

基于相对密度的不确定数据聚类算法

潘冬明 黄德才

(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 传统的基于相对密度的聚类算法有效地解决了密度聚类算法对参数敏感以及不能区分不同密度等级簇的问题。基于相对密度的不确定聚类算法,借用了相对密度算法的思想,根据不确定数据的特征,定义了不确定数据的距离公式、相对密度、核心点、密度可达等相关概念,从而提出了一种能够有效地处理不确定数据的新算法。数据仿真结果表明了该算法的有效性和可用性。

关键词 不确定数据,相对密度,聚类

中图法分类号 TP391 **文献标识码** A

Relative Density-based Clustering Algorithm over Uncertain Data

PAN Dong-ming HUANG De-cai

(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Traditional relative density-based clustering algorithm has advantage in handling shortcomings of user-defined parameters' sensitivity and distinguishing different hierarchy of density. This paper provided a new uncertain data clustering algorithm based on relative density, which defines distance formula, density ratio, core points and density-reachable, and can efficiently handle uncertain data. The simulation results illustrate the validity and availability of the algorithm.

Keywords Uncertain data, Relative density, Clustering

1 引言

随着计算机技术的飞速发展,数据库管理和数据挖掘领域中所采用的数据采集技术和数据处理技术也有了很大的提升。人们的注意力开始逐渐地从确定性数据的研究转向不确定数据的研究,针对不确定数据的挖掘课题已经成为国际学术界和工业界的研究热点。

传感器网络(Wireless Sensor Network)、无线射频识别(Radio Frequency Identification)等领域中产生出一种新的数据类,称为不确定数据。例如在传感器网络中,由于受到环境干扰、仪器精度等影响,传感器节点所监测到的数据往往是不确定的。处理这样的不确定数据,对数据挖掘技术的需求已十分迫切。但遗憾的是,在数据挖掘领域的大部分研究都是针对确定数据的,而适用于不确定数据的研究不多。文献[1]首先将不确定数据的研究作为一个新的研究方向提出,并依据经典的 K-means 聚类算法^[2]提出了针对不确定数据的 UK-means 聚类算法。UK-means 算法由于具有不适宜发现任意形状的簇、对噪声点敏感等缺点,因此实用性有限。文献[3]基于 DBSCAN 算法^[4],重新定义了距离公式,提出了 FD-BSCAN 算法,但该算法基于对对象连续分布的抽样,因此计算时间和计算精度都无法得到保证,从而会对聚类结果产生影响。本文根据确定性数据挖掘相对密度的思想,给出了一个新的不确定数据的距离公式,定义了不确定数据的相对密度、核心点和相对密度可达等相关概念,提出了一种针对不确

定数据的聚类算法 URDBC(Uncertain Relative Density Based Clustering)。该算法保留了相对密度的特性,可以发现不确定数据集中任意形状的簇,并有效地解决了不同密度等级的聚类问题。数据仿真不仅分析了 URDBC 算法参数 δ 的最佳取值区间,还说明了算法对不确定数据聚类的有效性和可用性。

2 相关工作和定义

2.1 不确定数据模型

不确定数据的表现形式可分为属性不确定性和存在不确定性。

(1)属性不确定性,也称值不确定性。在元组数目和数据模型已经确定的前提下,通常用不确定密度函数或其他统计量表示属性值中存在的 uncertain 信息。可描述为一个 D 维的数据点 X ,在任意维上的属性都存在不确定性,不确定性用 φ 表示,可记为 $(x_i, \varphi(x_i))$ 。

(2)存在不确定性,也称元组不确定性。数据元组存在与否具有一定的不确定性,一般情况下,假设各个元组之间是相互独立的,以便在不同的应用中进行数据处理。基于元组的不确定性,一般可采用点概率模型(point probability case)。在点概率模型中元组的属性值确定,而元组存在性不确定,用一个 $[0, 1]$ 之间的随机不确定值表示^[5-7]。点概率模型的应用很广泛,例如在 RFID 应用中,RFID 读卡器会存在漏读、多读的现象,所读数据的存在性并不确定。点概率模型可表示为:不确定数据集中的数据点 X , X 这个点的存在不确定性用 P

