第34卷第2期　　中 国 科 学 技 术 大 学 学 报　　Vol. 34, No. 2
2004年4月　JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA　Apr. 2004

文章编号:0253-2778(2004)02-0213-06

一种基于离群数据挖掘的数据抽查新方法*

耿焕同[1,2],于　琨[1],洪　流[1],蔡庆生[1]

(1. 中国科学技术大学计算机系,合肥 230026; 2. 安徽师范大学计算机系,芜湖 241000)

摘要:针对传统数据抽查方法很难保证数据抽查有效性的缺点,结合离群数据挖掘,给出了一种基于离群数据挖掘的数据抽查新方法. 通过实验表明,该方法既能克服了随机数据抽查难以保证抽查有效性的缺陷又能克服重点数据抽查对抽查者经验的依赖,从而保证了数据抽查的有效性和全面性.

关键词:离群数据;数据抽查;NCL_CLARA 聚类

中图分类号:TP18　　**文献标识码**:A

0　引言

数据挖掘(Data Mining)是从大量的数据中发现一些不易发现的知识. 一般说来,它可以分成四类[1,3]:相关、依赖关系的发现,类别的判定,类别的描述,离群数据的挖掘(Outlier Mining). 目前,离群数据(Outlier)的挖掘越来越引起数据库、机器学习、统计学等学者的兴趣,正成为数据挖掘和应用领域的研究热点. 如:信用卡欺诈、电子商务犯罪、数据异常检测等. 离群数据(Outlier)[1]是明显偏离其他数据,不满足数据的一般模式或行为,与存在的其它数据不一致的数据. 离群数据通常来源于测量错误、计算机录入错误、人为错误等. 在离群数据挖掘中,关键是解决两个核心问题[2,3]:一是如何有效和准确的从大数据集中发现和搜寻离群数据;二是对发现的离群数据进一步判断是正常数据(异常)还是错误数据(噪音).

数据抽查(Data Spot-checking)是主要对入库后数据的抽样检查,判断数据库中的数据与入库前的原始数据是否一致. 例如:在高考阅卷中,先进行人工阅卷和机器阅卷,然后录入考生原始各门成绩入库,最后各考生总分合成. 在此过程中,可能会出现如登分录入成绩出错,因此需在总分合成前对登分录入的成绩进行大量的数据抽查(如某考生的三门课成绩高,而一门的成绩极低;或一门课中人工分高而机器分低等情况都属异常需抽查),以确保录入成绩的正确性. 常见的数据抽查方法有随机抽查和重点抽查. 其中随机抽查虽然是一种比较简单且易实现的数据抽查方法,但很难保证抽查的有效性;而重点抽查则需抽查者事先设定一些的抽查条件,对满足抽查条件的数据进行重点检验,显然很难保证抽查的全面性,往往取决于抽查者的经验.

* **收稿日期**:2003-08-27
基金项目:国家自然科学基金(70171052,90104030)
作者简介:耿焕同,男,1973年生,博士生. 主要研究领域:人工智能、知识发现. E-mail: sdght@163.net

本文提出了一种基于离群数据挖掘的数据抽查新方法. 先利用基于距离方法对待抽查数据库中的数据进行离群数据挖掘,然后利用挖掘出的离群数据作为数据抽查的对象. 此方法既克服了随机抽查难以保证抽查有效性的缺陷又能克服重点抽查对抽查者抽查经验的依赖,兼顾了抽查的有效性和全面性.

1 离群数据发现过程

有效地发现离群数据的关键在于建立探测离群数据的有效数学模型即如何把隐藏在正常数据中的离群数据分离出来. 常见的方法有三种[4]:基于统计学的方法,基于偏移的方法和基于距离的方法. 基于统计的方法是对给定的数据集合假设了一个分布或概率模型,然后根据模型采用不一致检验来确定离群数据;基于偏离的离群数据的方法是模仿人类的思维方式,在观察一个连续序列后,迅速发现其中一项数据与其他数据明显不同. 本文采用基于距离的离群数据发现方法. 下面给出该方法的实现过程,如图 1 所示.

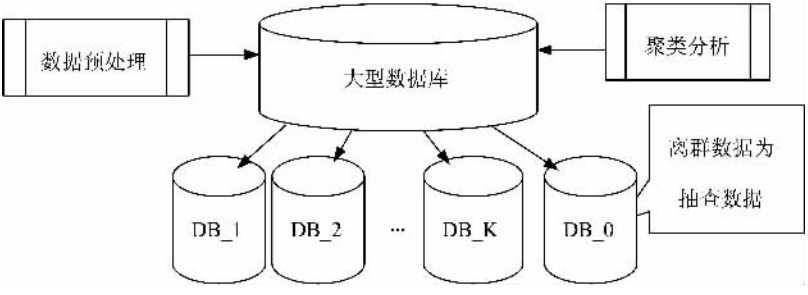

图 1 离群数据发现过程

Fig. 1 Outliers discovery process

具体发现过程是:首先、对大型数据库进行数据预处理,包括空值、数据规范化等处理;其次、对预处理后的数据库进行数据聚类分析,然后、生成多个聚类和一个特殊的类即所有离群数据组成的类. 可以看出基于距离的离群数据发现方法关键在于如何对待查数据库进行有效的聚类分析,使得 DB_1、DB_2、…、DB_K 聚类集中的数据为正常模式数据,而不在上述聚类中的出现的数据为离群数据.

