文章编号:1672 - 058X(2012)10 - 0096 - 05

一种基于相似度量的离群点检测方法*

孙启林, 方宏彬, 张　健, 刘明术
(安徽大学 数学科学学院,合肥 230039)

摘　要:离群点检测在是数据挖掘的重要领域,广泛应用在信用卡欺诈检测、网络入侵检测等重要方面,文中在结合层次聚类和相似性,给出高维数据的相似度量函数与类密度的概念,并基于类密度重新定义高维数据的离群点,从而提出一种基于相似度量的离群点检测算法;实验表明:算法对高维数据中的离群点检测有一定的价值。

关键词: 离群点;网络入侵;数据挖掘;层次聚类;相似性度量
中图分类号:TP390　　**文献标志码**:A

离群点检测是数据挖掘中重要领域,离群点(或异常点)的最早是由 Hawkings[1] 提出的本质性定义:异常点如此不同于数据集中的其他数据,以至于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。离群数据是数据集中偏离大部分数据的数据,它们的表现与大多数常规对象有着明显的差异,以至于让人怀疑它们可能是由另外一种完全不同的机制所产生的。离群点可能是由于度量或执行错误产生的,也有可能是由于固有数据变异产生的,或其他原因。早期对数据集进行预处理时,通常把离群点当作噪声,或修正离群点的值以减少其对正常数据的影响,或在挖掘过程中排除离群点。离群点检测是要在数据中发现极少的数据,但可以使人们发现有价值的东西。离群数据并不等同于错误数据,离群数据中可能蕴含着极为重要的信息,例如在信用卡欺诈检测、网络入侵检测、疾病诊断、通信欺诈分析、故障检测、灾害预测、恐怖活动防范等诸多领域中,离群点都是数据分析的主要对象[2,3],离群点检测最早出现在统计学领域[4]。后来,Knorr 等将其引入到数据挖掘领域[2]。离群点检测的方法主要有 5 类:基于统计的方法;基于深度的方法;基于聚类的方法;基于密度的方法;基于距离的方法。

目前,结合聚类与基于距离的离群点检测方法较为普遍,在高维数据以及大规模数据的聚类分析算法主要有子空间聚类和基于对象相似性的聚类算法两种,子空间聚类算法的主要代表有 CLIQUE[5] 和 PROCLUS[6] 等,子空间聚类是实现高维数据集聚类的有效途径,它是在高维数据空间中对传统聚类算法的一种扩展,其思想是将搜索局部化在相关维中进行。

而基于对象相似性的聚类算法主要有基于 SL 树的图分割算法和 HETIS 算法[7] 等,由于在高维空间中导致分辨率能力下降的主要原因之一是高维空间中点的噪音[9] 以及稀疏性[10] 的存在,如果在高维空间用距离来度量对象的相似性,会产生更多的噪音。

在此结合层次聚类和相似性,提出基于相似度的离群点检测方法。首先给出高维数据的相似度量函数

收稿日期:2012 - 03 - 15;**修回日期**:2012 - 04 - 21.
* **基金项目**:安徽省教育厅自然科学基金项目(05010428).
作者简介:孙启林(1982-),男,安徽合肥人,硕士研究生,从事金融数据挖掘研究.

与类密度概念,接着基于类密度概念重新定义高维数据的离群点,并在此基础上设计一种新的高维数据离群点检测算法,该算法首先对高维数据进行聚类,根据类密度来检测离群点。

1　相关概念的介绍

1.1　相似度量函数

在表示数据之间相似性方面,传统度量方式大多以数据之间距离来表示数据之间的相似性表示。一般情况下,在低维空间用距离来度量能效果较好,但在高维空间中效果并不好。如果将低维空间中基于距离问题的解决方法推广到高维空间,将会引起难以预料维度灾难问题[9]。最近几年有很多学者有提出相似度函数,如文献[2]。

$$H\sin(X,Y) = \frac{\sum_{i=1}^{d} \frac{1}{1 + |x_i - y_i|}}{d} \tag{1}$$

在 Hsim 函数的基础上提出出另外一种新函数:

$$F\sin(X,Y) = \frac{\sum_{i=1}^{d} \frac{m_i}{|x_i - y_i| + m_i}}{d} \tag{2}$$

下面验证 Fsin 是否是相似度函数:由(1) $0 \leqslant F\sin(M,N) \leqslant 1$ 显然;(2) 当且仅当 $F\sin(M,N) = 1$ 时,满足 $M = N$;(3) $F\sin(M,N) = F\sin(N,M)$;以上 3 个条件都满足由文献[8]相似性度量函数的定义可得知 Fsin 是相似性度量函数。其中 $X = (x_1, x_2 \cdots, x_d)$ 和 $Y = (y_1, y_2, \cdots, y_d)$,$m_i$ 表示第 i 列元素的平均值的绝对值,这样相似度量不仅和 X 和 Y 的差,还与数据维数的中心有关。

