基于频数的孤立点检测研究

朱东生,吴庆波,谭郁松
(国防科学技术大学 计算机学院,湖南 长沙 410073)

摘 要:基于距离的孤立点检测算法在很多领域都有重要应用,效率不高却限制了孤立点检测算法的广泛应用。针对这个问题,文中通过分析基于索引的检测算法和基于单元的分析算法,受频繁项集挖掘算法的启发,应用统计学原理,提出了改进的基于距离的孤立点检测算法(Index Unit Based-on-Distance Outlier Mining,IU-BDOM)。在待挖掘数据集中,出现次数越少的项集越可能是孤立点,即频数越低越可能是孤立点,本算法在检测孤立点时,从频数最小的项开始检测,从而节省了挖掘频数很高的肯定不是孤立点的数据所带来的开销。为了进一步加快速度,实现算法的并行性,采用超立方体代替传统的超球体来统计数据集中每一个对象 o 的邻居个数,把距离的计算分散到不同的维度上独立进行,并且给予不同维度不同的权重,此外,利用 Greenpulm 分布式数据库,并行了挖掘任务,极大地提高了挖掘效率,并通过实验证实了这种改进的有效性。

关键词:孤立点检测;频繁项集;基于距离;Greenplum
中图分类号:TP912.3 **文献标识码**:A **文章编号**:1673-629X(2013)05-0010-04
doi:10.3969/j.issn.1673-629X.2013.05.003

Research on Frequency-based Outlier Mining

ZHU Dong-sheng,WU Qing-bo,TAN Yu-song
(College of Computer,National University of Defense Technology,
Changsha 410073,China)

Abstract:Distance-based outlier detection algorithm in many fields has important applications,but the efficiency is not high which limited the widely used outlier detection algorithms. For this problem,through analysis of the index detection algorithm and cell-based analysis algorithms,inspired by frequent itemsets mining algorithm,using statistical principles,proposed an improved distance-based outlier detection algorithm (Index Unit Based-on-Distance Outlier Mining,IU-BDOM). Data to be excavated concentrated,appears more times the more less of the item sets may be an outliers,i.e. the frequency is the more low,the more likely is an outliers. The present algorithm in the detection of the outliers,from the frequency of the minimum of the items start detection,thereby saving the excavation frequency number overhead high certainly not an outliers. In order to further accelerate the speed and realize the parallelism of the algorithm,the number of neighbors used the hypersphere to statistics hypercubes instead of the traditional centralized each object o,the distance independently calculated dispersed into different dimensions,and give different weights to different dimensions,in addition,the use of distributed database of Greenpulm,parallel mining tasks and greatly improve the efficiency of mining,and the effectiveness of such an improved is confirmed by experiment.

Key words:outlier detection;frequent itemsets;distance-based;Greenplum

0 引 言

孤立点的检测是数据挖掘的重要组成部分,广泛应用在噪声检测、医疗、生态系统失调、公共卫生、贷款审批、气象预报、网络入侵检测等领域。孤立点的检测算法有多种,其中基于距离的孤立点检测算法是使用最为广泛的。

基于距离的孤立点检测算法主要有三种[1]:

1)基于索引的孤立点检测。对要挖掘的数据集建立索引,因为要对多维数据同时建立索引,经常使用的是 R-Tree 索引或者 K-D-Tree 索引。

2)嵌套循环的孤立点检测。不使用索引结构,对每一个点都计算它到其他点的距离,通过调整顺序达到最小化 I/O 的目的,从而提高算法效率。

3)基于单元的算法。通过划分单元,排除不是孤立点的点,尽可能地减少了需要检测的点数,以达到加快算法执行速度的目的。

三种算法各有优缺点,下面具体介绍和文中相关

收稿日期:2012-08-31;**修回日期**:2012-12-03
基金项目:国家核高基计划项目(2012ZX01040001)
作者简介:朱东生(1988-),男,硕士研究生,研究方向为海量数据挖掘。

的基于索引和基于单元的孤立点挖掘算法。

1　相关研究

1.1　孤立点定义

到目前为止,对于孤立点的定义还没有一个可以被研究人员广泛接受的表述[1~4]。Hawkins D M 对孤立点的定义是一个接受度比较广泛的表述:一个孤立点是一个观测值,它与其他观测值的差别如此之大,以至于怀疑它是由不同的机制产生的[3]。在基于距离的孤立点检测中,孤立点的定义常用如下形式:如果数据集 DB 中对象至少有 p 部分与对象 o 的距离大于 $d_{\min}$,则称对象 o 是以 p 和 $d_{\min}$ 为参数的基于距离的孤立点[1]。

