A Model of Clustering Uncertain Data

Zengfang Yang [1], Hewen Tang [2]

[1]*School of Information Technology and Engineering, Yuxu Normal University, Yuxu 653100,China*

[2] *School of Information Technology and Engineering, Yuxu Normal University, Yuxu 653100,China*

Email:yzf@yxnu.net, thw@yxnu.net

Abstract：In this paper we presented a model clustering uncertain data, first, generate M possible worlds by using Monte Carlo simulation data uncertainty, and then ,using DBSCAN algorithm to establish local clustering model on each possible world, finally, the M local clustering models combined into a global clustering model. Furthermore, new approach offers the possibility of creating the local clustering models in parallel, with this parallelism we wish can achieve better time performance and clustering quality.

Key words：Monte Carlo method; DBSCAN algorithm; uncertain data; possible worlds

一种不确定性数据聚类模型

杨增芳 [1]，唐合文 [2]

[1] *玉溪师范学院信息技术工程学院，玉溪，中国，653100*

[2] *玉溪师范学院信息技术工程学院，玉溪，中国，653100*

Email:yzf@yxnu.net, thw@yxnu.net

摘 要：本文提出了用确定的聚类算法来聚类不确定性数据的方法模型，首先用蒙特卡洛方法模拟数据不确定性，生成 M 个可能世界，然后，采用 DBSCAN 算法对每个可能世界建立各自的局部聚类模型，最后，把 M 个局部聚类模型合并成为最终的一个全局聚类模型。同时，本文提出了以并行的方式来创建局部聚类模型的可能性，通过这种并行性，希望能获得较好的时间性能和聚类质量。

关键词：蒙特卡罗方法；DBSCAN算法；不确定性数据；可能世界

1 引言

不确定数据是在传感器网络WSN、无线射频识别RFID、GPS定位、隐私保护等应用中涌现出来的一类数据，其特点是每个数据对象不是单个数据点，而是按照概率在多个数据点上出现。近年来，随着数据库技术的发展以及人们对数据采集和处理技术理解的不断深入，不确定性数据(uncertain data)得到了广泛的重视，在许多现实的应用中，例如金融、保险、电信、军事、物流等领域不确定信息普遍存在，不确定性数据在知识发现中扮演着关键角色，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。

目前数据挖掘领域的绝大部分研究成果都是针对“确定”数据的，对不确定性数据聚类的研究成果不多。文献[1]首先将不确定性数据挖掘作为一个新的研究方向提出来，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。该方法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[2]。况且K-means聚类方法具有不适宜发现非凸状簇、对噪声和离群点敏感等缺点，因此文献[1]提出的方法实用性有限。文献[2]中，Kriegel在著名的基于密度的聚类方法DBSCAN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。FDBSCAN算法聚类过程同DBSCAN算法非常相似，只是在相似性度量上重新定义了距离公式，采用的不是基于欧几里德距离期望，而是反映概率密集程度的距离分布函数，针对不确定数据，这种度量使得聚类结果的精度更高，但该方法的计算是基于对象连续分布的抽样，因此计算精度和时间都无法

项目支持：本文由科技基础性工作专项专题“云南农业生物资源调查 GPS/GIS 建立”（编号：2006FY11070901）支持。

978-1-935068-18-1 © 2010 SciRes.

保证，有可能对聚类结果产生影响。

DBSCAN[3]是一种传统的基于密度的聚类方法，由于它具有适用于各种形状簇、对噪声和离群点不敏感等优点，但DBSCAN没有考虑数据的不确定性，因此，本文中引入Monte Carlo方法来对数据的不确定性进行模拟生成可能世界，可能世界内的数据是确定的，这样，就可以在可能世界上来使用经典的BDSCAN算法进行聚类。

本文组织如下，在第2节中，首先对可能世界模型进行介绍，然后对Monte Carlo的思想和方法进行介绍。第3节中，提出一种基于Monte Carlo的不确定聚类模型。第4节分析本文所提出的方法模型的时间性能，第5节对本文的主要工作进行总结，同时，提出了今后需要研究解决的问题。

