文章编号:1672-3961(2012)04-0035-06 DOI:10.6040/j.issn.1672-3961.2012.04.007

一种基于近似 EMD 的 DBSCAN 改进算法

张宏兵[1],陆建峰[1*],汤九斌[2]
(1. 南京理工大学计算机科学技术学院, 江苏 南京 210094; 2. 中国电信江苏公司, 江苏 南京 210037)

摘要:DBSCAN(density-based spatial clustering of applications with noise)算法是基于密度的经典聚类算法,但是该算法应用于高维数据时,常用距离函数不能很好地反映出数据点之间的关系,从而可能导致聚类簇不够精确。如果能在高维空间中采用合适的距离度量,将会改善聚类结果。针对上述问题,提出利用近似 EMD(earth mover's distance,堆土机距离)作为距离测度,通过迭代搜索的方法找出所有直接密度可达对象实现聚类。实验结果表明:在高维文本数据的聚类中,和原来算法相比,改进算法的正确率提高了 6%,两者在时间上相差不大;而对低维的 Iris 数据,改进算法通过 EMD 改善了实体间的相似性度量,减少了划分为噪声点的数据点个数,平均正确率提高了 10%。实验结果表明了改进算法对高维数据的有效性,并可以改善聚类性能。

关键词:聚类;DBSCAN 算法;近似 EMD;高维数据
中图分类号:TP311 **文献标志码**:A

An improved DBSCAN algorithm based on the approximate EMD

ZHANG Hong-bing[1], LU Jian-feng[1*], TANG Jiu-bin[2]
(1. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China;
2. Jiangsu Corporation of China Telecom, Nanjing 210037, China)

Abstract: The DBSCAN algorithm is one of the classic clustering algorithms based on the density. When this algorithm was applied to high-dimensional data, the distance measures in common use could not reflect the relationships between instances well, which would lead to the inaccurate clustering. If appropriate distance measures were adopted in high-dimensional space, the clustering result would be improved. To solve the above problem, the approximate EMD (earth mover's distance) instead of the common distance was used as the distance measure, and the clustering was achieved by finding all density-reachable objects with the method of iterative search. The experimental results showed that the performance of improved algorithm was 6% higher than that of the original algorithm for the high-dimensional text clustering, while there is no obvious difference in time cost. For low-dimensional Iris data, the proposed algorithm could improve the similarity measure between the instances, reduce the number of data points classified as noise points, and boot the performance with 10%. The experimental results also indicated that the proposed algorithm could reveal its effectiveness for high-dimensional data, and could improve the clustering performance.

Key words: clustering; DBSCAN algorithm; approximate EMD; high-dimensional data

0 引言

随着网络的迅速发展和全球信息化程度的提高,各种资源成爆炸式增长。如何在这些繁琐的没有规律的数据中找到有价值的信息并对这些信息进行分类、融合已经成为目前数据挖掘、信息搜索和知识管理等研究领域的重要课题。数据挖掘[1-3]技术为数据的处理提供了切实有效的方法,已成为数据库技术中一个研究热点。其中高质量的聚类方法可

收稿日期:2012-05-06
基金项目:江苏省自然基金资助项目(BK2009489);江苏省青蓝工程资助项目
作者简介:张宏兵(1987 -),男,江苏东台人,硕士研究生,主要研究方向为文本挖掘. E-mail:iamzhanghongbing@126.com
*通讯作者:陆建峰(1969 -),男,江苏南京人,教授,博士生导师,主要研究方向为人工智能和图像图形技术等. E-mail:lujf@njust.edu.cn