1.2　孤立点算法分析

基于索引的孤立点检测算法[5~10]基本思想是:对给定的数据集 DB,首先在所有的属性上建立多维索引,再通过索引搜索每个对象 o 在半径 $d_{\min}$ 范围之内的点。如果在该范围内发现了足够多的点说明对象 o 不是孤立点,否则对象 o 是孤立点。具体如算法 1 的描述。

算法 1:

输入:数据集 DB,邻居个数 M,最小距离 $d_{\min}$

输出:孤立点的集合

1 初始化孤立点集合 S 为空集

2 初始化临近对象计数器 counter = 0

3 for element in DB loop

4 统计以 element 为中心,以 $d_{\min}$ 为半径的范围的邻居个数计入 counter

5 if counter <= M then

//孤立点加入孤立点集合中

6 S . add(element)

7 end if

8 end loop

9 return S

算法 1 在最好的情况下复杂度为 $O(k*M*n)$,在最坏的情况下复杂度为 $O(k*n^2)$,其中 n 是数据集中的记录数,k 是数据维数。复杂度不考虑建立索引所花费的时间。该算法的优点是对维度的扩展性好,可以处理高维数据,但是数据量的伸缩性差,对于大数据时代,这样的算法显然难以满足工业界的需求。

基于单元的孤立点检测算法基本思想是:按照给定的最小距离 $d_{\min}$,结合数据集的维数 k, 把数据集划分到一个个的单元中,通过统计每个单元及其两层环绕单元中对象的个数排除掉不可能是孤立点的对象。最后对剩余的对象进行孤立点检测。

算法描述如算法 2。

算法 2:

输入:数据集 DB,最小距离 $d_{\min}$,邻居个数 M

输出:孤立点数据集

1 初始化孤立点数据集 S 为空集

2 根据数据集中的维数 k,以长度 $d_{\min}/(2*k^{0.5})$ 为边长划分单元,记单元的集合为 D

3 统计单元中对象个数 counter_0,第一层单元中对象个数 counter_1,第二层单元中对象个数 counter_2

4 for element in D loop

5 if counter_0 + counter_1 < M then

6 if counter_0 + counter_1 + counter_2 < M then

7 for object in element loop

8 S . add(object)

9 end loop

10 else

11 for object in element loop

12 统计 object 中满足条件的对象个数 counter

13 if counter < M then

14 S . add(object)

15 end if

16 end for

17 end if

18 end if

19 end loop

20 return S

算法 2 的时间复杂度为 $O(c^k + n)$[1],其中 c 是依赖单元数目的常数,k 是数据维数,n 是数据集中数据的记录数。算法对数据量的大小有良好的伸缩性,但是随着数据维数的增加,该算法的执行时间急剧增加。Knorr 和 Ng 通过试验进行了验证[11],只有当维数不超过 4,即 $k \leqslant 4$ 时,基于单元的孤立点检测算法才会优于基于索引的挖掘算法。综上,基于单元的孤立点检测算法的优点是数据量的伸缩性好,可以很好地支持海量数据,缺点是维度的可扩展性差,如今的数据都是高维的,有的甚至达到上亿维度,所以,很有必要提出一种新的算法来解决这个问题。

孤立点的一个重要特征就是区别与其他的值[12],此外孤立点在数据集中总是少量的。针对基于索引的检测算法和基于单元的检测算法的不足,文中结合了单元划分和索引的优点,利用频繁项集中的频数,提出了改进的基于距离的孤立点检测算法(IU-BDOM)。

2　IU-BDOM 算法

2.1　准备工作

算法如果可以并行的计算,那么算法的执行速度要比串行执行快得多,比如在进行频繁项集挖掘时,算

法可以通过分布式系统把任务分布到各台机器,各个 CPU 上进行。而孤立点检测算法不论是基于索引还是基于单元几乎都是串行的,因此串行执行是孤立点检测算法执行效率低的一个重要因素。此外,数据集中的大部分数据对象都不是孤立点,对它们进行的计算都是无用的,虽然基于单元的检测算法进行了过滤,但是这种过滤并不能充分过滤掉不是孤立点的数据对象。针对不是孤立点的过滤,文中采用了频繁项集的做法和统计的做法,即孤立点是特别的、不常见的,它们出现的频数是最低的。

通过使用频数进行过滤,把大部分的数据对象过滤掉,只针对最可能是孤立点的那些数据对象进行检测。这样可以大大减少距离计算复杂度,加快算法的执行效率。该算法还克服了基于单元的检测算法对非数值型数据支持度不好的问题。为了进一步加快速度,实现算法的并行性,采用超立方体代替传统的超球体来统计数据集中每一个对象 o 的邻居个数。使用超立方体有 4 个好处:

1)可以把距离的计算分散到不同的维度上独立进行;

2)简化了距离的计算复杂度;

