基于聚类高维空间算法的离群数据挖掘技术研究

项响琴[1,2],汪彩梅[2]
(1.安徽大学 计算机科学与技术学院,安徽 合肥 230039;
2.合肥学院 网络与智能信息处理中心实验室,安徽 合肥 230601)

摘 要:离群数据挖掘是数据挖掘领域的一个研究分支,而聚类算法分析则是进行离群数据挖掘的重要研究方法之一。文中首先分析研究离群数据挖掘方法,对多个离群数据挖掘算法进行分析比较,讨论各自的优点和不足,同时针对高维空间数据的特点,分析挖掘高维空间数据中的离群点方法。其次对聚类分析算法进行讨论,分析一种基于网格和基于密度的聚类方法——聚类高维空间算法(CLIQUE算法),运用它可以更好地挖掘高维空间中的离群数据。提出了CLIQUE算法的有待改进的思想,为以后的研究指明方向。

关键词:数据挖掘;离群点;聚类分析;CLIQUE算法
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2010)01-0124-04

Study of Outlier Data Mining Based on CLIQUE Algorithm

XIANG Xiang-qin[1,2], WANG Cai-mei[2]
(1.Dept. of Computer Science & Engineering, Anhui University, Hefei 230039, China;
2.Laboratory of Network and Intelligent Information Management, Hefei University, Hefei 230601, China)

Abstract: As a branch of data-mining, outlier mining is a promising prospect, and clustering analysis is a kind of technology in spatial outlier mining. Analyse the clustering arithmetic, compare some arithmetic of clustring, and discuss the strongpoint and shortpoint of them. Research the spatial data and outlier attributes in high dimensional space. And analysing the CLIQUE algorithm to detect the outlier in high dimensional space, this approach can find the outliers in high-dimensional space effectively. In conclusion, the main trends of spatial outlier mining are forecaste.

Key words: data mining; outlier; clustering analysis; CLIQUE algorithm

0 引 言

数据挖掘也称KDD(Knowledge Discovery in Databases),是从大量数据源中发现正确的、新颖的、潜在的、有用的并能够被理解的知识的过程[1]。KDD的方法可以是数学的,也可以是非数学的;可以是演绎的,也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持和过程控制等,还可以用于数据自身的维护。数据挖掘是一门交叉性学科,涉及到人工智能、数据库、统计学、机器学习、知识获取等多个领域,因此,数据挖掘的研究引来了多门学科专家学者的兴趣和关注。

数据挖掘的一般操作流程是:根据挖掘算法(如聚类算法、分类算法、关联规则、线性回归等)生成数据挖掘模型,应用挖掘模型对相关数据进行挖掘,最终得到需要的挖掘结果,也就是发现隐藏在数据中的知识。

数据挖掘(Data Mining)技术是当前多个领域研究的热点课题,文中所讨论的离群数据挖掘(Outlier Mining)是数据挖掘的一个重要研究分支,它是从大量的数据集中挖掘出明显不满足一般规律或模式、行为异常的少量特殊数据。然而在数据挖掘研究的开始阶段,离群点常会被认为是不正常数据或"噪音",最终将其隔离,或修改其部分信息而让其接近于常规模式,忽视了对其研究的价值。后经研究发现,其实这些稀有的数据让人们从中获取更有价值的信息,比常规事件更令人感兴趣,如金融和通信领域诈骗分析、网络入侵检测、过程控制中的故障分析与诊断、恶劣天气预测、信用卡恶意透支等方面。目前,离群数据挖掘正逐渐引起人们的重视。

离群数据挖掘[2],可以理解为进行常规数据挖掘时,发现常规知识的副产品。通常人们是运用常用数据挖掘技术,挖掘出常规数据同时,就可得到非常规的

收稿日期:2009-05-04;**修回日期**:2009-08-09
基金项目:安徽省自然科学基金项目(KJ2009B122,KJ2008B03)
作者简介:项响琴(1976-),女,讲师,硕士研究生,研究方向为数据挖掘与智能软件。

