一种基于属性的异常点检测算法*)

刘洪涛[1,2] 童德利[1] 陈世福[1]
(南京大学计算机软件新技术国家重点实验室 南京210093)[1]
(大庆油田有限责任公司储运销售分公司 大庆163159)[2]

摘 要 异常数据检测是数据挖掘研究的热点之一。本文在对现有异常点检测算法分析的基础上,提出了一种基于属性的异常点检测算法。简要地介绍了异常检测的现状,对基于属性的异常检测算法进行了详细分析,包括算法设计基础、算法描述、复杂度分析等。并通过与基于距离的异常点检测算法进行实验比较,表明了算法的优越性。

关键词 数据挖掘,异常数据,异常点检测

The Research of Algorithm of Attribute-Based Detection of Outlier Data

LIU Hong-Tao[1,2] TONG De-Li[1] CHEN Shi-Fu[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)[1]
(Storage, Transportation & Sales Sub-Company of the Daqing Oilfield Co. Ltd, Daqing 163159)[2]

Abstract Outlier data detection is an important part of data mining. It is a hotspot in data mining research. Based on the analysis of the exsiting algorithms of outlier data detection, this paper put forward a new outlier detection algorithm based on attribute. We introduce the status quo of outlier detection briefly, and analyze the algorithm of outlier detection based on attribute particularly. This paper shows the design basis of the new algorithm, the depiction of the new algorithm and the analysis of the complexity of the new algorithm and so on. Compared with another algorithm based on distance by experiment, the new algorithm has an obvious superiority in detection precision and time consumption.

Keywords Data mining, Outlier data, Outlier detection

1 前言

数据库中包含了大量的数据,由于各种原因,其中存在着一些噪声数据、特殊数据或者不完整的数据对象,它们与数据的一般行为或模型不一致,这些数据被称为异常数据,又称为异常点或离群数据(outlier)。Hawkins[1,2]给出了异常点的本质性的定义:异常点是在数据集中与众不同的数据,使人怀疑这些数据并非随机偏差,而是产生于完全不同的机制。异常数据可能来源于测量错误、计算机录入错误、执行错误、人为错误等。异常数据还有可能就是数据的真实性质的反映,这些数据比一般数据所包含的信息更有价值,这些数据更需要保留与研究。如通过对电信服务、信用卡服务中的异常数据的分析,可发现不正常的消费模式,这种分析的方法就是异常数据的检测。

目前比较成熟的异常点检测技术在应用中都有各自的优势,但同时也存在不足之处[3],例如基于统计的异常检测技术的优势是能根据数据分布函数确切地检测出异常数据,但是我们很难事先知道数据的分布特征;基于距离的算法与基于统计的算法相比,它更接近异常本质的定义,但是基于距离算法的参数很难设置;基于偏离的异常点检测算法计算性能较好,但它对现实复杂数据的效果不太理想;基于密度的异常检测算法能够检测出基于距离异常算法所不能识别的一类异常数据--局部异常,但它不能检测出全部的异常,它检测的只是局部异常;实际数据往往具有较大的噪声,因此异常模式通常只存在于低维子空间中,而在全维空间中难以确定,且以前算法在维数较高时,性能急剧下降。因此 Aggarwal 和 Yu[4]提出一个高维数据异常检测的方法,但是在高维数据中寻找异常模式却是非常困难的。

从现有的异常点检测技术可以看出,虽然它们对于解决异常点检测有很大的帮助,但是都或多或少地存在某些方面的不足,因此,基于距离的方法和基于密度的方法的融合,我们提出了基于属性的异常点检测算法,它解决了异常点检测中的一些实际问题,弥补了一些现有异常点检测算法的不足。基于属性的异常点检测技术不但检测的效果提高了,而且大大地简化了参数的设置,方便了用户的使用,也扩大了使用范围。

2 基于属性的异常点检测算法

2.1 算法的设计基础

我们可以对异常点数据挖掘进行如下描述[3]:已知 N 个数据对象,异常点挖掘的目标是从该 N 个数据对象中发现一部分与其余数据有明显不同的例外数据。基于属性的异常点检测技术将从以下三个方面进行研究:(1)数据集中异常数据的定义;(2)异常数据的挖掘方法;(3)检测到的异常数据的分离。

基于属性的异常点检测算法的基本思想是:按照数据对

*)刘洪涛 硕士生,主要从事工作流、数据挖掘等研究。童德利 硕士生,主要从事数据挖掘、机器学习、人工智能等研究。陈世福 教授,博士生导师,主要从事机器学习、人工智能等研究。

