文章编号: 1673- 0062(2009) 04- 0070- 05

基于距离的异常成绩检测方法

李 萌, 阳小华

(南华大学 计算机科学与技术学院, 湖南 衡阳 421001)

摘 要: 数据质量对于学生成绩具有十分重要的意义. 本文将教育学原理与数据清洗技术相结合, 提出了基于距离的异常成绩检测方法. 在理论上论证了方法的合理性, 并通过实验验证了方法的有效性. 本文的工作不仅对于提高成绩管理系统的运行质量有直接的作用, 而且为将数据质量研究应用于教育信息化领域提供了很好的开端.

关键词: 数据质量; 数据清洗; 异常成绩检测

中图分类号: TP31 文献标识码: B

An Approach of Distance- based Detection to Grade- outliers

LI Meng YANG Xiao-hua

(School of Computer and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract Information quality has fundamental significance to student grade. Having combined the theory of pedagogy with data cleansing technique, we propose an approach of distance- based detection of grade- outliers in this paper. The rationality of the approach has been argued and its availability has been verified through our experiment. The results of this paper not only have direct effect on improving the quality of grade management system, but also provide a fine start to apply the study of information quality to educational informationization field.

Key words: information quality; data cleansing; detection of grade outliers

当今社会信息化程度不断提高, 各种信息系统广泛应用于各个领域, 积累了大量的数据. 为了使信息系统能够有效可靠地支持组织的工作运行, 要求系统的数据必须准确的反映现实世界的状况. 而在实际运行的系统中, 数据重复、数据缺失、数据不一致等问题普遍存在, 数据质量(Information Quality)问题日益突出, 引起了学术界和企业界的高度重视[1].

为了适应高等教育改革发展的需要, 教育信息化已经取得了很大的发展, 各个高校的教务管理基本上实现了信息化. 与其它领域类似, 在教务管理信息系统中也出现了许多的数据质量问题,

收稿日期: 2009- 06- 03
基金项目: 湖南省教育厅十五规划 2003基金资助项目 (XJK03CG021)
作者简介: 李 萌 (1975-), 男, 湖南衡阳人, 南华大学计算机科学与技术学院助教, 硕士. 主要研究方向: 教育信息化工程.

给学校的管理带来了不便. 由于成绩与学生的评优评先、升留级、毕业、学位等密切相关, 其数据质量尤其引人注目. 学生成绩的产生通常要经过评分、汇总、登分几个步骤. 教师阅卷评分时, 有时会出现误判; 成绩汇总时, 有时会发生计算错误; 在登分时, 有时会出现输入错误. 这样产生的不真实成绩, 会引起学生的疑问, 带来大量的成绩复核, 从而影响成绩的权威性, 给教学管理造成不良的影响.

孤立点产生原因很多[2], 主要有 3种. 1) 数据来自不同的类. 如: 使用信用卡欺诈的人群, 与合法人群相比, 属于完全不同的类; 2)数据固有变异; 3)数据度量与采集错误. 如: 人员操作失误、测量仪器的缺陷或故障导致部分数据成为孤立点. 因为人员操作失误是造成数据集中孤立点的主要原因之一, 因此不真实成绩极有可能成为成绩集合中的孤立点, 即异常成绩. 通过对异常成绩的检测及时发现不真实成绩, 从而改善学生成绩的数据质量.

本文尝试将异常数据检测技术应用于高校学生成绩管理, 以改善教务管理数据质量, 提高教务管理水平, 对于推进教育信息化建设有很好的理论意义和很强的实际价值.

1　异常数据检测

异常数据 (Outlier)又称为离群点、孤立点、野点等, 相关文献中多采用孤立点或离群点.

异常数据检测通常需要回答 3个问题[3]: 1) 何谓异常数据, 即孤立点定义; 2)有效检查异常的方法, 即算法的选择; 3)异常数据的意义, 即检测结果的合理解释. 用户真正感兴趣的是第 3点, 它也正是异常检测的目的.

异常数据检测始于 20世纪 80年代. 根据处理的数据对象所含维度的数量, 可将其分为低维数据异常检测、高维数据异常检测. 对于前者研究人员提出了基于分布、基于深度、基于距离、基于密度等多种异常检测方法. 学生成绩属于低维数据.

1.1　基于分布的方法

首先对给定的数据集假设一个分布或概率模型 (如正态或泊松分布), 然后对数据集中的每个点进行不一致检测, 如果与分布不符, 就认定其为孤立点.

该方法存在两个缺点, 1) 目前大多数不一致检测方法只能对单变量进行测试, 尽管有些方法可以对多变量进行测试, 但多变量分布往往是无法预知的; 2)需要对每个点进行测试, 代价很大.