本文受水利部公益性行业科研专项(201401044)资助。

潘冬明(1989—),男,硕士生,主要研究方向为数据挖掘,E-mail:panwinterming@163.com;黄德才(1958—),男,博士,教授,博士生导师,主要研究方向为数据挖掘、人工智能等,E-mail:hdc@zjut.edu.cn(通信作者)。

表示,则该模型表示为 (X, P) ,其中 P 在 $[0, 1]$ 之间随机产生。本文讨论的是存在不确定性的情况。

2.2 不确定数据聚类

文献[8]中指出,基于密度的 DBSCAN 算法,当所选的 $MinPts$ 较大, ϵ 较小时,就有可能把原本较为稀疏的 A、B 两个类判别为孤立点。同时,对于全局恒定的 $MinPts$ 和 ϵ ,高密度等级的类簇会被完全包含在相连的低密度等级的聚类结果中。如图 1 所示。

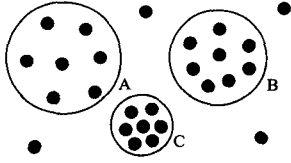


图 1 不同参数时的聚类结果

基于上述密度聚类存在的问题,考虑到数据的不确定性特征,本文提出的基于相对密度的不确定聚类算法重新定义了距离公式,改进了传统的相对密度的计算方法,使之可以很好地适应不确定数据。

定义 1(对象 x 关于对象 y 的不确定近邻距离) 对象 x 关于对象 y 的不确定近邻距离由式(1)表示:

$$d_{-p}(x, y) = \frac{\sqrt{\sum_{i=1}^n (x_i - y_i)^2}}{1 - |p_x - p_y|} \quad (1)$$

由式(1)可以看出,两个对象的不确定性的差值越大,则它们的不确定近邻距离就越远,当差值无限接近于 1 时,不确定近邻距离趋于无穷大;两个对象的不确定性的差值越小,则它们的不确定近邻距离就越近,当差值为 0 时,它们之间的不确定近邻距离就等于传统的欧氏距离。

定义 2(对象 x 的不确定 k 近邻) 设有对象 x 、不确定集合 D 且 $x \notin D$, 计算 x 到集合 D 任一点 o_i 的距离 d_i , 并将 d_i 按非递减方式排序, 可得到与 x 最近的前 k 个对象 $o_1, o_2, \dots, o_k$ 。对象 x 的不确定 k 近邻表示为 $N_{-k}(x)$ 。

$$N_{-k}(x) = \{o_i | o_i \in D, i=1, 2, \dots, k\} \quad (2)$$

定义 3(对象 x 的不确定近邻密度) 对不确定数据集 $D, x \in D, o \in N_{-k}(x)$, x 的不确定近邻密度记为 $Pnnd_{-k}(x)$, 定义如下:

$$Pnnd_{-k}(x) = 1 / \left(\frac{\sum_{o \in N_{-k}(x)} d_{-p}(x, o)}{\sum_{p \in P_{N_{-k}(x)}} p} \right) \quad (3)$$

从式(3)容易看出,对象 x 的不确定近邻密度即为对象 x 到其不确定邻居 $N_{-k}(x)$ 的平均距离的倒数,那么样本点的不确定近邻密度越大,表示不确定近邻之间分布越紧密;不确定近邻密度越小,表示不确定近邻之间的分布越稀疏。

定义 4(对象 x 关于其不确定 k 近邻 $N_{-k}(x)$ 的不确定相对密度) 对不确定数据集 $D, x \in D$, x 关于其不确定邻居 $N_{-k}(x)$ 的不确定相对密度记为 $Prd_{-k}(x)$, 定义如下:

$$Prd_{-k}(x) = \frac{\sum_{o \in N_{-k}(x)} \left(\frac{Pnnd_{-k}(o)}{Pnnd_{-k}(x)} \right)}{\sum_{p \in P_{N_{-k}(x)}} p} \quad (4)$$

不确定相对密度反映的是对象 x 的密度与其不确定近邻密度之间的差距: $Prd_{-k}(x)$ 的值越接近 1, 说明对象 x 与其不确定近邻在分布上越均匀, 它们就越相似, 可以归结为属于同一个簇。

