基于密度的不确定性数据概率聚类

许华杰[1,2] 李国徽[2] 杨 兵[2] 杜建强[3]
(上海第二工业大学计算机与信息学院 上海 201209)[1] (华中科技大学计算机学院 武汉 430074)[2]
(江西中医大学计算机学院 南昌 330006)[3]

摘 要 近期传感数据监测和移动对象跟踪等许多从自然界直接采集数据的新应用引发了不确定性数据管理这一新的研究课题。这些应用中相关数据的不确定性为传统的数据处理方法提出了新的挑战。探讨的重点是不确定性数据的聚类。提出了一个针对不确定性数据的基于密度的聚类算法,根据不确定性数据内在的概率分布信息进行概率聚类,并采用 R 树索引和概率阈值索引提高算法的效率。仿真试验表明,提出的算法在有效性和效率方面均优于当前主要的基于密度的不确定性数据聚类算法。

关键词 基于密度的聚类,不确定性数据,R 树

Probabilistic Density-based Clustering of Uncertain Data

XU Hua-jie[1,2] LI Guo-hui[2] YANG Bing[2] DU Jian-qiang[3]
(School of Computer and Information, Shanghai Second Polytechnic University, Shanghai 201209, China)[1]
(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)[2]
(School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330006, China)[3]

Abstract Recently, many new applications such as sensor data monitoring and mobile object tracking raise up the issue of uncertain data management. The intrinsic uncertainty of the data in such applications offers new challenges for traditional data processing methods. The focus of the paper is clustering of uncertain data. A probabilistic density-based clustering algorithm for uncertain data was proposed based on the probability distribution of uncertainty, with R-tree and probability threshold index for efficiency. Simulations show that the proposed algorithm outperforms other density-based clustering algorithms for uncertain data in effectivity and efficiency.

Keywords Density-based clustering, Uncertain data, R-tree

1 引言

随着传感技术、无线通信技术和定位技术的发展,对面向自然界的应用的需求越来越大,与之相关的技术研究引起了工业界和学术界的广泛重视。面向自然界的应用往往具有很大的数据量,且由于测量和采样等误差以及网络传输的延迟导致这些应用所涉及的数据往往在一定程度上具有某些不确定性。例如在无线传感器环境监测应用中,无线传感器网络极度受限的系统资源(如网络带宽和电能供给)只能够实现数据以离散的方式进行采集,自然界变化的连续性与数据采样的离散性之间的矛盾决定了从外部世界获得的数据本质上是随时间增长的不确定性数据,因此在对相关数据进行处理时必须考虑数据的不确定性,才有可能获得正确的处理结果,这对传统的数据处理方法提出了新的挑战。

国内外学术界对不确定性数据处理方法的研究方兴未艾,具有代表性的研究成果主要包括不确定性数据的概率查询技术[1,2]、概率索引技术[3]和数据广播技术[4]等,现有的研究成果主要是从数据库和数据查询的角度出发。面向自然界的应用往往伴随着巨大数据量,对数据挖掘技术的需求尤为迫切。但遗憾的是数据挖掘领域的绝大部分研究成果都是针对"确定"数据的,适用于不确定性数据挖掘的成果不多。这里所说的"不确定性数据"指的是数据对象的存在是确定的、但其取值具有一定的不确定性,有别于一些文献中提到的概率数据库(probabilistic database)[5]中的数据对象本身的存在就是概率性事件的"不确定性数据"。以移动对象聚类为例,如图 1 所示,其中图 1(a)表示的是根据真实环境中移动对象当前位置聚类的结果。由于数据采样的离散性对象当前的位置往往无法立即获得,根据最近一次采样记录的对象位置聚类的结果如图 1(b)所示,可见由于对象的移动根据记录数据(从某种意义上说部分"过时")进行聚类的结果与实际结果(图 1(a))有明显区别。图 1(c)所示的是根据最近一次采样记录的对象位置并考虑由于对象移动所带来的对象位置不确

到稿日期:2008-06-19 本文研究得到上海第二工业大学科研启动基金项目和国家高技术研究发展计划(863 计划)项目(No. 2007AA01Z309)资助。

许华杰 博士,讲师,研究方向为无线传感器网络、移动数据管理、不确定性数据处理,E-mail:hjxu@smail.hust.edu.cn;李国徽 博士,教授,博士生导师,研究方向为移动实时数据库、流数据处理、无线传感器网络;杨 兵 博士,研究方向为无线传感器网络;杜建强 教授,研究方向为软件工程。