象属性逐个判断数据点是否是异常点。根据输入的预期异常点数目，利用距离函数 F 计算数据对象间的属性距离值 d，根据异常属性的定义检测并标记出数据对象的异常属性，根据数据对象属性的异常标记分离并输出异常数据。下面给出算法的一些定义：

引理1 数据集 T，N 为数据对象的数目，对象 o 为异常点是这样定义的：以数据对象 o 为邻域中心，以 d 为邻域半径内所包含的数据对象最大个数为 k，$k<<N$，k 为异常数据参数（人为设定），d 为半径参数（人为设定）。其中包含在 d 邻域内的数据对象 q 满足这样的要求：$q\in T$ 且 $F(o,q)<=d$，$F(o,q)$是对象 o 和对象 q 的距离函数。

引理2 数据集 T，N 为数据对象的数目，L 为对象的属性个数，对象 o 的 i 属性为异常属性是这样定义的：以对象 o 的属性 i 为中心，d_i 为邻域半径，该邻域内所包含的数据对象最大个数 k，$k<<N$ 且 k 为输入的异常属性参数。当对象 o 的 i 属性的 d_i 邻域所包含的数据对象数目大于 k 时，对象 o 的 i 属性就为非异常属性。其中包含在 d_i 邻域内的数据对象 q 满足这样的要求：$q\in T$ 且 $F_i(o,q)<=d_i$，$F_i(o,q)$为对象 q 的 i 属性和对象 o 的 i 属性的属性距离函数。对于邻域半径 d_i 是这样定义的：d_i 的数值等于数据集 T 中除去数据对象 o 的所有数据对象的 i 属性值的平均，它是由算法自动计算的一个参数。

引理2是基于属性的异常点检测技术的理论基础。

2.2 算法描述

为了对算法进行描述，我们首先给出如下两个定义：

定义1（数据异常信用度） 设数据集 T，M 定义为异常领域所包含的对象数，称 k 为该数据集的数据异常信用度。其中 k=异常领域对象数 M / 数据集 T 中的对象总数 N。

定义2（异常标记） 设数组 *Array*，数组的大小为数据集 T 中的对象数目，当对象 o 的 i 属性为异常时，就将数组 *Array* 的相应元素的值置为异常标记值，例如"0"代表正常，"1"代表异常。数组 *Array* 可以是一维，也可以是多维的。

基于属性的异常检测技术的算法描述如下：

算法：基于属性的异常点检测

```
输入：数据异常信用度参数 k 或 M(M=k * N)，数据集 T
输出：C 个异常点数据。
step1: for(i=0;i<=N;i++)
       outlier [i]=0;
       num=0;
step2: for ( i=0; i<=L; i++ )
step3: for ( j=0; j<=N; j++ )
        if ( outlier [j] = 1)
             continue;
        else {
             d_i=对象 j 的 i 属性值;
             tmpsum=0;
             for( l=0; l<=N; l++ )
               if ( l!=j ) tmpsum = tmpsum + 对象 l 的 i 属性值;
             d_i=tmpsum / (N -1) ;
           for ( l=0; l<=N; l++ ){
             d_li= 对象 l 的 i 属性值;
             tmp = d_i- d_li;
             if ( tmp >= - d_i AND tmp <= d_i) num ++;
                  if ( num <= M )
                    outlier [j] = 1;
                  else
                    break;
                  }
                 }
step4: for(i=0;i<=N;i++)
       if(outlier [i]= =1) 输出 T [i];
```

算法第一步对异常标记数组和异常数据计数变量进行初始化；第二步与第三步是算法的核心步骤，这两步实现了按照属性对数据对象逐个进行异常判断，其中第二步是数据对象的属性循环，遍历数据对象的属性；第三步用以遍历数据对象，从而对数据集中数据对象在指定属性上的异常情况进行检测，并根据检测结果对异常标记数组赋值；在算法的第四步，根据异常标记数组的值，输出异常数据。

在该算法第三步遍历数据对象时，首先判断该对象是否已被置为异常，若是，则跳过该对象，否则，判断该对象在指定属性上是否为异常。

算法中数据异常信用度参数由相关人员输入，可以输入具体的数目，也可以输入异常数据的百分比，这时算法根据数据异常信用度公式 $M=k*N$ 自动计算出 M 的值。

我们还可以定义异常标记数组为二维数组，以数据对象数和属性数为下标，用于记录每个数据对象的每个属性的异常情况，这样可以让我们十分清晰地了解到出现异常的对象的哪些属性不符合数据的一般模式，但同时也增加了算法的复杂性。