以将大量信息分成若干个簇（Cluster），然后对感兴趣的簇进行后期操作。因此聚类技术已经成为数据挖掘技术的核心。文本聚类作为聚类的一个重要分支，在大规模文本集的组织与浏览等方面都具有重要的应用[4]。文本聚类中的文本表示模型通常采用向量空间模型[5]，许多聚类算法可以应用其中，如*K*-MEANS[6]、CLARANS[7]、DBSCAN、CURE[8]、BIRCH[9]、Wavecluster[10]等。其中DBSCAN算法能发现任意形状的簇并且能够较好地处理含有噪声的数据，因此广泛地应用于分析遥感、地理信息系统（GIS）等空间数据的分析。

1 经典DBSCAN算法

1.1 相关概念

DBSCAN算法是一个基于高密度连接区域的密度聚类算法，它能够发现任意形状的簇，并能有效地处理噪声。为了便于理解，给出下列定义[11]。

定义1 密度：是指以某点为圆心，以 ε 为半径的圆内包含的点的个数。

定义2 邻域：空间中任意一点的领域是以该点为圆心，以 ε 为半径的圆内包含的点的集合，记作 $N\varepsilon(p)=\{q\in D|\beta\mathrm{dist}(p,q)\leqslant\varepsilon\}$。

定义3 核心点：如果空间中某一点的密度大于给定的阈值MinPts，则称该点为核心点。

定义4 边界点：如果空间中某一点的密度小于给定的阈值MinPts，则称该点为边界点。

定义5 直接密度可达到：存在点 p 与点 q。若满足以下条件：

（1）点 p 处于点 q 的邻域中，即 $p\in N\varepsilon(p)$；

（2）点 q 是核心点，即 $N\varepsilon(p)\geqslant$MinPts。

则称点 p 从点 q 直接密度可达。

定义6 密度可达到：存在点 p 和点 q，若满足在 $(p_1,p_2,\cdots,p_n)$ 中，$p_1=p,p_n=q$，且有 p_i 从 p_{i+1} 直接密度可达，则称点 p 从点 q 密度可达到。

定义7 密度相连：存在点 p 和点 q，若满足 $\exists o\in p_i$，使得点 p 和点 q 都从点 o 密度可达到，则称点 p 和点 q 是密度相连的。

定义8 簇和噪声：基于密度可达到的最大的密度相连对象的集合称为簇，不在任何簇中的对象被认为是“噪声”。

DBSCAN算法一般步骤可以参考文献[12]，核心计算就是要求出两个数据对象之间的距离，然后和 ε 比较。可以看出DBSCAN算法中距离度量的准确性会对聚类结果会产生很大的影响[13-15]。

1.2 DBSCAN算法的优缺点

DBSCAN算法虽然可以发现任意形状的聚类，但是还存在许多问题。第一，对Eps和MinPts两个输入参数比较敏感，只能凭借经验对特定的数据指定这两个数据的值，选取困难[16]。若选取不当，则聚类效果不理想；第二，在聚类过程中，DBSCAN算法一旦找到一个核心对象就会以此对象为中心向外扩展，经过若干次迭代后核心对象将不断增多，如何存储这些信息成了难题。针对以上缺点，已有许多优秀的改进算法，比如Jin R等人提出的并行计算方法[17]、于亚飞等人提出的基于数据划分方法[18]、周水庚等人提出了一种基于密度的快速聚类算[12]和OPTICS算法[19]。数据挖掘中可以用蚁群算法寻找最优路径[20]，可以将博弈树搜索方法[21]用到其中，也可以利用概率论的知识来进行搜索[22-23]。基于上述研究，本研究提出了利用EMD作为距离测度来聚类的方法，以解决较高维聚类不精确的问题，从而达到更好的聚类结果。

2 算法改进

衡量实体之间的距离，可以利用欧氏距离、绝对值距离、切比雪夫距离、马氏距离等，但是对于高维数据而言，这些距离都不能很好地反映实体之间的关系。因为数据在高维空间中所呈现出的稀疏性和空空间现象[24]，使得上述距离不能有效消除“维度灾难”的影响，从而得不到精确的结果。在高维空间中，即使是最为相似的两个数据对象，其属性在某些数据维上仍然可能存在较大的差别。传统的距离度量没有考虑到这些维度上的偏差，对计算高维数据对象之间的相似性产生了很大的干扰。EMD的提出，精确了距离的表示，提高了计算的稳定性，使得实体之间关系的表示更加明确[25]。