定性(假设对象短期内的运动近似于直线运动)进行聚类的结果,可以看到所得到的结果与实际结果(图 1(a))基本相同。

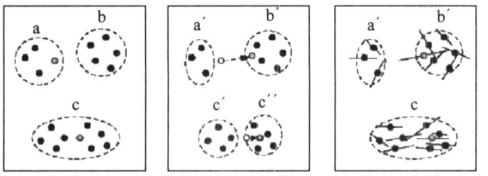

图 1　移动对象聚类示意图

目前国际上对不确定性数据聚类的研究成果不多,文献[6]首先将不确定性数据的数据挖掘作为一个新的研究方向提出来,并以数据聚类为例根据经典的 K-means 聚类方法提出针对不确定性数据的 UK-means 聚类方法。但该方法只简单地将中心点与数据对象点距离的期望值应用到 K-means 方法中,很多情况下这么做是不合适的[7]。况且 K-means 聚类方法具有不适宜发现非凸形状簇、对噪声和离群点敏感等缺点,因此文献[6]提出的方法实用性有限。文献[7]在著名的基于密度的聚类方法 DBSCAN[8]的基础上考虑数据的不确定性,提出针对不确定性数据的 FDBSCAN 聚类方法,但该方法的计算基于对对象连续分布的抽样(离散化,类似于 Monte Carlo 方法),因此计算精度和时间都无法保证,从而有可能对聚类结果产生影响。由于 DBSCAN 聚类方法具有适用于各种形状簇、对噪声和离群点不敏感等优良特性,本文基于 DBSCAN 方法提出一种采用不确定性数据索引技术、基于密度的不确定性数据概率聚类方法。

2　基于密度的不确定性数据概率聚类

为了说明方便,本文以移动对象聚类为例提出基于密度的不确定性数据概率聚类,但本文所提出的方法也适用于其他不确定性数据聚类应用语境。计算对象间的距离采用的是最常见的欧几里德几何距离,但方法对其他类型的距离也适用。所考察区域的移动对象用数据集 $D = \{O_1, O_2, \cdots, O_n\}$ 表示,其中每个不确定性对象 $O_i (1 \leq i \leq n)$ 用以最近一次记录到的该对象的位置为质心的不确定区域表示,对象当前的实际位置以一定的概率密度 pdf 在该区域内分布。对于任意两个不确定性对象 O_1 和 O_2,虽然它们当前的实际位置无法确定,但是它们之间距离的最大值 $d_{\max}(O_1, O_2)$ 和最小值 $d_{\min}(O_1, O_2)$ 是很容易计算出来的,分别代表 O_1 的不确定区域内的点与代表 O_2 的不确定区域内的点之间距离的最大值和最小值。文献[6]提出的方法简单地用两个不确定性对象之间距离的期望值代替实际值,从而丢失了对象位置在其不确定区间内的分布信息。文献[7]提出的方法虽然利用到概率分布信息,但是在对对象的不确定区间进行代价较高的离散化抽样计算后,只简单地将计算得到的核心对象概率和密度可达概率是否大于 0.5 分别作为对象是否是核心对象和是否可达的判断标准,这对正确聚类结果的获得都是不利的。本文提出的聚类方法充分利用对象位置在其不确定区间内的概率分布信息定义及计算核心对象概率和密度可达概率,并采用概率索引技术提高聚类算法的效率。

2.1　相关定义

定义 1(对象的 (ε, p) 邻居)　一个不确定性对象 O_i 的 (ε, p) 邻居用 $N_i(\varepsilon, p)$ 表示,定义为满足以下条件的对象:

$$N_i(\varepsilon, p) = \{O_j \in D \mid P(dis(o_j, o_i) \leq \varepsilon) \geq p, o_j \in O_j, o_i \in O_i\}$$

其中,o_j 和 o_i 表示所对应的不确定性对象当前的实际位置,几何上分别表示落在不确定性对象 O_j 和 O_i 不确定区域内的点;ε 是距离阈值;p 是概率阈值;$P(dis(o_j, o_i) \leq \varepsilon) \geq p$ 表示 o_j 与 o_i 之间的距离小于 ε 的概率大于 p。

定义 2(概率核心对象)　对于不确定性对象 O_i,若 $|N_i(\varepsilon, p)| \geq MinPts$,则对象 O_i 是关于 $\varepsilon, MinPts, p$ 的概率核心对象。

