维普资讯 http://www.cqvip.com

一种基于主成分分析的异常点挖掘方法*)

王洪春[1,2] 彭 宏[1]
(华南理工大学计算机科学与工程学院 广州 510641)[1]
(重庆师范大学数学与计算机科学学院 重庆 400047)[2]

摘 要 在对现有异常点挖掘算法分析的基础上,给出了一种异常点挖掘的新方法—基于主成分分析方法,该方法先用基于密度的聚类算法进行聚类,然后把不包含在任何聚类中的周围稀疏的样本对象用主成分分析(PCA)方法进行检验,确定是否为异常点,并通过实验数据验证了算法的可行性和有效性。

关键词 异常点,主成分分析,数据挖掘,聚类算法

A Outlier Mining Algorithm Based on Principal Component Analysis

WANG Hong-Chun[1,2] PENG Hong[1]
(Dept. of Computer and Engineering, South China University of Technology, Guangzhou 510641)[1]
(Dept. of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047)[2]

Abstract Based on the analysis of the existing algorithms of outlier mining, a new outlier mining algorithm is put forward based on principal component analysis, it clustering firstly with Density-based algorithm, and determine outliers according to principal component analysis within the sparse samples around the clusters that it don't contain any clusters, experimental results show the feasibility and effectiveness of the new algorithm.

Keywords Outlier, Principal component analysis(PCA), Data mining, Clustering algorithm

1 引言

在数据库中包含着少数的数据对象,它们与数据的一般行为或特征不一致,这些数据对象叫做异常点(Outlier),也叫做孤立点。在数据挖掘中,异常点由于只占数据集的较小部分,通常被作为聚类过程的副产品,当作噪声处理或不被过多的关注。因此,起初的许多数据挖掘算法通常被设计得比较健壮以包容异常点。但是一个人的噪声可能是另一个人需要的信号,所以异常点检测和分析是数据挖掘中一个重要方面,也是一个非常有趣的挖掘课题[1],它用来发现"小的模式"(相对于聚类),即数据集中间显著不同于其它数据的对象。

异常点出现的原因很多,但可归纳为3类:1)数据变量固有变化引起,即观测值在样本总体中发生了变化,这种变化是样本总体自然发生的特征,是不可控的,并且从侧面反映了数据集的数据分布特征;2)测量错误引起,由于测量仪器的一些缺陷导致部分测量值异常而成为异常点;3)执行错误引起,如黑客网络入侵、系统机械故障的出现导致数据集出现异常点[1]。

异常点挖掘具有广泛的应用,如电信和信用卡欺骗、贷款审批、药物研究、医疗分析、消费者行为分析、气象预报、金融领域客户分类、网络入侵检测等[2]。

2 异常点与异常点挖掘

(1)异常点

Hawkins给出了异常点的本质性的定义[1,3,4]:异常点是在数据集中与众不同的数据,使人怀疑这些数据并非随机偏差,而是产生于完全不同的机制。异常点的识别和检测是一种十分重要的数据挖掘类型,被称之为异常点挖掘。异常点挖掘可以定义为[5]:给定含有 n 个数据点或对象的集合及预期的异常点数目 k,发现与剩余的数据相比是显著相异的、异常的或不一致的 k 个对象。

(2)异常点挖掘

目前,异常点数据挖掘算法很多,归纳起来大致有以下四类:基于统计(statistical-based)的方法、基于距离(distance-based)的方法、基于偏差(deviation-based)的方法、基于密度(density-based)的方法。

其中基于统计的方法一般假设给定的数据集服从一个随机分布(如正态分布等),用不一致性测试(discordancy test)识别异常。存在的问题是,在许多情况下,用户并不知道这个数据分布,而且现实数据也往往不符合任何一种理想状态的数学分布;即使在低维(一维或二维)时的数据分布已知,在高维情况下,估计数据点的分布是极其困难的。

基于距离的异常点挖掘算法可描述为在数据对象集合 S 中至少有 p 个对象和对象 O 的距离大于 d,则对象 O 是一个带参数 p 和 d 的基于距离的异常点[6]。它又分为三个基本类型:基于索引(index-based)的算法、基于嵌套循环(nested-loop)算法、基于单元(cell-based)的方法。但它们分别存在输入参数很难确定,并且对于不同参数,结果有很大不稳定性;不能给定异常的程度;算法的复杂度较高等缺点。

