基于信息论的高维海量数据离群点挖掘

张 净[1,2] 孙志挥[1] 宋余庆[3] 倪巍伟[1] 晏燕华[3]

(东南大学计算机科学与工程系 南京 210096)[1] (江苏大学电气与信息工程学院 镇江 212001)[2]

(江苏大学计算机科学与通信工程学院 镇江 212001)[3]

摘 要 针对高维海量数据集离群点挖掘存在"维数灾难"的问题,提出了基于信息论的高维海量数据的离群点挖掘算法。该算法采用属性选择,去除冗余属性降维。利用信息熵作为离群点判断的度量标准,消除距离和密度量纲的弊端。在真实数据集上的实验结果表明,算法对高维海量数据离群点挖掘是有效可行的,其效率和精度得到了明显提高。

关键词 离群点挖掘,信息论,属性选择,熵,互信息

中图法分类号 TP311 **文献标识码** A

Outlier Mining of the High-dimension Datasets Based on Information Theory

ZHANG Jing[1,2] SUN Zhi-hui[1] SONG Yu-qing[3] NI Wei-wei[1] YAN Yan-hua[3]

(Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)[1]

(College of Electronic and Information Engineering, Jiangsu University, Zhenjiang 212001, China)[2]

(College of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang 212001, China)[3]

Abstract Phenomena of "curse of dimensionality" deteriorate lots of existing outlier mining algorithms validity. Concerning thw problem, the outlier mining algorithm of high-dimension and large datasets based on information theory was proposed. This algorithm used the concept of information entropy and the mutual information in the information theory, carried on the feature selection after using estimated mutual information value objective basis entropy power sorting, and eliminated redundant attribute for dimensionality reduction. Outlier mining using information entropy as a measure standard to judge eliminated the drawbacks of distance and density metric. The experimental result in the real data sets indicates that the algorithm for outlier mining in high-dimensional mass data is effective and feasible, its efficiency and accuracy are significantly improved.

Keywords Outlier mining, Information theory, Feature selection, Entropy, Mutual information

1 引言

近年来随着应用的不断深入,基于数据挖掘概念的离群点挖掘算法取得了不少重要的成果,例如:Knorr 等人提出的基于距离的算法 FindALLOutsD[1]、面向大规模数据离群点挖掘算法 FindOut[2]、Breuning 等人提出的带离群度的离群点挖掘算法 LOF(local outlier factor)[3]和 Papadimitirou 等人提出的 LOCI(local correlation integral)[4]等。这些算法对于一般低维度数据具有良好的性能。但当数据维度较高时,这些算法在"维度灾难"的存在下,准确性和效率受到了严重的影响。

目前,一些研究采用属性选择降维来解决高维问题。属性选择是指从已知的特征属性中,按照某一准则寻找特征子集,通过删除一些冗余的、不相关的属性降低数据集的维数、提高运行效率,同时改善算法性能,提高挖掘准确性和效率。在离群点挖掘算法中目前使用最多的属性选择方法就是子空间数据挖掘,例如:Aggarwal 等人提出了基于空间投影的离群点挖掘算法 EvolutionaryOutlierSearch[5],倪巍伟等人提出了基于局部信息熵的加权子空间离群点检测算法[6],但其子空间的选择仍然是一个复杂的问题。

信息论是一门将信息作为研究对象,以揭示信息的本质特性和规律为基础,应用数学方法研究信息存储、传输、处理、控制和利用等一般规律的科学。信息熵可以度量一个系统的无序和杂乱程度。熵值越大,系统中的数据越无序,系统越"杂乱";反之,熵值越小,系统中的数据越有序,系统越"纯净"。信息熵完全建立在原始数据的基础上,客观性比较强,其值只依赖于数据对象每个属性的概率,属性的取值可以是数值型,也可以是非数值型(如字符型)。利用信息熵度量、识别数据中无序的数据点,可以客观地识别出数据中的离群点。

