ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2012/06(10)-0903-09
DOI: 10.3778/j.issn.1673-9418.2012.10.005

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

基于数据场的改进DBSCAN聚类算法*

杨 静[1,2+],高嘉伟[1,2],梁吉业[1,2],刘杨磊[1,2]
1. 山西大学 计算智能与中文信息处理教育部重点实验室,太原 030006
2. 山西大学 计算机与信息技术学院,太原 030006

An Improved DBSCAN Clustering Algorithm Based on Data Field*

YANG Jing[1,2+], GAO Jiawei[1,2], LIANG Jiye[1,2], LIU Yanglei[1,2]
1. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China
2. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
+ Corresponding author: E-mail: morgan1127@126.com

YANG Jing, GAO Jiawei, LIANG Jiye, et al. An improved DBSCAN clustering algorithm based on data field. Journal of Frontiers of Computer Science and Technology, 2012, 6(10): 903-911.

Abstract: DBSCAN (density based spatial clustering of applications with noise) algorithm is a typical density-based clustering algorithm. The algorithm can discover the arbitrary-shaped clusters. However, the clustering results depend on the two parameters *Eps* and *MinPts* which are chosen by users. And for some datasets with large density differences, either the clustering results may have the incorrect cluster number, or the algorithm may label part of the data as noise. Using the advantages that data field can commendably describe the data distribution and reflect the data relationship, this paper proposes a new clustering algorithm called improved DBSCAN algorithm based on data field. The algorithm introduces the concept of average potential difference and dynamically determines *Eps* and average potential difference of each class during the clustering process. In this way, it can receive better clustering results for some clusters with large density differences. Experimental results indicate that the proposed algorithm performs better than DBSCAN algorithm.

Key words: DBSCAN algorithm; data field; clustering

* The Special Prophase Project on the National Grand Basic Research 973 Program of China under Grant No. 2011CB311805 (国家重点基础研究发展规划(973)前期研究专项); the Key Problems in Science and Technology Project of Shanxi Province under Grant No. 20110321027-01 (山西省科技攻关计划项目); the Construction Project of the Science and Technology Basic Condition Platform of Shanxi Province under Grant No. 2012091002-0101 (山西省科技基础条件平台建设项目).
Received 2012-05, Accepted 2012-07.

摘 要:DBSCAN(density based spatial clustering of applications with noise)算法是一种典型的基于密度的聚类算法。该算法可以识别任意形状的类簇,但聚类结果依赖于参数*Eps*和*MinPts*的选择,而且对于一些密度差别较大的数据集,可能得不到具有正确类簇个数的聚类结果,也可能将部分数据错分为噪声。为此,利用数据场能较好描述数据分布,反映数据关系的优势,提出了一种基于数据场的改进DBSCAN聚类算法。该算法引入平均势差的概念,在聚类过程中动态地确定每个类的*Eps*和平均势差,从而能够在一些密度相差较大的数据集上得到较好的聚类结果。实验表明,所提算法的性能优于DBSCAN算法。

关键词:DBSCAN算法;数据场;聚类

文献标识码:A **中图分类号:**TP18

1 引言

聚类是数据挖掘的主要方法之一,其主要目的是根据数据对象之间的相似性将数据集分割成不同的类簇,使得类内相似性最大,类间相似性最小[1]。目前,根据聚类技术进行分类,主要存在的聚类算法大致可分为以下几类:

(1)基于划分方法的聚类算法;

(2)基于层次方法的聚类算法;

(3)基于网格的聚类算法;

(4)基于密度的聚类算法。

DBSCAN(density based spatial clustering of applications with noise)算法[2]是一种典型的基于密度的聚类算法。该算法将类簇定义为密度可达的数据对象组成的集合,可以发现任意形状的类簇,且对噪声数据不敏感。该算法需要输入类的半径*Eps*和类内最小数据对象个数*MinPts*两个参数,而且其聚类结果的优劣依赖于用户对这两个参数的选择。DBSCAN算法给出利用*K*-dist图来粗略估算*Eps*的方法,并在二维属性下,取*MinPts*=4;也提出交互式给定*Eps*和*MinPts*的方法[2]。然而在聚类过程中,*Eps*和*MinPts*都是恒定的,这样对于密度差别较大的数据集,便得不到很好的聚类结果。

数据场理论[3]是一种利用物质粒子间的相互作用及其场描述方法来刻画抽象数域空间的理论。该理论可以用势值较好地描述数据分布,并反映数据间多对一的作用关系,这对聚类算法中一些参数的确定有着较好的指导作用[4-8]。

本文借鉴数据场思想,以DBSCAN算法为原型,提出了一种基于数据场的改进DBSCAN聚类算法。该算法模拟数据场中利用等势线划分数据的聚类过程,引入新的参数平均势差$\nabla\varphi$,用*Eps*和$\nabla\varphi$来指导聚类过程。在聚类过程中,通过计算得到*Eps*和$\nabla\varphi$,不仅避免了用户指定*Eps*的问题,而且适用于密度不均的数据的聚类。

