一种改进的DBSCAN密度算法

于亚飞，周爱武
（安徽大学 计算机科学与技术学院，安徽 合肥 230039）

摘 要：DBSCAN算法是一种基于密度的聚类算法，算法存在许多优点，也存在一些不足。比如对输入参数Eps敏感，DBSCAN由于采用全局Eps值，所以在数据密度不均匀和类间距离相差比较大的情况下，聚类质量会受到很大影响。文中主要针对算法输入参数Eps以及数据密度不均匀问题加以改进，提出了一种新的数据分区方法，通过对k-dist图纵坐标距离值单维度聚类，然后对比横坐标实现分区，使每个分区的数据尽可能均匀。实验证明，改进算法明显缓解了全局Eps导致的聚类质量恶化问题，聚类结果更加准确。

关键词：DBSCAN算法；Eps；数据分区；K-dist图
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1673-629X(2011)02-0030-04

An Improved Algorithm of DBSCAN

YU Ya-fei, ZHOU Ai-wu
(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: The algorithm of DBSCAN is an algorithm based on density, including both many points and also shortages. For example the algorithm is sensitive to the input parameters, because the algorithm uses the global Eps, therefore in the case of uneven data and the larger distance between classes, the clustering quality will be greatly affected. Mainly improved the choice of Eps, and solved the problem of uneven data. Proposed a new method of data partition, by clustering the value of k-dist vertical axis, the algorithm completed partition. Each data partition was uniform. Experimental results show that improved algorithm eases the problem of deterioration clustering quality significantly. The improved algorithm has a more accurate result of clustering.

Key words: DBSCAN; Eps; data partition; K-dist

0 引 言

数据挖掘就是对观测到的数据集（经常是很庞大的）进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据[1]。聚类分析是数据挖掘的一个重要方向。聚类是在预先不知道数据样本有多少类的情况下，使所有数据组成不同的类，类内元素相似性最大，类间元素相似性最小。聚类有很多种算法，传统的聚类方法包括划分方法，层次方法，基于密度的方法，基于网格的方法和基于模型的方法[2]。

DBSCAN[3]是一种经典的基于密度的聚类算法，可以在带有噪声的环境下发现任意形状的类，所以在图像处理[4]等许多领域有着广泛的应用。但是算法本身也存在许多问题。文中主要针对DBSCAN算法存在的问题加以改进。

收稿日期：2010-06-01；**修回日期**：2010-09-17
基金项目：安徽省教育科研重点项目（KJ2009A57）
作者简介：于亚飞（1986-），男，硕士生，研究方向为数据库与web技术、数据挖掘；周爱武，副教授，研究方向为数据库与web技术、数据仓库与数据挖掘、信息系统安全。

1 DBSCAN算法

1.1 DBSCAN算法

DBSCAN算法是将密度足够大的数据组成类。DBSCAN需要由用户主观来选择参数从而影响了最终的聚类结果，对于数据量为n的样本集合，DBSCAN的计算复杂度为$O(n^2)$。一般采用空间索引的方法降低时间复杂度，复杂度为$O(n\log n)$。

DBSCAN算法用到的定义如下：

定义1（数据点的Eps邻域）以数据样本中任意一点为圆心，Eps为半径的球形区域内包含的点的集合，叫做该数据点的Eps邻域。

定义2（数据点的密度）数据样本中任意一点的Eps邻域内包含的点数，叫做该数据点的密度。

定义3（核心数据点）核心数据点是指在Eps半径范围之内包含等于Minpts或大于Minpts个点的数据样本中任意一点。

定义4（边界数据点）边界数据点是指在某个核心数据点的邻域内，但自身不是核心数据点的数据样本

中任意一点。

定义5(直接密度可达)已知 Eps ,Minpts ,对于点 x 和点 y,如果 y 是核心点,而且 x 属于 y 的 Eps 邻域,则点 x 从点 y 直接密度可达。

定义6(密度可达) 如果对于给定的 Eps,Minpts 存在点链 $X1, X2, X3 \cdots Xn$, 其中 $X1 = X, Xn = Q$,而且 Xi 从 $Xi + 1$ 直接密度可达,那么点 X 从点 Q 密度可达。

