An Outlier Detection Algorithm for Incremental Data and its Application in Auditing of Social Security

Shaobin Huang[1], Tianyang Lv[1], Ronghua Chi[1], Yong Xia[1]
[1]*College of Computer Science and Technology, Harbin Engineering University*
Email:raynor1979@163.com, Harbin, Heilongjiang, 150001

Abstract: It is an important direction to detect outliers efficiently from the incremental data. The method can be widely applied in auditing, stream media analysis etc. Since the network-based auditing system can extract data from the audited enterprise or government in real time, it is crucial to discover outliers efficiently as conditates of audit doubts from the incremental data. However, most of the outlier analysis algorithms are designed to handle the static data sets. The paper presents a density-based local outlier analysis algorithm which is based on the clustering result of historical data and can efficiently detect outliers from the incremental data. The algorithm is applied to audit the real social security data. The experiments show the effectiveness and the efficiency of the proposed method.

Keywords: Outlier; Incremental data mining; Local density; Social-security audit

一种增量离群点识别算法及其在社会保障审计中的应用

黄少滨[1]，吕天阳[1]，迟荣华[1]，夏　勇[1]
[1]哈尔滨工程大学 计算机科学与技术学院
Email: raynor1979@163.com 黑龙江 哈尔滨 150001

摘　要：针对增量数据的数据挖掘技术是当前数据挖掘研究的重要方面，其中离群点发现技术对审计等诸多问题有较大的价值。由于联网审计能够实时采集被审计单位的增量数据，因此从中高效的发现离群点具有重要的审计意义。但是，现有的多数离群点分析算法针对静态数据，在处理增量数据集时复杂度较高。为此，提出基于密度的增量式离群点识别算法，根据历史数据的聚类分析结果，实现对增量数据的高效的离群点识别。实验表明，新算法在保证识别效果的前提下，极大的提高了效率，对社会保障数据的审计也证实其有效性。

关键词：增量式数据挖掘；离群点识别；局部密度；社会保障审计

1 引言

当前，数据挖掘技术已经取得了非常大的进展，在商业、工业、医学、科研等领域有着成功的应用[10]。如何对复杂、增量的数据进行有效的数据挖掘是一个较重要的研究方向。与此同时，数据挖掘技术在新领域中的应用也是一个有意义的研究课题。

其中，离群点识别是数据挖掘一个独立研究领域[10]，其目标是发现有“特异行为”的数据。Hawkins 将离群点定义为：离群点是在数据集中偏离大部分数据的数据，使人怀疑这些数据的偏离并非由随机因数产生，而是产生于完全不同的机制[2]。

由于审计疑点通常对应着异常数据，因此可以采用离群点识别技术在缺少审计专家指导的情况下发现审计疑点。在此方面国内一些研究者仅是提出一种设想或框架[8]。美国审计署则在实际工作中较为全面的采用数据挖掘技术发现审计问题[9]，但是并不针对增量数据。

本文在前期对审计署金审一期重点项目“社会保障联网审计”课题研究的基础上，研究针对社会保障数据的离群点识别技术。由于联网审计能够实时采集被审计单位的增量数据，因此从中高效的发现离群点具有重要的审计意义。获得增量数据中新离群点的方法通常有两

本文受国家科技支撑计划 2009BAH42B02，国家自然科学基金 60903080、60873038，黑龙江省哲学社会科学研究项目 08E061，中央高校基本科研业务专项资金项目 HEUCF100603 资助。

978-1-935068-18-1

种[1]：一是对整个数据集重新进行离群点分析，但是处理增量数据集时复杂度较高；二是增量式离群点分析。

为此，本文提出基于密度的增量式离群点识别算法 IncreLOF。首先采用依据 LOF 算法思想改进的基于密度的 DBSCAN 算法对历史数据集进行聚类分析；根据历史数据的聚类分析结果，实现对增量数据的高效的离群点识别。实验表明，新算法在保证识别效果的前提下，极大的提高了效率，对社会保障数据的审计也证实其有效性。

全文结构如下：第 2 节介绍相关工作；第 3 节给出基于密度的增量式离群点识别算法 IncreLOF 的基本概念和算法流程；第 4 节分析 IncreLOF 算法的特点；第 5 节实验与分析。

2 相关工作

早期的离群点识别算法依据对整个数据集分布的“全局”性假定发现“全局性”的离群点。由于数据分布情况的复杂性，上述“全局性”假定经常遇到困难。图 1 给出了一个实例。