本文组织结构如下:第2章介绍DBSCAN算法和数据场的相关内容;第3章详细描述算法;第4章通过实验验证算法的有效性;最后总结全文。

2 DBSCAN算法和数据场

2.1 DBSCAN算法

DBSCAN算法[2]是一种经典的基于密度的聚类算法,该算法计算每个数据对象的*Eps*邻域,通过把密度可达的数据对象聚成一个类簇来得到聚类结果。DBSCAN算法可以自动确定类簇的个数,发现任意形状的类簇,且对噪声数据不敏感。DBSCAN中的定义如下[2]:

定义1(数据对象p的*Eps*邻域)数据对象$\forall p\in D$的*Eps*邻域$N_{Eps}(p)$定义为以p为核心,*Eps*为半径的d维超球体区域内包含的点的集合,即$N_{Eps}(p)=\{q\in D|dist(p,q)\leqslant Eps\}$,其中,$D$为$d$维空间上的数据集,$dist(p,q)$表示$D$中点$p$和$q$之间的距离。

定义2(核心数据对象)给定*Eps*和*MinPts*,对于数据对象p,如果p的*Eps*邻域内包含的对象个数满足$\left|N_{Eps}(p)\right|\geqslant MinPts$,则称$p$为核心数据对象。

定义3(直接密度可达)给定*Eps*和*MinPts*,对于数据对象$p,q\in D$,如果p满足$p\in N_{Eps}(q)$且$\left|N_{Eps}(q)\right|\geqslant MinPts$这两个条件,则称$p$是从$q$关于(*Eps*,*MinPts*)直接密度可达的。直接密度可达不满足对称性。

定义4（密度可达）给定*Eps*和*MinPts*，对于数据对象 $p, q \in D$，如果有一个数据对象序列 $p_1, p_2, \ldots, p_n \in D$，其中 $p_1=q, p_n=p$，并且 p_{i+1} 是从 p_i 直接密度可达的，则称 p 是从 q 关于 (*Eps*, *MinPts*) 密度可达的。密度可达也不满足对称性。

定义5（密度相连）给定*Eps*和*MinPts*，对于数据对象 $p, q \in D$，如果存在一个数据对象 o 使得 p 和 q 都是从 o 密度可达的，则称 p 和 q 是关于 (*Eps*, *MinPts*) 密度相连的。密度相连满足对称性。

当给定*Eps*和*MinPts*时，DBSCAN算法的简要流程如下[2]：选择任一未划分的数据对象，判断其是否为核心数据对象，若是，寻找所有与其密度可达的数据对象，将这些数据对象标记为一类；若不是，则进行噪声数据对象判断，若是噪声，则对其进行标记，若不是噪声，则不对该对象进行处理。如此重复，直至所有的数据对象都被划分。

DBSCAN算法聚类结果的优劣不仅依赖于用户对参数*Eps*和*MinPts*的选择，而且由于参数*Eps*和*MinPts*在聚类过程中不发生改变，这样使得算法对于密度不均的数据集，可能会出现以下两种情况：把一些数据错分为噪声；得到错误的聚类个数。这样就得不到很好的聚类结果。

2.2　数据场

场的概念最早是1837年由英国物理学家法拉第提出的，是物体间非接触相互作用的传递媒质。如今在物理学中，场论已经成为比较成熟的理论和方法。随着场论思想的发展，人们将其抽象为一个数学概念，用来描述某个物理量或数学函数在空间内的分布规律。数据场理论[3]借鉴物理学中场的思想，将物质粒子间的相互作用及其场描述方法引入到抽象的数域空间中。该理论克服了传统算法中只考虑数据对象间一对一作用关系的弊端，用势函数来描述数据对象间的相互关系，认为任一数据对象的状态是其他数据对象共同作用的结果，并利用数据势场中等势线（面）的自然嵌套结构来实现数据对象的划分。

考虑到高斯分布的普适性及短程场更适合反映数据分布，数据场将势函数定义如下[3]：已知数据空间 $\Omega \subseteq R^d$ 上的对象集 $D=\{x_1, x_2, \ldots, x_n\}$ 及其产生的数据场，$m_i \geqslant 0$ 是对象 $x_i (i=1, 2, \ldots, n)$ 的质量，即对象 x_i 对其他对象的影响程度，一般情况下认为各个对象的影响程度是相等的，可以令 $m_i=1$；$\sigma \in (0, +\infty)$ 是用来控制对象间的相互作用力程，称为影响因子。则任一场点 x 处的势值可以定义为所有对象在该点处产生的势值的叠加：$\varphi(x)=\sum_{i=1}^{n}\left(m_i \cdot \mathrm{e}^{-\left(\frac{\|x-x_i\|}{\sigma}\right)^2}\right)$。