2 聚类分析

聚类分析方法是统计学的一个分支和一种无教师监督的机器学习方法. 聚类分析方法是将多维空间 X 中给定一个有限的取样点集聚成多类,使得类间的相似性尽量少,而类内的相似性尽量大. 目前聚类分析主要集中在基于距离的聚类分析;典型代表有 K-means 方法和 K-medoid 方法等[4],这些方法的基本思想是是对分布在 m 维空间上的一组对象 $X_1,\cdots,X_n$ 进行分析,并给出聚类的个数 K,并任意构造 K 个初始聚类,然后通过不断的调整 K 个聚类的中心,重新修改各聚类的内容,使其相应的统计误差 $\sum_{j=1}^{K}\sum_{i=1}^{n} d(X_i,Q_j)$ 取得最小值时停止,则最终得到 K 个聚类(其中:Q_j 为聚类的中心,X_i 为 j 聚类中的元素,$d(X_i,Q_j)$ 为对象 X_i 与聚类中心对象 Q_j 之间的距离). 这些算法存在不足:一是用户事先要给出聚类数目 K,且 K 值的大小对聚类的结果十分敏感;二是不适合处理聚类对象较多的情况,该算法的时间复杂

度为 $O(K(n-K)^2)$,其中:n 为对象数. 因此为了更高效地处理大型的数据集合,本文提出了一种在大数据集合中聚类学习的新方法 NCL_CLARA,该方法结合一种将聚类学习的新方法 NCL[5]和基于选择的方法 CLARA 相结合的算法模型[6].

同时特别要注意的是,在聚类分析时,把不属于任何一个聚类的数据作为单独一类即离群数据. 因此需引入一个阀值 T(介于 0 到 1 之间),当某一数据与当前所有聚类的中心点的规格化距离都大于阀值(T), 则把它放入单独的一类(0 类)中.

3　离群数据发现的主要算法

3.1　数据相似度

根据数据记录的相似性判断数据是否为离群数据是基于距离聚类离群数据发现的关键,也是体现离群数据发现过程正确性和有效性的保证. 下面给出基于属性-值对的数据记录相似度的计算公式:设 A、B 为两数据记录,它们的相似度 $\mathrm{sim}(A,B)$ 可用下面公式度量:

$$\mathrm{sim}(A,B) = 1 - \sqrt{\sum_{i=1}^{n} w_i \times d(A_i,B_i)^2}.$$

其中:$\sqrt{\sum_{i=1}^{n} w_i \times d(A_i,B_i)^2}$ 为 A、B 间的欧氏距离,记 e_i 为数据记录的第 i 个属性字段;A_i、B_i 为数据记录 A、B 中属性的值,w_i 为第 i 个属性的权重,$d(A_i,B_i)$ 为两属性的距离(规格化),分几种情况讨论:

(Ⅰ)若该属性为数值型,则 $d(A_i,B_i) = \dfrac{|A_i - B_i|}{\mathrm{ValueRange}(e_i)}$;其中:定义 $\mathrm{ValueRange}(e_i) = \max(e_i) - \min(e_i)$ 为 e_i 属性的值域;

(Ⅱ)若该属性为符号型,符号相同时 $d(A_i,B_i) = 0$,不同时 $d(A_i,B_i) = 1$;

(Ⅲ)若该属性为模糊型数据[7],设该属性的模糊等级有 m 级,记 $\mathrm{grades} = \{g_1,\cdots,g_m\}$,且设模糊属性在各等级中的隶属度的值分别表示为:$\mathrm{gradevalue} = \{v_1,\cdots,v_m\}$;其中 v_i 为该属性属于等级 g_i 的隶属度的值,则 $d(A_i,B_i) = \dfrac{\sum_{j=1}^{m} |v_{jA_i} - v_{jB_i}|}{m}$,其中:$v_{jA_i}$、$v_{jB_i}$ 分别为数据记录 A、B 的 e_i 属性在该属性模糊等级 j 中的隶属度的值.