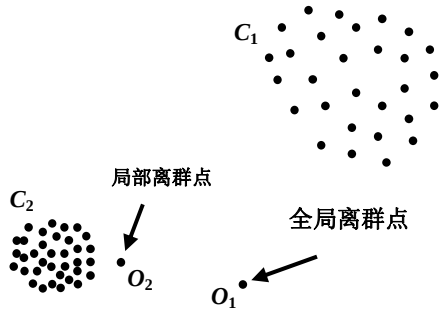

图 1 局部离群点示例

Breuning 等提出局部离群点的概念：如果一个对象相对于它的局部邻域是远离的，特别是关于邻域密度是远离的，即为局部离群点(local outlier)。还提出每个对象并不仅是二分属性，是离群点或不是离群点，而是具有离群度[3]，称为局部离群点因子 LOF。离群点就是有高 LOF 值的对象。

DBSCAN[4]算法的主要思想是：如果一个对象 p 在其半径为 ε 的邻域内包含至少 MinPts 个对象，那么该区域是密集的，并创建以 p 为核心对象的新簇；DBSCAN 算法反复合并从这些核心对象直接密度可达的对象；当没有新对象被添加到任何簇时，聚类结束。其中，给定对象半径 ε 内的区域称为该对象的 ε-邻域；给定一个对象集合 D，如果 p 在 q 的 ε-邻域内，并且 q 是一个核心对象，则认为对象 p 从对象 q 出发是直接密度可达的(directly density-reachable)。

由于 DBSCAN 算法不能有效的处理簇之间密度变化较大的情况，文献[5]引入 k 邻域半径、k 邻域点集等 LOF 算法中局部密度的思想改进了算法，使 DBSCAN 能够按数据对象所处区域的局部密度调整参数进行聚类，从而产生更为可靠的聚类结果。

文献[6]较先提出针对增量数据的 DBSCAN 改进算法，较系统地研究了插人单个数据对象 p，对其 ε-邻域对象密度的影响，并将这种影响各自分为 4 类：噪声、创建新的聚类、归入某一聚类、合并相邻聚类。

3 基于密度的增量式离群点识别算法

为实现增量式的离群点分析，本文首先采用 DBSCAN 改进算法[5]将历史数据集进行预先聚类，进而提出一种增量式离群点识别算法 IncreLOF。

IncreLOF 算法分两个子部分：（1）对增量数据集算法进行增量聚类，确定增量数据中的数据对象是否属于某一聚类；（2）对未被归入任何聚类的对象，计算其局部离群因子 LOF，判定 LOF 高于预先设定阈值的对象为离群点。

不妨设，D 为数据对象集，N 为数据对象的总数，a 为任意对象且 a∈D，p 为插入的新对象。相关概念如下：

定义 1 k-距离：a 的 k-距离定义为 a 到其 k 最近邻的最大距离，记为 k-distance(a)。

定义 2 k-距离邻域(k-distance neighborhood)：a 的 k-距离邻域定义为以 a 为中心，半径为 k-distance(a)的空间区域。

定义 3 k-邻域点集(k-distance points set)：a 的 k-邻域点集记作 Nk-distance(a)，或简记作 Nk(a)，它包含所有与 a 的距离不超过 k-distance(a) 的对象，即 Nk(a)={q|a∈D∧dist(a,q)≤k-distance(a)}。

定义 4 k-邻域半径(k-radius)：a 的 k 邻域半径定义如下：

$$k-radius(p)=\frac{1}{k}\sum_{q\in N_k(p)}dist(p,q) \qquad (1)$$

对象 a 的 k 邻域半径表征对象 a 的 k 距离邻域中的数据对象到 a 的距离的均值。

定义 5 核心对象(core point)：若 a 为核心对象，则有

$$k-radius(p)\le\frac{1}{k}\sum_{q\in N_k(p)}k-radius(q) \qquad (2)$$

本文用对象 p 的 k-邻域点集 Nk(p)的概念定义处理增量数据时用于更新的种子对象。

定义 6 用于更新的种子对象(Seed-Objects for the update)：UpdSeeds(p)={q|q 是 D∪{p}中的核心对象，且 p∈Nk(q)}。

在处理增量数据 p 时，将根据 UpdSeeds(p)的情况确定处理 p 的方法。当 UpdSeeds(p)为空时，p 为噪声对象。当 UpdSeeds(p)不为空时，其中的核心点数据对象有三种情况：一是 UpdSeeds(p)中的核心对象属于某一聚类，且在 D 中是核心对象；二是 UpdSeeds(p)中的核心对象属于某一聚类，且在 D 中是不是核心对象，由于 p 的插入改变了其的 k-邻域半径而成为核心对象；三是 UpdSeeds(p)中的核心对象在 D 中为噪音对象，由于 p 的插入成为核心对象。算法将 p 归入 UpdSeeds(p)中与 p 距离最近的核心对象所在的簇，对于后两种情况还要将核心对象的 k-邻域点集归入核心对象所在簇。