2.1 相关概念

EMD可以理解为从一种分布变化为另一种分布的最小代价。给定两个分布，定义一个两者之间非相似性的量化标准，使之最大可能地近似视觉感知上的非相似性，这是非常有用的。定义地面距离（Ground distance）为两个分布之间的距离，即聚集成分布的各个基本特征点之间的距离。计算EMD的方法是从一个运输问题中提炼出来的，其描述如下：假设有几个供应商，每个人都有一定数量的货物，需要供货给几个消费商，每个消费商都有一个购货能力，要指出供应商与消费商之间一个单位货物的成本给定。传输问题本质上就是寻找最小代价货

物流,使货物从供应商流向消费商且能满足消费商的要求,这样运输问题可以形式化为线性规划问题,线性规划问题就能够很容易地解决了。

定义 F 是特征空间,存在一个距离映射 $d:F\times F\rightarrow R^{+}$。给定:

$$P=\{(p_1,w_{p1}),\cdots,(p_m,w_{pm})\},\tag{1}$$

$$Q=\{(q_1,w_{q1}),\cdots,(q_n,w_{qn})\},\tag{2}$$

其中 $p,q\in F$,是特征空间的一个点,且 ω_i 是点 i 的权值。

定义矩阵 $D=[d_{ij}]$ 表示地面距离矩阵,其中,d_{ij}是点 p_i 和点 q_j之间的地面距离。需要寻找一个流矩阵 $F=[f_{ij}]$, 其中 f_{ij}是点 p_i 和点 q_j 之间的流,使得全局代价函数 $\mathrm{WORK}(P,Q,F)=\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij}$最小,且必须要满足如下条件:

$$f_{ij}\geqslant 0,\ 1\leqslant i\leqslant m,\ 1\leqslant j\leqslant n;\tag{3}$$

$$\sum_{j=1}^{n}f_{ij}\leqslant w_{pi},\ 1\leqslant i\leqslant m;\tag{4}$$

$$\sum_{i=1}^{m}f_{ij}\leqslant w_{qi},\ 1\leqslant j\leqslant n;\tag{5}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}=\min\left(\sum_{i=1}^{m}w_{pi},\sum_{j=1}^{n}w_{qj}\right)。\tag{6}$$

条件(3)指明流是从 P 流向 Q 的,条件(4)限制了每个供应商的供应量,条件(5)限制了每个消费商的消费量,条件(6)使得尽可能地移动最大数量的货物。一旦运输问题得到解决,就可得到了矩阵 F,于是 EMD 距离就定义为所有运输工作的规格化值:

$$\mathrm{EMD}(P,Q)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}}。\tag{7}$$

EMD 具有许多优点:(1)EMD 是连续的,即使存在特征分布的微小变化也不会引起 EMD 值的波动;(2)EMD 是用来测算特征分布距离的,特征分布的紧凑性和灵活性为它本身带来了优异的特性,避免了属性相似性度量的量化问题;(3)当 2 个特征分布不平衡时,EMD 会自然地被应用于局部匹配,从而不影响相似性的比较。虽然 EMD 有许多优点,但是复杂的计算、过分追求高精度却成了这个度量最大的缺点,漫长的计算时间是人们所不能容忍的。