数据,即离群数据。

1　离群数据挖掘

离群数据是大量的数据集中明显不满足一般规律或模式、行为异常的少量特殊数据,这些数据可能来自于人为度量过程错误、机器执行错误或特殊事件发生等。对离群点检测是为了发现数据集中极少数的一些数据,然而研究人员发现,在日常事务处理过程中,就是这些极少数离群数据的挖掘常常比其他常规数据的挖掘更有价值。因为这些数据可能就是一些非正常信息的真实反映,一万个正常的记录很可能只覆盖一条规则,而十个离群点很可能就意味着十条不同的规则[2],如在欺诈探测中,孤立点可能预示着欺诈行为。所以,离群点的检测,往往可以发现一些真实的、但又出乎意料的知识[3]。

近年来,从事数据挖掘的研究人员提出了大量离群数据挖掘方法,下面对其中几种方法进行讨论。

1.1　基于统计的离群数据挖掘

统计的方法对给定的数据集合假设了一个分布或概率模型(如正态分布),然后根据模型采用不一致性检验来确定孤立点[2]。基于统计的方法[4]是出现最早的离群点检测方法,它基于对小概率事件的判别来实现对数据样本异常的鉴别。其主要思想是假定数据集服从某种分布或概率模型,通过不一致检验把那些严重偏离分布曲线的记录视为离群点。

基于统计的方法检测出来的离群点很可能被不同的分布模型检测出来,可以说产生这些离群点的机制可能不唯一,解释离群点的意义时经常发生多义性,这是基于统计方法的一个缺陷;其次,基于统计的方法在很大程度上依赖于待挖掘的数据集是否满足某种概率分布模型[4],模型的参数、离群点的数目等对基于统计的方法都有非常重要的意义,而确定这些参数通常都比较困难。

1.2　基于距离的离群数据挖掘

基于距离的离群点最早是由 Knorr 和 Ng[4]提出的,他们把所有记录看作高维空间中存在的点,而离群点则被定义为数据集中与大多数点之间的距离都大于某个阈值的点,通常被描述为 DB(pct,dmin)。数据集 T 中一个记录 o 称为离群点,当且仅当数据集 T 中至少有 pct 部分的数据与记录 o 的距离大于 dmin 的。基于距离的离群数据挖掘算法,即使数据集不满足任何特定分布模型,它仍能有效地发现离群点,特别是当空间维数比较高时,算法的效率比基于密度的方法要高得多[5]。

目前,已开发了若干种高效的基于距离的离群数据挖掘算法[2],简单描述如下:

(1)基于索引检测算法:给定一个数据集,采用多维索引结构(如 R 树或 k－d 树),来查找每个对象 o 在半径 d 范围内的邻居。设 M 是一个孤立点的 d－领域内的最大对象数目,一旦对象 o 的 $M+1$ 个邻居被发现,o 就不是离群点。

(2)基于嵌套－循环检测算法:与前一算法相比,该算法避免了索引结构的构建,试图最小化 I/O 的次数,是将内存的缓冲区划分为两半,数据集合分为若干个逻辑块。

(3)基于单元(cell－based)检测算法:把数据集划分为单元,逐个单元的检测,而非逐个对象的检测。Knorr 和 Ng 通过试验证明,当 $k \leqslant 4$ 时此算法优于基于嵌套 k－循环检测算法。经研究表明,此三种方法对于高维空间中的大数据集,算法的效率都不高。

1.3　基于密度的离群数据挖掘

M. M. Breunig 等人提出基于密度的挖掘离群点方法[6]。基于密度的离群数据挖掘算法一般都建立在距离的基础上,某种意义上可以说基于密度的方法是基于距离的方法中的一种,但基于密度的异常观点比基于距离的异常观点更贴近 Hawkins 的异常定义,因此能够检测出基于距离的异常算法所不能识别的一类异常数据——局部异常。

