文章编号:1671-9352(2008)11-0058-03

基于网格模型的孤立点检测算法

闫宗奎,石冰
(山东大学计算机科学与技术学院,山东 济南 250101)

摘要:为了从数据集中快速有效地发现孤立点,提出了一种基于网格模型的孤立点检测方法,给出了数据空间的网格划分,定义了网格内孤立点存在性阈值,提出了基于网格的孤立点检测算法,在保证算法有效性的前提下,降低了算法的时间复杂度。

关键词:数据挖掘;孤立点;网格模型
中图分类号:TP311 **文献标志码**:A

An outlier-analysis algorithm based on the grid model

YAN Zong-kui, SHI Bing
(Computer Science and Technology School, Shandong University, Jinan 250101, Shandong, China)

Abstract: To find the outlier in a data set more quickly and efficiently, an outlier-analysis based on the gird model was provided. This algorithm gives the way to partition the data space by the grid model, defines the boundary value of judging if there is an outlier existing in one grid, and gives the algorithm, which can be used in detecting the outliers correctly with less time.

Key words: data mining; outlier; grid model

0 引言

近年来,在数据挖掘领域,发现数据对象的共同特征的研究已经取得了很大进展,人们称这些共有的数据特征为大模式(如分类[1],聚类[2],关联规则[3]等)。

本文主要讨论如何发现数据集中的孤立点,即小模式。孤立点就是数据集中被认为与其他数据对象不相似或不一致的数据。孤立点本身往往隐藏着非常重要的信息,孤立点检测是知识发现和数据挖掘中的活跃领域,被广泛应用于信用卡欺诈、入侵检测、天气和气候侦测等领域。

当前主要的研究方法:

基于统计的方法:对给定的数据集合假定了一个分布或概率模型,根据模型采用不一致性检验来确定孤立点。其主要缺点是要求预知数据的分布,而在大多数情况下数据分布是未知的。另外,这种方法不能有效地发现多维数据空间的孤立点。

基于距离的方法:这种方法又可细分为基于索引的算法[4]、基于单元的算法[5]和嵌入-循环算法[6]。

此外,一些聚类算法也提供了对孤立点的处理,如CURE[2],但其往往受到算法目的的影响而限制了对孤立点的检测。

本文采用了一种新的方法:首先对数据模型和标准模型进行网格化处理得到二维网格模型并取得相应的采样信号。对每一个网格,通过对这个采样信号和规则采样信号进行处理得到一个孤立点存在性阈值,以此来判断该网格内是否存在孤立点。如果存在,则在该网格内通过一般的孤立点检测算法找出孤立点;否则,处理下一个网格。

这种方法优势在于,孤立点在数据空间的分布一般是稀疏的,大部分的数据区域并不包含有孤立

收稿日期:2008-09-12
作者简介:闫宗奎(1983-),男,硕士研究生,研究方向为数据挖掘,信息检索. Email:yanzongkui@163.com
石冰(1957-),男,教授,主要研究数据库理论、数据挖掘、信息检索. Email:shibing@sdu.edu.cn

点,本算法通过对每个网格的预处理,避免了对大部分数据区域不必要的孤立点分析。缺点在于需要一个标准模型或近似标准模型来进行比对。但在大部分的实际应用中,这个近似模型都是可以获得的。比如在检测气候异常时,可以以过去某些正常年份的数据建立标准数据模型,以此作为判断依据。

1 一种基于单元的孤立点检测算法

基于单元的孤立点算法[7]是基于距离的孤立点算法的一种。最初的基于单元的孤立点算法具有较好的时间复杂度,但容易对边界单元格中的孤立点发声误判。下面介绍一种改进的基于单元的孤立点检测算法。它提出了动态调整 M 值边界函数,从而解决了边界单元格的孤立点误判问题。

算法描述:

(1) 对该网格进行单元格划分。

(2) 将每个单元格 C_{ij} 的对象数目 $\mathrm{Count0}_{ij}$ 设置为 0。

(3) 对于网格内的每个对象 P,将 P 分配到相应的单元格 C_{ij} 内,在 C_{ij} 内存储 P,再将 $\mathrm{Count0}_{ij}$ 加 1。

(4) 判断每个单元格内 $\mathrm{Count0}_{ij}$ 是否大于 M,若成立将 C_{ij} 置为 red。

(5) 对于每一个 red 的单元格,若它的第一层邻居不是 red,则置其为 pink。

(6) 对每一个既不是 red 也不是 pink 的单元格 C_{ij}:

a. 计算 C_{ij} 及 $L_1(C_{ij})$ 中所有对象数目 $\mathrm{Count1}_{ij}$ 的值

b. 若 $\mathrm{Count1}_{ij}$ 的值大于 M_{c1},将 C_{ij} 置为 pink

c. 否则,计算 C_{ij} 及 $L_1(C_{ij})$,$L_2(C_{ij})$ 中所有对象数目 $\mathrm{Count2}_{ij}$ 的值,若 $\mathrm{Count2}_{ij} \leqslant M_{c2}$,将 C_{ij} 中的所有对象标记为孤立点;否则,对于 C_{ij} 内的每一个对象 P:

ⅰ 将 $\mathrm{Count2}_{ij}$ 的值付给 Count_p

ⅱ 对于 $L_2(C_{ij})$ 中每一个对象 Q,计算 PQ 之间的距离 dist,若 $\mathrm{dist} \leqslant D$,则 Count_p 加 1;若 $\mathrm{Count}_p > M_{c2}$,$P$ 不是孤立点;否则标记 P 是孤立点。