2.3 算法复杂度分析

基于属性的异常点检测算法的计算复杂度与数据集中的数据量、数据对象的属性个数、数据异常信用度等有关。下面我们根据以上参数来分析算法的复杂度。

假设数据集中的数据对象总数为 N（在算法中，此参数由系统自动计算），数据对象的属性个数为 L（在算法中，此参数也是由系统自动计算），数据异常信用度参数 k（该参数一般情况下由实验人员给定），我们分三种情况讨论算法的计算复杂度：正常情况，最优情况，最坏情况。

在正常情况下，基于属性的异常点检测算法的算法复杂度 $F(L,N,k)=L*(Q*(N+P))$，$1=<Q<=N$，$k=<P<<N$。在最坏的情况下，就是系统要为每个数据对象的每个属性都进行检测，计算邻域所包含的对象时也检测了所有的数据对象，这时 $Q=N$，$k=N$，此种情况下，算法复杂度 $F(L,N,k)=L*(N*(N+N))=2*L*N^2$。在最优情况下，系统只需要对数据对象的每个属性计算一次，这时 $Q=1$，$P=k$，此种情况下，算法复杂度 $F(L,N,k)=L*(N+P)$。

2.4 实验结果和分析

我们采用 Weka[5] 这个数据挖掘工具作为算法的实验环境。Weka 是一个开放的数据挖掘实验和应用系统，用于对数据挖掘算法进行训练和测试。其独特之处在于用户可以在该环境下训练和测试自己的算法。系统提供对数据集进行算法训练，创建交互挖掘的环境，用户可以动态地选择数据集的数据结构和数据挖掘算法，对挖掘的结果进行分析，也可以对结果进行可视化，即在多个抽象层对数据的不同维度进行可视化分析。

我们所用到的实验数据集来源于 UCI[6]，采用相同的数据对基于属性的异常检测算法与基于距离的算法进行了比较实验，实验结果如表1所示。

表中实验一采用 Ecoli 数据集，共有8个属性。数据总数为114例，其中8个数据为异常数据。实验二采用的是 wine 数据集。共有185个数据，其中有7个异常数据。

从实验结果我们看出，基于属性的异常检测算法的检测准确率较高，达到了百分之八十以上，而且检测时间也比较短，与基于距离的异常检测算法相比，准确率和检测时间都比较理想，并且对数据的顺序敏感性较低，而基于距离的异常点检测算法对数据的顺序敏感性较强。

结束语 异常点检测是数据挖掘研究的热点之一，它有着广泛的应用，如欺诈检测，用异常点检测来发现不寻常的信

用卡使用或者电信服务；预测市场动向，在市场分析中分析客户的流失等异常行为；或者在医疗分析中发现对多种治疗方式的不寻常的反应等等。通过对这些数据进行研究，发现不正常的行为和模式，实现异常数据挖掘功能。

表1 基于属性的异常点检测算法与文[7,8]中基于距离的异常点检测算法比较

实验	算法	参数设定	检测结果	消耗时间
实验一	基于距离	distance=0.25,p=30	检测出异常数据54个	42分钟
		distance=0.38,p=18	检测出异常数据2个	26分钟
		distance=0.35,p=20	检测出异常数据37个	38分钟
	基于属性	p=10	检测出异常数据8个，其中一个检测错误	2分钟
实验二	基于距离	distance=0.25,p=20	检测出异常数据6个	20分钟
		distance=0.25,p=30	检测出异常数据7个	38分钟
		distance=0.30,p=20	检测出异常数据3个	14分钟
	基于属性	p=10	检测出异常数据6个	1分钟

本文提出的基于属性的异常检测算法实现了在一个未知的数据集中进行异常点数据检测的功能。它通过分析数据对象的各个属性，对数据进行异常检测，然后利用异常标记数组对数据集进行数据分离，并进行输出。实验表明，基于属性的异常检测算法比现有的一些算法在执行时间、检测精度等方面具有明显的优势，目前我们在算法的准确率、易用性等方面正在进行进一步的研究。