本文通过对离群点数据本质和属性选择过程的分析,提出了一种基于信息论的高维海量数据离群点挖掘算法。该算法通过计算出的互信息和熵权进行属性选择,并利用网格消除海量数据的影响,不再以距离作为数据点之间关系的度量标准,通过熵值进行离群点判断,可以有效地挖掘出高维海量

到稿日期:2010-11-02 返修日期:2011-01-25 本文受国家自然科学基金(40871176,60841003)资助。

张 净(1975-),女,硕士,副教授,主要研究方向为数据挖掘、知识发现,E-mail:jszj08062000@yahoo.com.cn。

数据集中的离群点。

2 相关工作

在数据挖掘领域虽然已提出了一些基于信息论的特征选择方法[7,8]，但基于信息论的离群点挖掘算法研究处于起步阶段[6,9-12]。

2006年，何等提出了基于信息熵的快速贪婪算法(GreedAlg)[10]，由于使用贪婪算法的策略，其计算过程中容易陷入局部最小，而且无法进行高维数据的离群点挖掘。2008年，倪等人提出的基于局部信息熵的加权子空间离群点检测算法(SPOD)[6]，产生子空间的选择问题，在其利用距离和密度定义时，维数不断增加后，子空间维数也会受到影响，高维数据越来越"稀疏"，影响算法的挖掘结果。同年，于等人提出的基于信息熵的相对离群点检测算法(ENBROD)在高维时给出每个对象所对应的相对离群点因子，计算量较大[11]，事先人为设置参数也影响了算法的运行效果。2010年，Feng等人提出的IE-based的离群点挖掘算法针对高维数据会有较高的时间复杂度，没有采取有效的属性选择方法[12]。

3 相关定义

假设 d 维数据集 D 的属性集为 $A=\{A_1,A_2,\cdots,A_d\}$，数据集中的点为 x_i，相关定义如下：

定义1(属性值频度) 设数据对象集为 D，属性维数为 d，数据对象 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}$，第 j 维属性为 A_j，则属性值关于对象集 D 的频度为：

$$FreqScore(A_{ij})=\sum_{j=1}^{|D|}f(A_{ij})$$

$$Freq(A_{ij})=FreqScore(A_{ij})/\sum_{j=1}^{|D|}FreqScore(A_{ij})$$

属性值频度越大，说明取该属性值的对象越多，其中，$f(A_{ij})$ 是 x_i 的第 j 个属性在数据集中出现的次数。对象 x_i 的属性频度值越低，其离群程度越高，反之则不然。

其计算复杂度为 $O(d*|D|)$，比Greedy算法的复杂度更低，仅需对整个数据集扫描一次，比起Greedy算法的 n 次扫描，其I/O开销也更少。可见这种方式的优点就是计算简单。

定义2 设 X,Y 是随机变量，其取值集合为 $S(X)$ 和 $S(Y)$，$P(x)$ 是 X 的概率函数，$P(y)$ 是 Y 的概率函数，这里的 i 是对象数 $i=1,2,\cdots,|D|$，j 代表属性维，$j=1,2,\cdots,d$，互信息定义为：

$$I(x:y)=\sum_{X,Y}P(xy)\log\frac{P(x|y)}{p(x)}\cong\sum_{ij}p(ij)\log\frac{p(ij)}{p_x(i)p_y(j)}$$

定义3 多维属性 $\hat{x}=\{X_1,\cdots,X_m\}$ 和 $\hat{y}=\{Y_1,\cdots,Y_m\}$ 的互信息为：

$$I(\hat{x}:\hat{y})=\sum_{\hat{x},\hat{y}}P(\hat{x}\hat{y})\log\frac{P(\hat{x}\hat{y})}{P(\hat{x})P(\hat{y})}$$

利用信息论中的互信息度量属性之间的相关程度，尽管互信息很难被精确地计算，但是算法的最终目标是为了挖掘离群点，旨在得到数据点之间的相互关系，可以通过数值计算直接得到结果，故而在不影响离群点挖掘结果的前提下，借用文献[13]的方法，对互信息进行估算。将数据集中的数据看成数据空间的点，通过点数的简单计算代替概率参与互信息计算。