处理增量数据的算法流程参见图 2。

```
算法：  IncreCluster
输入：  增量数据集 D'
输出：  增量数据集中的聚类及噪声数据
算法流程：
1.For 数据集 D'中的每个对象 p
2. 取得插入 p 后更新的种子对象 UpdSeeds(p)
3. If UpdSeeds(p)为空
4.      将 p 标记为噪声；
5. Else
6.      For UpdSeeds(p)中的每个数据对象 q
7.          If q 是噪声 或 q 是非核心对象
8.              将 q 的 k-邻域点集 Nk(q)中的元素用 q 的类号标记；
9.          End If
10.         End For
11.         将 p 用距离最近的 q 的类号标记；
12.     End If
13.End For
```

图 2 增量数据聚类处理流程

在增量式离群点检测算法中，算法只需分析初始数据集和增量数据集中所有非聚类中的对象，对象计算其的 LOF 值，在给定的阈值条件下确定其是否是离群点。参照文献[3]中的定义，给出了局部离群点因子(LOF)的定义及其计算方法。

定义 7：对象 p 相对于对象 o 的可达距离(reachability distance)。对象 p 相对于对象 o 的可达距离定义为对象 p 与对象 o 之间的实际距离与对象 o 的 k-距离之间的较大者，即

$$\text{reach-dist}_k(p, o) = \max\{\text{k-distance}(o), \text{dist}(p, o)\}$$

定义 8：对象 P 的局部可达密度(local reachability density)。对象 p 的局部可达密度定义为对象 p 与它的 k-邻域点集的平均可达距离的倒数，记作 lrdk(p)，即

$$lrd_k(p) = 1 \Big/ \left(\frac{\sum_{o \in N_k(p)} reach - dist_k(p,o)}{|N_k(p)|} \right) \quad (3)$$

定义 9：对象 p 的局部孤立因子(local outlier factor，LOF)。对象 p 的局部孤立因子定义为 p 和它的 k 最近邻的局部可达密度的比率的平均值，记作 LOFk(p)，即

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \quad (4)$$

根据定义 7 和定义 8，数据对象 p 的局部可达密度越小，并且 p 的 k 最近邻居的局部可达性密度越大，LOF(p)越高。

局部离群因子的定义给出了离群点的本质描述，它有一如下性质：在簇内的对象的 LOF 值约等于 1，在簇边缘的对象的 LOF 值略大于 1，而离簇的距离越远，对象的 LOF 的值越大。因此，可通过计算 LOF(p)来判断数据对象 p 是否是局部离群点。

算法设置一个 LOF 阈值 μ 来判断离群点，如果对于数据对象 p 的局部离群点因子 LOF(p)，有 LOF(p)≥ μ，则 p 为离群点，否则 p 为非离群点。

发现增量数据中离群点的算法流程参见图 3。

```
算法：  GetIncreOutlier
输入：  增量数据集 D'，离群点阈值 μ
输出：  增量数据集中离群点
方法：
1. For 数据集 D'中的每个对象 p
2.IF p 未标记类标号
3.      计算的数据对象 p 的局部离群点因子
4.      If LOF(p)≥ μ
5.          将 p 标记为离群点；
6.      End If
7.End If
8. End For
```

图 3 增量数据中离群点发现流程

4 讨论

978-1-935068-18-1 © 2010 SciRes.

首先，探讨增量离群点识别算法的结果与非增量离群点识别算法的结果是否一致。IncreLOF 算法将 LOF 算法思想用于 DBSCAN 聚类算法，使对初始化数据集聚类的结果更为准确，并且在聚类的过程中，取得了各数据对象的 k-邻域点集，生成了各个噪声数据的 LOF 值。在增量聚类过程中，算法处理增量数据对象插入时影响的邻域对象，同时处理插入数据对象的 k-邻域点集。可以看出 IncreLOF 算法与 LOF 算法在原理和计算方法上是一致的。因此 IncreLOF 算法在理论上是正确的。

