

基于数据集对象平均离群因子的离群点选择算法

朱付保,徐显景,白庆春,朱颢东

(郑州轻工业学院 计算机与通信工程学院,河南 郑州 450002)

摘 要: 针对传统的局部离群点检测算法中存在离群点判定的主观性过强的问题,通过研究局部离群点检测相关算法,提出了一种基于数据集对象平均离群因子的离群点选择算法.该算法首先求得各数据对象的邻域对象,进而根据邻域数据集求出各自的离群因子,在进行离群点选择时计算出数据集的平均离群因子,将每个对象的离群因子与平均离群因子进行比较判断对象是否为离群点.理论分析和实验对比结果均表明,该算法在进行离群点选择时可以有效地避免离群点选择的主观性,更好地提高检测的准确率.

关键词: 离群点;局部偏离因子;加权属性;数据挖掘

中图分类号: TP315

文献标识码: A

文章编号: 1000—7180(2016)01—0131—04

Local Outlier Selection Algorithm Based on Average Outlier Factor

ZHU Fu-bao, XU Xian-jing, BAI Qing-chun, ZHU Hao-dong

(School of Computer and Communication Engineering, Zhengzhou University of
Light Industry, Zhengzhou 450002, China)

Abstract: Aiming at the shortage of subjectivity existed in the traditional local outlier detection algorithm, we studied the relevant algorithm of local outlier detection, and a local outlier selection algorithm based on the average outlier factor was proposed. This algorithm calculates the neighborhood of each data object, and figures out the outlier factor according to the dataset of its neighborhood. Comparing the outlier factor of each object with the average outlier factor to determine whether the object is an outlier. Theoretic analysis and experimental results show that this algorithm can avoid the subjectivity of selection and improve the accuracy of selection.

Key words: outlier; local deviation factor; weighted attribute; data mining

1 引言

离群点检测问题在数据挖掘领域作为一个重要课题而被广泛关注^[1],离群点检测的目的是从数据集中挖掘出异常数据,进而对异常数据的形成机理进行分析研究以获取有趣的知识,离群点检测算法在诸如网络入侵、商业欺诈和数据预处理等领域有着广泛的应用^[2].离群点检测算法种类多样^[3],总的可以分为基于聚类的、基于密度的、基于距离的、基

于分布的和基于深度的检测方法.

LOF(Local Outlier Factor)算法^[4]是基于密度的离群点检测算法中最具代表性的方法,该算法运用广泛^[5],它通过计算每个对象的局部离群因子 LOF 来判断对象的离群程度,进而从数据集中选取离群因子最大的前 N 个作为离群点对象输出.在局部离群点的检测算法中主要解决的问题是如何确定对象的局部邻域范围和对象相对于局部邻域的离群程度,文献^[6]中提出了改进的 SLDF 算法,该算法

收稿日期: 2015—03—26; 修回日期: 2015—05—06

基金项目: 国家自然科学基金项目(61201447);河南省科技攻关项目(122102210492);河南省教育厅科学技术研究重点项目(13A520368,13A520367);河南省高等学校青年骨干教师资助计划项目(2014GGJS-084);郑州轻工业学院校级青年骨干教师培养对象资助计划项目(XGGJS02)

在原有算法基础上改进了空间点对象的距离度量方式,用属性加权距离来衡量对象间的远近程度,这样更符合现实世界中同一对象的不同属性具有不同的重要程度这一事实,但是在该算法中将离群点的个数 M 作为参数事先设定,这样难免存在一定的主观性,甚至会致使算法检测结果存在一定程度的偏差.

本文通过研究离群点检测相关算法,对原有的离群点评定方式进行改进,不再以给定的离群点个数按照对象偏离因子从大到小选择离群点,而是将对象的局部偏离因子与整个数据集中对象的平均偏离因子进行对比,高于平均偏离因子的对象作为离群点输出,利用整体性来判断个体的离群性,这样可以减少离群点选择过程中的主观性.

2 改进的局部离群点选择算法

LOF 算法是基于密度的局部离群点检测算法,该算法计算出每个对象的局部偏离程度,依局部偏离程度的大小来判断对象的离群程度,而非一般的是与否的评判.偏离程度越大说明该对象的近邻分布越稀疏,则为离群点对象的可能性就越大;偏离程度越小说明该对象周围分布密集,则该对象为离群点的可能性就越小.在 LOF 算法及其改进算法中,通常根据离群程度从大到小选择指定数量的对象作为离群点,本文采用了一种有别于传统离群点选择的方法,将每个对象的偏离因子与数据集中所有对象的平均偏离因子进行对比,进而选择出非主观设定数量的离群点.该算法的有关定义如下所示.