定义7(密度相连) 如果在给定 Eps,Minpts 的情况下,存在点 p,使得点 x 和点 y 都从 p 密度可达,则点 x 和点 y 是密度相连的。

定义8 事先给定 Eps 和 Minpts,基于密度聚类中的一个聚类就是可以密度连接所能包含的最多数据点的集合。不属于任何聚类的数据点的集合称为噪声。

假定输入参数为 Eps 和 Minpts,DBSCAN 的算法描述如下:

(1)输入聚类数据,然后任意选取一个数据点 x,检查数据点 x 的 Eps 邻域。

(2) 如果 x 是核心点而且没有被划分到某一个类,则找出所有从 x 密度可达的点,最终形成一个包含 x 的类。

(3) 如果 x 不是核心点,则被当做噪声处理。

(4)转到第一步,重复执行算法;如果数据集合中所有的点都被处理,则算法结束。

1.2 DBSCAN 算法的优缺点

DBSCAN 是一种基于密度的聚类算法,可以发现任意形状的聚类,不受噪声的影响。但算法需要事先确定 Eps,Minpts 两个全局变量。一般事先确定 Minpts 值比较容易。在数据样本不多的情况下,Minpts 在二维空间中的聚类中一般取 4[5]。另外取数据集合的 1/25 作为 Minpts 的值也是一种有效的方法[6]。确定 Minpts 之后,算法通过一个启发式方法来确定 Eps 参数值。首先在给定 k 数值的情况下,将每个点映射为该点与其第 k 个最邻近点间的距离,即 k-dist。然后绘制排序 k-dist 图。图中横坐标表示各个点,纵坐标表示每个点对应的 k-dist。接着在排序 k-dist 图中寻找第一个凹陷,也叫做阈值点。因为通常来说,阈值点所对应的数据点处于类的边界附近,所以当 k 等于 Minpts 时,所求的 Eps 值就是阈值点所对应的 k-dist 值[5]。另外算法可以通过用户对数据中噪音水平的估计更加客观的确定 Eps 的值[7]。

聚类之前建立 R * 树和绘制 k-dist 图都是非常耗时的工作。另外为了有更好的聚类效果,用户必须通过反复试验选择合适的 k-dist 值。

难以发现密度相差较大的类是 DBSCAN 算法的另一个缺点。由于参数 Eps 和 Minpts 是全局唯一的,所以 DBSCAN 只能发现密度近似类。此外,如果类间距离差别比较大,算法结果也会受到很大的影响,容易产生偏差。

2 改进的 DBSCAN 算法

近年来已经出现了很多的 DBSCAN 改进算法,比较著名的是 OPTICS 算法[8]。由于 DBSCAN 算法使用了全局性的参数 Eps,因此当各个类的密度不均匀,或者类间的距离相差很大时,聚类的质量较差。文献[9]提出了一种基于密度标记的聚类算法 DTBC, DTBC 算法对所输入的参数不敏感,比较适合处理密度不均匀的聚类问题[9]。文献[10]中算法根据基于网格与基于密度的聚类算法间的等效规则计算各个密度层次的密度阈值,解决了 DBSCAN 算法参数选取困难和难以发现密度相差较大的簇的问题[10]。文献[11]中提出了"分而治之"和高效的并行方法改进 DBSCAN 算法,使聚类效果明显改善[11]。周水庚等人提出了基于数据分区的 PDBSCAN 算法。PDBSCAN 算法有效解决了密度不均匀问题,但是当类之间存在包含和交叉关系的时候,比如互相缠绕的螺旋状类和互相包含的环状类时,PDBSCAN 方法难以见效[12]。

文中主要讨论二维空间数据的分区问题。提出一种新的数据分区方法,根据 K-dist 图的思想,计算每个数据第 K 个最近邻居之间的距离,然后对距离聚类,实现数据分区。因为原数据集的密度分布情况,和 K-dist 距离图比较相似,密度大的数据 K-dist 距离普遍较小,稀疏的数据 K-dist 距离普遍比较大。一般 K 取整个数据集的 1/25[6]。因为 K-dis 距离是基于密度概念,跟每个类的位置无关。所以当类之间存在包含和交叉关系的时候,基于 K-dist 距离的分区方法可以适用。

建立 K-dist 图,横坐标为每个数据点,分别用数据点的输入顺序表示,即 1,2,3,4,5,6$\cdots n$。自然数列与原数据集形成一个映射。纵坐标为每个数据点和它第 K 个最近邻居的距离。对 K-dist 图纵坐标距离值单维度聚类。