2 相关工作

2.1 可能世界模型

定义与应用场景相匹配的数据模型是不确定性数据管理的首要任务。在不确定性数据管理领域，最常用的模型是可能世界模型(possible world mode1)[6][7],该模型从一个不确定性数据库演化出很多确定的数据库实例(称为可能世界实例)，而且所有实例的概率之和为1。不确定性数据的种类较多，例如关系型数据、半结构化数据、流数据、移动对象数据等，尽管存在许多与数据类型紧密相关的数据模型，但是这些模型最终都可以转化为可能世界模型。

考虑一个例子，假设在一个客户数据库中有客户基本数据和其它若干属性数据CID(表示客户编号)、REGION（表示客户所生活的地区）以及CITY（表示客户所居住的城市）。假设我们所知道是客户所生活的地区而不知道准确的居住城市，那么用属性PROB来表示元组存在的概率，如表1所示，表中的数据具有存在级不确定性。

Table 1. Uncertain data
表1. 不确定数据

CID	REGION	CITY	PROB
1	西南	成都	0.3
2	华东	上海	0.7
3	华北	天津	0.6

元组之间可能独立也可能存在依赖关系，首先假设各个元组之间独立，则共有2^3=8个可能世界实例，各实例的概率等于实例内元组的概率乘积与实例外元组的不发生概率的乘积，如图1所示。例如，可能世界实例{1,2}的发生概率为0.3×0.7×(1-0.6)=0.084。某些场景下，元组之间并非独立，而是存在依赖关系，这种依赖关系可以用规则描述，假设规则为1⊕3，即元组1与元组3不能够同时发生，但可以同时不发生[10]。总共有6个可能世界实例，如图1所示。可能世界实例{1}的发生概率为0.3×(1-0.7)=0.09，可能世界实例{2}的发生概率为(1-0.3-0.6)×0.7=0.07。

元组独立：
PW={{},{1},{2},{3},{1,2},{1,3},{2,3},{1,2,3}}
P(PW)={0.084,0.036,0.196,0.126,0.084,0.054,0.294,0.126}
依赖规则1⊕3：
PW={{},{1},{2},{3},{1,2},{2,3}}
P(PW)= {0.03,0.09, 0.07,0.18, 0.21,0.42}

Figure 1. Example possible world
图 1. 可能世界样例

2.2 蒙特卡罗方法

蒙特卡罗（Monte Carlo）[11]即随机模拟方法，是用计算机模拟随机现象，通过仿真试验，得到实验数据，再进行分析推断，得到某些现象的规律或某些问题的求解的方法。例如在许多工程、通讯、金融等技术问题中，所研究的控制过程通常伴有随机的、不确定的因素，若要从理论上很好地挖掘出实际规律，必须把这些因素考虑进去。理想化的方法是在相同条件下进行大量重复试验，采集试验数据，再对数据进行统计分析，得出其规律。但是这样做耗时耗力，尤其当一个试验周期很长，或是一个破坏性的试验时，通过试验采集数据几乎无法进行，此时蒙特卡罗方法就是最简单实用的方法。下面介绍蒙特卡罗方法产生随机数的基本原理及步骤。

2.2.1 蒙特卡罗方法基本原理

蒙特卡罗方法以随机模拟和统计试验为手段，是一种从随机变量的概率分布中，通过随机选择数字的方法产生一种符合该随机变量概率分布特性的随机数值序列，作为输入变量序列进行特定的模拟试验、求解的方法[11]。在应用方法时，要求产生的随机数序列应该符合该随机变量特定的概率分布。若需要产生各种特定的、不均匀的概率分布的随机数序列，其可行的方法是先产生一种均匀分布的随机数序列，然后再设法转换成特定要求的概率分布的随机数序列，以此作为数字模拟试验的输入变量序列进行模拟求解。