定义 5(不确定核心对象) 给定阈值 $0 < \delta < 1$ 和不确定

数据集 $D, x \in D$, 若 $|Prd_{-k}(x) - 1| < \delta$, 则称该对象 x 为不确定核心对象。

定义 6(对象 y 关于不确定核心对象 x 不确定直接相对密度可达) $x, y \in D, D$ 为不确定数据集, $y \in N_{-k}(x)$, 则称对象 y 关于不确定核心对象 x 不确定相对密度可达。

定义 7(不确定相对密度连接) 给定阈值 $0 < \delta < 1$ 和不确定数据集 D , 当存在一个对象链 $x_1, x_2, \dots, x_n, x = x_1, y = x_n$ 时, 对于任意一个 $x_i (i=1, 2, \dots, n)$, 若存在 x_i 到 x_{i+1} 不确定相对密度可达, 则称 x 关于 y 不确定相对密度连接。

定义 8(不确定核心集合) 不确定核心集合, 记为 $P_{-CoreSet}$ 。不确定核心集合中的对象须满足以下条件:

(1) 给定阈值 $0 < \delta < 1$, 对任意的 $x \in P_{-CoreSet}$, 都有 $|Prd_{-k}(x) - 1| < \delta$ 。

(2) 对任意的 $x, y \in P_{-CoreSet}$, 对象 x 关于 y 不确定相对密度连接。

定义 9(对象 x 关于类 C 的不确定相对密度) 给定不确定数据集 $D, C \subseteq D, x \in D$, x 关于类 C 的不确定相对密度记为 $Prd_C(x)$, 定义如下:

$$Prd_C(x) = \frac{\sum_{o \in C} \left(\frac{Pnnd_{-k}(o)}{Pnnd_{-k}(x)} \right)}{\sum_{p \in P_C} p} \quad (5)$$

对象 x 的不确定相对密度 $Prd_C(x)$ 反映了对象 x 的近邻密度与类 C 中成员对象的平均近邻密度之间的差别。

3 聚类算法

基于相对密度的不确定聚类算法描述: 首先在不确定数据集 D 中找到任意一个核心对象 x , 求出 x 的不确定核心集合, 得到初始类 C ; 然后通过初始类 C 进行扩展, 通过判断不确定核心对象的不确定近邻是否为不确定核心对象来扩展不确定核心对象集合, 迭代至没有可扩展的点为止。通过最终的不确定核心对象集合和该集合中对象的不确定直接相对密度可达的点得到最终的一个簇。再从未归类的点中重新找不确定核心对象, 重复上述迭代过程, 直至没有任何对象可以归入任何的簇为止。

基于相对密度的聚类算法 URDBC 描述如下:

输入: 不确定数据集 D , 不确定近邻个数 k , 阈值 $0 < \delta < 1$

输出: 已聚类的不确定数据集 D

伪代码描述:

URDBC(Set UncertainSet, int k , float δ)

{

计算任意两点之间的不确定距离, 得到不确定距离的矩阵 Distance_Matrix

通过 Distance_Matrix 得到集合中每个点的不确定近邻

由 Distance_Matrix 和 Neighbors[i] 得到每个点的不确定近邻密度由 $Pnnd_{-k}(i)$ 得出每个点的不确定相对密度

//迭代过程

REPEAT

在未得到标识的对象中找到一个核心对象点, 得出 point 的核心对象集合

初始化类 C , 给类 C 中对象分配类标识 clusterID

扩展类 C , ExpandCluster(UncertainSet, C , $P_{-CoreSet}$, k , δ);

}

ExpandCluster(Set UncertainSet, Cluster C , Set $P_{-CoreSet}$, int k , float δ)

{

While(不确定核心集合 CoreSet 不为空){

从不确定核心集合 P_CoreSet 中取出一个核心对象 point

在未标识的对象中找到 point 的核心对象集合 Point_CoreSet

While(Point_CoreSet.size()>0)

{

newPoint=Point_CoreSet.get(i);

if (|Prd_k(x)-1|<δ)//进行阈值检查

把 newPoint 加入核心集合;