3.2　NCL_CLARA 聚类算法

聚类分析对象是待抽查大型数据库中的数据记录,本文按数据记录间的相似度进行聚类分析,把数据记录的每一个属性看成多维空间中的一维,每一维都有它的上下界,数据记录的所有属性形成了一个多维空间(进行规格化处理). 每一条数据记录是这个多维空间中的一个点,数据记录间的相似度反映了数据记录在多维空间中的距离,相似性越大,距离越短,属于同一聚类的可能性越大. 该 NCL_CLARA 算法结合了一种将聚类学习的新方法 NCL 和基于选择的方法 CLARA 相结合的算法模型的优点. 详细的 NCL、CLARA 算法描述参见文献[5,6],虽然 NCL 算法的时间复杂度基本上为 $O(nm)$, 但常数因子非常大,特别对数据集合非常大,该算法的实现的效率不是很好;因此结合基于选择的方法 CLARA 的优点,提出了 NCL_CLARA 方法. 其主要思想:不考虑整个待抽查数据集合,随机的选择待抽查数据集

中的一小部分作为数据集的样本,然后通过 NCL 聚类算法找出该待抽查数据子集中的各聚类中心点,再以各聚类中心点对原始待抽查数据库中所有的数据记录进行归类,重复上述过程多次,找出该 $\sum_{j=1}^{K}\sum_{p\in C_j} d(p,Q_j)$ 统计量为最小的作为聚类结果. 待抽查数据库中基于数据记录相似度的分区聚类算法 NCL_CLARA 可形式的描述三阶段为:

第一阶段:随机的从原始待抽查大型数据库中抽取一定比例的待抽查数据子集(SPer 为比例,一般取 5% 到 10% 之间);采用一种无监督学习的聚类算法 NCL 来找出该待抽查数据子集 SubsetData 中各聚类的中心点;

第二阶段:按上阶段找出的聚类中心点集对原始待抽查数据库中所有的数据记录实行归类,要求满足阈值(T), 否则单独作为离群数据(记为 0 类);

第三阶段:按聚类的结果求出除单独一类以外所有数据记录的平均相似度,要求越大越好,重复一定的次数 RNum(一般介于 5 和 10 之间);找出具有平均相似度最大的聚类结果作为分区的结果.

3.3 抽查数据库聚类分析及离群数据的生成

为了保证离群数据发现的有效性和规范性,在数据聚类分析前,需对数据库中的数据进行预处理,包括数据清理、数据集成和变换、数据规约等;然后对预处理过数据库进行聚类分析,本文采用在大数据集合中聚类学习的新方法 NCL_CLARA,当然可采用其他的聚类方法. 下面给出基于 NCL_CLARA 聚类算法分析的主要处理流程:

```
WHILE 重复次数 < RNum DO
    SubsetData = RandomSelectDataRecord( DB,SPer )
      //从待抽查数据库 DB 中随机选出一定比例的数据记录,作为子集
    Core[ ] = NCL( SubsetData )
      //调用 NCL 聚类算法求出待抽查数据子集的各中心点;
    FOR 待抽查数据库 DB 中的每一个数据记录 R DO //对数据记录进行归类
        I = CalculateMaxSameDegree( R,Core[ ],T )
          //将 R 与各中心点求相似度,返回相似度最大的聚类 I
          IF I 为空 THEN InsertCluster( R,0 ) //0 类为离群数据类
                    ELSE InsertCluster( R,I ) //将 R 放入聚类 I 中;
    AvgSameDegree = CalculateAvgSameDegree( DB,Core[ ]);
    //计算在聚类中所有数据记录的平均相似度,
Return ClusterResultSet = Max { AvgSameDegree }.
//返回平均相似度最大的作为聚类的分类结果
```

将上述聚类分析结果中 0 类作为离群数据库,并作为待抽查大型数据库中数据抽查的对象. 从算法可以看出基于离群数据挖掘的数据抽查新方法既克服了随机抽查难以保证抽查有效性的缺陷又能克服重点抽查对抽查者抽查经验的依赖性,兼顾了数据抽查的有效性和全面性. 而且此算法的时间复杂度为 $O(\mathrm{RNum} * K * n)$,其中:RNum 为重复次数,K 为聚类数,一般地,K 和 RNum 都远远的小于原始待抽查大型数据库的记录数 n.

4　实验结果及说明

为了进一步说明基于离群数据挖掘的数据抽查新方法的有效性,本文利用省秋季高考登分录入的成绩数据库,高考阅卷的基本过程如下:先进行人工阅卷和机器读卡阅卷,然后录入各考生各门课程原始成绩到数据库,最后合成各考生总分. 在此过程中,可能会出现阅卷登分和登分录入出错,因此需在总分合成前对录入的成绩进行大量的数据抽查(如某考生的三门课成绩高,而一门的成绩极低;或一门课中人工分高而机器分低等情况都属异常数据需抽查),以确保录入成绩的正确性和高考的公平性.