1.4　基于深度的离群数据挖掘

基于深度的离群点检测算法[7,8]的主要思想是,先把每个记录标记为 k 维空间里的一个点,然后根据深度的定义,给每个点赋予一个深度值;再根据深度值按层组织数据集,深度值较小的记录是离群点的可能性比深度值较大的记录大得多,因此算法只需要在深度值较小的层上进行离群检测,不需要在深度值大的记录层进行离群检测。基于深度的方法比较有代表性的有 Struyf 和 Rousseeuw 提出的 DEEPLOC[9]算法。虽然,理论上基于深度的识别算法可以处理高维数据,然而实际计算时,处理高维数据显得很吃力。

1.5　基于聚类的离群数据挖掘

将物理或抽象的集合分组成为由类似的对象组成的多个类的过程被称为聚类。由聚类所产生的簇是一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中的对象相异。针对目前的多种离群数据检测算法,应依据数据集中信息的性质和特点,择优进行选取。在实际应用中,还可以先采用聚类算法对数据集进行聚类,然后再采用孤立点检测算法来发现离群点。

聚类(Clustring)分析是数据挖掘的重要手段之一。聚类就是将数据对象分组成为多个类或簇,由聚

类所生成的簇是一组数据对象的集合，在同一簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。在实际应用中，可以将一个簇中的数据对象作为一个整体对待。目前有很多的聚类算法，如CLARANS, DBSCAN, STING, BIRCH, ROCK 等。它们具有一定的异常处理的能力，但是主要目标是产生聚类，即寻找性质相同或相近的记录并归为一个类，不属于任何一个类的记录，就是要挖掘的孤立点。所以说，离群点可以认为是聚类分析的副产品。

早期的离群检测多见于统计领域，其实，离群数据挖掘与聚类分析密切相关，有一些典型的具有离群检测功能的聚类算法，如 CLARANS, DBSCAN, OPTICS 等。聚类的目的主要在于寻找类别，离群点是它们的一个附属物，正是抓住了这一点，利用上述提到的聚类算法，作为离群点检测的前期工作，通过聚类算法找出不属于任一常规模型的孤立点。但多数聚类算法只能处理低维的空间数据，处理高维的空间数据就显得力不从心了。在高维空间中聚类数据对象是非常有挑战性的，特别是有些数据集可能非常稀疏而且高度偏斜。

针对这一点，下面介绍基于密度和基于网格的聚类算法 CLIQUE，来对高维空间数据集中孤立点进行检测。

2 高维空间的离群数据挖掘

2.1 高维空间数据

高维数据的特性不同于低维数据[10]，高维空间中的数据分布得比较稀疏，这使得高维空间中数据之间的距离尺度及区域密度不再具有直观的意义，所以高维空间中的离群点的检测也不同于传统的离群点发现。空间数据与其他类型数据的本质区别是其空间属性。空间属性包括空间位置、距离、几何形状、大小等内容，并且可引伸为空间个体之间的相互关系，如拓扑关系、方位关系、度量关系等，从而使得空间数据比其他类型的数据要更为复杂。

为解决高维空间离群数据的发现问题，尝试把高维空间的数据投影到低维子空间中来发现数据异常点。这样就不再平等地看待各个维，而是通过某些标准选择出一些维，在这些维中组成的自空间中寻找近邻。随着维数的增加，对维数进行组合得到的子空间个数成指数级增长（数据集为 n，可能的组合有 $2n$）。要构造好的高维空间离群点检测算法，必须充分利用高维数据自身的特性，处理好数据集的主体聚类性与稀疏性之间的矛盾，以提高异常检测算法的效率。

2.2 高维空间数据的离群数据挖掘方法

Aggarwal 和 Yu（SIGMOD'2001）提出一个高维数据异常检测的方法[11]。他把高维数据集映射到低维子空间，根据子空间映射数据的稀疏程度来确定异常数据是否存在。