单维度聚类,即将所有点的第 K 最近距离集合聚类。用最简单的 K-Means 算法实现聚类。单维度聚类之后,对比横坐标,就可以使原数据实现分区。当然 K-Means 算法本身存在许多不足,合理选取初始聚类中心,才能使每个分区的数据尽可能均匀。

分类完成之后,对于每一个数据分区,重新计算所有 K-dist 值进行排序。然后,按原 DBSCAN 启发式方法寻找 K-dist 排序图的第一个凹陷(阈值点),从而确定每个分区的 Eps 值。

排序用到 MATLAB 函数-sort(-x) 形成降序序列,x 为 k-dist 值矩阵。

输入参数 Eps 和 K，运行 DBSCAN 算法得到一个分区的聚类结果。然后对所有分区的结果进行合并即可。

k-dist 图具体建立过程：

(1)运行距离矩阵输出程序，输出数据集中所有点之间的距离，大小为 $N*N$ 的 cid 矩阵。每一行代表一个数据点与其他所有数据点的距离。

(2)运行 matlab 函数[y,l] = min(cid,[ ],2)，求出矩阵 cid 每一行元素的最小值，第一次找到的是全 0 数据，因为矩阵 cid 是对角线全为 0 的矩阵。

(3 然后运行大数替换程序[cidn] = DBS(cid, n, l)，n 为数据点个数。最小值元素用一个大数 100 替换，得到矩阵 cidn。同时保持矩阵中的数据位置不变。

(4)对矩阵 cidn 求[y,l] = min(cidn,[ ],2)。会得到所有数据点的最近邻居距离。

(5)然后执行 cid = cidn，把 cid 重新赋值。返回第 3 步，再次输入替换矩阵程序，循环执行。找到第二个最近距离。

(6)循环执行第 3 到 5 步，求出所有数据的第 K 个最近邻居距离，建立 k-dist 图。

3 实验分析

实验使用 MATLAB 环境，采用二维数据，且数据类型为实型。数据属性分别对应平面直角坐标系的横轴和纵轴。分别利用原 DBSCAN 算法和改进后的算法进行聚类。

3.1 一般数据

利用人工方法构造规则的 130 个数据，一般 K 取样本数据的 1/25 比较合适，所以 K = 5。

根据 K-dist 图建立过程，计算所有数据的第 5 个最近邻居之间的距离。然后对距离进行聚类，选取最远的两个点 0.0860 和 0.9000 作为初始聚类中心，执行 K-Means 算法进行聚类。距离分成两类，第一类 90 个点，第二类 40 个点。对比横坐标，确定 2 个分区，分别为 90 个点和 40 个点。

首先分析第一个分区 90 个点，K = 4。计算所有数据的第 4 个最近邻居之间的距离，绘制分区 K-dist 图。对排序后的 K-dist 图查找第一个凹陷(阈值点)，确定 Eps 值为 0.1503。

输入参数 Eps 和 K 值，运行 DBSCAN 算法，数据分为 3 类，加号 40 个点，圆形 30 个点，五角星 20 个点。实验结果如图 1 所示。即左侧的三类数据。

然后分析第 2 个分区，40 个点。因为点比较少，所以直接用 K-Means 聚类更方便。如果用 DBSCAN 算法，K = 2。计算所有数据的第 2 个最近邻居之间的距离。在排序之后的 K-dist 图中查找第一个凹陷(阈值点)，确定分区 Eps 值为 0.5036。结果分为 2 类，加号 20 个点，圆形 20 个点。即右侧的两类数据。

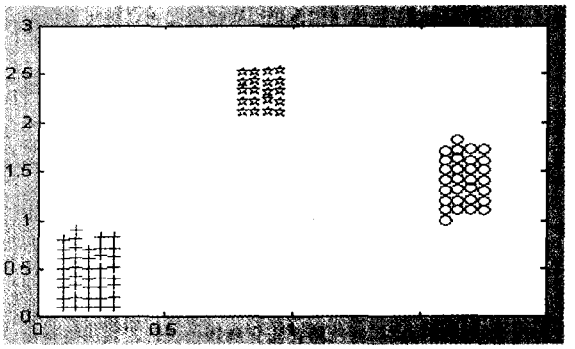

图 1 第一分区 90 个点聚类图

每个分区聚类完成后，对 2 个分区的聚类结果进行合并。聚类结果符合数据实际分布，聚类效果较好。

如果开始不分区，要先建立全局排序的 K-dist 图，然后根据启发式方法查找 Eps 值。。

K-dist 距离 0.8043 是第一个凹陷 选取 Eps = 0.8043，从聚类图 2 中可以看到，结果有明显偏差。因为 Eps 太大，所以左边的 2 类被合并了。如图 3 所示。第一类圆形，第二类五角星，第三类加号，第四类向下的三角形。