定义 3(直接概率密度可达)　若对象 O_i 为概率核心对象,且对象 $O_j \in N_i(\varepsilon, p)$,则称对象 O_j 是从对象 O_i 出发关于 $\varepsilon, MinPts, p$ 直接概率密度可达的。

定义 4(概率密度可达)　对于对象 O_i 和对象 O_j,若存在一个对象队列 $O_1, \cdots, O_m$,其中 $O_1 = O_i$ 且 $O_m = O_j$,$1 \leq k \leq m$,O_{k+1} 是从 O_k 出发直接概率密度可达的,则称 O_j 是从 O_i 出发关于 $\varepsilon, MinPts, p$ 概率密度可达的。

定义 5(概率密度连接)　对于对象 O_i 和对象 O_j,若存在一个对象 O_k,O_i 和 O_j 都是从 O_k 出发概率密度可达的,则称 O_i 关于 $\varepsilon, MinPts, p$ 概率密度连接 O_j。

引入概率阈值 p 的目的是利用小概率事件发生的可能性很小、通常可以被忽略这一特性建立概率阈值索引(见本文后面部分)。p 的取值是计算精度和效率之间的折衷。

2.2　基于密度的不确定性数据概率聚类算法

在以上定义的基础上提出的基于密度的不确定性数据概率聚类算法如下:

PDBSCAN 聚类算法

输入:$D = \{O_1, O_2, \cdots, O_n\}, \varepsilon, MinPts, p$

输出:簇集 $C = \{C_1, C_2, \cdots, C_m\}$

(1) 对 D 中的对象建立 R 树索引,在 D 中任意选定一个对象 O_i 作为起始对象;

(2) 设 O_i 是当前对象,通过 R 树索引裁剪掉与 O_i 的距离不可能小于 ε 的对象,即从 R 树索引的根节点出发,若分枝节点所代表的 MBR 与 O_i 的最小距离大于 ε,则以该分枝节点为根的子树所包含的所有对象均可以裁剪,通过 R 树索引可以排除大部分与 O_i 的距离大于 ε 的对象;

(3) 对于剩下的可能成为对象 O_i 的 (ε, p) 邻居的每个对象 O_j,计算其与 O_i 的最小距离 $d_{\min}(i, j)$ 和最大距离 $d_{\max}(i, j)$ 并将结果分别保存到全局矩阵 $M_{\min}$ 和 $M_{\max}$ 中;

(4) 对于可能成为对象 O_i 的 (ε, p) 邻居的每个对象 O_j,根据其 $d_{\min}(i, j)$ 和 $d_{\max}(i, j)$ 建立概率阈值索引 PTI,PTI 的创建方法见下一节;

(5) 在概率阈值索引 PTI 上以 p 为概率阈值、$Q = [0, \varepsilon]$ 为查询范围做范围查询,将满足查询条件的对象的标识加入到候选邻居集 $CN(\varepsilon, p)$ 中;

(6) 若 $|CN(\varepsilon, p)| \geq MinPts$,则对象 O_i 是关于 $\varepsilon, MinPts, p$ 的概率核心对象,将其加入到核心对象集 $CORE$;否则在 $D/CORE$ 中任选一个对象作为当前对象 O_i 并返回(2);

(7) $CN(\varepsilon, p)$ 中所包含的对象是 O_i 关于 $\varepsilon, MinPts, p$ 的直接概率密度可达对象,将这些对象标识为与 O_i 同一个簇,从 $CN(\varepsilon, p)$ 中任选一个对象作为当前对象 O_i 并将 $CN(\varepsilon,$

p)中的对象和 O_i 标识为“已聚类”,令 $CN(\varepsilon, p) = \varnothing$ 并返回(2);

(8) 对于本身不具备成为概率核心对象的条件而又无法从任何其他对象概率密度可达的对象将其标识为“离群点”,当 D 中的所有或规定比例的对象都做了标识,算法终止。

PDBSCAN 聚类算法有以下特点:①对概率核心对象和概率密度可达的计算并不像有的文献那样生硬地将两个不确定性对象(区域)之间的距离用单个值(如距离的期望值)来代替,而是利用两个不确定性对象之间的距离的最小值和最大值作为限定范围,并考虑不确定性在该范围上的概率分布;②算法在判断概率核心对象和概率密度可达时考虑不确定性对象概率分布,允许用户在计算精度和效率之间进行权衡,设置概率阈值 p,而不是简单地将概率是否大于 0.5 作为核心对象和密度可达的判断标准;③通过 R 树和概率阈值索引 PTI 这两种索引方法提高计算效率。

