

基于聚类的离群点分析方法^{*}

邓玉洁, 朱庆生

(重庆大学 计算机学院, 重庆 400044)

摘 要: 对于离群点的形成,不同的属性起着不同的作用,离群点在不同的属性域中,会表现出不同的离群特性,在大多数情况下,高维数据空间中的对象是否离群往往取决于这些对象在低维空间中的投影。针对如何将离群点按照形成原因分类的问题,引入离群属性和离群簇等概念,以现有离群挖掘技术为基础,提出了基于离群分类来进行离群点分析的方法,实现了基于聚类的离群点分类算法 CBOC (cluster-based outlier classification),以揭示离群点的内涵知识。实验表明了该方法在实际应用中的有效性。

关键词: 离群分析; 离群分类; 离群属性; 离群簇; 内涵知识

中图分类号: TP311; TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)03-0865-04

doi:10.3969/j.issn.1001-3695.2012.03.018

Analysis algorithm for cluster-based outlier

DENG Yu-jie, ZHU Qing-sheng

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: For the formation of outliers, different attributes play different roles. Outliers in the different attribute space will show the different characteristics. In most cases, whether the high dimensional data is the outlier usually depends on the projection of these objects in low-dimensional space. In order to classify the origin of outliers, this paper defined some concepts such as outlier attributes and outlier cluster and proposed the approaches to outlier analysis by classified outliers based on the existing outlier mining technology. Furthermore, it gave an effective CBOC algorithm to provide outliers' intensional knowledge, whose validity in practical applications is finally verified by the experimental results.

Key words: outlier analysis; outlier classification; outlier attributes; outlier cluster; intensional knowledge

0 引言

数据挖掘是从大量数据中发现知识的方法,它一般分为四类:相关、依赖关系的知识发现;类别的判定;类别的描述;离群数据挖掘^[1]。离群数据挖掘是数据挖掘的基本任务之一。离群数据是指数据集中那些远离常规对象的数据,表现为与多数常规对象有明显差异,以至于被怀疑可能是由另外一种完全不同的机制产生^[2]。

离群点可能是在数据形成过程中出现的错误所导致,如人为错误、测量设备故障或存在噪声、过时数据、人工输入数据丢失等,这类数据一般为错误数据,可以剔除或以某种方法进行替换;但离群点也可能表明的一些并非错误的特殊情况,而数据变量固有变化引起,是自然发生的,反映了数据集的数据分布特征,如气候变化、顾客的新的购买模式、基因突变等。或是数据来源于异类,如信用卡欺诈检测、通信盗用、网络入侵检测、疾病诊断、天文观察乃至所有的科学研究领域。这类离群点可能会蕴涵着极为重要的信息,是数据分析的主要对象。显然,对这一类离群点应该进一步分析,以获取从常规模式不能得到的新知识。

离群点挖掘通常可以看做三个子问题:

a) 离群点的定义,即什么样的数据是离群点;

b) 离群点检测,即如何有效地从数据集中挖掘出离群对象形成离群数据集;

c) 离群点分析,即对挖掘出的离群数据进行解释与分析,合理解释离群挖掘的结果。

目前对离群点检测的研究十分活跃,已经有许多离群数据检测算法。离群点的检测方法大致可以把它们归纳为五类:基于统计分布的、基于深度的、基于距离的、基于密度的和基于聚类的^[3~5]。

相对于离群数据的检测技术而言,对离群数据集的解释与分析方面的研究则略显滞后。文献[6~10]在这方面进行了一些工作,但这些方法对离群数据的产生原因等方面尚不够深入。本文对离群点进行分类讨论,定义了平凡离群点、非平凡离群点和噪声等概念,提出了基于聚类的离群点分类算法 CBOC,并基于 UCI KDDCUP99 数据离群分析数据集实现了该算法。实验表明该算法能较好地按离群点的产生原因分类。

1 相关研究

1.1 离群特性及分类

到目前为止,离群点还没有一个正式的、为人们普遍接受的定义。根据离群点检测方法的不同,每种方法都给出了自己的离群点的定义。本文中所指的离群点是基于聚类的离群点,

收稿日期: 2011-08-12; **修回日期:** 2011-09-20 **基金项目:** 国家自然科学基金资助项目(61073058)