C.append(N_k(newPoint));//根据密度可达性,把 new-

Point 的 k 近邻扩展进 C

}

}

}

4 实验与分析

4.1 参数选择

定理 1^[8] 设 D 是一个不确定数据对象集,分别用 $dist_min$ 、 $dist_max$ 表示 D 中对象的最小不确定 k 近邻距离和最大不确定 k 近邻距离。令 $\epsilon = \frac{dist_min}{dist_max}$, 对于对象 $p \in D$, 若 $N_k(p) \subseteq D$, 而且 $\forall q \in N_k(p)$, 也都有 $N_k(q) \subseteq D$, 则 $\epsilon \leq Prd_k(p) < 1$ 。

定理说明:若存在一个类 D , 考虑类内深处的一个对象 p , 即 p 的 k 个不确定近邻 o 都在 D 中, 且 o 的 k 个不确定近邻也在 D 中, 这样对象 p 的不确定相对密度 $Prd_k(p)$ 有了一定的取值范围。通过以上的定理可知, 如果 $dist_min$ 和 $dist_max$ 越接近, 则 $Prd_k(p)$ 的值就越接近于 1, 为参数 δ 的取值提供了参照, 避免了取值的盲目性和试探性。

实验结果采用聚类准确率和聚类质量^[9]作为评价标准: 聚类准确率采用的方法是比较加入了不确定性后的数据聚类结果与已知结果作比较。聚类质量反映的是一个聚类结果中每个类的质量情况: 一个类的质量与该类内元组的平均不确定性成正比, 与类的半径成反比。聚类结果中若生成的类的质量越高, 则类内数据点越紧密, 说明聚类效果越好; 类的质量越低, 类内数据点越稀疏, 聚类效果越差。

实验数据集采用 DS1, DS2 两个人工数据集和 Iris, heart 两个 UCI 数据集。分别与 FDBSCAN 做比较, DS1, DS2 参数设置为 $k=4, \delta=0.3$; Iris, heart 的参数设置为 $k=5, \delta=0.3$ 。比较结果如图 2—图 4 所示。

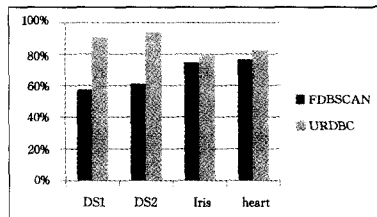


图2 聚类准确率

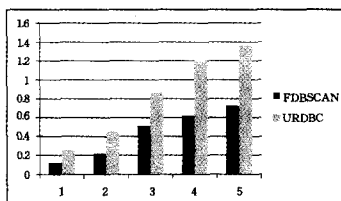


图3 DS1 聚类质量

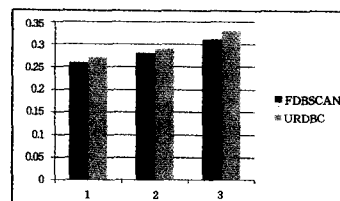


图4 Iris 数据集聚类质量

根据图 2—图 4 对数据集 DS1, DS2 的结果分析, URDBC 聚类的准确率和聚类质量都明显优于 FDBSCAN 算法, 从 Iris 和 heart 数据集的准确率和聚类质量来看, URDBC 也占有优势。DS1, DS2 数据集效果显著的原因是在人工设定明显地设置了密度等级, 而 URDBC 能够很好地处理不同密度划分的情况, 在这点上较 FDBSCAN 算法有明显的优势; Iris 和 heart 数据集中的点在分布上并没有特别的密度等级的差异, 但是 URDBC 在聚类的准确率和聚类质量上还是有优于 FDBSCAN 算法。

4.2 参数 δ 的参考取值范围

为了给出参数 δ 的一个最佳取值区间, 本实验对数据集 DS1, DS2 在不同 δ 的情况下进行实验观察, 以准确率为 Y 轴, δ 为 X 轴, 设置 δ 取值从 0.1 到 0.8, $k=4$; 取两个数据集在 Y 轴的值的平均值, 结果如图 5 所示。