在二维时，近似表示为：$P_x(i)\approx n_x(i)/n$，$P_y(j)\approx n_y(j)/n$，$P(ij)\approx n(ij)/n$。P_x 和 P_y 可以看成是在该维的投影。$P(ij)$ 表示在 x 维和在 y 维投影中相交的部分，可扩展到多维。I 越大，表示 $\hat{x}$ 和 $\hat{y}$ 之间的相关程度越高。$\hat{x}$ 和 $\hat{y}$ 是对象的属性。

定义4(信息熵) 熵是信息论中用来描述信息和随机变量不确定性的重要工具。设 X 是一个随机变量，其取值集合为 $S(X)$，$P(x)$ 是 X 的概率函数，则信息熵 $H(X)$ 可定义为：

$$H(X)=-\sum_{x\in S(x)}p(x)\log(p(x))$$

定义5 多维变量 $\hat{x}=\{X_1,\cdots,X_d\}$ 的信息熵为：

$$H(\hat{x})=-\sum_{x_1\in S(X_1)}\cdots\sum_{x_n\in S(X_d)}p(x_1,\cdots,x_d)\log p(x_1,\cdots,x_d)$$

，当数据集中的对象属性之间彼此独立时，

$$\begin{aligned}H(\hat{x})&=-\sum_{x_1\in S(X_1)}\cdots\sum_{x_n\in S(X_d)}(p(x_1)\cdots p(x_d))\log(p(x_1)\cdots p(x_d))\\&=H(X_1)+\cdots+H(X_d)\end{aligned}$$

此时，属性取值的联合概率成为各个属性概率的乘积，所以计算数据集中对象的信息熵就简化为计算属性的信息熵之和。各属性的信息熵计算中主要是利用其相应属性取值的概率参与计算，如果属性取值是连续型随机变量，则离散化后计算熵值。

从概念上来说，如果一个数据集的所有属性之间都是相互独立的，那么它的内在维就等于嵌入维，但是如果两个或多个属性之间存在相关性，内在维数就小于嵌入维数。属性的不确定性越大，熵值也就越大，所需要的信息量也就越大，使得对应对象成为离群点的可能性就会增大；熵值越小，不确定性就越小，较为稳定，使得对应对象成为离群点的可能性就会减少。信息熵标志着所含信息量的多少，是对系统不确定性程度的描述。

必须指出，通过互信息的计算可以去除冗余属性，使得属性之间相互独立，起到降维的作用，但是属性之间独立性的判定需要设置参数。为了避免如投影后进行子空间的选择的降维方法中人为因素的干扰，应通过熵权的计算，较客观地依据互信息值进行属性选择。

定义6 r_{ji} 称为第 j 个对象在 i 属性上的值，且 $r_{ji}\in[0,1]$，则在 n 个对象 d 维属性中，第 i 维属性的熵定义为：

$$H_i=-k\sum_{j=1}^{n}p_{ji}\ln p_{ji}\quad i=1,2,\cdots,d$$

式中，$p_{ji}=\frac{r_{ji}}{\sum_{j=1}^{n}r_{ji}}$，$k=\frac{1}{\ln d}$。

根据数据处理方便的需要，选择 $0\leqslant H_i\leqslant 1$。

定义7 第 i 维属性的熵权 ω_i 定义为：

$$\tilde{\omega}_i=\frac{1-H_i}{d-\sum_{i=1}^{d}H_i}$$

式中，$0\leqslant\tilde{\omega}_i\leqslant 1$，$\sum_{i=1}^{d}\tilde{\omega}_i=1$。

根据熵的权值确定不同属性对对象挖掘的贡献程度。根据计算出的熵权自动排序，可以减少人为干扰。根据权值进一步降低属性维数。所选出的相互独立的特征必定是熵权比较大的，同时与其他属性的互信息值比较小的那些属性值。对于具有相关性的属性，利用均值代替相应属性数值，对于熵权的值相近的属性也用均值进行代替，不影响挖掘结果。如

