2
均值 μ	$\mu_x=15,\mu_y=15$	$\mu_x=25,\mu_y=25$
方差 σ^2	15	15

进行 100 次蒙特卡罗实验，统计实验结果均值如图 10、图 11 所示。

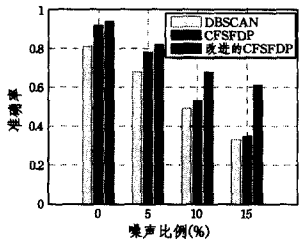

图 10 不同噪声比例下算法的准确率对比

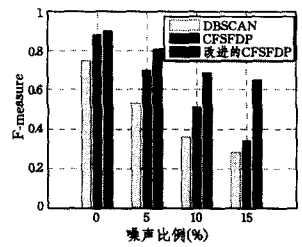

图 11 不同噪声比例下算法的 F-measure 对比

由图 10、图 11 可以看出，两种算法的准确率及 F-measure随着引入噪声的比例的增加而降低。但无论引入噪声还是不引入噪声，改进的 CFSFDP 算法的准确均比 DBSCAN 及标准的 CFSFDP 率更高，且随着噪声比例的增加，改进的 CFSFDP 的优势更加明显。

这说明相比 DBSCAN，改进的 CFSFDP 算法对噪声的抗干扰能力更强。这主要是由于改进的 CFSFDP 算法无需设置其他参数即可进行自动聚类，而 DBSCAN 算法对参数设置更加敏感，*Eps* 及 *minPts* 等参数设置不当更容易导致其准确率降低。相比于标准的 CFSFDP，改进的算法能够通过搜索“拐点”找到差异度最大的一组点，避免了人工选择簇中心点的过程受到噪声干扰的问题，因此具有更强的鲁棒性。

3.4 计算量分析

为考察改进算法的复杂度，分析数据集维度对聚类算法计算量的影响，采用 4 类数据集进行 100 次蒙特卡罗实验。CFSFDP 算法在进行聚类中心点选择时需通过人工判别，这使得算法运行时间带有一定客观性。因此，仅对 DBSCAN 及改进的 CFSFDP 算法的平均运行时间进行对比，仿真结果如图 12 所示。

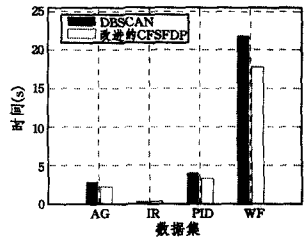

图 12 聚类运算时间对比

由图 12 可以看出，改进后的算法的计算量较 DBSCAN 略低，这主要是由于 DBSCAN 算法需要根据参数遍历点集并标注核心点、边界点或噪声点，而 CFSFDP 改进算法仅需要计算一次两点间的距离即可，在判断聚类中心时，取前 m 个点计算。另外，DBSCAN 算法在确定参数的过程也较为繁琐，首先确定 $minPts=k$，然后根据不同的 $Dist_{minPts}$ 确定最优 *Eps*。上述因素都降低了 DBSCAN 的运算效率。

尽管改进的算法引入了聚类中心判别和选择的过程，但其计算量并未大幅增加。同时，应注意到算法的计算量不仅与样本数目有关，还会因样本维度的增加而急剧增加。由此可见，该算法适用于对较低维度数据集（如空间位置数据）的处理。另外，值得说明的是，该统计是在本文既定的仿真环境下给出的蒙特卡罗实验结果，MATLAB 仿真带有一定条件性，并不能如实反映应用中的运行时间，仅供相同条件下不同算法参考对照。

结束语 本文介绍了一种基于密度峰值的快速聚类算法，针对该算法存在的无法进行聚类中心自动判别的问题，提出了通过判断分离度变化趋势来确定聚类中心数目的策略，并将该策略与 CFSFDP 算法结合。这一改进使得该算法无需输入其他参数及人工辅助即可实现自动聚类。仿真实验采用 5 类数据集，分别从准确度、鲁棒性、计算量 3 个方面对算法进行了分析。结果表明，相比于 DBSCAN 及 CFSFDP 算法，基于聚类中心选择策略的 CFSFDP 改进算法具有较高的准确率和鲁棒性且能够适应不同的噪声环境。计算量方面，该算法比 DBSCAN 算法略高，适用于较低维度的数据处理，对具有自然簇属性、任意形状簇的数据的聚类问题具有一定的参考意义。

参 考 文 献

[1] Jain A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666

[2] Kriegel H P, Kröger P, Sander J, et al. Density-based clustering [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(3): 231-240

[3] Achtert E, Goldhofer S, Kriegel H P, et al. Evaluation of clusterings—metrics and visual support[C]// 2012 IEEE 28th International Conference on Data Engineering (ICDE). IEEE, 2012: 1285-1288

[4] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis(344 ed)[M]. John Wiley & Sons, 2009

[5] Li Cui-xia, Shi Wei-hang, Li Zhan-bo. Density Based Weighted Fuzzy Clustering Algorithm[J]. Computer Science. 2012, 39(5): 180-182(in Chinese)

（下转第 280 页）

1316,1333(in Chinese)
韩俊英,刘成忠. 自适应混沌果蝇优化算法[J]. 计算机应用,2013,33(5):1313-1316,1333

[10] Liu Cheng-zhong, Han Jun-ying. Adaptive fruit fly optimization algorithm based on bacterial migration[J]. Engineering and Computer Science,2014,36(4):690-696(in Chinese)
刘成忠,韩俊英. 基于细菌迁徙的自适应果蝇优化算法[J]. 计算机工程与科学,2014,36(4):690-696