作者简介: 邓玉洁(1986-),女,重庆綦江人,硕士研究生,主要研究方向为数据挖掘(deng_senvy@gmail.com);朱庆生(1956-),男,安徽当涂人,教授,博导,博士,主要研究方向为数据挖掘、图像处理、数字农业。

即将不属于任何聚类的点定义为离群点。

在文献[6]中,Knorr 等人根据属性域子空间对离群数据特性进行了分析,把基于距离的离群点分为平凡离群点和非平凡离群点,其中非平凡离群点又分为强离群点和弱离群点。设 p 是属性域空间 A_p 上的离群点,若存在非空子空间 $B \subset A_p$,在 B 中 p 也是离群点,则称 p 为平凡离群点,否则称 p 为非平凡离群点;此时若 B 中无其他任何离群点,称 p 为强离群点;若 $L(A')$ 中存在另一离群点 p' ,且 $p' \neq p$,则称 p 为弱离群点。

1.2 关键域子空间

任何知识都是对象与属性的统一,对离群集的分析应以属性域为背景,探讨离群对象特性。Chen 等人^[7]发展了 Knorr 的思想,引入了离群模式的概念,将其作为离群内涵知识并提出相关的查找算法。该文把属性空间分割为标志子空间、指示子空间和观测子空间,以此建立了一种关于离群数据分类与特征相似性的模型。这些文献中重点讨论了非平凡离群点特性及其发现算法,但忽略了可能具有重要价值的平凡离群点。

连凤娜^[8]改进了 Knorr 提出的 FindNonTrivialouts 算法,认为造成离群点离群的是最小属性集,揭示离群点的内涵知识,并提出了一种基于关联空间的离群挖掘算法。即首先对数据集进行关联规则挖掘,然后在存在关联规则的属性空间(称之为关联空间)中检测强离群点和弱离群点。该算法剪除了无关联规则的属性子空间,降低了算法的计算量,但很难有效地确定关联规则挖掘的参数。

金义富等人^[10]将对数据离群具有重要贡献的属性集定义为离群对象的关键域子空间 KAS。即若对象 $O_i = (O_{i1}, O_{i2}, \dots, O_{id})$ 为 L^d 中的离群点,但其在子空间 L' 中不是离群点, $0 < r < d$,找出所有这种在 L' 中具有最大维数的子空间 L' ,称 $L^d - L'$ 为 O_i 的一个关键域子空间(KAS)。得到离群点的关键域子空间后,将具有相同关键域子空间的离群点定义成一个离群簇。文中重点讨论了根据离群点的关键域子空间进行离群簇分析及离群趋势分析,但在讨论关键域子空间搜索时,是通过检测离群点在所有属性子空间内的离群情况,从而试探性地确定离群对象的关键域子空间,若对 d 维数据集进行检测,则需要检测 $2^d - 2$ 次。随后金义富等人提出了基于统计和基于显著性子空间的关键属性域搜索算法,但是这些方法仍然不适用于高维数据集。

2 问题定义

本文中的离群点是基于聚类的离群点,即将不属于任何聚类的点定义为离群点。笔者认为,数据在某一维或者某个属性空间中不属于任何聚类,是造成该数据点离群的原因。所以找到离群点相对于每个聚类的离群属性,能有效地解释离群点的离群原因。

设四元组 (X, A, B, F) 表示一个离群检测与分析系统。其中: $X = \{x_1, x_2, \dots, x_n\}$ 为对象集, $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ 为第 i 个数据对象, n 为数据对象的规模, d 为属性的维数;属性集 $A = \{A_1, A_2, \dots, A_d\}$ 是一有界域集合; B 为决策属性集,值域为 $\{0, 1, \dots, k\}$,属性值为 0 表示对应的数据对象为离群数据类,为 $i(i \neq 0)$ 表示常规数据对象,并且该数据属于第 i 个聚类, k