其次，探讨 IncreLOF 算法的时间复杂度。仅考虑算法最主要的步骤：

在处理历史数据中，找 k 最近邻居并计算 k-距离。计算每个对象之间的距离需要((N/2) ×(N-1))次的比较，时间复杂度为 O(N2) ，这一步的时间复杂度是 O(k×N2)；计算 k-邻域半径。计算每个点的 k-邻域半径要计算 k×N 次时间复杂度为 O(k×N)；计算 LOF 值至少需要 k×N 次的比较，其时间复杂度为 O(k×N)。总的复杂度为 O(k×N2)。

在增量数据聚类阶段，发现 k-距离邻域发生变化的点的算法复杂度为 O(N)；若 k-距离邻域发生变化的对象的个数为 m，计算 LOF 变化的点的算法复杂度为 O(m×N×k)，当 m、k<<N 时，总的算法复杂度为 O(k×N)。

在增量数据离群点识别过程中，IncreLOF 算法继承了已有的聚类成果，只对增量数据集中的“噪声”数据进行计算，鉴于离群点的本质决定了其数据量必然稀少，因此该过程的复杂度必然较低。

其中，k 是每个对象的最小邻居数，N 是数据大小。

5 实验与分析

首先，采用 UCI 数据集[7]中的 Yeast 数据集对算法进行实验验证。Yeast 数据集共含数据 1484 条，有 6 维属性，属性值类型为实型。实验将 Yeast 数据集的前 1084 条数据作为初始数据，将后 400 条数据分为 200 条和 200 条两部分作为增量数据进行实验，实验中 k=10。

IncreLOF 算法与 LOF 算法在每个增量数据集上执行时间如图 3 所示。IncreLOF 算法在处理增量数据时，只对增量数据进行处理，可以极大地节省时间，提高效率。

IncreLOF 算法实验结果与 LOF 算法实验结果对照如表 1 所示。可以看出 IncreLOF 算法所得到的离群点结果部出现在 LOF 算法的结果中，并且当阈值达到某一值时，两个算法结果一致。由此可见，IncreLOF 可以有效的监测出数据集中的离群点。

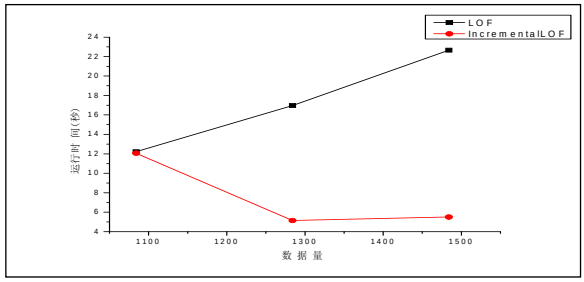

图 4 IncreLOF 与 LOF 效率比较

Table 1. Comparison of the experimental result of IncreLOF and LOF

表 1 IncreLOF 算法结果与 LOF 算法结果对照

离群点 算法 个数 阈值	IncreLOF	LOF	两者相同的离群点数量	相同条数占 LOF 算法结果百分比
2.0	9	9	9	100.0%
1.7	17	19	17	89.5%
1.5	53	58	53	91.3%

其次，选取具有一定规模的实际社会保险参保数据集作为实验数据。在采集的 A 省 B 市社保数据库中，取单位实缴信息表 AB16 中费款属期在 2006 年 1 月至 2006 年 12 月总计 12 个月的数据为实验数据，数据共计 12212 条。选取 2006 年 1 月至 2006 年 8 月总计 8 个月的数据 8796 条作为初始数据集，选取 2006 年 9 月数据 665 条、2006 年 10 月数据 812 条、2006 年 11 月数据 752 条和 2006 年 12 月数据 1187 条作为四个增量数据集进行增量式离群点识别。

IncreLOF 算法与 LOF 算法在每个增量数据集上执行时间如图 4 所示。从中可以清楚地看到：(1) IncreLOF 算法对于增量数据集上的离群点分析运行时间远远小于 LOF 算法(2) IncreLOF 的运行时间取决于每次增量数据集的数据量大小。

IncreLOF 算法实验结果与 LOF 算法实验结果对照如下如表 2 所示。由表可见 IncreLOF 与 LOF 算法在高阈值情况下，可识别出绝大部分相同的离群点。在低阈值情况下会出现一定差别，通过跟踪差别数据对象，发现造成差别主要原因是：LOF 算法使用的是对整个数据集进行离群点分析，而增量离群点识别算法 IncreLOF 的初始化数据集和各增量数据集都是整个实验数据集