2.2　近似的 EMD

Ling H 和 Okada K 根据小波分析的一些原理,对 EMD 做出了相应的一些近似。如果用 L_1 地面距离来代替欧氏距离,EMD 算法能将算法复杂度控制在 $O(n^2)$[26]。将高维空间建模成一个具有权重的图,利用光谱图论的知识将小波分析应用其中。设 S 是一个紧凑空间,满足 $S\subset\mathbf{R}^{\mathrm{d}}$,在 S 上有两个概率分布 P_1 和 P_2,其概率密度分别是 p_1 和 p_2,存在具有连续性的函数映射 $d;S\times S\rightarrow\mathbf{R}^{+}$。定义 s 为一个固定参数,满足 $0<s\leqslant 1$。于是近似问题就会转变成找出 $S\times S$ 空间上联合概率密度 q 的下确界,其中 q 类似离散集里面流的概念。

$$\mu d=\inf_{q}\int d(x,y)^{s}q(x,y)\,\mathrm{d}x\mathrm{d}y。\tag{8}$$

联合分布概率必须满足条件(9) 的约束:

$$p_1(u)-p_2(u)=\int q(u,y)\,\mathrm{d}y-\int q(x,u)\,\mathrm{d}x。\tag{9}$$

设$p:=p_1-p_2$,满足$\int p=0$,它是不同于属性集的一个概率密度。将式(8) 与式(9) 合并得:

$$\mu d=\sup_{f}\int f(x)p(x)\,\mathrm{d}x。\tag{10}$$

这个上确界是基于空间 S 上所有连续函数 f 的,且 f 要满足 $f(x)-f(y)\leqslant d(x,y)_s$, $x,y\in S$。当 $S=\mathbf{R}^{\mathrm{d}}$ 并且 $d(x,y)=\|x-y\|$ 时,公式(10)就是要最大化势函数和直方图的内积,这在小波分析中是很容易做到的,因为其为此提供了一个详尽的公式:

$$\mu^{*}(p_1,p_2)=\sum_{\lambda}2^{-j(1+n/2)}|p_{\lambda}|。\tag{11}$$

P 是 n 维差分直方图,p_λ 是小波系数。这样利用小波分析,就可以找到 C_L 和 C_U,使得 $\mu^{*}C_L\leqslant\mu_d\leqslant\mu^{*}C_U$。为了能够得到近似值,只要利用公式(11)去计算每个独立属性就可以了。一旦计算完成,我们会把得到的结构存放到一个稀疏向量中,留给聚类算法使用。这样 EMD 就可以用在 DBSCAN 算法中了。

3　基于近似 EMD 的 DBSCAN 算法

3.1　算法步骤

改进算法利用 EMD 代替其他距离计算实体之间的距离,通过迭代搜索所有直接密度可达到的对象,继而找到各个簇所包含的所有密度可达到的对象。

3.1.1　算法 1

求近似 EMD 的算法:

(1) 给定空间 S 上的两个分布 P_1,P_2,P_1,P_2 分别是由各个样本的属性集组成的,权值为对应的属性值。对空间 S 上的数据进行抽样,使数据离散化;

(2) 建立一张权值图 G;

(3) 为图 G 建立小波分析;

(4) 把 p_1,p_2 作为输入参数,利用公式(11)得出结果 $\omega d_1,\omega d_2$;

（5）计算出近似值 $D^* = L_1 - \mathrm{dist}(\omega d_1, \omega d_2)$，返回结果。

3.1.2 算法2

改进的 DBSCAN 算法：

（1）检查数据库中尚未检查过的对象 p，如果 p 未被处理，利用算法1返回的结果检查其 ε 邻域 $N\varepsilon(p)$，若 $N\varepsilon(p)$ 包含的对象个数不小于 MinPts，建立新簇 C，将 $N\varepsilon(p)$ 中所有点加入 C；

（2）对 C 中所有尚未被处理的对象 q，利用算法1返回的结果检查其 ε 邻域 $N\varepsilon(p)$，若 $N\varepsilon(p)$ 包含至少 MinPts 个对象，则将 $N\varepsilon(p)$ 中未归入任何一个簇的对象加入 C；