定义 1 属性加权距离

设包含 N 个对象的数据对象集合 $O = \{O_1, O_2, \dots, O_N\}$, 每个对象 O_i 具有 M 个可度量特征属性,即 $O_i = \{O_{i1}, O_{i2}, \dots, O_{iM}\}$, 对象的 M 个可度量特征属性对应的权值 $W = \{w_1, w_1, \dots, w_M\}$, 其中 $w_m \in [0, 1]$ 对应的是对象的第 m 个特征属性的权值.

设 $O_i, O_j \in O$, w_m 是第 m 个特征属性对应的权值,则数据对象 O_i 和 O_j 的加权属性距离为

$$d(O_i, O_j, W_m) = \sqrt{\left| \sum_{m=1}^M w_m^2 (O_{im} - O_{jm})^2 \right|} \quad (1)$$

定义 2 对象 O_i 的邻居指的是在特定条件 C 下与对象 O_i 存在邻接关系 δ_c 的对象,即 $\forall O_i \in O, \exists O_j \in O \setminus \{O_i\}$, 对象 O_i 与 O_j 满足邻接关系 δ_c .

定义 3 对象 O_i 的 K -距离

对于任意正整数 K , 对象 O_i 的 K -距离 $K\text{-dist}(O_i)$ 为对象 O_i 与数据集 O 中某个对象 o 两者

之间的可度量距离,此时 O_i 需满足两个条件:(1) 至少存在 K 个数据对象 $o' \in O \setminus \{O_i\}$, 且 $d(O_i, o', W_m) \leq d(O_i, o, W_m)$; (2) 至多存在 $K-1$ 个数据对象 $o' \in O \setminus \{O_i\}$, 且 $d(O_i, o', W_m) < d(O_i, o, W_m)$.

定义 4 对象 O_i 的 K -距离邻域 $N_{K\text{-dist}(O_i)}$, 指的是在已知 O_i 的 K -距离 $K\text{-dist}(O_i)$ 的情况下, 所有与数据对象 O_i 的距离小于或等于 $K\text{-dist}(O_i)$ 的数据对象的集合, 即:

$$N_{K\text{-dist}}(O_i) = \{o' \mid d(O_i, o', W_m) \leq K\text{-dist}(O_i)\} \quad (2)$$

定义 5 对象 O_i 的局部偏离率 $S(O_i)$ 反映的是对象 O_i 相对于特定邻域内数据对象集合的偏离程度, 偏离率越大表明数据对象 O_i 的周围对象分布越稀疏, 则该数据对象成为离群点对象的可能性就越大; 偏离率越小则说明该数据对象周边数据分布越稠密, 则该点是数据集中离群点对象的可能性就越小. 对象 O_i 的局部偏离率定义为

$$S(O_i) = \frac{\sum_{o' \in N_{K\text{-dist}}(O_i)} (d(O_i, o', W_m) - \mu)^2}{|N_{K\text{-dist}}(O_i)|} \quad (3)$$

式中, $|N_{K\text{-dist}}(O_i)|$ 是数据对象 O_i 的 K -距离邻域内的数据对象的个数, μ 表示的是对象 O_i 到其邻域内其他数据对象加权距离的平均值, μ 的定义如下:

$$\mu = \frac{\sum_{o' \in N_{K\text{-dist}}(O_i)} d(O_i, o', W_m)}{|N_{K\text{-dist}}(O_i)|} \quad (4)$$

定义 6 对象 O_i 的局部偏离影响率 $S'(O_i)$ 从一定程度上反映出了其邻域内数据对象对其偏离程度的影响, $S'(O_i)$ 的定义如下:

$$S'(O_i) = \frac{\sum_{o' \in N_{K\text{-dist}}(O_i)} S(o')}{|N_{K\text{-dist}}(O_i)|} \quad (5)$$

式中, $|N_{K\text{-dist}}(O_i)|$ 表示的是给定数据对象 O_i 的 K 距离邻域中数据对象的个数, $S(o')$ 表示的是数据对象 O_i 的局部偏离率, 公式中分子部分表示的是对象 O_i 的邻域内所有数据对象的局部偏离率之和.