果简单地直接使用均值法，会弱化甚至掩盖掉其中的异常值，对挖掘结果产生不利影响。简单的例子就可以看出其影响，比如一组数据(100,99,98,1)，用其均值代替这组数据就变成(74.5,74.5,74.5,74.5)，看不出任何异常。数据对象的离群性很大程度上是由于其对应属性的离群性质所决定的，在离群点挖掘过程中，很多离群点都是由于其属性上的异常造成该数据对象表现出离群性质[14]。

计算所有属性值的频度，并根据定义4得到数据对象的信息熵。变量的不确定性越大，熵也就越大，所需要的信息量也就越大；熵值越小，不确定性就越小。数据点成为离群点说明其不属于任何聚类，归属性不确定，与聚类的信息熵值很大。信息熵标志着所含信息量的多少，是对系统不确定性程度的描述。可通过计算熵值确定出离群点。对熵值进行自动排序，熵值大的就是离群点。

4 算法

4.1 数据预处理

不同属性(如均值、方差等)之间在数值上存在很大的悬殊。为了避免这种悬殊对挖掘结果的影响，需要对原始数据进行预处理，包括属性值的归一化、属性值约减(去除冗余属性)和离散化等。

4.1.1 属性的归一化

首先消除量纲的影响。将数据经过标准差变换后，使每个数据对象的均值为0，标准差为1。由于信息熵计算中对数函数的出现，为了防止熵值计算中对数计算出现无穷大的情况，必须进行极差变换，将数据值映射到0.1～0.9之间，变化后，每个变量的取值范围为[0.1,0.9]，且消除了量纲的影响，不影响最终的挖掘结果。式(1)和式(2)分别给出了标准差和极差的变换公式。

$$x''_{ik}=\frac{x'_{ik}-\bar{x}_k}{S_k} \tag{1}$$

式中，$S_k^2=\frac{1}{n}\sum_{i=1}^{n}(x'_{ik}-\bar{x}_k)^2$

$$\bar{x}_k=\frac{1}{n}\sum_{i=1}^{n}x'_{ik}$$

$$x_{ik}=0.8\frac{(x''_{ik}-x''_{k\min})}{(x''_{k\max}-x''_{k\min})}+0.1 \tag{2}$$

式中，$x''_{k\max}$和$x''_{k\min}$分别表示最大值和最小值。

4.1.2 属性选择和离散化

通过对数据的处理和分析可以看出，属性的离群性使得数据对象表现出离群性质，属性之间由于多样化和细致性的划分使得多属性之间具有相关性，这种相关性表现出存在冗余属性，针对高维的降维处理可以作为一种降维条件，属性相关性不会影响数据对象的离群性，不会影响离群点的挖掘效果。

由于冗余属性和属性相关性的存在，可以通过属性选择降维，在保持挖掘效果不受影响的前提下，删除一些不相关或者不重要的属性，同时合并一些具有相关性的属性，降低维数，便于离群点挖掘，采用信息论的方法进行属性选择。

4.1.3 算法复杂度

ITfOM算法的流程分为两部分，属性选择后的降维部分和网格中选取代表点参与离群点挖掘，在不借助数据集索引结构的情况下，该算法与LOF，SPOD具有同一级别的复杂度$O(|D|^2)$。

基于信息论的离群点挖掘算法 Information Theory for Outlier Mining(ITfOM)

Step1 数据初始化，分类数据数值化，连续型数据离散化。

Step2 数据归一化，为了保证后面熵值计算的正确性，使得数据归一化在[0.1,0.9]之间。

Step3 针对选择出的属性参考 Koufakou[13]的方法利用定义1进行属性值频度计算。

Step4 利用定义2和定义3对所有属性两两进行互信息计算，并对计算出的互信息进行排序，再根据定义6和定义7计算出属性的熵权，参考熵权后由领域专家选取阈值，使得所有属性之间相互独立，若没有冗余属性，进一步降低属性维数。对于具有相关性的属性，利用均值代替相应属性数值，缩减其维数。