为了说明单数据对象产生的数据场中场强与距离的分布关系，不妨令 $m_i, \sigma=1$，由此可得到图1。由图1可知，在距离数据对象0.705处，场强达到了最高值，即在以场源对象为中心，半径为0.705的球面上存在很强的指向场源对象的作用力。结合数据具有自组织聚集的特性，可以认为在数据对象0.705邻域内的数据对象同属于一类的可能性较大，详细说明见文献[3]。在下文计算*Eps*和 $\nabla\varphi$ 时会用到这一内容，文中不再赘述。

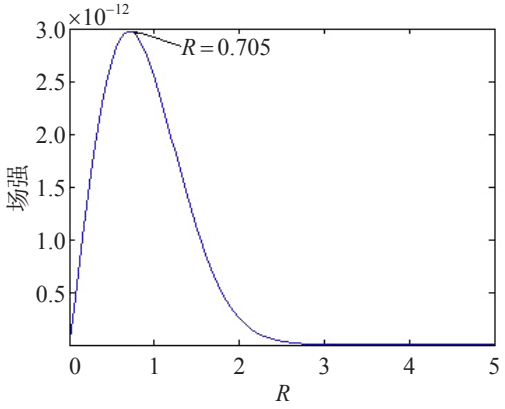


Fig.1　The distribution relationship between field intensity function and distance in the data field of single point

图1　单数据对象数据场中场强函数与距离的分布关系

数据场理论具有较好的刻画数据分布，反映数据间多对一的作用关系的优势。该优势对聚类算法中一些参数的确定有着较好的指导作用。

3　基于数据场的改进DBSCAN聚类算法

3.1　相关定义及计算

（1）数据对象 x_i 的 R 邻域 $N_R(x_i)$

对于数据对象 $x_i \in D$，其 R 邻域 $N_R(x_i)$ 是以 x_i 为核心，R 为半径的 d 维超球体区域内包含的点的集合，即 $N_R(x_i)=\{x_j | dist(x_i, x_j) \leqslant R, x_i, x_j \in D\}$。其中，

D 为 d 维空间上的数据集，$dist(x_i, x_j)$ 表示 D 中数据对象 x_i 和 x_j 之间的距离。

(2)*Eps* 的计算

对于数据对象 $x_i \in D$，由 x_i 确定的 *Eps* 计算方法如下：先计算 $N_{0.705\sigma}(x_i)$ 中每个数据与其第 *MinPts* 近的数据对象的距离，最后将这些距离的平均值赋给 *Eps*。

(3) $\nabla\varphi$

为了将数据场与 DBSCAN 算法融合，本文引入平均势差 $\nabla\varphi$ 来辅助 DBSCAN 聚类。定义 $\nabla\varphi$ 是邻域 R 内任意两个数据对象的势差的平均值。从等势线的分布可以看出，如果两个数据对象间的势差较小，那这两个数据对象同属一类的可能性较大。

对于数据对象 $x_i \in D$，由 x_i 确定 $\nabla\varphi$ 的计算方法如下：先计算 $N_{0.705\sigma}(x_i)$ 中任意两个数据的势差的绝对值，然后将这些势差的平均值作为 $\nabla\varphi$ 的值，其形式化描述为 $\nabla\varphi=\frac{1}{M^2}\sum_{x_p\in N_{0.705\sigma}(x_i)}\sum_{x_q\in N_{0.705\sigma}(x_i)}\left|\varphi(x_p)-\varphi(x_q)\right|$，其中 M 表示集合 $N_{0.705\sigma}(x_i)$ 所包含的元素个数，即 $M=\left|N_{0.705\sigma}(x_i)\right|$。

(4)噪声

如果某个数据对象的 0.705σ 邻域内没有其他数据，或者满足由该数据确定的 *Eps* 和 $\nabla\varphi$ 这两个指标得到的数据集包含的数据个数小于 *MinPts*，那么则认为该数据是噪声。

3.2 算法描述

输入：数据集 $D=\{x_1, x_2, \ldots, x_n\}$，类内最小对象数 *MinPts*。

输出：聚类结果。

步骤1 初始化未处理数据对象集合 $U=D$，临时存放属于某一类的数据对象集合 $A=\varnothing$。

步骤2 计算各数据对象的势值 $\varphi(x_j)=\sum_{i=1}^{n}\left(m_i\cdot e^{-\left(\frac{\|x_j-x_i\|}{\sigma}\right)^2}\right)$，$x_i, x_j\in D$，其中 $m_i=1(i=1,2,\ldots,n)$。

步骤3 根据势值为极值的数据对象，寻找与其同属于一类的数据对象。

步骤3.1 确定 x_p，并将 x_p 放入集合 A。x_p 满足条件 $\varphi(x_p)=\max(\varphi(x_i))$，$x_i, x_p\in U$ 且 x_p 不为噪声。