基于偏差的方法不采用统计检验或对象间的距离度量值来确定异常对象,而是通过检查一组对象的主要特征来确定异常点,如果一个对象的特征与给定的描述过分"偏离",则该

*)基金资助:广东省科技攻关项目(2004A10202001),广州市科攻关项目(2004Z2~D0091)。王洪春 副教授,博士,主要研究方向:人工智能,数据挖掘。彭 宏 教授,博士生导师,主要研究方向:智能网络技术,数据挖掘。

维普资讯 http://www.cqvip.com

对象被认为是异常点。基于偏差的异常点挖掘方法主要有序列异常技术和 OLAP 数据立方体技术 2 种[1]。序列异常技术在对异常存在的假设太过理想化，对现实复杂数据效果不太好，而 OLAP 数据立方体技术当存在许多涉及多层概念层次的维时，人工探测变得非常困难。

基于密度的方法是给每个数据赋予一个异常因子的属性作为数据异常程度的度量，但由于聚类密度如果存在不同就会出现问题，为了解决这个问题，基于密度模型的局部异常点挖掘算法被提出。根据局部异常点的定义及其特征，可通过局部异常点因子 LOF(1ocal outlier factor)的计算来确定异常点，只要一个对象的 LOF 远大于 1，它可能就是一个异常点，需要引起数据使用者注意。簇内靠近核心点的对象的 LOF 接近于 1，那么不应该被认为是局部异常。而处于簇的边缘或是簇的外面的对象的 LOF 相对较大。它不能检测全部异常点，只能检测局部异常点，对于高维数据，存在计算时间复杂度高的缺点。

从上面的分析可以看出，现有的异常点挖掘法，虽然它们对于解决异常点的挖掘有很大的帮助，但是都或多或少地存在某些方面的不足，因此，我们把基于统计的方法和基于密度聚类的方法融合，提出了基于主成分分析的异常点挖掘算法，它既不需要先假定数据集的分布和预期的异常点的个数，也不会由于数据集的密度不同出现问题，因此它解决了异常点挖掘中的一些实际问题，弥补了一些现有异常点检测算法的不足。

3 主成分分析

主成分分析是一种统计相关分析技术，主成分概念首先由 Karl Parson 在 1901 年引进，当时只是针对非随机变量来讨论的(求拟合直线或超平面)。1933 年 Hotelling 将这个概念推广到随机变量。

在多数实际问题中，不同指标之间有一定相关性。由于指标较多及指标间有一定的相关性，势必增加分析问题的复杂性。主成分分析就是设法将原来指标重新组合成一组新的互相无关的几个综合指标来代替原来指标。同时根据实际需要从中选取几个较少的综合指标来尽可能多地反映原来的指标的信息。

因此，主成分分析是考察多个数值变量间相关性的一种多元统计方法，它是研究如何通过少数几个主成分来解释多变量的方差-协方差结构。通过导出几个主成分，使它们尽可能多地保留原始变量的信息，且彼此间不相关。

其具体步骤为：

(1)将原始数据进行标准化处理；

(2)计算样本相关矩阵 R；

(3)求相关矩阵 R 的特征值与特征向量，并计算贡献率；

(4)选择主成分；

(5)对所选主成分进行解释。

4 基于主成分分析的异常点识别

当把目光放到样本点空间，由于数据样本在某个方向上数据的变异信息是各个样本点在这个方向上提供变异信息的总和，而变异信息可用各个样本偏离中心位置的偏差(即方差)来表示，若某个样本点提供的变异信息量远大于其他样本点提供的变异信息量，说明这个样本点严重偏离数据集的重心，这正是要寻找的异常点。

具体方法为：

(1)输入数据样本并进行标准化

输入样本数据 $X=\{x_1,x_2,\cdots,x_n\}$，其中 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})\in R^p$，对 x_{ij} 进行标准化 $y_{ij}=\frac{x_{ij}-\bar{x}_j}{S_j}$，得到标准化矩阵 $Y=(y_{ij})_{n\times p}$，这里 $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$，$i=1,2,\cdots,n,j=1,2,\cdots,p$。其中 n 为样本数据的个数，p 为样本属性的个数。标准化的目的是为了消除量纲的不合理影响。

(2) 寻找异常点的可疑点

由于异常点都远离正常点，并且只占数据集的较小部分，为了减少计算，于是我们先对标准化后的样本运用基于密度的聚类算法 DBSCAN(Density-based Spatial Clustering of Application with Noise)[7]进行聚类，提取不包含在任何聚类中的样本对象(通常聚类分析中认为的噪声数据)，这些样本点就是异常点的可疑点。