在实验时,选取了有 26 个属性,有 18 个整型、8 个符号型;库中有 22 万条数据,不考虑属性的权重差别即各属性的权重相同;同时人为的加入一些考生成绩、考场成绩错位的情形,以验证抽查的有效性;在 NCL_CLARA 聚类算法中,设定 SPer = 6%,RNum = 7;在 *K*_means聚类算法中,初始聚类数 $K = 100$,阈值(T)的大小反映数据间的差异大小,其选取参见表 1.

表 1　在不同阈值(T)下的要抽查数据的分布情况

Tab. 1　Needed SelectiveExamination Data Distributing of Different Threshold(T)

阈值(T)	0.2		0.4		0.6		0.8	
NCL_CLARA 和 *K*_means 聚类	NCL_ CLARA	*K*_means	NCL_ CLARA	*K*_means	NCL_ CLARA	*K*_means	NCL_ CLARA	*K*_means
同一考生各门成绩间	5 024	5 523	2 134	2 412	506	581	220	240
同一考生同一门主、客观成绩间	4 012	4 201	1 956	1 985	423	524	258	327
考生间错位	30	30	30	30	30	30	30	30
考场间错位	60	60	60	60	60	60	60	60
其他情形	1 300	1 327	978	934	476	454	212	257
合计	10 426	11 141	5 158	5 421	1 495	1 649	780	914

实验步骤如下:

(Ⅰ)先进行数据的预处理,主要是数据规格化操作;

(Ⅱ)分别采用 *K*-means 聚类和 NCL_CLARA 聚类的离群数据挖掘方法选取要抽查的数据;

(Ⅲ)根据抽查数据的反馈结果调整算法的参数,重复(Ⅱ)、(Ⅲ)步,直到未发现问题结束. 实验运行环境为:前台为 VC 开发的测试程序,后台用 MSSQL Server 数据库服务器来存放成绩数据库,800 MHz 的 CPU,256 M 内存和 Windows 2000 操作系统. 实验结果如表 1.

通过实验表明,利用基于距离的离群数据挖掘方法,能有效的、科学的、客观的从庞大的待抽查数据集中发现可疑的要抽查的数据;并且基于 NCL_CLARA 聚类的抽查数据量要小于 *K*-means 聚类的数据抽查量,这是因为 NCL_CLARA 聚类分析更符合待抽查数据集的分布情况. 从实验的结果来看,既能发现了同一考生各门成绩间或同一门的主客观成绩间的高低显著差别,也能发现考生间、考场间登分错位等异常数据. 有效地保证了普通高考的国考性、公正性和客观性.

5 结束语

本文提出了一种基于离群数据挖掘的数据抽查新方法,从实验结果可知,此方法很好既克服了随机抽查难以保证抽查有效性的缺陷又能克服重点抽查对抽查者抽查经验的依赖,兼顾了抽查的有效性和全面性. 为了进一步保证抽查的有效性,在实际应用中可以将三种方法结合起来. 与此同时,算法中经验参数的选取有待进一步的研究.

参 考 文 献

[1] Barnett V, Lewis T. Outliers in Statistical Data, 3rd ed.[M]. New York :John Wiley, , 1994.

[2] Edwin M Knorr, Raymond T Ng. Vladimir Tucakov. Distance-based outliers: algorithms and applications[J]. VLDB Journal, 2000,8(3-4): 237-253.

[3] Edwin M Knorr, Raymond T Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets[A]. VLDB[C], New York:1998, 392-403.

[4] 范明,孟小峰等译. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001,223-261.

[5] 朱明,王俊普. 一种聚类学习的新方法[J]. 模式识别与人工智能,2000,13(3):262-265

[6] Ng Raymond T, Han Jiawei. Efficient and effiective clustering methods for spatial data mining[A]. VLDB[C], Santiago, Chile:1994,144-155.

[7] Jeng B C, Liang T P. Fuzzy indexing and retrieval in case-based systems[J]. Expert Systems with Applications,1995,8(1):135-142.

A New Method of Data Spot-checking Based on Outliers Mining

GENG Huan-tong[1,2], YU Kun[1], HONG Liu[1], CAI Qing-sheng[1]

(1. *Department of Computer Science, USTC, Hefei* 230026, *China*)

(2. *Department of Computer Science, Anhui Normal University, Wuhu* 241000, *China*)

Abstract: A new method of data spot-checking based on outlier mining is proposed, which promises a solution to the lack of validity using traditional data spot-checking method. The experiments show that the new algorithm can overcome not only the random data spot-checking deficiency lacking in validity and but also the selective data spot-checking dependence on personal experience, thus ensuring the validity and completeness of data spot-checking.

Key words: outlier; data spot-checking; NCL_CLARA clustering