步骤3.2 计算由 x_p 确定的 *Eps* 和 $\nabla\varphi$。

步骤3.3 对于集合 A 中的任一元素 A_i，$i=1, 2, \ldots, |A|$，计算 A_i 确定的 *Eps* 邻域 $N_{Eps}(A_i)$，对于 $\forall x_q\in N_{Eps}(A_i)$，若 x_q 满足条件 $\exists x_s\in N_{Eps}(x_q)\cap A$ 使得 $\left|\varphi(x_s)-\varphi(x_q)\right|\leqslant\nabla\varphi$，则将 x_q 放入 A 中。

步骤4 将 A 中数据对象划分为一类，$U=U-A$，若 $U\neq\varnothing$，令 $A=\varnothing$，转步骤3，否则输出聚类结果。

4 实验结果

为了验证算法的有效性，本文分别在人工数据集和UCI数据集上进行了对比实验。

4.1 在人工数据集上的实验结果

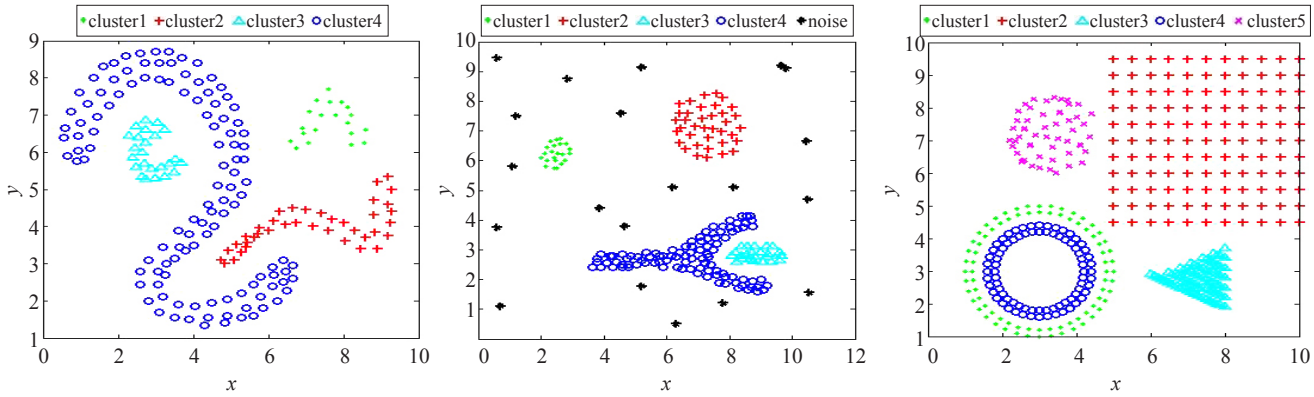
测试数据集如图2所示，其中Dataset1和Dataset2来自文献[2]，不同类的密度差距较小；Dataset3是人工构造的数据集，不同类的密度差距较大。Dataset1包含200个数据，由4个不规则的类构成，有S型、V型、C型及对勾型；Dataset2也包含200个数据，包含4个类，1个Y型、1个椭圆形及2个大小不同的圆形类，并且数据集含有10%的噪声数据；Dataset3包含420个数据点，数据分布呈现5类，包含3个不同大小、形状、密度的类和2个相嵌套的环形类。

本文通过将提出的算法与标准 *K*-means 算法[9]、数据场聚类算法[3]和DBSCAN算法[2]进行比较，从而验证算法的有效性。

(1)Dataset1的实验结果

如图3所示：*K*-means算法不能有效地区分这些不规则的S型、V型、C型、对勾型的类。利用等势线嵌套结构得到聚类结果的数据场聚类算法（自适应得到 $\sigma=4.25$）可以较好地划分这些不规则的类。其中图3(b)的坐标是标称距离[3]，下文中由数据场聚类算法得到的实验图的坐标也都是标称距离，以下不再赘述。DBSCAN算法在选择合理的 *Eps* 和 *MinPts* 的情况下，可以较好地划分这些不规则的类。本文算法只需输入 *MinPts*，就可以将这些不规则的类簇合理地区分，其聚类效果与数据场聚类算法和DBSCAN算法相当。