Step5 利用定义4和定义5计算对象点之间的信息熵后进行离群点挖掘，对熵值进行自动排序，熵值大的就是离群点。

5 实验

对基于信息论的离群点挖掘算法的性能进行实验分析，实验平台配置如下：intel 1.6GHz/512MB，Windows 2000(Server版)，数据库采用 SQL2000，代码用 Visual C++(6.0)实现。

实验中使用的真实的网络数据包括两部分：背景数据和合成数据。背景数据来自网络入侵检测数据集 KDDCUP-99，其中的每一条数据都是从原始网络连接记录抽取出的特征值组成的特征值向量，该数据集中的数据对象可划分为5种，包括各种入侵、攻击和正常的连接，其中包含的41个网络连接记录的特征属性中34个是数值属性，7个是分类属性。每一个都是在混有模拟入侵的原始网络数据中提取出来的一个特征向量，一个连接记录就是一个从源地址到目的地址的TCP数据包的序列，所有TCP数据包被Bro程序重组成了一条条连接记录，每一条连接被标上了正常或某一种特定类型的攻击。所提取的特征包括了单个TCP连接中的基本特征，对于提供的每一个TCP/IP连接，除了一些基本属性(例如协议类型、传送的字节数等)外，还利用领域知识扩展了一些属性(例如登录失败的次数、文件生成操作的数目等)，某些属性是在过去2秒钟之内信息的基础上得到的。

一条连接记录的完整形式如下所示：

0,tcp,http,SF,223,185,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,4,4,0.00,0.00,0.00,0.00,1.00,0.00,0.00,71,255,1.00,0.00,0.01,0.01,0.00,0.00,0.00

为了进行实验，对此数据集中的数据进行了适当的预处理(离散化和归一化)，利用定义1计算出属性值频度。利用定义2和3计算互信息后降维，在属性与属性之间计算互信息，在高维利用定义3计算，实验时，参与计算的点数实际上就是在各维的投影数据，但是跟投影降维方法相比，近似计算时不需要人为选择投影空间。参考互信息的属性选择算法选取了其中的7个相互独立的属性维(见表1)，考虑到分布式数据，加上一维作为站点标记维，不参与计算，只是作为标记。降维时的处理对结果还是有影响的，但是兼顾到进行离群点挖掘时的速度和效率，选择使用了7维属性参与离群点挖掘。

数据集总量选为约 60 万，挖掘数据中 97%采集的是 KDD-CUP-99 中的正常数据，3%是选取加入与主体数据有较大偏差的异常点，6000 条带有隐私的数据中 1000 条是异常数据。然后针对不同的属性值(离散的或连续的)分别进行预处理，所有的数据都经过标准化处理，然后利用定义 1 中的公式计算数据对象的属性值频度，随后分别使用算法 LOF、降维后使用 LOF[3]算法、子空间离群点检测算法 SPOD[6]和 ITfOM 算法对网络数据进行离群点挖掘实验。通过算法的参数影响、精确度和相应时间等主要指标来分析和评价所提出的基于信息论的高维海量数据集离群点挖掘算法 ITfOM。

表 1　参考熵权利用互信息选取出来的相互独立的属性维

合并以后的维数	原来列的位置
1	2,4,8,10,13,18,26－30,32－37
2	3,5,7,11,12,14－17
3	8,21,22,25,34,35
4	1,9,23,24,31,36,37
5	19,20
6	32
7	6,18