为聚类的个数; F 为分类规则,表示一种基于聚类的离群检测算法。

定义 1 投影。设聚类 C_i 第 j 维的所有分量的最大值为 $D_{jmax} \in A_j$, 则对 $\prod x_i \in C_i$, 都有 $D_{jmax} \geq x_{ij}$; 同理, 设聚类 C_i 第 j 维分量的最小值 $D_{jmin} \in A_j$, 对 $\prod x_i \in C_i$, 都有 $D_{jmin} \leq x_{ij}$, 则聚类 C_i 第 j 维的值域长度 $P(C_i, j) = D_{jmax} - D_{jmin}$, 将这个值域长度 $P(C_i, j)$ 称为聚类 C_i 在第 j 维上的投影(projection)。

定义 2 离群属性。设 x 是离群点, 若 x 在第 j 维上的分量 x_j 不在聚类 C_i 在第 j 维上的投影 $P(C_i, j)$ 中, 则称第 j 维是 x 相对于聚类 C_i 的离群属性(outlier attribute), 将 x 相对于聚类 C_i 的所有离群属性记为 $OA(x, C_i) (j \in OA(x, C_i))$ 。

定义 3 非平凡离群点。设 x 是离群点, 若在第 j 维上, x 相对于所有聚类都离群, 即 $j \in OA(x, C_1) \cap OA(x, C_2) \cap \dots \cap OA(x, C_k)$, 则称 x 为平凡离群点; 否则称 x 为非平凡离群点。

定义 4 强离群点。设 x 是非平凡离群点, 若存在聚类 C_i , 对任意一维子空间 $j(j \in d)$, x 在 j 上的分量 x_j , 都有 $x_j \in P(C_i, j)$, 则称 x 为弱离群点; 否则称 x 为强离群点。

定义 5 噪声。设 x 是离群点, 若在任意一维子空间 $j(j \in d)$ 上, x 相对于所有聚类都离群, 则把 x 当做噪声。

以上概念可通过图 1 来解释说明。图 1 中椭圆标示聚类, “☆”表示的离群点为噪声, 因为该点在一维子空间中的投影都是离群的; “◇”表示的离群点为平凡离群点, 因为该点在水平方向上的投影是离群的, 而在竖直方向上的投影却不离群; “△”表示的离群点为弱离群点, 因为该点在对应于聚类 A 的两个一维子空间中投影均不离群; 而“◎”表示的离群点为强离群点。

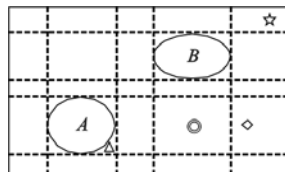


图1 不同类型的离群点

由图 1 可以看出, 不同离群点相对于聚类的投影具有不同的特征。从离群点相对于聚类的投影方面进行研究, 可以获得相当丰富的关于离群点的扩展知识。

定义 6 离群簇。设 x, y 是离群点, 若对任意的聚类 C_i 都有 $OA(x, C_i) = OA(y, C_i)$, 则认为这两个离群点属于同一个离群簇。

与 Knorr 在文献[6]中对离群点的分类进行比较, 本文对离群点的分类是与普通数据的聚类密切相关的, 并认为除噪声可能是错误数据之外, 大多数平凡离群点及所有非平凡离群点均蕴涵了知识。本文按离群点和聚类的关系对离群点进行分类, 通过不同的离群簇来进行离群点来源及含义分析, 可以不需要对问题领域的深入了解而获取扩展知识。

3 基于聚类的离群点分类算法

3.1 准备工作

基于聚类的离群点分类算法的方法是: 首先利用一种基于聚类的离群检测方法从给定的数据集中检测出全部离群对象和聚类, 然后判断单个离群对象在各个一维子空间内相对于各

个聚类投影的离群情况。全部检测完毕之后,再按定义对离群点进行分类。

在进行离群点的分类之前,需要进行以下准备工作:

a)数据准备。由于数据可能来自多个数据源,需要对复杂异构的数据进行必要的预处理等,以消除数据的不一致性和冗余,为离群挖掘提供规范集成的数据集。

b)离群挖掘。确定与数据集特性及规模相适应的离群数据挖掘算法。由于在进行离群点的分类时不仅需要离群数据,还需要使用普通数据的聚类结果,所以应该选用基于聚类的离群挖掘算法,以获得用于离群分析的离群数据集。