（3）重复步骤（2），继续检查 C 中未处理的对象，直到没有新的对象加入当前簇 C；

（4）重复步骤（1）~（3），直到所有的对象都归入了某个簇或者标记为噪声。

3.2 算法分析

从算法复杂度而言，改进算法的复杂度和原算法差不多。设 n 表示实体的个数，每个实体具有 m 个属性，其属性值作为对应权值。经典 DBSCAN 算法在计算实体之间距离的算法复杂度是 $O(m^2)$，故算法整体复杂度是 $O(nm^2\log n)$。新算法的复杂度主要表现在利用小波分析求近似 EMD 上。在得到 EMD 矩阵的时候，时间复杂度是 $O(m^2)$；利用算法1求出 EMD 距离时，时间复杂度是 $O(m^2)$，所以改进后的算法复杂度也是 $O(nm^2\log n)$。分析可知，改进算法利用复杂的距离来提高聚类的质量，在时间上是可以接受的。

4 实验结果

为了检验改进算法的有效性，对经典算法和改进算法进行了对比实验。实验采用 VC++实现，软件平台 Windows Xp，酷睿 T5450 的 CPU，2 G 内存，120 G 硬盘。

实验数据来自中国科学院计算技术研究所的文本数据，包括文艺 1 482 篇，历史 934 篇，计算机 2 715篇，经济 3 201 篇，体育 2 506 篇。将这些数据通过文本预处理程序后，取其中一维作为参照标准，随机选取 200 项数据。通过人工画出直方图，观测大概可以分成几部分，继而算出这几部分的平均 Eps 作为实验参数。从预处理后的数据上可以看出平均 EMD 大约是平均欧氏距离的 5 倍，所以选定 Eps = 0.3 和 MinPts = 6 用欧氏距离的经典 DBSCAN 算法来聚类，选定 Eps = 1.5 和 MinPts = 6 用改进后的算法来聚类，记录聚类结果，并比较两种算法的性能。表1和表2分别显示了原始算法和改进算法的聚类结果和正确率，每一行表示该类在聚类算法后归入到对应类别文档的数目。其中正确率 w 可以通过公式（12）来计算：

$$w = \frac{\text{聚类结果中所属类别正确个数}}{\text{实际分类中类别的总个数}} \times 100\% 。\quad (12)$$

表1 经典 DBSCAN 算法聚类结果（Eps = 0.3，MinPts = 6）

Table1 Clustering results of classic DBSCAN algorithm

文献总数/篇	聚类结果/篇					正确率/%
	文艺	历史	计算机	经济	体育	
文艺（1 482）	1 143	173	23	97	46	77.1
历史（934）	98	697	44	46	49	74.6
计算机（2 715）	181	231	2 191	89	23	80.7
经济（3 201）	109	147	146	2 631	168	82.2
体育（2 506）	97	180	115	112	2 002	79.9

表2 改进 DBSCAN 算法聚类结果（Eps = 1.5，MinPts = 6）

Table 2 Clustering results of improved DBSCAN algorithm

文献总数/篇	聚类结果					正确率/%
	文艺	历史	计算机	经济	体育	
文艺（1 482）	1 235	129	36	43	39	83.3
历史（934）	76	745	34	41	38	79.8
计算机（2 715）	176	120	2 321	51	47	85.5
经济（3 201）	170	101	77	2 836	17	88.6
体育（2 506）	168	131	56	29	2 122	84.7