的一部分，在对数据集分析的过程中，一部分在之前数据集上被判定为离群点的数据对象，随着新数据对象的不断插入，周围数据对象的不断增加，密度的不断变化，或是 LOF 值降低，低于所设定的阈值，或被包含进已有的某个聚类，成为已有聚类的数据对象，或被新生成的聚类包含，成为新的聚类的数据对象，从而成为非离群点。在 LOF 算法中，这部分数据对象被判定为离群点，在 IncreLOF 算法中，由于这部分数据对象被包含入聚类当中，这部分数据对象在现有数据集当中被判为非离群点。

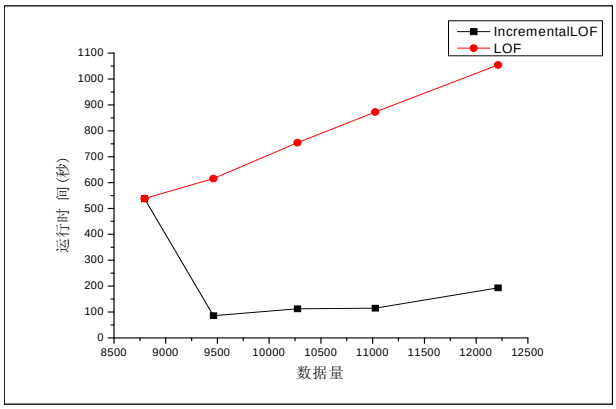

图 4 IncreLOF 与 LOF 效率比较

Table 2. Incremental Outlier Detection Result in Comparison with the result of LOF

表 2 增量离群点识别结果与 LOF 算法结果对照

离群点 算法 个数 阈值	IncreLOF	LOF	两者相同的离群点数量	相同条数占 LOF 算法结果百分比
400.0	27	27	27	100.0%
300.0	37	37	37	100.0%
200.0	40	39	39	97.5%
100.0	54	62	54	87.1%
50.0	69	83	69	83.1%

6 总结

本文分析基于局部密度离群点数据挖掘算法的思想和优点，结合聚类算法 DBSCAN 和离群点识别算法 LOF 提出一种增量式基于局部密度的离群点挖掘算法，并在实际的社保数据库上进行应用。证明了增量式离群点分析算法大大提高了数据分析的效率，其离群点分析的结果可以作为审计人员进行社会保险审计的疑点和依据，能够提高对社会保障审计的效率，弥补传统审计方法的一些重要缺欠，提高对社会保障基金管理的监管力度；拓展了数据挖掘技术的应用领域，对数据挖掘算法的研究成果有助于解决现有技术的一些缺欠，对相关领域有借鉴意义。

References (参考文献)

[1] Ma Shuai, Tang shiwei. An Incremental Clustering Algorithm for the Topology Adjustment of Location Databases. Journal of software，2004,15(9):1351-1360P.
马帅，唐世渭，杨冬青.一种用于位置数据库结构调整的增量聚类算法.软件学报，2004,15(9):1351-1360 页

[2] Hawkins D (1980). Identification of Outliers. Chapman and Hall, London. Reading, 1980.

[3] M M Breunig, H P Kriegel, R T Ng.LOF: identifying density-based local outliers. Proceedings of 2000 ACM SIG-MOD International Conference on Management of Data.2000: 93-104P

[4] Ester M, Kriegel HP, Sander J, Xu X. A density based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han JW, Fayyad UM, eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996.226-231P

[5] Ni Weiwei, Sun Zhihui, Lu Jieping. k-LDCHD—A Local Density Based k-Neighborhood Clustering Algorithm for High Dimensional Space. Journal of Computer Research and Development, 2005,42(5): 784-791.
倪巍伟，孙志挥，陆介平. k-LDCHD——高维空间 k 邻域局部密度聚类算法.计算机研究与发展,2005,42(5): 784-791.

[6] Ester M, Kriegel H-P, Sander J. Incremental Clustering for Mining in A Data Warehousing Environment. Proceedings of 24th International conference on Very Large Data Base, New York: Morgan Kaufmann Publishers Inc,1998. 323-333P

[7] Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.,1998.

[8] Chen Danping. Applying the Data mining Techniques in Modern Auditing. Journal of Nanjing Audit University, 2009 (2).
陈丹萍. 数据挖掘技术在现代审计中的运用研究. 南京审计学院学报. 2009 年 02 期.

[9] www.gao.gov/new.items, GAO-04-548 Data Mining: Federal Efforts Cover a Wide Range of Uses

[10] Han Jiawei, Micheline Kamber. Data Mining: Concepts and Techniques. China Machine Press, 2001.

978-1-935068-18-1 © 2010 SciRes.