3.2 CBOC 算法描述

基于聚类的离群点分类算法(CBOC)分为离群属性搜索阶段、离群分类阶段和离群簇分析阶段。在离群属性搜索阶段,通过一遍扫描离群点来获得各个离群对象在各一维子空间内相对于各个聚类投影的离群情况;在离群分类阶段,利用属性搜索阶段得到的各个离群点的离群情况,按照定义将离群点分为噪声、平凡离群点、强离群点和弱离群点四类;在离群簇分析阶段,将平凡离群点、强离群点和弱离群点按定义分成各个离群簇,以便对各个离群簇进行分析。

3.2.1 离群属性搜索

在离群分类阶段需要用到各个离群对象在各一维子空间内相对于各个聚类投影的离群情况。聚类在一维子空间上的投影是由聚类中数据点在该维上的最大最小值决定,可以先将聚类向各个一维子空间投影,然后再判断离群点是否在这个投影范围内。

算法描述如下:

```
输入:聚类结果  $C = \{C_1, C_2, \dots, C_k\}$ , 属性集  $A = \{A_1, A_2, \dots, A_d\}$ , 离群对象  $O = \{O_1, O_2, \dots, O_m\}$ 。  
输出:离群属性集  $OA(O_i, C_j)$ , 其中  $O_i \in O, C_j \in C$ 。  
for  $i = 1, 2, \dots, k$   
  for  $j = 1, 2, \dots, d$   
    计算聚类  $C_i$  在属性维  $j$  上的投影  $P(C_i, j)$ , 即:  
     $P(C_i, j) = \max(x_j) - \min(x_j)$ , 其中  $x \in C_i$ ;  
  end for  
end for  
for  $i = 1, 2, \dots, m$   
  for  $j = 1, 2, \dots, k$   
    for  $h = 1, 2, \dots, d$   
      if 离群点  $O_i$  在属性维  $h$  上的分量  $O_{ih}$  不属于聚类  $C_j$  在属性维  $h$  上的投影  $P(C_j, h)$  范围内 then  
        将属性维  $h$  加入到集合  $OA(O_i, C_j)$  中;  
      end if  
    end for  
  end for  
end for
```

3.2.2 离群分类

得到各个离群点相对于各个聚类的离群属性后,根据离群属性将离群点分为噪声、平凡离群点、强离群点和弱离群点四类。由于本文将噪声认定为由于错误产生的数据,所以先将噪声从离群数据中去除。对于剩下的普通离群点,再按照定义进行分类。

算法描述如下:

```
输入:离群属性集  $OA(O_i, C_j)$ , 聚类结果  $C = \{C_1, C_2, \dots, C_k\}$ , 离群对象  $O = \{O_1, O_2, \dots, O_m\}$ 。  
输出:噪声集  $OutlierNoise$ , 平凡离群点集  $OutlierTrivial$ , 强离群点集  $OutlierStrong$ , 弱离群点集  $OutlierWeak$ 。  
for  $i = 1, 2, \dots, m$   
  count = 0;
```

```
  for  $j = 1, 2, \dots, k$   
    if 离群点  $O_i$  相对于聚类  $C_j$  的离群属性  $OA(O_i, C_j)$  为整个属性集  $A$  then  
      count = count + 1;  
    else  
      将离群点  $O_i$  加入到普通离群点集  $OutlierOrdinary$  中;  
    end if  
  end for  
  if count = d, 即离群点  $O_i$  相对于每个聚类的离群属性均为整个属性集  $A$  then  
    将离群点  $O_i$  加入到噪声集  $OutlierNoise$  中;  
  end if  
end for  
for each  $O_i$  in  $OutlierOrdinary$   
  for  $j = 1, 2, \dots, k$   
    if 离群点  $O_i$  相对于聚类  $C_j$  的离群属性  $OA(O_i, C_j)$  为空 then  
      将离群点  $O_i$  加入到弱离群点集  $OutlierWeak$  中;  
      break;  
    end if  
  end for  
end for  
if 离群点  $O_i$  相对于每个聚类都有相同的离群属性 then  
  将离群点  $O_i$  加入到平凡离群点集  $OutlierTrivial$  中;  
else  
  if 离群点  $O_i$  不是弱离群点 then  
    将离群点  $O_i$  加入到强离群点集  $OutlierStrong$  中;  
  end if  
end if  
end for
```