为了能更好地展示改进算法的性能，本研究将两个算法对应的正确率放在同一幅图中进行了对比，如图1所示。

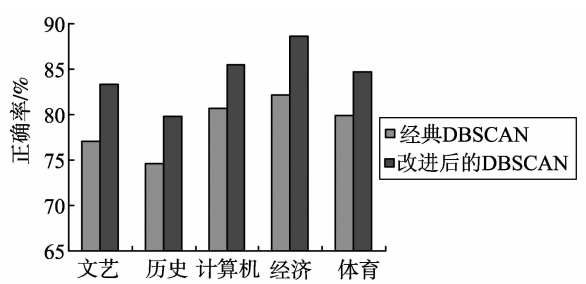

图1 改进前后算法正确率对比图

Fig. 1 Performance comparison between original and improved algorithm

实验结果表明，经典算法利用欧氏距离来聚类的平均正确率是 78.9%，基于 EMD 的改进算法的平均正确率为 84.4%。由图1可知，在文本聚类中改进的 DBSCAN 算法较原算法的聚类质量是显著提高的。

时间分析上，本研究同样选取 Eps = 0.3 和 MinPts = 6 作为经典 DBSCAN 的参数，选取 Eps = 1.5 和 MinPts = 6 作为改进算法的参数，从其中分别选取 3 000 个、5 000 个、7 000 个、9 000 个数据点，其

实验结果如图 2 所示。

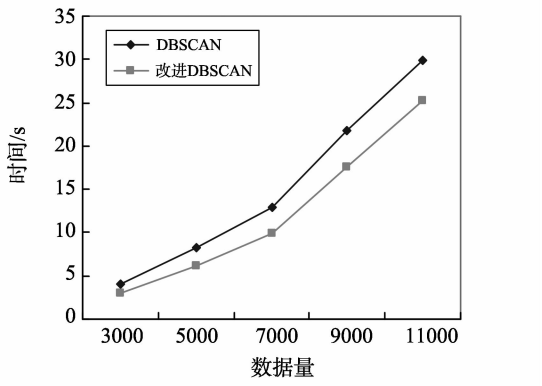

图 2　改进前后算法时间比较

Fig. 2　Time comparison between original and improved algorithm

由图 2 可以看出,两个算法的执行时间近似抛物线,验证了理论结果。从图 2 中可以直观地看出,改进的 DBSCAN 算法执行时间只略少于对应的改进前的算法时间。由于本研究着重研究利用 EMD 距离来代替传统距离提高聚类质量,而在时间上有稍许提高,整体效果还是比较乐观的。

为了体现改进算法的普遍性,本研究还选取了 Iris 数据集作为实验数据集。Iris 数据集以鸢尾花的特征作为数据来源,数据集包括 150 个数据,分为 3 类,每类 50 个数据,每个数据包含 4 个属性,是数据挖掘中公认的数据集。分别计算出 3 类的平均欧氏距离和 EMD 距离,作为 Eps 的选取标准。实验中经典 DBSCAN 算法选定 Eps = 0. 42, MinPts = 4; 改进算法选定 Eps = 2. 1, MinPts = 4, 实验结果如表 3 和表 4 所示。

表 3　改进前算法结果(Iris, Eps = 0. 42, MinPts = 4)

Table 3　Results of classic DBSCAN algorithm

类别	实验聚类总个数	其中正确聚类个数	其中非正确聚类个数	正确率/%
C1(50)	48	48	0	96
C2(50)	35	35	0	70
C3(50)	34	32	2	64
噪声点	33	—	—	—

表 4　改进后算法结果(Iris, Eps = 2. 1, MinPts = 4)

Table 4　Results of improved DBSCAN algorithm

类别	实验聚类总个数	其中正确聚类个数	其中非正确聚类个数	正确率/%
C1(50)	49	49	0	98
C2(50)	43	43	0	86
C3(50)	39	38	1	76
噪声点	19	—	—	—

该实验表明,改进后的算法的正确率有一定提高,相比之下噪声点的个数明显减少(噪声点的定义见 1. 1 的定义 8)。这是因为 Iris 数据集每个数据只有 4 个属性,在低维空间中,每维数据之间的差距很小,利用欧氏距离就能够表示出实体之间的相似性,因此聚类正确性已经很好了。利用 EMD 距离主要是减小数据的稀疏性来提高实体之间的相似性,这就减少了实验中噪声点的个数。