参考文献

1 李炎，李皓，等. 异常检测算法分析. 计算机工程，2002，028(006)：5～6，32
2 李之棠，刘颉. 入侵检测中的模糊数据挖掘技术. 计算机工程与科学，2002，024(002)：18～21
3 Han Jiawei，Kamber M. Data Mining Concept and Technique . 北京：高等教育出版社，2001
4 魏藜，宫学庆，等. 高维空间中的离群点发现. 软件学报，2002，013(002)：280～290
5 Witten Ian H，Frank Eibe. Data mining：practical machine learning tools and techniques with Java implementations. Morgan Kaufmann，1999
6 Collections of datasets. http://www.cs.waikato.ac.nz/ml/weka/
7 Knorr E，Ng R. A unified notion of outliers：Properties and computation. In：Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining(KDD'97)，Newport Beach，CA，Aug. 1997. 219～222
8 Knorr E，Ng R. Algorithms for mining distance-based outliers in large datasets. In：Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)，New York，Aug . 1998. 392～403
9 Petrovskiy M I. Outlier Detection Algorithms in Data Mining Systems. Programming and Computing Software New York，2003，029(004)：228～237
10 Kollios G，Gunopulos D，Koudas N，Berchtold S. Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets. IEEE Transactions on Knowledge and Data Engineering，2003，015(5)：1170～1187
11 Arning A，Agrawal R，Raghavan P. A linear method for deviation detection in large database. In：Proc. 1996. Int Conf. Data Mining and Knowledge Discovery (KDD'96)，Portland，OR，Aug. 1996. 164～169
12 李翠平，李盛恩，王珊，杜小勇. 一种基于约束的多维数据异常点挖掘方法. 软件学报，2003，014(009)：1571～1577
13 黄守坤. 异常数据挖掘及在经济欺诈发现中的应用. 统计与决策，2003(004)：32～33
14 黄莹. 基于数据挖掘的异常检测模型. 电子工程师，2003，029(006)：11～13
15 孔学峰. 数据挖掘及其在信用卡风险控制中的应用. 中国金融电脑，2003(010)：21～22，33
16 宋世杰，胡华平，胡笑蕾，金士尧. 数据挖掘技术在网络型异常入侵检测系统中的应用. 计算机应用，2003，023(012)：20～23

(上接第163页)

参考文献

1 Martin J H. A Computational Model of Metaphor Interpretation. Boston，Academic Press，1990
2 Nehaniv C L. Computation for Metaphors，Analogy，and Agent. Springer，1999
3 Wilks Y. Making preferences more active. Artificial Intelligence，1978，11
4 Fass D. Collative Semantics：A Semantics for Natural Language：[PhD thesis]. New Maxico State University，CRL Report No. MCCS-88-118，1988
5 Carbonell J G，Minton S. Metaphor and Commonsense reasoning. In Hobbs & Moore，Formal Theories of the Commonsense World，Norwood，NJ：Ablex，1985. 405～426
6 Gentner D. Structure-mapping：A theoretical framework for analogy，Cognitive Science，1983，7：155～170
7 Holyoak K，Thagard P. Analogical mapping by constraint satisfaction. Cognitive Science，1989，13：295～355
8 Indurkhya B. Metaphor and Cognition：Studies in Cognitive Systems. Kluwer Academic Publishers，Dordrecht：The Netherlands，1992
9 Veale，Tony. Metaphor，Memory and Meaning：Symbolic and Connectionist Issues in Metaphor Interpretation. [PhD. Dissertation]. 1995
10 Kintsch W，Bowles A. Metaphor comprehension：What makes a metaphor difficult to understand? Metaphor and Symbol，2002，17：249～262
11 Mori T，Nakagawa H. A formalization of metaphor understanding in situation semantics. In：Barwise，J. et al. eds. Situation Theory and its Applications，vol. 2，CSLI Lecture Notes 26，Stanford，CA：CSLI Publications. 1991
12 van Dijk T A. Formal semantics of metaphorical discourse. Poetics 1975，4：173～198
13 Hall R P. Computational approaches to analogical reasoning：A comparative analysis. Artificial Intelligence，1989，39：39～120
14 Indurkhya B. Approximate semantic transference：A computational theory of metaphors and analogies. Cognitive Science，1987，11：445～480
15 Isabel D'Hanis. A logical approach to the analysis of metaphors. In：Magnani. L，ed. Logical and computational aspects of Model-based Reasoning，Kluwer Academic，Dordrecht，2002. 21～37
16 Steinhart E C. The Logic of Metaphor：Analogous Parts of Possible Worlds，Kluwer Academic Publishers，2001
17 周昌乐. 认知逻辑导论. 清华大学出版社，2001
18 周昌乐. 心脑计算. 北京：清华大学出版社，2002
19 胡壮麟. 认知隐喻学. 北京：北京大学出版社，2004