文献[6]提出，为了使算法在高维空间上也有效，离群点发现算法必须满足以下要求：

（1）能够有效地解决高维空间中数据的稀疏问题；

（2）能够解释是什么原因造成了数据的异常；

（3）能够找到合适的衡量办法，给出 k 维子空间中离群点的物理意义；

（4）对高维空间中的数据仍然是计算高效的；

（5）判断一个点是否为离群点时，要考虑到数据点的局部行为。

同时，依据这些标准，文献[4]提出了一种新的方法，通过观察数据投影后的密度分布来发现离群点。该方法用演化计算来寻找最优的子空间，并根据数据的特点对选择、交换、变异算子进行了调整，能够比较有效地找到高维空间中的离群点。

3 基于聚类高维空间的离群数据挖掘

3.1 聚类分析

聚类（Clustring）分析是数据挖掘的重要手段之一。聚类就是将数据对象分组成为多个类或簇，由聚类所生成的簇是一组数据对象的集合，在同一簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。

常见的聚类分析方法有：

（1）划分的方法：给定一个 n 个对象或元组的数据库，一个划分方法构建数据的 k 个划分，每个划分表示一个聚簇，并且 $k < n$。也就是说，它将数据划分为 k 个组，同时满足如下的要求：(a)每个组至少包含一个对象；(b)每一个对象必须属于且只属于一个组。

为了达到全局最优，基于划分的聚类会要求穷尽所以可能的划分。实际上，绝大多数应用采用了以下两个比较流行的启发式方法：(a)k－平均算法，在该算法中每个簇用该簇中对象的平均值来表示；(b)k－中心点算法，在该算法中，每一个簇用接近聚类中心的一个对象来表示。这些启发式聚类方法在中小规模的数据库中发现球状簇很适用。

（2）层次的方法：层次方法对给定的数据对象集合进行层次的分级。根据层次的分解如何形成，层次的方法可以分为凝聚的和分裂的。凝聚的方法，也称为自底向上方法，一开始将每个对象作为单独的一个组，然后相继地合并相近的对象和组，直到所有的组合合并为一个，或者达到一个中止条件。分裂的方法，也称

为自顶向下的方法,一开始将所有的对象置于一个簇中。在迭代的每一步中,一个簇被分裂为更小的簇,直到最终每个对象在单独的一个簇中,或者达到一个终止条件。

(3)基于密度的方法:绝大多数划分方法基于对象之间的距离进行聚类。这样的方法[12]只能发现球状的簇,而在发现任意形状的簇上遇到了困难。随之提出了基于密度的另外一类聚类算法,主要思想是:只要邻近区域的密度,即邻近区域的数据点的数目超过了一个阈值,就继续聚类。也就是说,对给定类中的每个数据点,在一个给定范围内的区域中必须至少包含某个数目的点。基于密度的方法[13]主要思想是将记录之间的距离和某一给定范围内记录数这两个参数结合起来,从而得到"密度"的概念,然后根据密度判定记录是否为离群点。

(4)基于网格的方法:它把对象空间量化为有限数目的单元,形成了一个网格结构。这种方法的主要优点是处理速度很快。

(5)基于模型的方法:它为每个簇假定了一个模型,寻找数据对给定模型的最佳拟合。一个基于模型的算法可能通过构建反映数据点空间分布的密度函数来定位聚类。它也基于标准的统计数字自动决定聚类的数目,从而产生健壮的聚类方法[1]。

3.2　基于聚类高维空间(CLIQUE)算法

CLIQUE(Clustering In QUEst) 聚类算法综合了基于密度和基于网格的聚类方法[2]。它对于大型数据库中的高维数据的聚类非常有效。CLIQUE 作为一种基于网格和密度的算法,可以对高维数据进行全面聚类和子空间聚类,并且有着良好的可伸缩性和数据处理能力。CLIQUE 算法把数据空间分割成网格单元,将落到某个单元中的点的个数当成这个单元的密度,可以指定一个阈值,当某个单元中的点的个数大于这一阈值时,就说这个单元格是稠密的[14]。

CLIQUE 算法采用关联规则挖掘中的先验性质:如果一个 k 维单元是密集的,那么它在 $k-1$ 维空间上的投影也是密集的。也就是说,给定一个 k 维的候选密集单元,如果检查它的 $k-1$ 维投影单元,发现任何一个不是密集的,那么就知道第 k 维的单元也不可能是密集的[14]。因此,可以从 $k-1$ 维空间中发现的密集单元,来推测 k 维空间中潜在的或候选的密集单元。