2.3 概率阈值索引 PTI

在 PDBSCAN 聚类算法中,利用 R 树索引虽然能够将大量与当前对象 O_i 之间的距离不可能小于 ε 的对象排除,大大减少了需要考虑的对象数量,但是实际上 R 树索引只利用了不确定区域本身,没用到数据在不确定区域上分布的概率密度信息,这将导致大量与 O_i 之间的距离接近 ε 的对象不能及时排除。当对象分布密度较高时,这一现象尤为明显。概率阈值索引 PTI 基于 R 树的思想,利用各个对象的不确定性概率分布信息对包含对象的 MBR 的边界进行收缩,从而进一步降低需要访问该 MBR 所在节点的机会。

PDBSCAN 聚类算法的步骤(4)首先根据可能成为对象 O_i 的 (ε, p) 邻居的每个对象 O_j 的不确定区域以及对象位置在不确定区域上的概率分布函数很容易分别得出它们与当前对象 O_i 之间距离的范围 $[d_{\min}(i,j), d_{\max}(i,j)]$ 及其上的概率分布函数。对于这些对象根据其 $[d_{\min}(i,j), d_{\max}(i,j)]$ 建立一维 R 树索引,设当前考虑的 R 树节点所表示的 MBR(用 M 表示)中包含 3 个对象的 $[d_{\min}(i,j), d_{\max}(i,j)]$,分别表示为 A,B 和 C,如图 2 所示。设 M 中第 k 个对象不确定区间的概率密度函数为 $f_k(y)$,则可定义 M 的 x-bound 为一对直线,分别是 left-x-bound(表示为 $M.lb(x)$)和 right-x-bound(表示为 $M.rb(x)$)。对于包含在当前 MBR 中的任一区间 $[L, R]$,若 $L \le M.lb(x)$,则 $\int_{-\infty}^{M.lb(x)} f_k(y)\,dy \le x$;若 $R_i \ge M.rb(x)$,则 $\int_{M.rb(x)}^{+\infty} f_k(y)\,dy \le x$。left-x-bound 和 right-x-bound 之间的区间为 x-bound 区间。由定义可知,在 MBR 内的每个区间都保证最多有 x 的概率在 left-x-bound 的左边或在 right-x-bound 的右边。未压缩前的 MBR 的边界可以看作是 0-bound,因为它们保证 MBR 中的节点以概率 1 包含在该 MBR 内。$x \in [0,1]$,其值越大,x-bound 区间就越小。图 2 中画出了当前 MBR 中的 0.2-bound 和 0.3-bound。

我们要求在 R 树的一个 MBR 节点中,x-bound 是唯一的且左右 x-bound 尽可能收缩到 MBR 的中心。在 R 树中保存 x-bound 信息的目的是为了压缩 MBR,以尽量避免访问不满足要求的 MBR 中的对象。如图 2 所示,假设要处理一个概率阈值为 0.3 的查询 Q,如果没有 x-bound,由于 Q 与 MBR 相交,因此当前 MBR 不能被裁剪,需要依次检查 MBR 中包含的对象并计算其满足 Q 的概率(积分运算),计算的代价非常高。然而计算出结果后却发现实际上 MBR 中各对象 A,B 和 C 都不满足 Q。如果通过 x-bound 该 MBR 就能够及早被裁剪掉,避免了对其所包含的节点的检查和计算。这是由于 Q 不在 0.3-bound 区间内且与 0.3-bound 不相交,根据 x-bound 的定义,当前 MBR 中包含的所有对象满足 Q 的概率必然小于 0.3。因此,对于概率阈值为 x 的 PTRQ,只需要检查查询区间是否与 MBR 中的 x-bound 相交,就可以判断该 MBR 中是否有满足这一概率阈值查询的对象,从而可以确定是否需要对该 MBR 进行深入访问。