1.2　聚类中的类密度的定义

假设有 N 个 d 维数据集 $X_i(i = 1,2,\cdots,N)$按某种规则聚类成 k 个类 $LC = \{C_1, C_2, \cdots, C_k\}$,则 C_i 的类密度为 $dest(C_i) = \frac{|C_i|}{N}$其中 $|C_i|$ 为 C_i 类中的数据集个数。

1.3　基于类密度的离群点定义

离群点的定义并没有精确的定义,一般的学者认为,离群点就是数据集的行为和表现偏离大部分数据集的数据,即大多数的常规对象有着明显的差异。离群点的定义:N 个 d 维数据经过相似聚类以后的类密度小于某个值的类中的数据为离群点。简言之,数据经过某种相似性聚类以后,某些类的个数少于事先给定的阈值,这些少数的类中的数据就认为是离群点。

2　基于相似度的类密度离群点检测算法

2.1　算法的设计:

假设有 N 个 d 维数据 $X_i(i = 1,2,\cdots,N)$,计算每两个对象的 Fsin 相似度函数,能得到对称的相似度矩阵。对于类间的相似度,按照公式 $s(P,Q) = \min\{s(x,y) \mid x \in P, y \in Q\}$[11] 来计算,$s(x,y)$ 表示两个类中的点 x 与 y 之间的相似度。接着将各个对象分别初始化一个簇,再根据相似度的大小合并这些簇,使之成为越来越大的簇,直到不能满足阈值为止。该算法能够使得到的每个簇里面各个对象之间相似度至少达到阈值的

大小,而不同簇之间对象相似度不能满足阈值条件。具体的算法:

第1步:对数据集分类:

(1) 计算每个对象的 $F\sin$ 的相似度函数值;

(2) 将每个对象分别初始化一个类;

(3) 找到类中相似度最大的两个类,当这两个类相似度大于给定的阈值 r,合并两个类;

(4) 重复(3),直到满足类相似度都小于 r;

(5) 输出各类。

第2步:找离群点,设定阈值 t,当某个类的类密度小于该阈值就为离群点。

具体的算法描述:

For $i=1:k$

$dest(C_i)=\frac{|C_i|}{N}$

If $dest(C_i)<t$

C_i 为离群点并输出

Endif

Endfor

2.2 算法复杂度分析

由文献[11]中的相似度聚类算法的时间复杂度为 $O(n^3)$,在 n 比较大的时候,需要较多的时间的开销,但能够聚类出较高的质量的类。在离群点检测的时间复杂度为 $O(n)$,可以根据阈值的大小来能对数据进行离群点的检测,其算法根据阈值很容易计算,从而找到离群点。

3 实验数据分析

实验程序是用 Matlab 7.0 来编写的,在 CPU 为 AMD Anthon 64 2.91GHz,2 G 内层的计算机上来实现的。实验的数据是 UCI 数据库中的 breast-cancer-wisconsin 数据,且其中的特殊符号都用0代替,在阈值 r 不变的情况下,来讨论离群点的检测的情况。

通过实验取固定阈值 r,在维数相同的情况下,随着数据量的增大,发现检测的离群点个数接近于数据量的个数,效果不佳。如表1所示:数据维数为11,$t=0.05$,$r=0.8$。

表1 $r=0.8$,$t=0.05$ 不同数据量该算法形成的相关情况

数据量	聚类数	离群点数	运行时间/ms
10	9	0	47
20	15	11	609
50	35	38	10 359
100	68	85	102 172
150	102	129	353 078
200	128	172	750 531

在试验中 $t=1/N$,其中 N 为数据个数,在维数相同的情况下,随着数据量的增加,会发现离群点点个数较稳定,效果好。如表 2 所示。

表 2　$r=0.8$,$t=1/N$,N 为选用数据集中的数据量

数据量	聚类数	离群点数	运行时间/ms
10	9	0	47
20	15	11	625
50	35	26	10 328
100	68	55	98 297
150	102	85	340 937
200	128	107	814 468

表 2 中,$r=0.8$,$t=1/N$,N 为选用数据集中的数据量,不同数据量该算法形成的相关情况。如表 1 和表 2 可以发现,在维数、阈值 r 相同的情况下,t 的选取对离群点的检测有很大的影响,因此,当数据量较大时候,t 要取的偏大一点,反之,t 取小点。这样离群点检测才会相对准确。