综上所述,改进后的 DBSCAN 算法在效率和正确率上是可行的。

5　结束语

聚类分析在数据挖掘中占有举足轻重的作用,高效而精确地得出簇会给今后寻找海量数据中的隐藏规律带来方便。本研究从经典 DBSCAN 算法出发,针对在较高维空间中提高簇的精确聚类问题,提出了一种基于近似 EMD 的改进算法。使用中国科学院计算技术研究所的文本数据和 Iris 数据集进行测试,结果表明改进的 DBSCAN 算法在聚类质量方面优于经典 DBSCAN 算法。近似 EMD 距离的提出,解决了高维空间中数据的稀疏性问题,减少了每维内数据的偏差,改善了实体间相似性的度量。在实际应用中,新的度量方法的应用将使得它们分析的结果更加精确。

改进算法最关键的就是如何快速的计算出 EMD。虽然近似的 EMD 度量在较高维数据聚类中比一般方法精确,但是繁琐的公式导致计算时间较长。如果是一味追求正确率而忽略了时间是得不偿失的。因此站在时间的角度,开发并行聚类算法和找到另外一种近似方法是我们今后研究的一个重要课题。

参考文献:

[1] TAN Pangning, STEINBACH MICHAEL, KUMAR VIPIN. Introduction to data mining[M]. Beijing: Posts and Telecom Press, 2006.

[2] 邓纳姆 M H. 数据挖掘教程[M]. 北京:清华大学出版社,2005.
DUNHAM MARGARET H. Data mining: introductory and advanced topics[M]. Beijing: Tsinghua University Press, 2005.

[3] 米哈尔斯基 R S,布拉特科 I,库巴特 M. 机器学习与数据挖掘:方法和应用[M]. 北京:电子工业出版社,2004.
MICHALSKI RYSZARD S, BRATKO IVAN, KUBAT MIROSLAV. Machine learning and data mining, methods and applications [M]. Beijing: Electronics Industry Press, 2004.

[4] STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques[R]. Technical Reprot Department of Computer and Information Science, Linkoping, 1995: 143-150

[5] FASULO D. An analysis of recent work in clustering al-

gorithms[R]. Techical Reprot UW-CSE-01-03-02, University of Washington, 1999: 176-186.

[6] 龚静,李安民. 一种改进的 *K*-means 中文文本聚类算法[J]. 湖南工业大学学报,2008,22(2):52-54.
GONG Jing, LI Anmin. Clustering algorithm of one improved K-means Chinese text[J]. Journal of Hunan University of Technology, 2008, 22(2):52-54.

[7] ESTER M, KRIEGELH P, XU X W. Knowledge discovery in large spatial database focusing techniques for efficient class identification [C]// Proceedings of the 4th International Symposium on Advances in Spatial Databases, LNCS 951. London: Springer, 1995: 67-82.

[8] 马帅,王腾蛟,唐世渭,等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报,2003,14(6): 1089-1095.
MA Shuai, WANG Tengjiao, TANG Shiwei. A fast clustering algorithm based on reference and density[J]. Journal of Software, 2003, 14(6):1089-1095.

[9] ZHANG T. BIRCH: an efficient data clustering method for very large databases[C]// Proceedings of the ACM SIGMOD Intel Conf on Management of Data. Montreal: ACM Press, 1996: 73-84.

[10] SHEIKHOLESLAMI G. Wave cluster: muli-resolution clustering approach for very large spatial databases [C]// Proceedings of the 24th VLDB Conference. New York, USA: Morgan Kaufmann, 1998: 428-439.

[11] HAN Jiawei, KAMBER M. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2000.

[12] 周水庚,周傲英,曹晶,等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展,2000,37(11): 1287-1292.
ZHOU Shuigeng, ZHOU Aoying, CAO Jing, et al. A fastdensity-based clustering algorithm [J]. Journal of Computer Research and Development, 2000, 37(11): 1287-1292.