定义 7 对象 O_i 的局部偏离因子 $SLDF(O_i)$ 的公式定义如下:

$$SLDF(O_i) = \frac{S(O_i)}{S'(O_i)} \quad (6)$$

式中, $S(O_i)$ 是对象 O_i 的局部偏离率, $S'(O_i)$ 是对象 $S(O_i)$ 的局部偏离影响率, 局部偏离因子 $SLDF(O_i)$ 反映的是数据对象 O_i 的离群程度. 偏离因子值越大, 则说明该对象在整个数据空间中其周边数据分布越稀疏, 成为离群点对象的概率就越大; 偏离因子

的值越小,说明对象在整个数据空间中周边数据分布越稠密,则该对象成为离群点的概率就越小。

定义 8 数据集对象的平均偏离因子 $\overline{\text{SLDF}}$, 平均偏离因子描述的是数据集中对象偏离程度的平均水平,其定义如下:

$$\overline{\text{SLDF}} = \frac{\sum_{i=1}^{O.Length} \text{SLDF}(O_i)}{O.Length} \quad (7)$$

式中, $O.Length$ 是数据集合 O 中数据对象的数量, $\overline{\text{SLDF}}$ 表示的是数据集的整体偏离水平,本文算法以此为基准来衡量特定数据对象是否为离群点。

离群点的选择不再以预先设定的离群点数量作为离群点选择的标准,在该算法中通过式(6)计算出数据集合 O 中每一个数据对象 O_i 的局部偏离因子 $\text{SLDF}(O_i)$,进而计算出数据集对象的平均偏离因子 $\overline{\text{SLDF}}$ 。对于数据集中任意数据对象 O_i ,如果 $\text{SLDF}(O_i) > \overline{\text{SLDF}}$,则将该数据对象作为离群点对象输出。

3 算法描述

通过对算法相关概念的定义说明,改进的局部离群点选择算法描述如下。

(1) 初始化数据集合 $O = \{O_1, O_2, \dots, O_N\}$, 针对数据对象的 M 个可度量特征属性,根据领域专家意见分别赋予不同的权值 $W = \{w_1, w_1, \dots, w_M\}$, $w_m \in [0, 1]$ 对应的是对象的第 m 个特征属性的权值;初始化数据对象邻域参数值 K , K 为正整数;初始化一个 $N \times N$ 的二维数组 Arr ,用以存储对象两两之间的加权距离。

(2) 利用式(1),根据数据对象的可度量特征属性及各属性的权重,计算出任意两个数据对象之间的加权距离,并将计算结果存入二维数组中对应位置:

```
For i=1 To O.Length
  For j=1 To O.Length
    Arr[i-1][j-1]=d(Oi, Oj, wM);
  End For
End For
```

(3) 计算数据集中每一个对象的 K -距离并暂存于数组 $\text{KDistance}[N]$ 中;

```
For i=1 To N
```

根据定义 3 依次计算出每个对象的 $K\text{-dist}(O_i)$ 。

```
KDistance[i-1]= K-dist(Oi);
End For
```

(4) 计算出每个对象的 K -距离邻域;

```
For i=1 To N
```

```
  For j=1 To N
```

```
    If(Arr[i-1][j-1]≤KDistance[i])
```

```
      Add Oj To NK-dist(Oi)
```

```
    End For
```

```
  End For
```

(5) 在求得每个数据对象的 K -距离邻域的基础上,根据式(3)~(5)分别求取每个数据对象的局部偏离率 $S(O_i)$ 和局部偏离影响率 $S'(O_i)$,进而根据式(6)求出每个数据对象的局部偏离因子 $\text{SLDF}(O_i)$ 。

(6) 计算数据集对象的平均偏离因子 $\overline{\text{SLDF}}$ 。

```
For i=1 To O.Length
```

```
  Count+= SLDF(Oi);
```

```
End For
```

```
 $\overline{\text{SLDF}}$ =Count/O.Length;
```

(7) 将数据集中每个对象的局部偏离因子同平均偏离因子进行比较,将大于平均偏离因子的数据对象作为离群点输出。

```
For i=1 To O.Length
```

```
  If SLDF(Oi) >  $\overline{\text{SLDF}}$ 
```

```
    OutPut(Oi);
```

```
End For
```