2.2.2 蒙特卡罗方法的基本步骤

蒙特卡罗方法的基本步骤所图2所示：

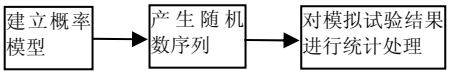

Figure 2. Monte Carlo basic step

图 2. 蒙特卡罗方法的基本步骤

首先，建立概率模型，即对所研究的问题构造一个符合其特点的概率模型(随机事件，随机变量等)；然后，产生随机数序列。在模型建立的情况下，要先产生随机抽样值，即在给定运行中各参数的统计分布规律的条件下，在计算机上产生符合其分布规律的随机数抽样值，这个过程称为伪随机数的模拟；最后，对模拟试验结果进行统计处理(计算频率、均值等特征值)，给出所求问题的解和解的精度的估计。

3 一种基于 Monte Carlo 的不确定聚类模型

本节中，以DBSCAN算法为基础，采用2.2节介绍的数据不确定性的Monte Carlo模拟方法，取样得到M个确定的数据库实例(称为可能世界实例)，然后在每个可能世界上进行DBSCAN聚类分析得到M个局部聚类模型，最后，采用文献[5]中的聚类合并技术把M个局部聚类模型合并成为最终全局聚类模型。完整的过程如图3所示。

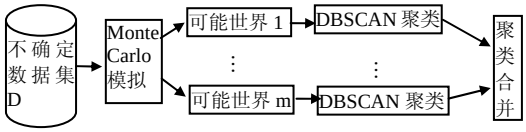

Figure.3 A model clustering uncertain data on possible world

图 3. 可能世界聚类模型

3.1 数据不确定性的 Monte Carlo 模拟

假设一个不确定数据库D有N个元组，这N个元组由不确定性定义来注释，生成M个可能世界。为了生成可能世界，我们借助第2.2节中介绍的Monte Carlo方法对其进行不确定性模拟，利用数据集D中的元组的不确定性定义来生成每个元组的随机值，由于Monte Carlo是在概率密度函数的基础上生成新的元组，因此M个可能世界的并集可以看作数据集D中所有原始点的概率密度函数代表。具体算法如下：

（1）确定每个待输入数据集的不确定性类型(比如基本数据的圆形正态模型，其它属性数据的一维正态模型)；

（2）取用服从数据集分布的随机采样来代替原输入数据；

（3）对每一次实现，存储其结果；

（4）随机抽取M组实验数据作为样本数据。

3.2 可能世界的 DBSCAN 聚类模型

生成了M个可能世界后，接下来就是在每个可能世界上来执行聚类分析。每一个可能世界都有各自的局部聚类模型，因此，可能产生M个局部聚类模型。由于Monte Carlo方法所生成的可能世界彼此独立，因此每一个可能世界都可以彼此并行来执行聚类处理，所以，这种方法有较高的并行度。在多内核和多CPU系统中，这种方法与经典的不确定聚类方法相比，可以实现高速处理。

3.3 聚类合并

由于我们的目标是对数据库D生成一个全局聚类模型，最后的聚类结果必须以某种方式从局部聚类模型中得到，为此，采用了文献[5]中所介绍的聚类合并方法，具体做法是：

（1）首先，建立M个局部相似图。每个局部相似图是这样构建的：点P_i是一个节点，一个聚类模型中，包含在同一簇中的所有点用一条权重为1的边连接起来，形成一个局部相似图。结果对M个局部聚类模型将产生M个局部相似图；