(2)Dataset2的实验结果

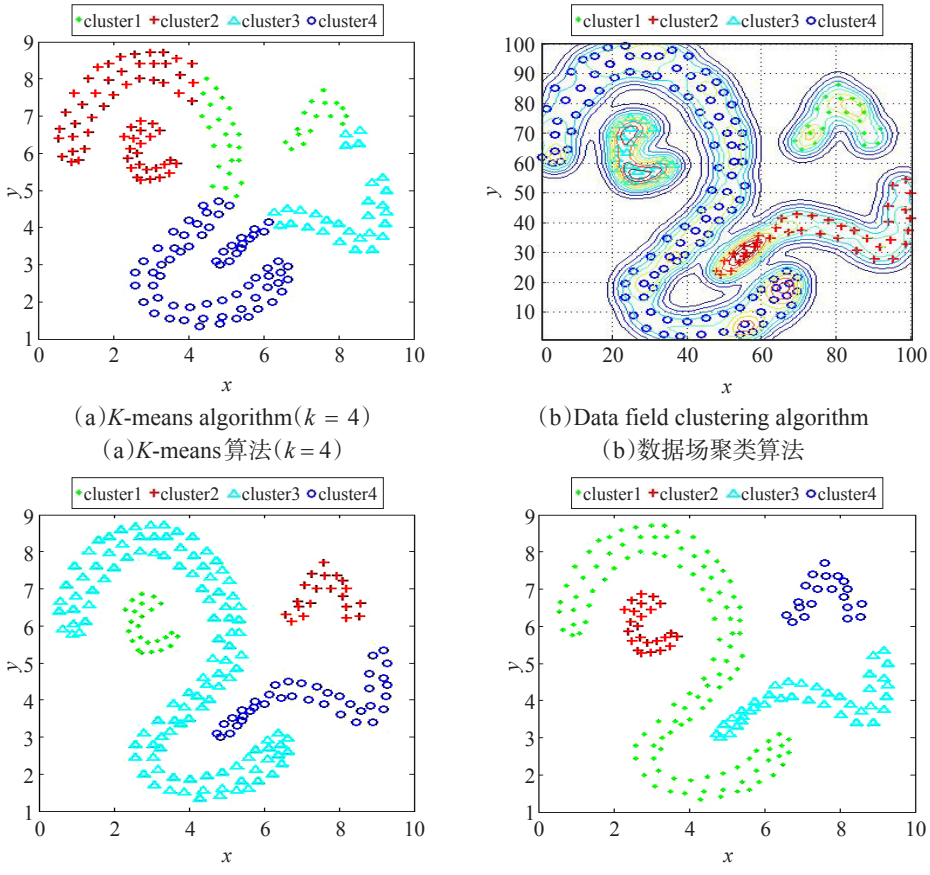


（a）Dataset1　（b）Dataset2　（c）Dataset3

（a）数据集1　（b）数据集2　（c）数据集3

Fig.2　Datasets

图2　测试数据集



（a）*K*-means algorithm（$k=4$）　（b）Data field clustering algorithm

（a）*K*-means算法（$k=4$）　（b）数据场聚类算法

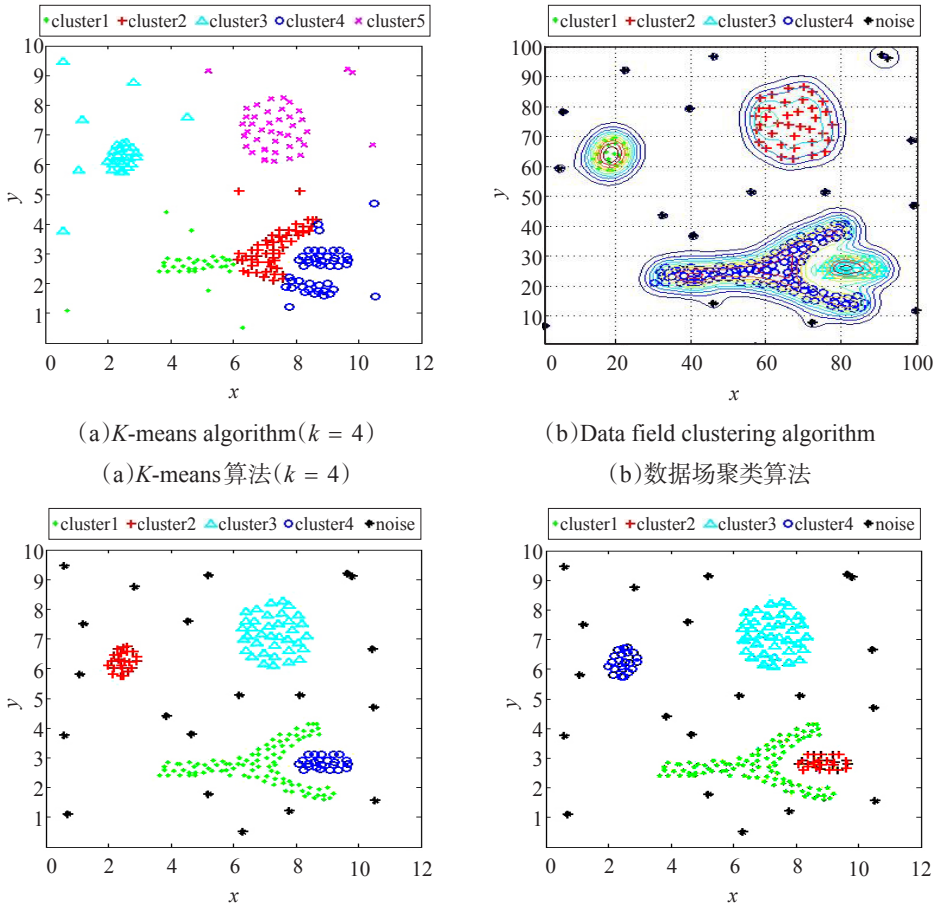
（c）DBSCAN algorithm（$Eps=0.75$，$MinPts=4$）　（d）Proposed algorithm（$MinPts=4$）