(3)主成分提取

① 计算标准化数据矩阵 Y 的协方差矩阵 V

$$V=\frac{1}{n-1}Y^TY$$

求 V 的特征值 λ_j，$j=1,2,\cdots,p$，并按 λ_j 的大小排列，$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$，并计算对应的特征向量 $a_1,a_2,\cdots,a_p$，这些特征向量对应于将样本数据集 Y 所在的空间进行坐标变换(平移或旋转)后，新坐标系的原点与样本数据集 Y 的重心位置重合于坐标系的第一轴、第二轴、……，而相应的方向对应于数据变异的第一大方向、第二大方向、……。

②求数据矩阵 Y 在沿 a_h 方向上的投影数据 t_h：

$$t_h=Ya_h=\sum_{j=1}^{p}a_{hj}Y_j\quad h=1,2,3,\cdots,p$$

这里，a_{hj} 是主轴 a_h 的第 j 个分量，Y_j 是数据矩阵 Y 的第 j 个列向量。

③计算累积方差贡献率

因为 $Var(t_i)=Var(Ya_i)=\lambda_i$，所以 $\sum_{i=1}^{m}Var(t_i)=\sum_{i=1}^{m}\lambda_i$

根据累积方差贡献率大小确定主成分的个数 m(通常按累积方差贡献占总方差中的比例=85%为标准)，于是确定的主成分为 $t_1,t_2,\cdots,t_m$，$m<p$。

(4)确定异常点

将异常点的可疑点中的每个点进行检验，具体为：定义第 i 个可疑的样本点对第 h 主成分 t_h 的贡献率为 T_{hi}^2：

$$T_{hi}^2=\frac{t_{hi}^2}{(n-1)s_h^2}$$

式中，n 为样本点总个数；t_{hi} 为第 i 个可疑点样本点在第 h 主成分上投影的坐标值；$S_h^2=Var(t_h)$ 为第 h 主成分的方差。测算第 i 个可疑的样本点对各成分的累积贡献率 T_i^2：

$$T_i^2=\frac{1}{n-1}\sum_{h=1}^{m}\frac{t_{hi}^2}{S_h^2}$$

根据 Tracy 等人证明的统计量[8]：

$$\frac{n^2(n-m)}{m(n^2-1)}T_i^2\sim F_a(m,n-m)$$

式中 $F_a(m,n-m)$ 为 F 分布，根据检验水平 α 的不同可由 F 分布临界值表查得。当 $\frac{n^2(n-m)}{m(n^2-1)}T_i^2\geqslant F_a(m,n-m)$ 或者 $T_i^2\geqslant\frac{m(n^2-1)}{n^2(n-m)}F_a(m,n-m)$ 时，可以认为在 $1-\alpha$ 的检验水平

维普资讯 http://www.cqvip.com

上，第 i 个可疑的样本点对成分 $t_1, t_2, \cdots, t_m$ 的贡献过大，即该样本点严重偏离数据集重心，可将该样本点视为异常点。

事实上也可以不先进行聚类，对样本数据中每个样本都直接利用主成分分析法进行检验，验证其是否是异常点，但是当样本量很大时计算量就会很大，由于绝大多数的样本点都不是异常点，于是很多的计算就会浪费在正常点的验证上，因此一般我们采取先聚类再对不包含在聚类中的数据检验的方式。如果检验不是异常点，就把它归入距离最近的一类中。这样，对于聚类来说，只把真正的异常点作为噪声处理，就不会把本来是正常数据误为噪声而去掉了。

另外，该方法还可以通过调节基于密度聚类算法[7]的参数 ε 和 $MinPts$ 的值来确定异常点的选取方法。如果想最大限度地检测到所有的异常点，我们可以减小参数的值；如果增大参数的值，那么随着参数值增大检测到的异常点为确实为异常点的可能性就更大。

5 实验验证

我们采用来源于 UCI 的 wine 数据集，该数据集包含 185 个数据，有 13 个属性，先利用 DBSCAN 算法进行聚类，可以把数据集分成 3 类，共包含 178 个数据，其它的 7 个数据不包含在任何一个类中，利用 PCA 方法检验，取 $\alpha=0.05$，7 个数据在 95% 的检验水平上都是异常点，这与文[9]中基于距离的异常点挖掘方法所得出的最好结果一致。

结束语 本文给出了一个基于主成分分析的异常点挖掘方法，这种方法把基于密度的聚类算法和基于统计的异常点挖掘方法结合了起来，同时又融合了数据维数消减的主成分分析方法。实验表明，该方法在异常点挖掘方面效果明显。