5.1　参数影响

算法中在进行属性选择时需要进行参数选择，数据集的大小对于挖掘结果也是有影响的，数据集不断增大时，借用文献[15]的方法，采用网格减少计算量，用网格中的点数作为阈值判断数据点性质，选用代表点简化计算，不影响计算结果[16]。参数值决定算法的最终挖掘结果。实验表明，互信息估算后选择的属性对于挖掘结果有很大的影响，而且通过熵权计算的结果可以参考互信息计算时属性选择的结果，这样，选择出来的结果较好。

表 2　属性选择对挖掘结果的影响

数据集(万)	属性个数	检测率比	时间比	属性个数	检测率比	时间比	属性个数	检测率比	时间比
10	37	1.309	2.21	20	1.145	1.11	10	1.019	1.02
20	37	1.303	3.08	20	1.113	1.39	10	1.018	1.05
30	37	1.208	3.36	20	1.101	1.52	10	1.016	1.06
40	37	1.119	4.42	20	1.104	1.94	10	1.01	1.07
50	37	1.102	4.87	20	0.992	2.51	10	0.992	1.09
60	37	1.101	5.29	20	0.998	3.17	10	0.989	1.11

从表 2 可以看出，随着数据量的不断增大，维数对于检测率和时间是有很大影响的，如果维数不进行缩减，则耗费的时间较长，网格划分的意义不大，很多点都是独自参与计算的，很多网格都是空的，维数灾难体现得较为明显，但是检测率还可以，等于是对每个点进行计算，属性维数在 20 维时，因为属性之间有相关性的较多，所以挖掘时间与 7 维相比还是较长，但是挖掘效率已经比较接近了，当维数缩减到 20 维和 10 维时，检测率已经比较接近了，甚至当数据量增大时，7 维数据的检测率还超过了 20 维和 10 维数据，这是因为在维数缩减时，20 维和 10 维数据还保留了一些相关维，维数之间还有相关性，在近似计算时，由于网格数增多，落入网格中的点数减少，密度阈值减少，稠密区域与聚类的差异性增大，增加了挖掘的难度，减少了检测率。数据维数缩减为 7 维时，尽管维数减少了很多，但是很多相关维数的特性通过代表属性被反映，离群点的性质是通过属性值表现出来的，离群点的特异性会使得其相应的属性值也具有异常性，就不可能与其他属性具有相关性，不可能被合并，所以缩减以后的维数不会改变将被挖掘的离群点的异常性，但是维数缩减以后，挖掘速度得到了很大程度的提高(表 2 中的检测率比和时间比都是与属性个数为 41 时的结果进行比较)。

5.2　算法的精确度

精度采用量度 Precision 对算法评价：

Precision＝the number of actual outlier/the number of outlier through routine calculation

从图 1 可知，属性个数对于算法精度值是有影响的，图中主要说明了 LOF 算法使用基于信息论的降维方法后精确度的提高，此降维方法可以适用于传统的所有离群点挖掘算法的预处理过程中，括号中的数值表示 ITfOM 算法每次选择的属性个数。分别采用 3 种算法挖掘离群点，实验情况说明如下：

1)直接使用 LOF 算法，计算每个点的 LOF 值，此时 LOF 算法的精确度是很低的。从图 1 中可以看出使用文中利用信息论降维的思想，计算互信息后分别选取 30 维、20 维、10 维和 7 维使用 LOF 算法挖掘离群点，挖掘精确度明显提高。此时，选取的 LOF 中的参数大小一致，$m=15$，$\sigma=0.5$，$\xi=3$。

2)对于 SPOD 算法，由于采用了子空间降维的思想，针对高维数据(41 维)，算法精确度受到高维的影响较小，但是在高维时数据量很大，精确度有所下降，依赖于选取的子空间，将属性分为离群属性和非离群属性，但是没有考虑到属性之间的关系，而且在高维空间领域中的点会受到“维数灾难”的影响。该算法中需要设置的参数较多，对算法精确度的影响较大(说明：实验中该算法没有运用文中的降维方法，只是根据原算法挖掘 41 维数据的离群点，所以精确度没有变化。数据量较大，参数选择对于算法有很大影响，所以精确度受到影响)。