（c）DBSCAN算法（$Eps=0.75$，$MinPts=4$）　（d）本文算法（$MinPts=4$）

Fig.3　The summary results of four algorithms on Dataset1

图3　四个聚类算法在Dataset1上的聚类结果

如图4所示：*K*-means算法不仅对不规则形状类的聚类效果不太理想，而且对噪声数据较为敏感，噪声数据会影响类中心的选择，导致聚类结果不佳。数据场聚类算法（自适应得到$\sigma=4.41$）可以较好地处



(a)*K*-means algorithm(k = 4)　(b)Data field clustering algorithm

(a)*K*-means算法(k = 4)　(b)数据场聚类算法

(c)DBSCAN algorithm(Eps=0.5, $MinPts$=4)　(d)Proposed algorithm($MinPts$ = 4)

(c)DBSCAN算法(Eps=0.5, $MinPts$=4)　(d)本文算法($MinPts$ = 4)

Fig.4 The summary results of four algorithms on Dataset2

图4 四个聚类算法在Dataset2上的聚类结果

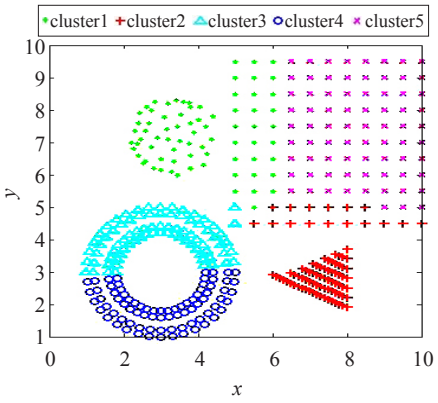
理不规则形状类,但对于一些相距较近的类不能很好地区分,也容易把一些距离类簇较近的噪声数据划分到类中。DBSCAN算法在选择合理的 *Eps* 和 *MinPts* 的情况下,既能较好地划分不规则形状类,也能较好地将噪声数据区分出来。本文算法只需输入 *MinPts* ,便可以在较好划分不规则形状类的同时,也可以较好地区分噪声数据,其聚类效果与DBSCAN算法相当。

(3)Dataset3的实验结果

如图5所示:*K*-means算法不能有效地区分环形数据和密度差别较大的数据集。数据场聚类算法(自适应得到 σ=4.82)不能有效区分环形数据,而且对于相距较近但密度差别较大的数据不能较好地区分,较容易将这些数据划分为一类。DBSCAN算法通过选择合适的 *Eps* 和 *MinPts* 可以得到图5(c)中的两种结果。其中图(c1)的聚类结果拥有正确的聚类个数,但是将右上角的密度较小的方形类以及左上角的圆形类中的一部分数据划分成噪声数据,由于在这种情况下,选择的 *Eps* 较小,DBSCAN算法将左上角同属于圆形类的数据划分成了2类;图(c2)的聚类结果有着较高的聚类精度,但是类簇个数是错误的,将5类问题错分为4类,将2个环形数据划分为一类。本文算法只需给定 *MinPts* 就可以较好地划分包含不同大小、形状且密度相差较大的数据集,从而得到较好的聚类结果。

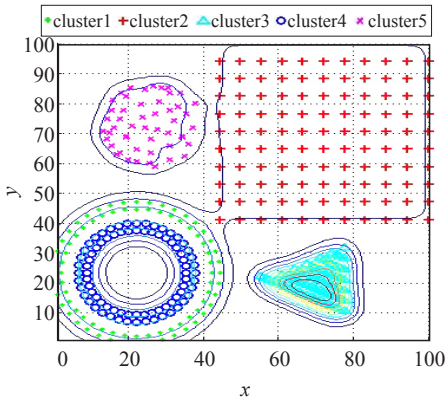
4.2 在UCI数据集上的实验结果

从UCI数据集中挑选了6组数据Iris、Wine、Der-



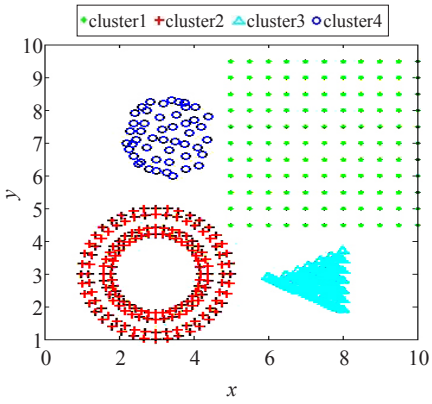
(a)*K*-means algorithm(k = 5)

(a)*K*-means算法(k = 5)