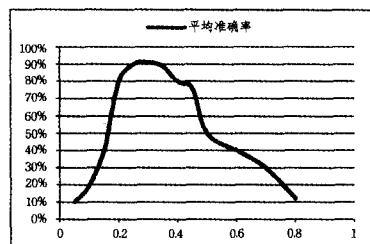


图5 参数范围

从图 5 中可以看出当 δ 的取值在 $[0.2, 0.4]$ 区间内, 聚类的准确率较好。最佳取值区间在实验仿真的时候有很大的参考价值。例如:

令 $\delta_0 = 0.2, \Delta = 0.025, \delta = \delta_0 + \Delta * i, i = 0, 1, 2 \dots 8$;

WHILE($\delta \leq 0.4$) {

URDBC(Set UncertainSet, int k, float δ);

$\delta = \delta_0 + \Delta * i$;

$i = i + 1$;

}

上述算法符合一般近似计算的思想, 得出一般基于相对密度的近似算法, 能够较快、较好地得到聚类结果。

结束语 近年来, 由于隐私保护、测量精度不够等原因, 不确定数据在诸多应用领域中出现, 人们也意识到对其研究的重要性, 不确定数据挖掘技术将扮演着十分重要的角色。本文基于相对密度的不确定聚类算法 URDBC, 借鉴了传统相对密度算法的思想, 提出了一种适用于不确定数据的聚类算法。算法能够在不确定数据集中发现任意形状的簇, 有很强的处理噪声的能力; 能够区分不同密度等级的簇, 适用于不确定数据的聚类研究。下一步工作将针对不确定混合数据的研究, 对距离公式做进一步的改进, 使其适用性更加宽泛。

参考文献

- [1] Chau M, Cheng R, Kao B, et al. Uncertain data mining: An example in clustering location data[M] // Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006: 199-204

(下转第 88 页)

果采用独立运行 20 次的平均值。算法收敛速度随迭代次数变化的进化曲线如图 1 所示。

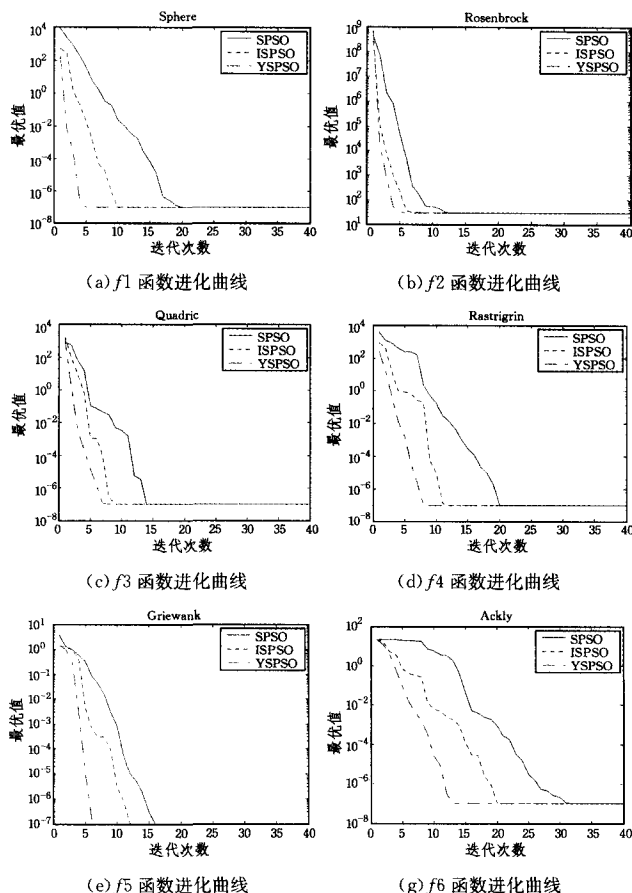


图 1 f_1 — f_6 在 3 个实验中的函数进化曲线

图 1(a)–(f) 是 3 个算法在 6 个基准测试函数上运行 20 次后取平均值生成的曲线,横坐标为算法迭代次数,纵坐标为算法适应度(函数最优值)。从图 1 中可以看出,与 SPSO 算法和 ISPSO 算法相比,YPSO 算法在收敛速度和精度上明显优于二者,尤其是进化后期收敛速度和精度具有明显提高,在迭代 15 次以内均已成功收敛,说明了位置更新式(6)的正确性和高效性;而且,YPSO 算法生成的曲线明显比较平滑,算法的全局搜索能力和摆脱局部极值能力均有明显提高,说明了引入的反向随机惯性权重的有效性。因此,YPSO 算法是一种收敛速度快、精度高、优化性能好的高效简化粒子群优化算法。