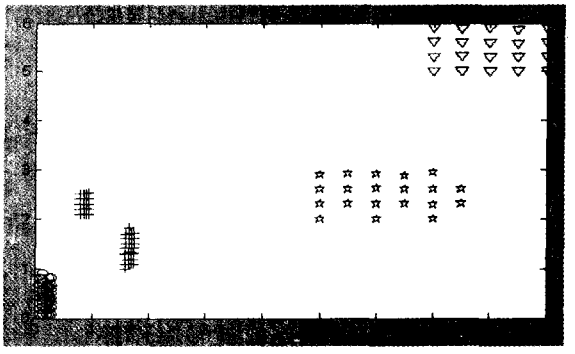

图 2 130 个点 Eps = 0.8043 聚类图

如果 Eps = 0.5590 ，从聚类图中可以看出右面 2 类被合并，成为噪声。噪声用实心点表示。如图 3 所示。第一类圆形，第二类五角星，第三类加号。

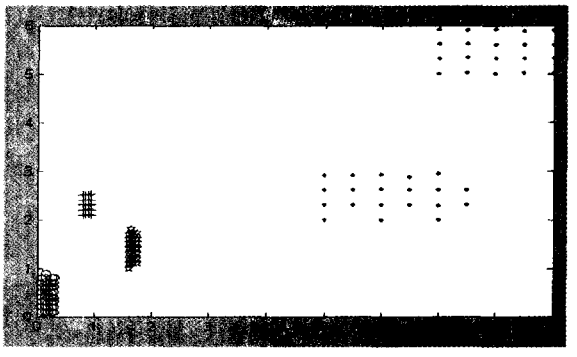

图 3 130 个点 Eps = 0.5590 聚类图

通过聚类图可以看出，改进算法由于运用了数据分区，使每个分区的数据尽可能均匀，然后选取分区

Eps 值进行聚类,所以聚类结果更接近实际数据分布。原算法由于采用全局 Eps 值,所以聚类质量难以保证,与实际数据分布相差很大。

3.2 环绕型数据

人工构造环绕型数据集 220 个点,$K=9$。对于环绕型数据,PDBSCAN 算法难以见效,因为无论是在 X 轴或者 Y 轴上分区,都是基于数据分布特性,跟类的相对位置有关。在直方图的同一矩形区域内会存在大量不同密度的点。改进方法基于密度概念,可以有效分区,使分区内数据尽可能均匀。

首先计算 220 个点的第 9 最近距离集合,建立 K-dist 图。不考虑噪声,选取 K-dist 最远的两个点作为初始聚类中心,即 0.1800 和 0.9000。分成 2 类,分别为 40 和 180 个数据。对比横坐标,确定 2 个分区,分别为 40 个点和 180 个点。

首先分析第一个分区 40 个点,因为点比较少,可以用 K-Means 算法直接聚类。分成 2 类,每一类 20 个点。

然后分析第二个分区 180 个点,计算所有点 K-dist 值,K 取 7。执行 DBSCAN 算法,最后分成 2 类,每一类 90 个点,如图 4 所示。2 类分别为五角星和实心点。