3)对于 ITfOM 算法，采用了降维的思想，算法精确度比较稳定，但是在选择属性时属性值的个数也会对其挖掘精确度造成影响，在选择的属性值为 30、20 和 10 时，由于属性之间还有相关性，还存在冗余属性，因此精确度受到了一定程度的影响。但是总的来说，精确度保持在 90%以上。不过由于数据集中数据量较大，在采用网格的方法近似时存在误差，影响了算法的精确度，所以图 1 中发现利用信息论的方法降维后应用于 LOF 算法，其精确度甚至会高于 ITfOM 算法，但是兼顾到文中的方法设置的参数少，人为干扰小等因素，综合来看，ITfOM 算法还是具有优势的。

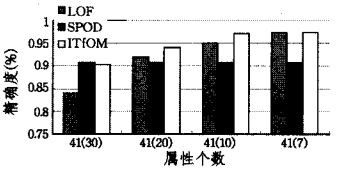

图 1　不同算法的精确度对比

对于 ITfOM 算法来说，主要的精度问题存在于数据集增大以后，被代表的点增多，稠密区域变大，很多点的特性都是通过代表点代替了，所以随着数据集的增多，精确度有所下降。

5.3　算法的响应时间

图 2 中可以看出 ITfOM 算法的挖掘时间有很明显的改善。由于在计算中仅仅计算数据对象的熵值，对象之间的关

(下转第 161 页)

[C]// Proceedings of the 2003 SIAM International Conference on Data Mining(SDM'03). San Francisco:CA,2003

[9] Minaei-bidgli B, Topchy A, Punch W F. Ensembles of Partitions via Data Resampling[C]// Proceedings International Conference on Information Technology, Coding and Computing (ITCC 2004). 2004:188-192

[10] Fern X Z, Brodley C E. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach[C]// Proceedings of the 20th International Conference on Machine Learning. 2003:186-193

[11] Fred A L. Finding Consistent Clusters in Data Partitions[C]// Proceedings of the 2nd International Workshop on Multiple Classifier Systems. Volume 2096 of Lecture Notes in Computer Science. Springer, 2001:309-318

[12] 阳琳赟,周海京,卓晴,等. 基于属性重要性的加权聚类融合[J]. 计算机科学,2009,36(4):243-249

[13] Strehl A, Ghosh J. Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions[J]. Journal of Machine Learning Research, 2003, 3(3):583-617

[14] Rodriguez J J, Kuncheva L I, Alonso C J. Rotation forest: A new classifier ensemble method[J]. IEEE Transactions Pattern Anal Mach Intell, 2006, 28:1619-1630

(上接第 151 页)

系通过预处理时属性频度和数据对象频度计算相联系,避免了相互之间大量的 LOF 计算,同时在挖掘过程中使用熵值作为度量标准,避免了数据集中离群点挖掘时高维空间距离度量的弊端。SPOD 算法在利用信息熵选取子空间后还要进行 LOF 计算,数据量较大时,耗费的时间较多。利用信息论作为度量标准得出的离群点易于解释,熵值比较大,不确定性比较大,成为异常也就很正常了。

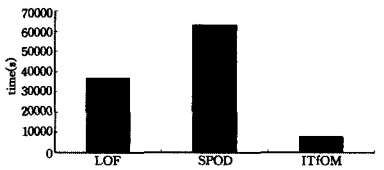

图 2 不同算法的执行时间对比

5.4 算法对数据集维数的伸缩性

为了测试算法对数据集维度的伸缩性,分别在计算熵权后,将参数设置为 35,28,18,9 和 7,按序提取属性,去除冗余属性,对 LOF 算法、SPOD 算法和 ITfOM 算法运行情况进行了分析。

从图 3 可以看出,由于 ITfOM 算法和 SPOD 算法对数据进行了降维处理,因此相应算法对高维有更好的伸缩性,而 LOF 算法没有对高维数据进行降维处理,算法对高维数据的挖掘精确性受到明显的影响。