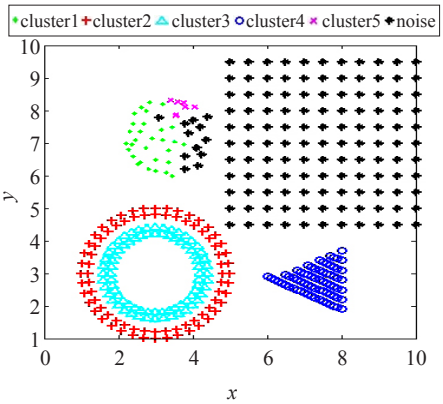
(b)Data field clustering algorithm

(b)数据场聚类算法



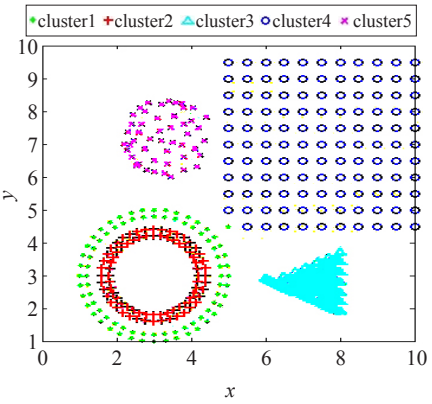
(c1)DBSCAN algorithm

(*Eps* = 0.36, *MinPts* = 4)



(c2)DBSCAN algorithm

(*Eps* = 0.5, *MinPts* = 4)



(d)Proposed algorithm(*MinPts* = 4)

(d)本文算法(*MinPts* = 4)

(c1)DBSCAN算法(*Eps* = 0.36, *MinPts* = 4)　(c2)DBSCAN算法(*Eps* = 0.5, *MinPts* = 4)

Fig.5　The summary results of four algorithms on Dataset3

图5　四个聚类算法在Dataset3上的聚类结果

matology、Breast、Contraceptive Method Choice和Pendigits，数据描述如表1所示。由于所选数据集的属性个数较多，不便于数据场等势线的可视化显示，在此仅将本文算法同标准*K*-means[9]和DBSCAN算法[2]进行比较。考虑到*K*-means算法初始类中心选择的随机性，文中*K*-means的聚类指标均选择20次聚类结果的平均值。表2～表7分别给出了在上述6个数据集上各聚类算法的性能比较。文中采用多个评价指标对聚类结果进行分析评价，指标有分类正确率(AC)、Rand Index(RI)和Adjusted Rand Index(ARI)。其中AC、RI和ARI的计算公式分别与文献[10-12]定义的一致，其值域分别为[0, 1]、[0, 1]和[−1, 1]，其取值越大，聚类效果越好。AC用来衡量聚类结果的正确率，RI和ARI均是用来衡量聚类结果与数据原始类标签之间的一致程度。

Table 1　The characteristics of datasets

表1　数据集描述

数据集	样本数	属性数
Iris	150	4
Wine	178	13
Dermatology	366	33
Breast	699	9
Contraceptive Method Choice	1 473	9
Pendigits	3 498	16

通过比较表2～表7可以看出，本文算法在Iris和Dermatology数据集上的聚类结果较好，均优于

Table 2 The summary results of three algorithms on Iris dataset

表2 在Iris下算法的性能比较

算法	AC	RI	ARI
K-means	0.845 0	0.844 1	0.662 1
DBSCAN	0.693 3	0.772 4	0.515 8
本文算法	0.960 0	0.949 5	0.885 8

Table 3 The summary results of three algorithms on Wine dataset

表3 在Wine下算法的性能比较

算法	AC	RI	ARI
K-means	0.939 9	0.886 9	0.771 8
DBSCAN	0.655 2	0.537 5	−0.012 8
本文算法	0.915 6	0.845 2	0.689 9

Table 4 The summary results of three algorithms on Dermatology dataset

表4 在Dermatology下算法的性能比较

算法	AC	RI	ARI
K-means	0.381 7	0.703 2	0.048 8
DBSCAN	0.525 1	0.650 6	0.163 5
本文算法	0.636 9	0.802 3	0.384 5

Table 5 The summary results of three algorithms on Breast dataset

表5 在Breast下算法的性能比较

算法	AC	RI	ARI
K-means	0.700 6	0.716 0	0.369 2
DBSCAN	0.646 1	0.605 5	0.241 2
本文算法	0.651 7	0.612 3	0.260 4

Table 6 The summary results of three algorithms on Contraceptive Method Choice dataset

表6 在Contraceptive Method Choice下算法的性能比较

算法	AC	RI	ARI
K-means	0.452 0	0.557 6	0.026 9
DBSCAN	0.427 0	0.384 7	−0.011 5
本文算法	0.448 7	0.487 2	0.007 3

Table 7 The summary results of three algorithms on Pendigits dataset

表7 在Pendigits下算法的性能比较

算法	AC	RI	ARI
K-means	0.692 2	0.909 0	0.533 4
DBSCAN	0.119 5	0.581 4	−0.000 1
本文算法	0.608 1	0.859 4	0.445 7

K-means算法和DBSCAN算法;在Wine和Contraceptive Method Choice数据集上的聚类结果和*K*-means算法相当,但优于DBSCAN算法;在Breast和Pendigits数据集上的聚类结果比*K*-means算法略差一些,但也优于DBSCAN算法得到的聚类结果。整体而言,本文算法的聚类结果在各个评价指标上均高于DBSCAN算法。

5 结束语

本文通过将数据场考虑全局信息的优势结合到DBSCAN算法中,并引入势差的概念,在聚类过程中动态地确定每个类的*Eps*,使得改进后的DBSCAN算法不仅可以较好地识别一些不规则的类簇,还可以较好地识别一些密度差别较大的数据集。由于算法中的*MinPts*仍需要人为给定,如何合理地给定*MinPts*,并减小时间复杂度,以使得算法更加完善将是下一个讨论的问题。

References:

[1] Xu Rui, Wunsch D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[2] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96), 1996: 226-231.