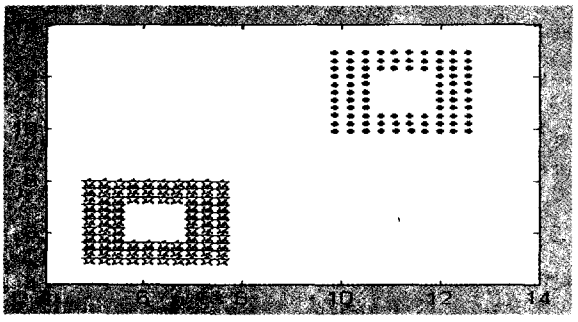

图 4　第二分区 180 个点聚类图

如果不分区,则要建立全局排序的 K-dist 图,然后确定全局 Eps。

首先选取 Eps=0.6798,结果如图 5 所示。左侧 2 类被合并为加号类,右侧 2 类被合并为五角星类,与实际数据分布差别很大。

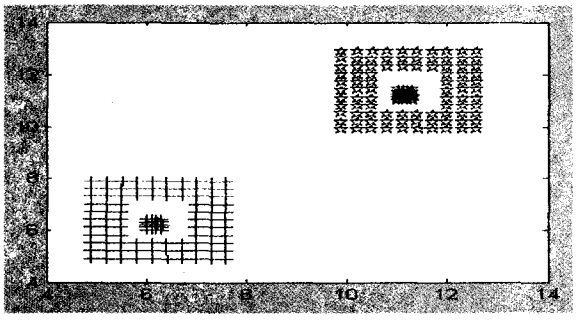

图 5　220 个点 Eps=0.6798 聚类图

然后选取 Eps=0.3000,噪声为三角形,结果如图 6 所示。外围密度比较小的 2 类数据被当成噪声,明显偏差。

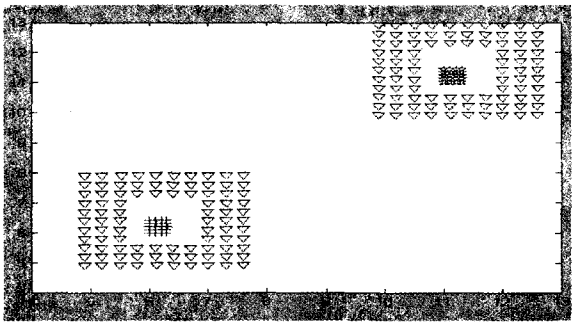

图 6　220 个点 Eps=0.3000 聚类图

通过聚类图可以看出,Eps 距离分区方法使环绕型数据很好的实现分区,聚类结果更接近实际数据分布。原算法由于采用全局 Eps 值,所以聚类质量难以保证。

4 结束语

DBSCAN 算法是一种基于密度的算法,可以发现任意形状的聚类,不受噪声影响。但是算法本身对输入参数 Eps 非常敏感,而且对于密度分布不均匀的数据集不适用。文中提出了一种新的数据分区方法,实验结果表明,改进算法有效解决了数据密度不均匀的问题,聚类结果更接近实际数据分布,准确性较高。

参考文献:

[1] Hand D, Mannila H, Smyth P. 数据挖掘原理[M]. 张银奎, 廖　丽, 宋　俊, 等译. 北京:机械工业出版社, 2003.

[2] 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.

[3] Ester, Martin, Kriegel H P, 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). Ortland, Oregon: [s. n.], 1996.

[4] 李莉平, 沈俊媛. 基于数据挖掘的 DBSCAN 算法及其应用[J]. 科技创业月刊, 2009(8):134-135.

[5] 孙凌燕. 基于密度的聚类算法研究[D]. 太原:中北大学, 2009.

[6] Daszykowski M, Walczak B, Massart D L. Looking for Natural Patterns In Data[J]. Chemometrics and Intelligent Laboratory Systems, 2001, 56:83-92.

[7] 任兴平, 何忠龙, 孟增辉. 改进 DBSCAN 算法中参数 Eps 值的确定[J]. 现代电子技术, 2007(11):120-121.

[8] Ankerst M, Breunig M, Kriegel H P, et al. Optics: Ordering points to Identify the Clustering Structure[C]//Proceedings of ACM SIGMOD International Conference on Management of Data. Philadephia: ACM Press, 1999:49-60.

(下转第 38 页)

对;在任务处于完成状态,表示 $u1$ 已经完成了对数据的录入,$u2$ 在校对时,除了删除和修改已有的数据项外还可以对 $u1$ 遗漏的数据项进行添加;在任务处于提交状态,用户 $u1$ 和 $u2$ 对数据项都只有读的权限。在数据录入任务不同状态下用户 $u1$ 和 $u2$ 的权限如表1所示。

表1 $u1$ 和 $u2$ 在数据录入任务不同状态下的对数据库数据项的操作权限

权限 \ 状态	用户u1 初始	执行	挂起	完成	提交	夭折	用户u2 初始	执行	挂起	完成	提交	夭折
读	×	√	√	√	√	×	×	√	√	√	√	×
添加	×	√	×	×	×	×	×	×	×	√	×	×
删除	×	√	×	×	×	×	×	√	√	√	×	×
修改	×	√	×	×	×	×	×	√	√	√	×	×