参 考 文 献

1 王宏鼎，童云海，等. 异常点挖掘研究进展. 智能系统学报，2006，1(1)：67～73

2 陈华，李继波. 异常(Outlier)检测算法综述. 大众科技，2005，9：96～97

3 李炎，李皓，等. 异常检测算法分析. 计算机工程，2002，28(6)：5～6，32

4 李之棠，刘颉. 入侵检测中的模糊数据挖掘技术. 计算机工程与科学，2002，24(2)：18～21

5 钱昌明，李国庆，等. 分类异常点检测算法及在 IDS 模型中的应用. 计算机应用研究，2006，23(4)：94～96

6 Han Jiawei, Kamber M. Data mining: concepts and techniques. New York: Morgan Kaufmann Publishers, 2001

7 Ester M, Kriegel H, Sander J, Xu X. A density—based algorithm for discovering clusters in large spatial database with noise. In: Proceedings of the 2nd conference on Knowledge Discovering in Databases and Data Mining, Portland, USA, 1996. 226～231

8 王惠文. 偏最小二乘回归分析及其应用. 北京：国防工业出版社，1999. 130～184

9 Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets. In: Proceedings of Very Large Data Bases(VLDB'98), New York, USA, 1998. 392～403

(上接第 184 页)

11 Krieger M J B, Billeter J B. The call of duty: Selforganised task allocation in a population of up to twelve mobile robots. Robotics and Autonomous Systems, 2000, 30: 65～84

12 Lumer E, Faieta B. Diversity and adaptation in populations of clustering ants. In: J.-A. Meyer, et al. eds. Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats, Vol. 3, MIT Press/Bradford Books, Cambridge, MA, 1994. 501～508

13 Dorigo M, Gambardella L M. Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem [J]. IEEE Trans. On Evolutionary Computation, 1997, 1(1): 53～66

14 Watkins C J C H. Learning with delayed rewards [A]: [PhD Thesis]. University of Cambridge, England, 1989

15 Dorigo M, Gambardella L M. A study of some properties of Ant-Q. In: Proc. PPSN IV-4th Int. Conf. Parallel Problem Solving from Nature. Berlin, Germany: Springer-Verlag, 1996. 656～665

16 Johnson D S, McGeoch L A. The travelling salesman problem: a case study in local optimization. In: Local Search in Combinatorial Optimization. E. H. L. Aarts, et al. eds. New York: Wiley and Sons, 1997

17 Stutzle T, Hoos H. Improvements on the Ant System: Introducing MAX-MIN Ant System [A]. In: Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms [C], Springer Verlag, Wien, 1997. 245～249

18 Stutzle T, Hoos H. Max-Min ant system [J]. Future Generation Computer System, 2000, 16: 889～914

19 Lin S, Kernighan B W. An effective heuristic algorithm for the traveling Salesman Problems [J]. Operations Research, 1973, 21: 498～516

20 Maniezzo V. Exact and approximate nondeterministic tree-search procedures for the quadratic assignment problem. INFORMS J. Comput, 1999, 11(4): 358～369

21 Reimann M, Doerner K, Hartl R F. D-ants: savings based ants divide and conquer the vehicle routing problems. Comput. Oper. Res, 2004, 31(4): 563～591

22 Sorges U, Gunes M, Bouazizi I. ara - the ant colony based routing algorithm for manets. In: Proc. of the 2002 ICPP Workshop on Ad Hoc Networks (IWAHN 2002), 2000. 79～85

23 Casta D, Hertz A. Ant Can Colour Graphs [J]. Journal of the operational Research Society, 1997, 48: 295～305

24 Fan X P, Luo X, Yi S, et al. Path planning for robots based on ant colony optimization algorithm under complex enviroment [J]. Control and Decision, 2004, 19(2): 166～170

25 Dorigo M, Di Caro G. The Ant Colony Optimization meta-heuristic. In: New Ideas Optimization. D. Corne, et al., eds. McGraw Hill, London, UK, 1999. 11～32

26 Dorige M, Caro G D, Gambardella L M. Ant Algorithms for Discrete Optimization [J]. Artificial Life, 1999, 5(3): 137～172

27 Bullnheimer B, Hartl R F, Strauss C. A New Rank-based Version of The Ant System: A Computational Study: [Technical Report POM-03/97]. Institute of Management Science, University of Vienna, Accepted for Publication in the central European Journal for Operations Resezrch and Econimics, 1997

28 Maniezzo V, Dorigo M, Coloni A. The Ant System Applied to the Quadratic Assignment Problem [A]: [Technical Report IRIDIA/94～28]. Universite de Bruxelles, Belgium, 1994