4 实验结果与分析

为了验证本文算法的有效性,测试数据是从加州大学提供的机器学习数据库 UCI 上下载的三个数据集 Ecoli 数据集、Statlog (Heart) 数据集和 Mushroom 数据集,算法的性能以离群点检测结果的正确率为评价指标。

实验环境: Microsoft Windows XP 操作系统, IBM Pentium(R) 2.6 GHz, 2 GB 内存, 250 GB 硬盘, .NET Framework 4.0。

算法在 Visual Studio 2010 软件开发平台上使用 C# 编程语言实现,同时借助于 Microsoft SQL Server 2008 实现对测试数据集的管理。

实验 1: 采用 Ecoli 数据集将本文算法与 SLDF 算法、LOF 算法进行对比试验,该数据集的数据记录是 336 条,数据类别是 8 类,其中 cp 类包含 143 条数据记录,本文将类别为 cp 的数据作为正常数据记录,将类别为 om 和 oml 的 25 条数据记录作为离群点数据组成新的数据集进行测试。根据数据集的大小,设定邻域参数值 K 为 30。同时,由于 LOF 算法和 SLDF 算法都是根据各数据对象的离群程度

从大到小排序,然后选取前 M 个数据对象作为离群点输出,因此在本实验中分别设置参数 M 为 20、23、25、27、30 做五次实验验证. 实验对比结果如图 1 所示.

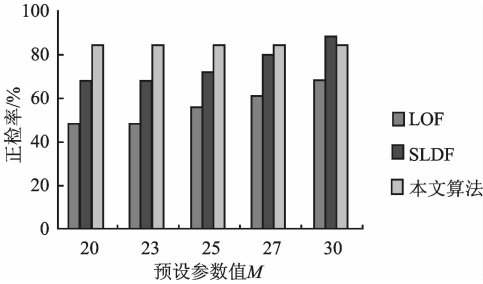


图 1 Ecoli 实验结果

实验 2:采用 Statlog(Heart)数据集进行对比实验,该数据集包含 270 条数据记录,选择其中的 150 条类别标识为 1 的作为正常数据,同时随机选择 20 条类别为 2 的数据作为离群点数据组成新的测试数据集. 根据测试数据集的大小设定邻域参数值 K 为 30. 同时,在本实验中分别设置参数 M 为 15、18、20、22、25 做 5 次实验验证. 实验对比结果如图 2 所示.

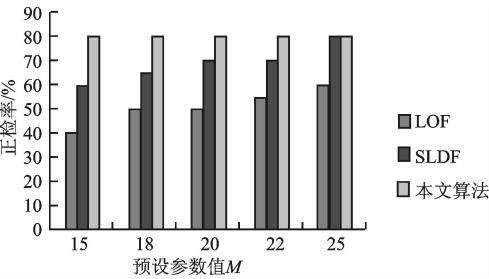


图 2 Statlog 实验结果

实验 3:采用 Mushroom 数据集来进行对比实验,该数据集包含 8 124 条数据记录,其中有一个类别属性用于标识 Mushroom 是否有毒,选择其中的 4 208 条可食用 Mushroom 作为正常数据记录,再随机选取 30 条有毒的 Mushroom 记录作为离群点数据. 根据测试数据集的大小设定邻域参数值 K 为 50. 同时,在本实验中分别设置参数 M 为 25、28、30、33 和 35 做 5 次实验验证. 实验对比结果如图 3 所示.

由以上实验对比结果可知,LOF 算法和 SLDF 算法受预设离群点参数 M 的影响比较明显,当预设的离群点数量不同时,这两种算法正确检测出的离群点数量也会随之发生变化,而本文算法由于采用的是基于平均偏离因子来选择离群点,所以避免了

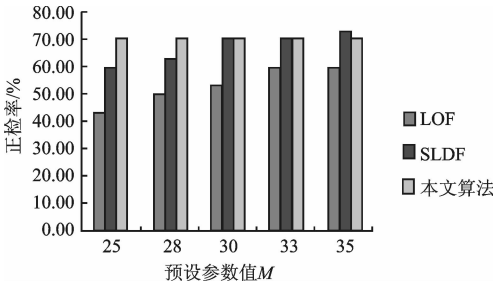


图 3 Mushroom 实验结果

预设离群点参数所带来的主观性影响,能够相对客观地根据整体的偏离因子来判断离群点.