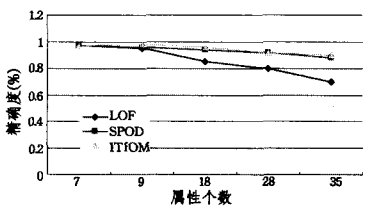

图 3 不同算法对数据集属性个数的伸缩性

结束语 本文研究了高维海量数据集的离群点挖掘问题,提出了基于信息论的高维海量数据的离群点挖掘算法。该算法采用了信息论中信息熵和互信息的概念,利用估算的互信息数值客观地依据熵权排序后进行属性选择,去除冗余属性降维。利用信息熵作为离群点判断的度量标准,可消除距离和密度量纲的弊端。实验结果表明,所提出的算法针对高维海量数据是切实有效的,效率和精度得到了明显提高。进一步提高算法的实现效率和精度以及实时挖掘高速数据并将其扩展到分布式环境,是下一步的研究内容。

参 考 文 献

[1] Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets[C]// A Gupta, O Shmueli, J Widom, eds. Proc of the 24th int'l conf on Very Large Databases. New York: ACM, 1998:392-403

[2] Yu D, Sheikholeslami S, Zhang A. FindOut: Finding Outliers in very large datasets[R]. 99-03. Univ. of New York, Buffalo, 1999:1-19

[3] Breunig M M, Kriegel H, Ng R T, et al. LOF: Identifying density-based local outliers[C]// Chen WD, Naughton JF, Bernstein PA, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data. Dallas: ACM Press, 2000:93-104

[4] Papadimitirou S, Kitagawa H, Gibbons P B, et al. LOCI: Fast outlier detection using the local correlation integral[C]// Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering. IEEE Computer Society Press, 2003:315-326

[5] Aggarwal C, Yu P. An effective and efficient algorithm for high-dimension outlier detection[J]. The VLDB Journal, 2005, 14(2):211-221

[6] 倪巍伟,陈耿,陆介平,等. 基于局部信息熵的加权子空间离群点检测算法[J]. 计算机研究与发展,2008,45(7):1189-1194

[7] Fleuret F. Fast binary feature selection with condition mutual information[J]. Journal of machine learning research, 2004, 5:1531-1555

[8] Wang G, Loehovsky F. Feature seleetion with conditional mutual information maximin in text Categorization[C]// Proeeeding of the Thirteenth ACM Conference on Information and know-ledge management. 2004:342-349

[9] Mao Ye, Xue Li, Orlowska M E. Projected outlier detection in high-dimensional mixed-attributes data set[J]. Expert Systems with Applications, 2009, 36(3)PART 2:7104-7113

[10] He Zeng-you, Xu Xiao-fei, Deng Sheng-chun. A fast Greedy Algorithm for Outlier Mining[C]// Proceedings of PAKDD'2006 (LNAI3918). 2006:567-576

[11] 于绍越,商琳. ENBROD:基于信息熵的相对离群点的检测方法[J]. 南京大学学报:自然科学,2008,44(2):1189-1194

[12] Jiang Feng, Sui Yue-fei, Cao Cun-gen. An information entropy-based approach to outlier detection in rough sets[J]. Expert Systems with Applications, 2010, 37(10):6338-6344

[13] Kraskov A, Stogbauer H, Grassberger P. Estimating Mutual information[J]. Physical Review, 2004, E69(7):69-84

[14] 周晓云,张净,孙志挥. 高维 Turnstile 型数据流聚类算法[J]. 计算机科学,2006,33(11):14-17

[15] 张净,孙志挥. GDLOF:基于网格和稠密单元的快速局部离群点检测算法[J]. 东南大学学报:自然版,2005,42(5):863-866

[16] 李存华,孙志挥. GridOF:面向大规模数据集的高效离群点检测算法[J]. 计算机研究与发展,2003,40(11):1586-1592