即数据录入任务 t =({o,x},{o,"read,add,delete,modify"},{o,"read"},{o,"read"},{o,"read"},{o,x}),

assigned_permissions(t)={read,add,delete,modify},

$u1$ 向转授权服务器提出转授权申请

delegate($u1$,$u2$,o,"read, delete,modify",1,1,t,$st1$),

delegate($u1$,$u2$,o,"read, delete,modify",1,1,t,$st2$),

delegate($u1$,$u2$,o,"read,add,delete,modify",1,1,t,$st3$),

delegate($u1$,$u2$,o,"read",1,1,t,$st4$),

其中 o 表示数据库的数据项对象,用 $st0$,$st1$,$st2$,$st3$,$st4$,$st5$ 分别表示任务的初始态、执行态、挂起态、完成态、提交态和夭折态。转授权服务器根据转授权规则对这些转授权的合法性进行判定,经判定,这些转授权合法,同意转授权。当数据录入任务运行到不同的状态时,$u2$ 就获得对数据库数据项对象的对应的权限,从而可以进行相应的操作。

4 结束语

转授权技术作为访问控制的重要内容,在工作流系统中得到广泛的应用,文中提出了一种基于任务状态的用户-用户部分权限转授权模型,可以实现多步和多重转授权。对该模型进行了形式化定义,同时提出了转授权规则、冲突消解和转授权撤销等问题,最后举例说明了该模型在工作流系统中的应用。该转授权模型实际上是将任务的权限按照任务状态不同分别授予合作用户各自相应的权限,实现了多个用户协作完成某个任务而不相互干扰和冲突。

参考文献:

[1] 范玉顺.工作流管理技术基础[M].北京:清华大学出版社,2001:31-32.

[2] Zhang X,Oh S,Sandhu R. PBDM: a flexible delegation model in RBAC[C]//SACMAT'03: Proceedings of the Eighth ACM symposium on Access Control Models and Technologies. New York:ACM Press,2003:149-157.

[3] Barka E,Sandhu R. A Role-Based Delegation Model and Some Extensions [C]//Proc. of 23rd National Information Systems Security Conference. Baltimore, MD, USA: NIST, 2000:101-114.

[4] Zhang Longhua, Ahn Gail-Joon, Chu Bei-Tseng. A Rule-based Framework for Role-based Delegation[C]//Proc. of the 6th ACM Symposium on Access Control Models and Technologies(SACMAT 2001). New York: ACM Press, 2001: 153-162.

[5] 张黎明,王小明,李 黎.几种基于角色的代理授权模型特征比较[J].微机发展,2004,14(11):126-129.

[6] 董光宇,卿斯汉,刘克龙.带时间特性的角色授权约束[J].软件学报,2002,13(8):1521-1527.

[7] 孙 波,赵庆松,孙玉芳.TRDM-具有时限的基于角色的转授权模型[J].计算机研究与发展,2004,41(7):1104-1109.

[8] 廖 旭,张 力.工作流管理系统中一种基于任务的委托模型[J].计算机工程与应用,2005,41(7):44-46.

[9] 洪 帆,段素娟,黎成冰.基于图的委托授权模型[J].北京邮电大学学报,2005,28(6):5-7.

[10] 张润莲,武小年,董小社.基于委托的分布式动态授权模型[J].计算机应用,2008,28(6):1365-1368.

[11] Barka E,Sandhu R. Framework for Role-Based Delegation Models [C]// Proc of 16th Annual Computer Security Application Conference. IEEE Computer Society. Washington,DC, USA:[s.n.],2000:168-176.

[12] 魏永合,王成恩,马明旭.工作流系统中的委托授权机制研究[J].计算机集成制造系统,2009,15(1):160-165.

(上接第33页)

[9] 高 昇.基于密度聚类算法的改进方法研究[D].大连:大连理工大学,2007.

[10] 谭 颖,胡瑞飞,殷国富.多密度阈值的DBSCAN改进算法[J].计算机应用,2008,28(3):745-748.

[11] 冯少荣,肖文俊.DBSCAN聚类算法的研究与改进[J].中国矿业大学学报,2008,37(1):105-111.

[12] 周水庚,周傲英,曹 晶.基于数据分区的DBSCAN算法[J].计算机研究与发展,2000,37(10):1153-1159.