CLIQUE 算法能自动地发现最高维的子空间,高密度聚类存在于这些子空间中。同时,CLIQUE 对元组的输入顺序不敏感,无需假设任何规范的数据分布。它随输入数据大小线性扩展,当数据的维数增加时具有良好伸缩性。

聚类高维空间 CLIQUE 算法的核心思想[2]:

(1)给定一个多维数据点的大集合,数据点在数据空间中通常不是均衡分布的。CLIQUE 区分空间中稀疏的和拥挤的区域,以发现数据集合的全局分布模式。

(2)如果一个单元中的包含的数据点超过了某个输入参数,该单元是密集的。在 CLIQUE 中,相连的密集单元的最大集合定义为簇。

聚类高维空间 CLIQUE 进行多维聚类一般分两步进行。首先,CLIQUE 对 n 维数据空间进行划分,划分为互不相交的长方形单元,识别其中的密集单元。再者,CLIQUE 为每个簇生成最小化的描述。对于每个簇,它确定覆盖相连的密集单元的最大区域,然后确定最小的覆盖。

3.3　CLIQUE 算法的局限性

CLIQUE 算法在实际运用中,也有其局限性:

(1)它根据用户输入的参数等宽分割每一维,这样会导致可能有某一聚类被人为地分割成多个区域,而在覆盖相连的密集单元时又将其相连,使得划分单元的数目增加,在高维情况下,相邻单元数量以指数级增长,在覆盖相连阶段又花费大量的时间;

(2)基于最小描述长度的剪枝算法,把在同一子空间中的密集单元分组,并且找出每一个子空间中密集单元选出的数据覆盖,覆盖大的子空间被选出,其余的被剪枝,在自底向上的算法中,为了发现一个 k 维的密集所有的子空间都应该被考虑。但是,如果这些子空间在被剪掉的空间中,那么这个密集就不可能被发现了,虽然提高了算法的运行效率,但同时也丢失了一些很重要的聚类[14]。

4　结束语

研究表明,离群点的检测在很多领域中有着非常重要的应用,随着研究的深入,它将在更多领域中发挥更重要的作用。

文中首先对离群数据挖掘方法进行分析,给出几种算法的适应性和对高维空间数据分析的不足,同时对聚类算法进行了讨论。讨论了聚类算法在离群数据挖掘中的使用。研究发现大多离群检测算法在高维空间、时间序列以及地理数据中的效率还比较低,算法还或多或少地存在各种不同的缺陷。

最后主要是对一种基于网格和密度的聚类 CLIQUE 算法进行分析与研究,介绍了 CLIQUE 算法的思路和实现步骤,同时也提出了 CLIQUE 算法的有待改进的思想,为以后的研究指明方向。

(下转第 131 页)

节越清楚;尺度半径越大,边缘越粗。还可以看出,多尺度形态滤波的效果要比单一结构元素的滤波效果要好,不但很好地去除了噪声,而且也抑制了边缘模糊的现象,更适合人眼视觉的识别。因此,在彩色图像去噪时,可选用对彩色图像颜色通道分别进行滤波去噪的多尺度形态学滤波,不但对噪声有较好的抑制作用,还可克服线性滤波边缘模糊和细节损失的缺陷。滤波器的性能比较见表 1。

表 1　滤波器的性能比较

滤波器 \ 性能参数	MSE1	PSNR1	MSE2	PSNR2
单尺度 se2	1.3333e-004	200.0519	2.7127e-005	215.9752
单尺度 se3	0.0019	173.6167	8.3076e-005	204.7829
单尺度 se1	0.0010	179.8198	0.0041	165.8647
多尺度 se1、se2、se3	0.0019	173.6167	8.3076e-005	204.7829