5 结束语

针对局部离群点检测算法中存在对离群点选择的主观性问题,本文提出了一种改进的局部离群点选择算法,该算法对离群点的判别方式进行了改进,对特定数据对象的判断借助于整体数据对象的离群因子来进行,从而有效地避免了因主观设定离群点数量而造成的检测结果的主观性,及其引发的对数据对象离群点的误判. 将本文算法在实验数据集上与 LOF 算法和 SLDF 算法进行对比分析,结果表明该算法具有较好地对离群点数据的客观评判,对离群点检测具有更高的准确率.

参考文献:

[1] 刘欢,吴介军,苏锦旗. 基于分化距离的离群点检测算法[J]. 计算机应用研究,2010, 27(9):3316-3318.

[2] Bokyoung Kang, Dongsoo Kim, Suk-Ho Kang. Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction[J]. Expert Systems with Applications,2012,39 (5):6061-6068.

[3] 姜立明,柴瑞敏. 基于单元格和属性权重的离群点检测[J]. 计算机应用与软件,2012,29(10):216-218.

[4] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density based local outliers[C]// Proc of ACM SIGMOD Conf. New York: ACM, 2000: 427-438.

[5] Yuxin Ma, Hongbo Shi, Hehe Ma, et al. Dynamic process monitoring using adaptive local outlier factor [J]. Chemometrics and Intelligent Laboratory Systems, 2013,127(15):89-101.

[6] 张天佑,王小玲. 基于空间局部偏离因子的离群点检测算法[J]. 计算机工程,2011,37(14):282-284.

4 结束语

本文设计的全差分 10 bit 低功耗逐次逼近转换器芯片,在 0.18 μm CMOS 工艺下,所占用的面积为 0.6 mm^2 ,电源电压为 1.8 V. 本文采用的上极板采样结构新颖,能简化逻辑和降低功耗(可以简化逻辑和降低功耗),使速度也有进一步的提升. 比较器使用降低回馈噪声设计,对电容阵列的排布进行了优化从而有效提高匹配性. 后仿真结果有效位数达到 9.5 位. 本模数转换器的主要应用在对速度和精度要求不高,但却对于功耗有一定要求的电子仪器设备中.

参考文献:

[1] Liu ChunCheng, Chang SoonJyh, Huang GuanYing, et al. A 10-bit 50-MS/s SAR ADC with a monotonic capacitor switching procedure[J]. IEEE Journal of Solid-State Circuits, 2010,45(4):731-740

[2] Zhang Hui, Qin Yajie, Yang Siyu, et al. Design of an ultra-low power SAR ADC for biomedical applications [C]// Solid-State and Integrated Circuit Technology

(ICSICT), 2010 10th IEEE International Conference on. Shanghai:IEEE,2010:460- 462

[3] 模拟电路版图的艺术,张为译. 电子工业出版社,2007.

[4] Rouse Ball W W, Coxeter H S M. Mathematical Recreations and Essays[M]. Thirteenth Edition, [s. l.]: Dover Publications, 1987.

[5] Liu C C, Chang S J, Huang G Y, et al. A 0.92 mW 10-bit50-MS/s SAR ADC in 0.13 μm CMOS process [C]// IEEE Symp VLSI Circuits Dig. Kyoto, Japan: IEEE,2009: 236-237.

[6] Draxelmayr D. A 6 b 600 MHz 10 mW ADC array in digital 90 nm CMOS[C]// 2004 IEEE International Solid-state CVircuits Conference Digest of Technical Papers. [s. l.]:IEEE,2004: 264-265.

作者简介:

宁可庆 男,(1983-),博士研究生,实验师. 研究方向为大规模集成电路设计与测试

史立轺 男,(1990-),硕士研究生. 研究方向为模拟集成电路设计.

戴 澜 男,(1975-),副教授. 研究方向为大规模集成电路设计. E-mail:dailan@ncut.edu.cn.

(上接第 134 页)

作者简介:

朱付保 男,(1974-),博士,副教授,硕士生导师. 研究方向为智能信息处理、空间数据库、地理信息系统.

徐显景 男,(1988-),硕士研究生. 研究方向为智能信息处理、空间数据挖掘、地理信息系统.

白庆春 女,(1990-),硕士研究生. 研究方向为智能信息处理、数据挖掘.

朱颢东 男,(1980-),博士,副教授,硕士生导师,在站博士后,研究方向为智能信息处理、智能计算.

E-mail:zhuhaodong80@163.com