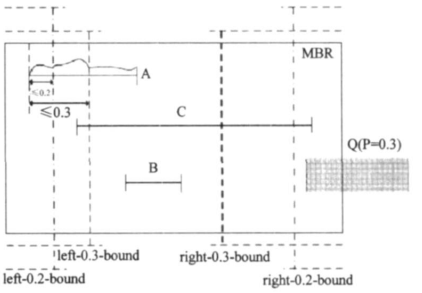

图 2 PTI 索引示意图

MBR 的 x-bound 构造过程如下:对于包含在 MBR 中的每个对象根据 x-bound 的定义单独计算其不确定区间的 x-bound,MBR 中包含的所有对象的 x-bound 区间的并就是整个 MBR 的 x-bound 区间。在建立 PTI 索引时 x 可以取一些常用的典型概率值,并分别计算所对应的 MBR 相应于这些 x 值的 x-bound 区间与 MBR 信息保存在 R 树中相同的节点中。x 的取值数量越多,对 MBR 的压缩就越精细,裁剪效果就更好。但是,相应地会增加 R 树节点保存 x-bound 信息的存储代价。利用 PTI 索引进行搜索的过程与 R 树搜索类似:根据用户指定的查询概率阈值 p 以及 PTI 中已有的 x 的取值选取相应的 x,让所选的 x 刚好不大于 p;从 PTI 索引的根节点开始判断所对应的 MBR 的 x-bound 是否与查询区间相交。如果不相交,则以该节点为根的子树都可以被裁剪,否则进一步搜索当前节点中所包含的子节点,直到叶子节点为止,最后返回满足查询条件的对象。限于篇幅,PTI 索引结构节点的添加和删除操作方法说明从略。

3 性能分析

由于不确定性数据聚类是一个新兴的研究领域,有影响的成果不多,比较具有代表性的有 UK-means 聚类方法[6]和 FDBSCAN 聚类算法[7]。UK-means 所基于的 K-means 聚类方法具有不宜于发现非凸形状簇、对噪声和离群点敏感等缺点,与 DBSCAN 聚类算法无可比性,因此在对比试验中 UK-means 聚类方法不予考虑。为了分析本文提出的 PDBSCAN 聚类算法的性能,进行了一系列的仿真试验,并将试验结果与文献[7]提出的 FDBSCAN 聚类算法进行比较。比较的性能指标是聚类的准确度和效率。仿真试验采用的数据集来自美国地理信息基准数据集 SEQUOIA 2000[9]。参数 ε 和 $MinPts$ 的取值根据文献[8]提出的启发式算法确定,而参数 p(概率阈值)设为 0.8。仿真试验在 Pentium IV 2.4GHz,512MB 的 PC 机上用 Visual C++ 实现。

首先考察 PDBSCAN 聚类算法的准确度。设最近一次采

样到当前时刻空间对象的最大移动距离为 d, d 值的大小反映了移动对象位置的不确定程度。设空间对象的不确定区间用最近一次采样得到的空间对象的位置为中心、以 d 为半径的圆表示,设对象位置在不确定区间中符合正态分布。试验中对于包含移动对象个数 $N = 5000$ 的数据集,针对不同的 d 值,分别采用本文提出的 PDBSCAN 聚类算法和 FDBSCAN 聚类算法(参数取值与文献[7]中相同)对不确定性对象进行聚类,采用 DBSCAN 聚类算法对当前时刻对象的"确定"位置进行聚类,设结果分别表示为 P, F 和 D。由于无法及时知道当前时刻移动对象的准确位置,D 实际上是无法获得的,在试验中只是起基准的作用。P 和 F 中与 D 相似程度越高,说明对不确定性数据聚类的准确度越高。比较两个聚类结果相似程度的指标采用的是广为使用的 Adjusted Rand Index (ARI)[10]。ARI 的值越大,说明两个聚类结果越相似。对于不同 d 值 P 与 D 之间(用 PDBSCAN 标识)、F 与 D 之间(用 FDBSCAN 标识)的 ARI 值如图 3 所示。由图 3 可见,随着 d 值增加,两种算法聚类的结果与理想的对精确数据聚类的结果之间的误差都有所增加,说明数据不确定程度增大导致聚类的准确性下降;对于相同的 d 值,PDBSCAN 聚类算法得到的结果比 FDBSCAN 聚类算法得到的结果更接近理想的实际结果(ARI 值更大),说明 PDBSCAN 聚类算法的有效性更佳。原因在于 FDBSCAN 聚类算法是通过对数据不确定区域的抽样(离散化)进行计算的,样本数量对计算精度影响很大;而本文提出的 PDBSCAN 聚类算法不存在这样的问题。

为了检验算法的效率,设对象的最大移动距离 $d = 25$m,采用 PDBSCAN 聚类算法和 FDBSCAN 聚类算法分别对具有不同移动对象数的数据集进行聚类,比较聚类所需的时间,结果如图 4 所示。从图中可以看出,在运行时间方面对于不同的数据规模采用本文提出的 PDBSCAN 聚类算法明显优于 FDBSCAN 聚类算法。原因在于 FDBSCAN 算法对数据不确定区域离散化带来了额外的时间花销,而 PDBSCAN 算法虽然直接基于不确定性数据在其不确定区域上的概率分布进行计算,但通过 R 树索引和概率阈值索引 PTI 预先对绝大部分不满足要求的对象进行排除,因此提高了聚类过程的效率。