3)可以对每一维的数据单独建立索引,简化了索引的复杂度;

4)可以对不同维的数据给予不同的权重,而不影响计算的复杂度。

同时,为了能够处理增量的问题,IU-BDOM 算法使用一个 Map 结构存储结果集,其中的 key 用于标识不同的孤立点数据项,value 记录孤立点数据项中满足条件的邻居个数。这样的数据结构在进行增量的处理时可以进行快速合并而不会引起误差放大。

2.2 IU-BDOM 算法描述

由于 IU-BDOM 只是对最可能是孤立点的那一部分数据进行分析,这就需要另一个参数 $q_{\max}$, 表示最感兴趣的孤立点的比例。算法的伪代码表示如算法 3。

算法 3:

输入:数据集 DB, 邻居个数 M,最小距离 d,感兴趣孤立点数 N

输出:孤立点集合

1 初始化孤立点集合 S 为空集,对每一维数据建立索引

2 对数据集 DB 中的每一个数据项 x, 统计它们出现的频数

3 根据各数据项出现的频数按照从小到大的顺序进行排序, 选出 N 个数据项 D

4 根据距离 d 和数据集的维数 k,确定每一维的距离(d/k)

5 for element in D loop

6 以 element 为中心,统计以 d/k 为边长的超立方体中的数据对象个数 counter

7 if counter < M then

8 S. add(element)

9 else

10 break

11 end if

12 end loop

13 return S

算法 3 在最坏的情况下的时间复杂度为 $O(k*q*n^2)$, 在最好的情况下的时间复杂度为 $O(k*q^2*n^2)$,其中 k 是数据集 DB 的维数, q 是用户感兴趣孤立点的比例,n 是数据集中的记录数。这里同样没有考虑建立索引使用的时间,即认为建立索引使用的时间和算法执行使用的时间相比可以忽略不计。

3 实验结果及分析

3.1 实验环境

实验基于 Greenplum 4.1.0.0 平台,使用一个 master 节点,两个 segment 节点,不使用镜像。实验在 Intel ® Pentium T2330 @1.6GHz 处理器,2G 内存的机子上运行,使用的操作系统是 RedHat Linux Enterprise5.0,测试代码使用 PLPGSQL 编写。

实验数据一个是从数据堂下载的真实数据(http://www.datatang.com/data/43314),一个是使用 MATLAB 按照高斯分布生成的数据,两个数据都是 12 维的。

3.2 实验结果

图 1 显示了在数据集的维数比较大时,基于单元的孤立点检测算法执行时间要多于基于索引的检测算法,这与 Knorr 和 Ng 的试验结果相吻合。而 IU-BDOM 算法随着数据集的增加有良好的伸缩性,基本保持了随着数据集的增加而线性增长。

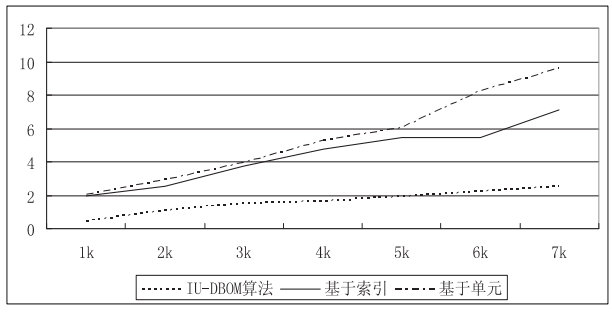

图 1 12 维数据不同数据集大小

图 2 显示在 2000 条数据集上不同维数的各种算法的执行时间,从图中可以看出基于索引的检测算法

基本保持了线性增长方式;基于单元的检测算法在低维度时具有良好的效率,但是随着维度的增加,执行时间快速增加;而 IU-BDOM 算法对维度的增加也基本上保持了线性增长,而且增长速度要低于基于索引的检测算法。这说明 IU-BDOM 算法在数据集的大小和维度上都具有良好的伸缩性。