4　结束语

图像去噪是图像处理和机器视觉的一个非常基础而又重要的课题。文中采用结合形态学滤波的多尺度方法对图像进行去噪,可克服线性滤波边缘模糊和细节信息损失的缺陷。试验结果表明该方法对噪声有较好的抑制作用,保持原图像的细节信息,具有一定的可行性和实用性。根据此方法的特点,该方法适用的图像类型是图像中的对象尺寸都比较大,没有细小的细节,对这种类型的图像除噪的效果会比较好。

为了能使从噪声污染的图像中恢复原始图像的结果达到最优,在确定结构元素半径时,可以采用优化方法。具体操作中为达到这一目的,可将图像和噪声视为随机过程,通过经验数据或统计优化分析得到优化结果。

参考文献:

[1] 徐　飞,施晓红. MATLAB 应用图像处理[M]. 西安:西安电子科技大学出版社,2002:101-124.

[2] 冈萨雷斯. 数字图像处理[M]. 第 2 版. 北京:电子工业出版社,2007:420-450.

[3] 邹永星,周仁魁,罗秀娟,等. 一种提取图像目标边缘的新方法[J]. 光电工程,2005,32(6):76-78.

[4] 崔　屹. 图像处理与分析——数学形态学方法及应用[M]. 北京:科学出版社,2000.

[5] 李　卓,郭立红. 多尺度形态学边缘检测算法[J]. 电子器件,2006,29(3):821-824.

[6] 何东健,耿　楠,张义宽. 数字图像处理[M]. 西安:西安电子科技大学出版社,2003:175-197.

[7] 商艳丽,王小鹏,夏志成. 一种基于多尺度形态学的彩色图像边缘检测方法[J]. 电子元器件应用,2007,9(8):68-70.

[8] 胡学龙,许开宇. 数字图像处理[M]. 北京:电子工业出版社,2006:170-182.

(上接第 127 页)

参考文献:

[1] 吕建军. 数据挖掘技术的应用研究[D],北京:中国农业大学,2002.

[2] Han Jiawei, Kamber M. 数据挖掘:概念与技术[M]. 范　明等译,北京:机械工业出版社,2001.

[3] 蔡江辉,张华煜. 离群数据挖掘方法研究[J]. 电脑开发与应用,2005,18(12):46-47.

[4] Knorr E M, Ng R T. Algorithms for Mining Distance-based Outliers in Large Datasets[C]//New York: Proc. of Int. Conf. Very Large Data-bases(VLDB'98). New York:[s. n.],1998:392-403.

[5] Wang W, Yang J, Muntz T R. Sting: A Statistical Information Grid Approach to Spatial Data Mining[C]//Jarke M, Carey M J, Dittrich K R, et al. Proc. of Bases. Athens: Morgan Kaufmann,1997.

[6] Aggarawal C C, Yu P S. Outliers Detection for High Dimen-sional Data [C]//In: Aref W G. Proceedings of the ACM SIGMOD International Conference on Management of Data. Santa Barbara, CA: ACM Press, 2001:37-47.

[7] Tukey J W. Exploratory Data Analysis[M]. MA: Addison Wesley and Sons, Inc., 1994.

[8] Preparata F, Shamos M. Computational Geometry: An Introduction[M]. [s. l.]: Springer-Verlag, 1988.

[9] Struyf A, Rousseeuw P J. High-dimensional Computation of the Deepest Location[J]. Computational Statistics and Data Analysis, 2000, 34:415-426.

[10] 魏　藜,宫学庆. 高维空间中的离群点发现[J]. 软件学报,2002,13(2):280-282.

[11] 熊君丽. 高维空间下基于密度的离群点探测算法实现[J]. 现代电子技术,2006(15):67-69.

[12] 崔贯勋,朱庆生. 一种改进的基于密度的离群数据挖掘算法[J]. 计算机应用,2007,27(3):559-560.

[13] 黄洪宇,林甲祥. 离群数据挖掘综述[J]. 计算机应用研究,2006(8):8-13.

[14] 刘嘉嘉. CLIQUE 算法改进及其在电子商务企业中的应用与研究[D]. 合肥:合肥工业大学,2007.