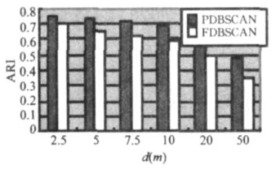

图 3 ARI 与最大移动距离 d 的关系

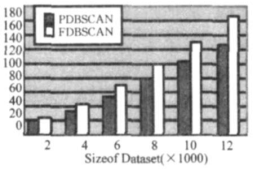

图 4 不同数据集规模所对应的聚类时间

结束语 随着传感器技术和无线通信技术的发展,对面向自然界的应用的需求越来越大,而从自然界采集到的数据内在的不确定性使得不确定性数据处理技术的研究成为当前科研的一个热点。本文分析了当前不确定性数据聚类的主要研究成果,并在此基础上提出基于密度的不确定性数据概率聚类算法 PDBSCAN,根据数据不确定区域的概率分布信息提高算法的准确性并通过 R 树索引和概率阈值索引 PTI 提高算法的效率。仿真试验表明,本文提出的方法在有效性和效率方面均优于当前主要的基于密度的不确定性数据聚类算法。概率阈值 p 的选取对聚类结果的影响有待于下一步的深入研究。

参 考 文 献

[1] Cheng R. Managing Uncertainty in Constantly-evolving Environments[D]. Purdue University, 2005

[2] Cheng R, Kalashnikov D V, Prabhakar S. Evaluating probabilistic queries over imprecise data[C]//The 2003 ACM SIGMOD International Conference on Management of Data. San Diego, 2003

[3] Cheng R, Xia Y, Prabhakar S, et al. Efficient indexing methods for probabilistic threshold queries over uncertain data[C]//The 30th International Conference on Very Large Data Bases. Toronto, 2004

[4] 许华杰,李国徽. 移动计算环境中易变数据的在线广播调度[J]. 计算机科学, 2009, 36(1)

[5] Dalvi N, Suciu D. Efficient query evaluation on probabilistic databases[C]//The 30th International Conference on Very Large Data Bases. Toronto, 2004

[6] Chau M, Cheng R, Kao B, et al. Uncertain Data Mining: An Example in Clustering Location Data[C]//The 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 2006

[7] Kriegel H-P, Pfeifle M. Density-based clustering of uncertain data[C]//The 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, 2005

[8] Ester M, Kriegel H-P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]//The 2nd International Conference on Knowledge Discovery and Data Mining. Portland, 1996

[9] Stonebraker M, Frew J, Gardels K, et al. The SEQUOIA 2000 Storage Benchmark[C]//The 1993 ACM SIGMOD International Conference on Management of Data. Washington, 1993

[10] Yeung K, Ruzzo W. An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data[J]. Bioinformatics, 2001, 17(9): 763-774

(上接第 55 页)

[6] Song R, Korba L, Yee G. AnonDSR: Efficient Anonymous Dynamic Source Routing for Mobile Ad-Hoc Networks[C]//Proc. ACM Workshop Security of Ad Hoc and Sensor Networks (SASN '05). 2005: 320-327

[7] Zhang Y, Liu W, Lou W. Anonymous Communications in Mobile Ad Hoc Networks[C]//Proc. INFOCOM. 2005: 1940-1951

[8] Boneh D, Franklin M. Identity-based encryption from the Weil pairing[C]//Advances in Cryptology-Crypto '01, LNCS 2139. Berlin: Springer-Verlag, 2003: 213-229

[9] Barreto P, Kim H Y, Lynn B. Scott. Efficient Algorithms for Pairing-Based Cryptosystems[C]//Proc. CRYPTO 02. Springer Verlag, August 2002: 354-368

[10] Fall K, Varadhan K. ns notes and documentation [EB/OL]. http://www-mash.cs.berkeley.edu/ns/, 2003

[11] Boneh D, Lynn B, Shacham H. Short signatures from the Weil pairing[C]//Advances in Cryptology-Asiacrypt 2001 Volume 2248 of Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2002: 514-532

[12] Bareeto P, Lynn B, Scott M. Efficient Implementation of Pairing-based Cryptosystems[J]. Journal of Cryptology, 2004, 17: 321-334