2 基于网格的孤立点检测算法

传统的检测算法需要对数据模型中的每一个对象进行检测,而孤立点往往只存在于数据空间的很小区域里,这就造成了在孤立点检测时大量的时间浪费在对非孤立点数据对象的分析上。下面的算法在一定程度上避免了这种不必要的时间开销。

2.1 算法描述

Step 1 对数据模型和标准模型分别建立网格模型 M, M'。

Step 2 对 M, M' 进行采样得到采样信号 F, F'。

Step 3 对 F, F' 采用下述方法获得检测向量 $\boldsymbol{w}, \boldsymbol{w}^*$:

对 F, F' 进行 n 级 DWT 变换,得到分级的小波子带系数 VLL_i, VLH_i, VHL_i, VHH_i $(i = 1, 2, \cdots, n)$,其中 LL, LH, HL, HH 分别为小波变换域的低低频子带,低高频子带,高低频子带,高高频子带,i 是小波分解级别。

对每个子带按公式(1)获得其相应的权值系数。

$$w_i = \frac{V'_{s,ix} - V_{x,ix}}{\alpha_s T_s}, \tag{1}$$

其中,α_s 是与子带 s 有关的权值因子,可由用户指定。T_s 是被选择子带的阈值。

重复步骤 1 和 2,直到遍历完所有的子带,得到检测向量 $\boldsymbol{w}, \boldsymbol{w}^*$。

Step 4 检测网格内是否存在孤立点。

$$\mathrm{corr}(\boldsymbol{w}^*, \boldsymbol{w}) = \frac{\sum_{i=1}^{N}(w_i^* - \overline{w^*})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^{N}(w_i^* - \overline{w^*})^2}\sqrt{\sum_{i=1}^{N}(w_i - \bar{w})^2}},$$

其中,$\boldsymbol{w}^*$ 和 $\boldsymbol{w}$ 分别是从数据模型和标准模型中提取出的判断向量,$\bar{w}$ 是向量 $\boldsymbol{w}$ 的均值,corr 取值在 $[-1, 1]$ 之间。如果这一相关值超过某一阈值,就判定该网格中存在孤立点,进入 Step5。

Step 5 通过第 1 章中提到改进的基于单元的孤立点检测算法找出孤立点。

2.2 算法分析

该算法首先通过对数据模型的网格化和采样处理,取得区域内存在孤立点的可能性,对存在孤立点可能性较大的区域,采用普通的孤立点检测算法检测可能存在的孤立点。从整体上来看,由于避免了对整个数据空间内每个数据点的判断,从而降低了算法的时间复杂度。对于每个单元格,由于只在检测到孤立点的情况下才会进行下一步处理,而在网格划分模型足够好的情况下,可以将大部分不含有孤立点的网格排除在外,从而降低算法的执行时间。在最坏的情况下,所划分的网格内都存在有孤立点,这就需要在网格划分时采用正确的网格模型进行预处理,避免这种情况发生。在实际应用中,虽然在网格化、采样等环节上会影响其执行效率,但由于孤立

点在数据空间中的数量较小,这种影响还是可接受的。

3　实验分析

本文采用太平洋近赤道地区的气候数据[8]作为实验数据集。数据集中存在 178 000 数据记录。通过调整网格内孤立点存在性判断阈值,得到表 1 所示的试验结果。

表 1　试验结果

Table 1　The experiment result

算法	阈值	时间/s	孤立点检测数目
基于单元格的检测算法	NA	41	137
基于网格的检测算法	0.85	65	136
	0.90	49	136
	0.95	30	128
	0.99	14	97

从实验结果可以看出,通过适当调整判断阈值,可以得到相对理想的检测结果。实验结果表明,本文的算法在降低了检测算法时间的同时,可以保证检测结果的有效性。

4　结语

本文给出了一种孤立点检测的改进算法,当数据量很大并且已存在或者可以构造出一个标准的数据模型时,本算法可以提高孤立点检测的效率,并通过实验结果验证了算法的有效性。本文的方法可以应用于异常气候监测、诈骗监测等领域。

参考文献:

[1] 陆声链,林士敏. 基于距离的孤立点检测研究[J]. 计算机工程与应用, 2004,(33):9-10.

[2] GUHA Sudipto. Cure: An efficient clustering algorithm for large databases[C]// SIGMOD Conference, New York: ACM Press, 1998: 73-84.

[3] AGRAWAL Rakesh. Fast discovery of association rules[C]// Advances in Knowledge Discovery and Data Mining. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996: 307-328.

[4] 尚俊平,邱保志,刘合兵. 一种基于距离的聚类和孤立点检测算法[J]. 河南科学, 2007,(06):975-978.

[5] 孙焕良,鲍玉斌,于戈, 等. 一种基于划分的孤立点检测算法[J]. 软件学报, 2006,17(05):1009-1016.

[6] JOHANNA H, ROCKE D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator[J]. Computational Statistics & Data Analysis. 2004, 44:625-638.

[7] 邵峰津,孙仁成,于忠清. 基于单元的孤立点发现改进算法[C]// 中国科协 2003 年学术年会论文集:上,2003:538.

[8] U S University of California, Irvine. El nino data [DB]. [2008-09-09] http:// kdd.ics.uci.edu/databases/el_nino/el_nino.html, 30 June 1998

(编辑:孙培芹)