[3] Li Deyi, Du Yi. Artificial intelligence with uncertainty[M]. Beijing: National Defence Industry Press, 2005.

[4] Gan Wenyan, Li Deyi, Wang Jianmin. A hierarchical clustering method based on data fields[J]. Acta Electronica Sinica, 2006, 34(2): 258-262.

[5] Yu Jianqiao, Zhang Fan. A PAM algorithm based on data field[J]. Computer Science, 2005, 32(1): 165-168.

[6] Wang Haijun, Deng Yu, Wang Li, et al. A C-means algorithm based on data field[J]. Geomatics and Information Science

of Wuhan University, 2009, 34(5): 626-629.
[7] Li Xue, Miao Duoqian, Feng Qinrong. Rough clustering algorithm based on data field[J]. Computer Science, 2009, 36(2): 203-206.
[8] Li Chunfang, Liu Lianzhong, Lu Zhen. Probabilistic neural network based on data field[J]. Acta Electronica Sinica, 2011, 39(8): 1739-1745.
[9] Jain A K. Data clustering: 50 years beyond *K*-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.
[10] Liang Jiye, Bai Liang, Cao Fuyuan. *K*-Modes clustering algorithm based on a new distance measure[J]. Journal of Computer Research and Development, 2010, 47(10): 1749-1755.
[11] Yu Zhiwen, Wong H-S, You J, et al. Hybrid cluster ensemble framework based on the random combination of data transformation operators[J]. Pattern Recognition, 2012, 45(5): 1826-1837.
[12] Bai Liang, Liang Jiye, Dang Chuangyin, et al. A novel attribute weighting algorithm for clustering high-dimensional categorical data[J]. Pattern Recognition, 2011, 44(12): 2843-2861.

附中文参考文献：

[3] 李德毅, 杜鹢. 不确定性人工智能[M]. 北京: 国防工业出版社, 2005.
[4] 淦文燕, 李德毅, 王建民. 一种基于数据场的层次聚类方法[J]. 电子学报, 2006, 34(2): 258-262.
[5] 余建桥, 张帆. 基于数据场改进的PAM聚类算法[J].计算机科学, 2005, 32(1): 165-168.
[6] 王海军, 邓羽, 王丽, 等. 基于数据场的C均值聚类方法研究[J]. 武汉大学学报: 信息科学版, 2009, 34(5): 626-629.
[7] 李学, 苗夺谦, 冯琴荣. 基于数据场的粗糙聚类算法[J]. 计算机科学, 2009, 36(2): 203-206.
[8] 李春芳, 刘连忠, 陆震. 基于数据场的概率神经网络算法[J]. 电子学报, 2011, 39(8): 1739-1745.
[10] 梁吉业, 白亮, 曹付元. 基于新的距离度量的*K*-Modes聚类算法[J]. 计算机研究与发展, 2010, 47(10): 1749-1755.

YANG Jing was born in 1988. She is a master candidate at School of Computer and Information Technology, Shanxi University. Her research interest is machine learning.
杨静(1988—),女,山西大同人,山西大学计算机与信息技术学院硕士研究生,主要研究领域为机器学习。

GAO Jiawei was born in 1980. He is a Ph.D. candidate and lecturer at School of Computer and Information Technology, Shanxi University, and the student member of CCF. His research interest is machine learning.
高嘉伟(1980—),男,山西太原人,山西大学计算机与信息技术学院博士研究生、讲师,CCF学生会员,主要研究领域为机器学习。

LIANG Jiye was born in 1962. He received his Ph.D. degree from Institute of Information and System Science, Xi'an Jiaotong University in 2001. Now he is a professor and Ph.D. supervisor at School of Computer and Information Technology, Shanxi University, and the senior member of CCF. His research interests include machine learning, computational intelligence and data mining, etc.
梁吉业(1962—),男,山西晋城人,2001年于西安交通大学信息与系统科学研究所获得博士学位,现为山西大学计算机与信息技术学院教授、博士生导师,CCF高级会员,主要研究领域为机器学习,计算智能,数据挖掘等。

LIU Yanglei was born in 1990. He is a master candidate at School of Computer and Information Technology, Shanxi University. His research interest is machine learning.
刘杨磊(1990—),男,山西运城人,山西大学计算机与信息技术学院硕士研究生,主要研究领域为机器学习。