（2）然后，利用[5]中的成对相似技术把各个局部相似图合成为一个全局相似图；

（3）最后再利用全局相似图来生成一个全局聚类模型。

4. 时间性能分析

本文提出的不确定性数据聚类模型的时间开销主要由以下三方面决定，一是生成可能世界，二是产生局部聚类，三是聚类合并。本文还提出了以并行的方式来创建局部聚类模型的可能性，通过这种并行性，希望改善时间性能。第一部分时间开销是生成可能世界，对于有n个数据对象的数据集合而言，生成一个可能世界需要遍历n个元组，那么进行m次蒙特卡罗迭代生成m个可能世界的时间开销则为 $o(m\times n)$，这里 $m << n$。第二部分时间开销是产生局部聚类，对每个可能世界采用DBSCAN方法进行聚类，若采用空间索引结构则时间复杂度为 $o(n\log_2 n)$，如果m个可能世界聚类可以分配到P

978-1-935068-18-1 © 2010 SciRes.

个线程上并行执行，那么生成m个局部聚类的时间开销 $\frac{m}{p}\times(n\log_2 n)$。第三部分时间开销是聚类合并，文献[8]中，两个聚类合并算法的时间复杂度为 $o(n^2)$，那么m个局部聚类进行合并的时间开销为 $(m-1)\times n^2$， 这里 $m << n$。

5 总结与展望

本文提出了一种用确定聚类方法来聚类不确定性数据的方法模型，第一步，用蒙特卡洛方法（Monte Carlo）来对不确定性数据进行模拟生成多个可能世界（可能世界内的数据是确定的）；第二步，对每个可能世界建立各自的聚类模型；第三步，把各个局部聚类结果最终合并成为一个全局聚类模型。

关于本文中的方法模型，还有许多需要研究解决的问题，例如，如何产生最优的可能世界数量M的问题是今后需要研究的问题。本文提出的方法与所采用的并行组件相关，因为它必须产生许多可能世界来近似确定概率密度，因此所需要的并行资源可能超过目前的CPU的并行处理技术，所以，有许多可能世界不得不以串行方式来执行，导致了算法速度的急剧下降，为了解决这个问题，今后将需要进一步对优化可能世界数量的策略进行研究。

References (参考文献)

[1] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data[C]. In PAKDD :volume 3918 of Lecture Notes in Computer Science, pages 199–204. Springer, 2006

[2] H.-P. Kriegel and M.Pfeifle. Density-based clustering of uncertain data[C]. In KDD ’05: Proceedings of theeleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 672–677, New York, NY, USA, ACM, 2005.

[3] M.Ester, H.-P.Kriegel, J.Sander, and X.Xu. A density_based algorithm for discovering clusters in large spatial databases with noise[C]. In E.Simoudis, J.Han, and U.Fayyad,editors, Second International Conference on Knowledge Discovery and Data Mining, pages 226–231, Portland, Oregon,1996. AAAI Press.

[4] R.Jampani, F. Xu, M. Wu, L. L. Perez, C. M. Jermaine, and P. J. Haas. Mcdb: a monte carlo approach to managing uncertaindata[C]. In J. T.-L. Wang, editor, SIGMOD Conference, pages 687–700. ACM, 2008.

[5] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation[J]. ACM Trans. Knowl. Discov. Data, 1(1):4, 2007.

[6] S.Abiteboul，P.Kanellakis，G.Grahne．On the representation and querying of sets of possible worlds[J]． ACM SIGMOD Record，1987，16(3)：34—48

[7] T.J.Green，V.Tannen．Models for incomplete and probabilistic information[J]． IEEE Date Engineering Bulletin， 2006， 29(1)：17-2

[8] C. Re, N. N. Dalvi, and D. Suciu. Query evaluation on probabilistic databases[J]. IEEE Data Eng. Bull., 29(1):25–31, 2006.

[9] C.C.Aggarwal, P.S.Yu. A survey of uncertain data algorithms and applications[R]. IBM Research Report RC24394, 2007.

[10] C. Re, N. N. Dalvi, and D. Suciu. Efficient top-k query evaluation on probabilistic data. In ICDE, pages 886–895,2007.

[11] G.Fishman. Monte Carlo: Concepts, Algorithms, and Applications[M]. Springer, 1996.