3.2.3 离群簇分析

离群分类完成后,普通离群点被分成了弱离群点、强离群点和平凡离群点三类。在进行离群簇划分时,本文将对同一聚类,具有相同离群属性的点划分到同一个簇。

以弱离群点为例:

```
输入:弱离群点集  $OutlierWeak$ 。  
输出:弱离群点簇  $OutlierWeakCluster$ 。  
将弱离群点集  $OutlierWeak$  中第一个离群点  $O_1$  加入到弱离群点簇  $OutlierWeakCluster$  中;  
for  $i = 2, 3, \dots, \text{length}(OutlierWeak)$   
  flag = 0;  
  for  $j = 1, 2, \dots, \text{length}(OutlierWeakCluster)$   
    if 离群点  $O_i$  相对于每个聚类的离群属性都与  $O_j$  相同 then  
      将离群点  $O_i$  加入到  $O_j$  所在的弱离群点簇中;  
      flag = 1;  
    end if  
  end for  
  if flag 等于 0, 即  $O_i$  不属于任何已有的弱离群点簇 then  
     $O_i$  形成一个新簇,加入到弱离群点簇  $OutlierWeakCluster$  中;  
  end if  
end for
```

CBOC 算法最后输出离群集 O 的不同分类的离群簇以及该簇对应的离群属性集 OA 。属于相同簇的离群点都是对相同的聚类有相同的离群属性。

4 算法分析及实验

4.1 算法效率

假设数据点个数用 n 表示,离群点个数用 m 表示,聚类个数用 k 表示,属性维数用 d 表示,弱离群点、强离群点和平凡离群点个数分别用 x 、 y 和 z 表示,并且 $x + y + z \leq m$ 。

在离群属性搜索阶段,首先一遍扫描数据集,获得各个聚类的一维投影,这一步需要时间为 $O(n \times d)$;然后计算各个离群点的离群属性集,需要时间为 $O(m \times k \times d)$ 。在离群分类阶段,首先去除噪声数据,这一步需要时间为 $O(m \times k)$;然后对普通离群点进行分类,循环次数最坏为 $(x + y + z)k$ 次,所以需要时间 $O((x + y + z)k)$ 。在离群簇分析阶段,分别对弱离群

点、强离群点和平凡离群点三类进行离群簇划分,需要时间分别为 $O(x^2)$ 、 $O(y^2)$ 、 $O(z^2)$ 。由上述可知,CBOC 算法总的最坏时间复杂度为 $O(n \times d + m \times k \times d + m^2)$ 。

CBOC 算法不需要对每个数据点都进行储存,只需要一遍扫描数据集,对聚类的一维投影进行储存。所以该算法存储空间要求主要在于对每个离群点的离群属性集的储存,需要的空间复杂度为 $O(m \times k \times d)$ 。

4.2 实验结果

本文的实验环境是 Core i5 CPU 2.40 GHz,2 GB 内存;操作系统为 Windows 7,在 MATLAB 环境下实现 CBOC 算法。

实验 1 所用数据集是由 MATLAB 软件生成的二维数据集,其数据分布如图 2 所示。

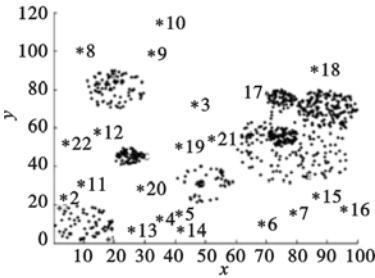


图2 二维数据集

对此数据子集首先进行基于聚类的离群点检测,得到 8 个聚类和 22 个离群点。对这 22 个离群对象进行了标记,并按 CBOC 算法进行了分类,结果如表 1 所示。

表 1 CBOC 对二维数据集分类结果

离群分类	离群簇
平凡离群点	{4,13,20},{8,16}
弱离群点	{1,2},{17}
强离群点	{3},{5,14},{6,7},{11},{12},{15},{18},{19},{21},{22}
噪声	{9,10}

CBOC 算法在对离群点进行分类时,是根据离群点相对于聚类的离群属性来划分的。由图 2 和表 1 可得,最后在同一离群簇中的离群点不但具有相同的离群属性,并且通常在空间分布中彼此邻近。但是由表 1 可以看出,虽然离群点 4、5 和 14 在空间上相距得非常近,但是它们的形成原因即离群属性却不同。离群点 4 相对于所有聚类都在 x 维上离群,即离群点 4 的 x 属性值不同于其他常规数据是其离群的主要原因;而离群点 5 和 14 在一维空间上均不离群,所以是其 x 属性值和 y 属性值共同决定了点 5 和 14 离群。