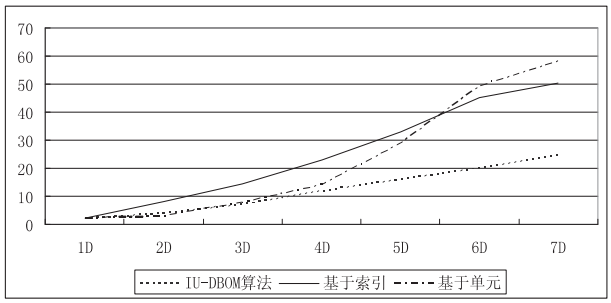

图 2　相同数据集大小不同维度

4　结束语

基于距离的孤立点检测算法在很多领域都有重要应用,但是现有的算法效率都不是很高,导致了工业界的应用并不是很广泛。针对这个问题,文中通过阅读相关的参考文献,受频繁项集挖掘算法的启发,提出了改进的基于距离的孤立点检测算法(IU-BDOM)。在待挖掘的数据集中,传统的做法是扫描整个数据集来挖掘孤立点,然后频数越低越可能是孤立点,本算法在检测孤立点时,从频数最小的项开始检测,从而节省了挖掘频数很高的肯定不是孤立点的数据所带来的开销。为了进一步加快速度,实现算法的并行性,文中采用了超立方体代替传统的超球体来统计数据集中每一个对象 o 的邻居个数,此外,利用 Greenpulm 这个分布式数据库,并行了挖掘任务,极大地提高了挖掘效率,通过上述实验证实,IU-BDOM 算法在数据集的大小和维度上都有良好的伸缩性并且效率很高。

尽管利用 IU-BDOM 算法可以改进孤立点挖掘的效率,但是处于大数据时代,如果每次挖掘进行时都需要大量重复地扫描全库,势必会带来不必要的开销,如果可以实现增量式的挖掘,那样整个算法的效率将大大提高,所以,下面的工作主要是在 IU-BDOM 上实现增量式的挖掘。

参考文献:

[1] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. Beijing: Higher Education Press, 2009: 335-345.

[2] Knorr E, Ng R, Tucakov V. Distance-based outliers: algorithms and applications [J]. The VLDB Journal, 2002, 8(3/4): 237-253.

[3] Hand D, Mannila H, Smyth P. Principles of Data Mining [M]. Beijing: China Machine Press, 2003: 272-276.

[4] Han J, Pei J, Yin Y. Mining Frequent Patterns without Candidate Generation [C]//Proc. of ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00). Dallas, TX: [s. n.], 2010.

[5] David C, Han Jiawei, Vincent T N, et al. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique [C]//Proc. of the 12th Int'l Conf. on Data Engineering. New Orleans, Louisiana, USA: [s. n.], 2011.

[6] Tan Pangning, Steinbach M, Kumar V. Introduction to Data Mining [M]. Beijing: Post & Telecom Press, 2010: 403-405.

[7] 黄德才,张良燕,龚卫华,等. 一种改进的关联规则增量式更新算法[J]. 计算机工程,2008,34(10):38-42.

[8] 商志会,陶树平. 一种高效的关联规则增量更新算法[J]. 计算机应用,2011,25(4):830-833.

[9] Zhu Honglei, Li Ming. An Incremental Updating Algorithm for Maintaining Discovered Association Rules [J]. Application Research of Computer, 2004, 21(9): 107-109.

[10] 梁之舜,邓集贤,杨维权,等. 概率论及数理统计[M]. Beijing: Higher Education Press, 2009: 150-153.

[11] Knorr E, Ng R. Algorithms for Mining Distance-based Outliers in Large Datasets [C]//Proc. of the 24th VLDB Conference. New York: [s. n.], 2012: 392-403.

[12] Bay S, Schwabacher M. Mining Distance-based Outlier in Near Linear Time with Randomization and Simple Pruning Rule [C]//SIGKDD'03. Washington DC: [s. n.], 2011.

(上接第 9 页)

ing and Comparing of Disaster Plans [C]//Proceedings of the Second International ISCRAM Conference. Brussels, Belgium: [s. n.], 2005.

[8] 刘　栋,陈　颖,沈　沉,等. 电力应急预案数字化方法研究[J]. 电力系统自动化,2009,33(21):48-52.

[9] 奚旦立,陈秀华. 突发性污染事件应急处置工程[M]. 北京:化学工业出版社,2009:11-13.

[10] 师立晨,刘　骥,魏利军. 重大危险源多米诺效应的后果分析[J]. 中国安全生产科学技术,2007,3(6):44-48.

[11] Jennex M E. Modeling Emergency Response Systems [C]// Proceedings of the 40th Hawaii International Conference on System Sciences. Hawaii: [s. n.], 2007.

[12] Dorasamy M, Raman M. Information Systems to Support Disaster Planning and Response: Problem Diagnosis and Research Gap Analysis [C]//Proceedings of the 8th International ISCRAM Conference. Lisbon, Portugal: [s. n.], 2011.

[13] Aedo I, Díaz P, Bañuls V A, et al. Information Technologies for Emergency Planning and Training [C]//Proceedings of the 8th International ISCRAM Conference. Lisbon, Portugal: [s. n.], 2011.

[14] 王晓明,何天平. 江苏省化工企业应急救援现状分析[J]. 中国安全生产科学技术,2008,4(5):126-129.

基于频数的孤立点检测研究

作者： 朱东生， 吴庆波， 谭郁松
作者单位： 国防科学技术大学 计算机学院,湖南 长沙 410073
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(5)