结束语 本文借鉴 SPSO 算法和 ISPSO 算法的思想,利用黄金分割法重新定义了粒子惯性、个体经验和社会经验对粒子位置更新的影响力关系,并引入反向随机惯性权重,提出了 YPSO 算法。通过对比实验证明了 YPSO 算法在收敛速度和精度以及摆脱局部极值的能力方面均有所提高,是一种高效的简化粒子群优化算法,为粒子群优化算法在相关领域的应用奠定了良好的基础。

参考文献

- [1] Kennedy J, Eberhart R. Particle swarm optimization[C]// Proceeding of IEEE International Conference on Neural Networks, IEEE, 1995:1942-1948
- [2] Shi Y, Eberhart R. A modified particle swarm optimizer[C]// Proceedings of IEEE International Conference on Evolutionary Computation, IEEE, 1998:69-73
- [3] 胡旺,李志蜀.一种更简化而高效的粒子群优化算法[J]. 软件学报, 2007, 18(4):861-868
- [4] 周昊天,吴志勇,田雨波.简化粒子群优化方法的改进研究[J]. 计算机工程与应用, 2012, 48(24):41-44
- [5] 黄太安,生佳根,徐红洋,等.一种改进的简化粒子群算法[J]. 计算机仿真, 2013, 30(2):327-330, 335
- [6] 赵志刚,黄树运,王伟倩.基于随机惯性权重的简化粒子群优化算法[J]. 计算机应用研究, 2014, 31(2):361-363, 391
- [7] 刘瑞芳,王希云.一种混沌惯性权重的简化粒子群算法[J]. 计算机工程与应用, 2011, 47(21):58-60
- [8] 李鑫滨,马阳,鹿鹭.一种基于校正因子的自适应简化粒子群优化算法[J]. 燕山大学学报, 2013, 37(5):453-459
- [9] 任伟建,武璇.一种动态改变学习因子的简化粒子群算法[J]. 自动化技术与应用, 2012, 31(10):9-11, 37
- [10] 郑春颖,王晓丹,郑全弟,等.自逃逸云简化粒子群优化算法[J]. 小型微型计算机系统, 2010, 31(7):1457-1460
- [11] 周丹,南敬昌,高明明.改进的简化粒子群算法优化模糊神经网络建模[J]. 计算机应用研究, 2015, 32(4):1000-1003
- [12] 熊众望,罗可.基于改进的简化粒子群聚类算法[J]. 计算机应用研究, 2014, 31(12):3550-3552
- [13] 刘瑞芳.混沌 w 的简化粒子群算法在机械设计中的应用[J]. 机械工程与自动化, 2010, 162(5):26-27
- [14] 田雨波.粒子群优化算法及电磁应用[M]. 北京:科学出版社, 2014:88-95, 169-245
- [1] MacQueen J. Some methods for classification and analysis of multivariate observations[J]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1(14):281-297
- [2] Kriegl H P, Pfeifle M. Density-based clustering of uncertain data[C]// Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005:672-677
- [3] Ester M, Kriegl H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[J]. Kdd, 1996, 96(34):226-231
- [4] Cormode G, McGregor A. Approximation algorithms for clustering uncertain data[C]// Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM, 2008:191-200
- [5] Kanagal B, Deshpande A. Online filtering, smoothing and probabilistic modeling of streaming data[C]// IEEE 24th International Conference on Data Engineering, 2008 (ICDE 2008). IEEE, 2008:1160-1169
- [6] Ré C, Letchner J, Balazinksa M, et al. Event queries on correlated probabilistic streams[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, 2008:715-728
- [7] Liu Q B, Deng S, Lu C H, et al. Relative density based k-nearest neighbors clustering algorithm[C]// 2003 International Conference on Machine Learning and Cybernetics, IEEE, 2003
- [8] 张晨,金澈清,周傲英.一种不确定数据流聚类算法[J]. 软件学报, 2010, 21(9):2173-2182

(上接第 74 页)