[11] Chang Peng, Li Shu-rong, Ge Yu-lei, et al. Fruit fly optimization algorithm with self adapting adjustment of iteration step value [J]. Computer Engineering and Applications,2014,1(1):1-6(in Chinese)
常鹏,李树荣,葛玉磊,等. 迭代步进值自适应调整的果蝇优化算法[J]. 计算机工程与应用,2014,1(1):1-6

[12] Ma Chao, Dong Ling. Fruit flies optimization algorithm (FOA) improved step length and its multiple function optimization method[J]. Mathematics Learning and Research,2013,1(13):90-92(in Chinese)
马超,董玲. 果蝇优化算法(FOA)步长改进及其多元函数最优化方法[J]. 数学学习与研究,2013,1(13):90-92

[13] Ning Jian-ping, Wang Bing, Li Hong-ru, et al. Research on and application of diminishing step fruit fly optimization algorithm [J]. Journal of Shenzhen University Institute of Technology,2014,31(4):367-373(in Chinese)
宁剑平,王冰,李洪儒,等. 递减步长果蝇优化算法及应用[J]. 深圳大学学报理工版,2014,31(4):367-373

[14] Xu Guo-bing, Han Wen-wen. Study on vibration responses of powerhouse structures based on FOA-GRNN[J]. Journal of Hydroelectric Power,2014,33(6):187-191(in Chinese)
徐国宾,韩文文. 基于 FOA-GRNN 的水电站厂房结构振动响应研究[J]. 水力发电学报,2014,33(6):187-191

[15] Li Shu-ling, Liu Rong, Liu Hong. Multi-label Learning for Improved RBF Neural Networks[J]. Computer Science, 2015, 42(4):316-320(in Chinese)
李书玲,刘蓉,刘红. 改进型 RBF 神经网络的多标签算法研究[J]. 计算机科学,2015,42(4):316-320

[16] Wang Yu-fei, Shen Hong-yan. Network security situation forecastbased on improved general regression neuralnetwork[J]. Journal of North China Electric Power University,2011,38(3):91-95(in Chinese)
王宇飞,沈红岩. 基于改进广义回归神经网络的网络安全态势预测[J]. 华北电力大学学报,2011,38(3):91-95

[17] Zhou Ping, Bai Guang-chen. Robust design of turbine-blade low cycle fatigue life based on neural networks and fruit fly optimization algorithm[J]. Journal of Air Power,2013,28(5):1013-1018(in Chinese)
周平,白广忱. 基于神经网络与果蝇优化算法的涡轮叶片低循环疲劳寿命健壮性设计[J]. 航空动力学报,2013,28(5):1013-1018

[18] Shen Zhang-quan, Zhou Bin, Kong Fan-sheng, et al. Study On Spatial Variety of Soil Properties by Means of Generalized Regression Neural Network[J]. Journal of Soil,2004,41(3):471-475(in Chinese)
沈掌泉,周斌,孔繁胜,等. 应用广义回归神经网络进行土壤空间变异研究[J]. 土壤学报,2004,41(3):471-475

[19] Pan Wen-chao. Application of fruit fly optimization algorithm to optimize the generalized regression neural network to enterprise operating performance evaluation[J]. Journal of Taiyuan University of Technology,2011,29(4):1-5(in Chinese)
潘文超. 应用果蝇优化算法优化广义回归神经网络进行企业经营绩效评估[J]. 太原理工大学学报,2011,29(4):1-5

[20] Lin Hai-ming, Du Zi-fang. Some Problems in Comprehensive Evaluation in the Principal Component Analysis[J]. Statistical Research,2013,30(8):25-31(in Chinese)
林海明,杜子芳. 主成分分析综合评价应该注意的问题[J]. 统计研究,2013,30(8):25-31

(上接第 258 页)

李翠霞,史苇杭,李占波. 一种基于密度的加权模糊均值聚类算法[J]. 计算机科学,2012,39(5):180-182

[6] Parimala M, Lopez D, Senthilkumar N C. A survey on density based clustering algorithms for mining large spatial databases [J]. International Journal of Advanced Science and Technology,2011,31(1):59-66

[7] Braune C, Besecke S, Kruse R. Density Based Clustering: Alternatives to DBSCAN[M]. Springer International Publishing,2015

[8] Kisilevich S, Mansmann F, Keim D. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos[C]// Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application. ACM,2010:591-598

[9] Kumar N, Sivasathya S. Density-Based Spatial Clustering with Noise-A Survey[J]. International Journal of Computer Science and Mobile Computing,2014,3(3):1004-1011

[10] Zhou H, Wang P, Li H. Research on Adaptive Parameters Determination in DBSCAN Algorithm[J]. Journal of Information & Computational Science,2012,9(7):1967-1973

[11] Smiti A, Elouedi Z. DBSCAN-GM: An improved clustering method based on Gaussian Means and DBSCAN techniques[C]// 2012 IEEE 16th International Conference on Intelligent Engineering Systems (INES). IEEE,2012:573-578

[12] Zhang Li-jie. Stable saturation clensity of DBSCAN algorithm [M]. Application Research of Computers,2014(7):1972-1975 (in Chinese)
张丽杰. 具有稳定饱和度的 DBSCAN 算法[J]. 计算机应用研究,2014(7):1972-1975

[13] Tran T N, Drab K, Daszykowski M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters[J]. Chemometrics and Intelligent Laboratory Systems,2013,120:92-96

[14] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496

[15] Gionis A, Mannila H, Tsaparas P. Clustering aggregation[J]. ACM Transactions on Knowledge Discovery from Data (TKDD),2007,1(1):341-352

[16] Powers D M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation[J]. Journal of Macline Learning Technologies,2008,2:2229-3981