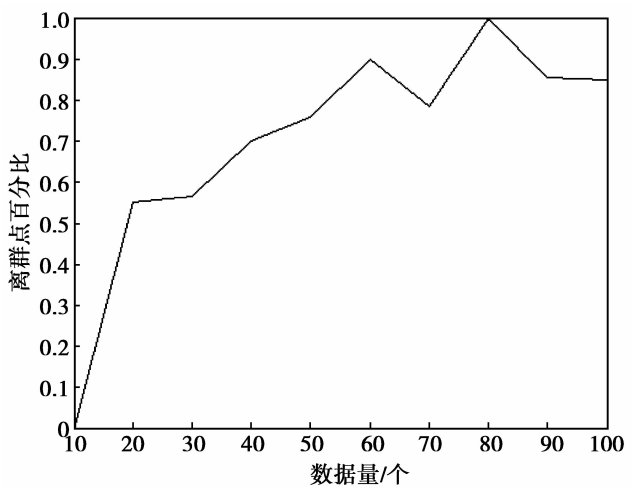

图 1　维数为 11、$r=0.8$、$t=0.05$ 不同数据量的离群点个数变化图

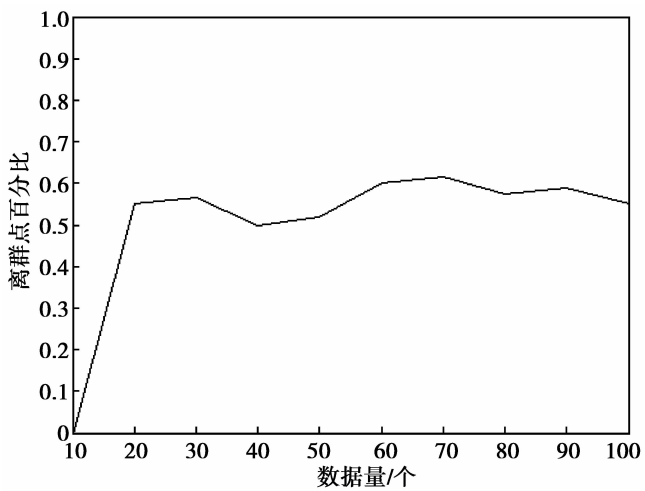

图 2　维数为 11、$r=0.8$、$t=1/N$ 不同数据量的离群点个数变化图

图 1 和图 2 分别表示数据量(单位:个)是 10、20、30、40、50、60、70、80、90、100,维数为 11,$r=0.8$ 的情况下,在 t 值分别取 0.05 与 $1/N$ 的情况下,离群点百分比的情况变化情况,$t=1/N$ 在离群点检测方面明显的优于 $t=0.05$ 的情况,显然数据集中的数据是否是离群点具有相对性。

4　结　语

结合相似度聚类算法和凝聚层次聚类的思想,提出一个相似度函数的模型,并提出基于相似聚类的离群点检测算法。实验表明,在阈值 t 的变化下,离群点随 t 值不同有可能不同,在大数据集下离群点数的比例随 t 值变小而趋向稳定。下一步将相似度函数进行优化,而聚类出质量较高的类,从而更精确的检测离群点。

参考文献:

[1] HAWKINS D. Identifications of Outliers[M]. London:Chapman and Hall,1980

[2] EKNORR R. Algorithms for mining distance-based outliers in large datasets[A]. In Proc of the24th VLDBConf[C]. NewYork: MorganKaufmann,1998:392-403

[3] HAN J W,DAMBER M. Data Mining:Concepts andTechnologies[M]. SanFrancisco:Morgan Kaufmann 2001

[4] ROUSSEEUW P J,LEROY A M. Robust Regression and Outlier Detection[M]. New York:John Wiley& Sons,1987

[5] RAKESH A,IJOHANNES G,DMIITRIOS G,et al. Automat ic Subspace Clustering of High Dimensional Data for Data Mining Application [C] //Proceedings of the 1998 ACMSIGMOD Internation a Conference on Management of Data,Seattle,Washington, 1998

[6] GGARWAL A,PROCOPIUC C,WOLF J L,et al. Fast al- gorithmsf or projected clustering [C] //Proc. of the ACMSIGMOD Conference Philadel Phia,P A,1999:61-72

[7] 杨风召. 高维数据挖掘技术研究[M] . 南京:东南大学出版社,2007

[8] XU Z S,XIA M M. Distance and similarity measures for hesitant fuzzy sets[J]. Information Sciences,2011. 2128-2138

[9] AGRAWAL R,GEHRKE J. GUNOPOLOS D,et al . Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In ACM SIGMOD Conference,1998

[10] 贺玲,吴玲达,蔡益朝. 高维空间中数据的相似性度量[J]. 数学的实践与认识,2006,36(9):189-194

[11] 黄斯达,陈启买. 一种基于相似性度量的高维聚类算法的研究[J]. 计算机应用与软件,2009:102-105

A Kind of Outlier Detection Algorithm Based on Similarity Measurement

SUN Qi-lin, FANG Hong-bin, ZHANG Jian, LIU Ming-shu

(School of Mathematical Science, Anhui University, Hefei 230039, China)

Abstract: Outlier detection is an important content in data mining and is widely used in the field of credit card fraud detection, network invasion detection and so on. According to hierarchical clustering and similarity, this paper presents the concept of high dimensional data similarity measurement function and class density, based on class density, the outlier of high dimensional data is redefined so that a kind of outlier detection algorithm based on similarity measurement is proposed. Experiment shows that this algorithm has certain value on outlier detection in high dimensional data.

Key words: outlier; network invasion; data mining; hierarchical clustering; similarity measurement

责任编辑:代小红