[13] YOSSI RUBENER, CARLO TOMASI, TEONIDAS J GUIBAS. The earth mover's distance as a metric for image retrival[J]. International Journal of Computer Vision, 2000, 40(2):99-121.

[14] CORTES C, PREGIBON D. Signature-based methods for data streams [J]. Data Mining and Knowledge Discovery, 5(3):167-182.

[15] LEVINA E, BICKEL P. The earth mover's distance is the mallows distance: some insights from statistics [C]// Proceedings of ICCV, Vancouver, Canada: Elsevier, 2001: 251-256.

[16] 蔡颖琨,谢昆青,马修军. 屏蔽了输入参数敏感性的 DBSCAN 改进算法[J]. 北京大学学报,2004, 40(3): 480-486.
CAI Yingkun, XIE Kunqing, MA Xiujun. An improved DBSCAN algorithm which is insensitive to input parameters[J]. Acta Scientiarum Naturalium University Pekinensis, 2004, 40(3):480-486.

[17] JIN R, YANG G, AGRAWAL G. Shared memory parallelization of data mining algorithms: techniques, programming, interface, and performance[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(1):271-289.

[18] 于亚飞,周爱武. 一种改进的 DBSCAN 密度算法[J]. 计算机技术与发展,2011,21(2):30-38.
YU Yafei, ZHOU Aiwu. An improved algorithm of DBSCAN[J]. Computer Technology and Development, 2011, 21(2):30-38.

[19] 廖旭,张力. 工作流管理系统中一种基于任务的委托模式[J]. 计算机工程与应用,2005,41(7):44-46.
LIAO Xu, ZHANG Li. A task-based delegation model for workflow management system[J]. Computer Engineering and Applications, 2005, 41(7):44-46.

[20] 沙露,鲍培明,李尼格. 基于蚁群系统的聚类算法研究[J]. 山东大学学报:工学版,2010, 40(3):13-18.
SHA Lu, BAO Peiming, LI Nige. The research of a clustering algorithm based on the ant colony system[J]. Journal of Shandong University: Engineering Science, 2010, 40(3):13-18.

[21] 张明亮,李凡长. 一种新的博弈树搜索方法[J]. 山东大学学报:工学版,2009, 39(6):1-7.
ZHANG Mingliang, LI Fanzhang. A new search method for a game tree[J]. Journal of Shandong University: Engineering Science, 2009, 39(6):1-7.

[22] 张新猛,蒋盛益. 一种基于相似度概率的不确定分类数据聚类算法[J]. 山东大学学报:工学版,2011, 41(3): 12-16.
ZHANG Xinmeng, JIANG Shengyi. An algorithm for clustering uncertain categorical data based on similarity probability[J]. Journal of Shandong University: Engineering Science, 2011, 41(3):12-16.

[23] 赵科军,王新军,刘洋. 基于结构化覆盖网的连续 top-k 联接查询算法[J]. 山东大学学报:工学版,2009, 39(5):32-37.
ZHAO Kejun, WANG Xinjun, LIU Yang. Algorithms of continuous top-k join query over structured overlay networks[J]. Journal of Shandong University: Engineering Science, 2009, 39(5):32-37.

[24] BEYER K, GOLDSTEIN J, RISHNAN R. When is nearest neighbor meaningful[C]// Proceedings of ICDT Conference Proceedings. Jerusalem, Israel: [s. n.], 1999: 217-235.

[25] RUBNER Y, TOMASI C, GUIBAS L J. A metric for distributions with applications to image databases[C]// Proceedings of the Sixth International Conference on Computer Vision. Washington DC, USA: [s. n.], 1998: 59-66.

[26] LING H, OKADA K. An efficient earth mover's distance algorithm for robust histogram compadrison[J]. IEEE TPAMI, 2007, 29(5):840-853.

（编辑:胡春霞）