根据离群分类可以看出,弱离群点是相对于某个聚类离群的,由于弱离群点与该聚类非常接近,所以弱离群点可能是局部离群点,或者是在离群挖掘时将常规数据错误检测成了离群点。对弱离群点进行进一步讨论,有助于优化离群检测的效果。

实验 2 所用数据集是 UCI KDDCUP99 数据集的一个子集。UCI KDDCUP99 数据集是网络入侵数据,共有 41 个属性,其中数值属性 34 个,分类属性 7 个,包含的记录数为 4 898 431,而其中被标志为攻击(attack)的占 80.1%。本文从 corrected 数据集中除去 normal 数据和重复的数据后得到 29 346 条记录作为一个测试子集。对此数据子集首先进行了离群检测,共发现聚类 6 个、离群对象数为 517 个。

本文对这 517 个离群对象按 CBOC 算法进行了分类,得到噪声 0 个,强离群点 172 个,弱离群点 186 个,平凡离群点 159 个。通过与 UCI KDDCUP99 数据集中原已标志的攻击类型(attack_type)进行比较,结果如表 2 所示,总体正确率为 0.94,说明 CBOC 算法能较好地按离群点的产生原因分类。

表 2 CBOC 对 KDDCUP99 子集入侵分类结果

攻击类型	数据个数	强离群点 个数	弱离群点 个数	平凡离群点 个数	正确率
buffer_overflow	22	1	4	17	0.77
ipsweep	155	151	4	0	0.97
land	9	0	0	9	1.00
loadmodule	2	0	0	2	1.00
multihop	15	10	3	2	0.67
nmap	73	0	0	73	1.00
perl	2	0	0	2	1.00
pod	44	5	0	39	0.89
portsweep	173	4	165	4	0.95
rootkit	13	1	10	2	0.77
teardrop	9	0	0	9	1.00

5 结束语

本文根据离群属性对基于聚类的离群点进行分类并设计了分类算法 CBOC。该算法可以将相对于同一个聚类具有相同离群属性的离群点划分到同一个离群簇,反映了离群点的离群特性和产生原因。该算法只需一遍扫描数据,在实际数据挖掘过程中,可以根据用户对数据集问题领域的背景知识,去除一些对数据离群无显著贡献的属性,从而提高算法效率。实验结果证明了这种思路的正确性,为基于聚类的离群点分析问题提供了新的解决思路。

参考文献:

[1] 黄洪宇,林甲祥,陈崇成,等.离群数据挖掘综述[J].计算机应用研究,2006,23(8):8-13.

[2] HAWKINS D. Identification of outliers[M]. London: Chapman and Hall, 1980.

[3] 薛安荣,姚林,鞠时光,等.离群点挖掘方法综述[J].计算机科学,2008,35(11):13-27.

[4] 陈晓云,马良斋.基于属性权重的局部离群点挖掘算法研究[J].微计算机信息,2010,26(33):9-11.

[5] 项响琴,汪彩梅.基于聚类高维空间算法的离群数据挖掘技术研究[J].计算机技术与发展,2010,20(1):124-131.

[6] KNORR E M, NG R T. Finding intensional knowledge of distance-based outliers[C]//Proc of the 25th International Conference on Very Large Data Bases. New York:Morgan Kaufmann,1999:211-222.

[7] CHEN Zhi-xiang, TANG Jian, FU A W C. Modeling and efficient mining of intentional knowledge of outliers[C]//Proc of the 7th International Database Engineering and Applications Symposium. [S. l.]: IEEE Computer Society, 2003:1-10.

[8] 连凤娜.离群点挖掘及其内涵知识发现研究[D].厦门:厦门大学,2008.

[9] 金义富,朱庆生.一种离群数据集延伸知识发现框架[J].华南理工大学学报,2008,36(9):31-36.

[10] 金义富,杨俊杰.离群数据关键域子空间快速搜索技术[J].计算机工程与应用,2011,47(17):145-147.