1.2　基于深度的方法

将每个数据点映射到 k 维空间的一个点, 并由深度定义计算相应深度值, 其中深度值较小的数据点成为孤立点的可能性较大. 该方法在处理 4维及以上数据时, 效率低.

1.3　基于距离的方法

基于距离方法认为数据点与其邻居之间不应相距过远. 该方法存在多种孤立点定义, 其中常用的有:

1)*DB* 孤立点, 在包含有 N 个数据点的数据集 D 中, o是孤立点, 仅当 D 中至少有 pct部分对象与 o的距离大于最小距离 d. 换句话说, 如果 o 在 d 范围内的邻居不多于 $k = N * (1 - pct)$ 个, 则 o是一个带参数 pct和 d的 *DB* 孤立点.

实际上, 对于恰当定义的 pct和 d, 一个可以被给定的不一致检验测出的孤立点同样可以利用 *DB*(pct, d)检测出[4]. 同时, 它克服了基于分布的检测仅能检测单个属性的缺点.

2)*WK* 孤立点, 数据集中那些与其 k 个最近邻距离之和最大的 m 个对象.

在数据集 D 中, 数据点 o与其 k 个最近邻距离之和称为 o的权, 记为 *WK*(o), D 中的孤立点是 *WK*(o)最大的 m 个数据点. 显然, *WK*(o)度量了 o邻域的稀疏程度.

基于距离的孤立点定义包含并扩展了基于分布的思想[4], 尽管该方法无需事先知道数据集的分布, 并且理论上可以处理任意维的数据, 但仍存在以下缺陷, 1)算法参数的选择. 如: 不同聚类密度的数据集往往最小距离 d 差异很大, 因此很难事先确定; 2)它只能发现全局 (*global*)孤立点.

1.4　基于密度的方法

其孤立点定义是建立在距离基础之上, 它将点之间的距离与某给定范围内点的个数相结合, 构成密度概念[5]. 它不但可以发现全局孤立点, 而且能够检测局部 (*local*)孤立点. 同时, 前述几种孤立点定义都是二值的, 即一个点要么是孤立点、要么不是孤立点, 而基于密度的定义通过局部离群因子 (LOF)描述了数据点的孤立程度.

1.5　距离公式

度量距离的常用公式有: 绝对距离 (曼哈顿距离)、欧氏距离、兰氏距离等.

绝对距离 $d_{ij} = \sum_{k=1}^{m} | x_{ik} - x_{jk} |$;

欧氏距离 $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$;

兰氏距离 $d_{ij} = \frac{1}{m}\sum_{k=1}^{m}\frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$ $i, j = 1, \cdots, n$;

2 异常成绩检测

设学生 S的成绩集合为 C. 从成绩管理的角度,可将 C 划分为真实成绩和不真实成绩两个不相交的子集 C_{true} 和 C_{false}. 其中 C_{false}是指与实际成绩不符的成绩. 如:由于教师录入错误将实际成绩 68写成了 69,此时只要与实际成绩存在差异,无论差距大小,该成绩都是一个不真实成绩,属于 C_{false}. 从时间的角度,又可将 C 看作由 C_{old} 和 C_{new} 两个子集组成,其中 $C_{old} = \{c_{old_1}, c_{old_2}, \cdots, c_{old_n}\}$ 是经过复核的历史成绩集合, c_{old_1}是课程 $k1$的成绩; $C_{new} = \{c_{new_1}, c_{new_2}, \cdots, c_{new_m}\}$为新录入待检验的成绩集合. 同时,由于 C_{old} 中的成绩均已经过成绩管理部门和学生复核认可,因此认定为真实成绩,即 $C_{old} \subseteq C_{true}$,尽管其中可能存在不真实成绩. C_{new} 中可能存在的不真实成绩属于 C_{false}.

2.1 异常成绩定义

基于距离的异常成绩是成绩集合中那些与大部分成绩相距过远的成绩. 设 $\forall c_{new_1} \in C_{new}$, $\forall c_{old_j} \in C_{old}$, c_{new_1} 与 c_{old_j} 的距离为 d_{ij}.

1)*DB* 异常成绩,当 C_{old} 中 d_{ij} 大于 d的成绩个数不小于 $N^* \ pct$时 (N为 C_{old}成绩总数), c_{new_1}是一个 *DB*(pct, d)异常成绩,即 if C_{old}中 $d_{ij} > d$的个数 $N_i \geqslant N^* \ pct$, c_{new_1} 是 *DB* 异常成绩;

2)*WK* 异常成绩,将 c_{new_i} 放入 C_{old},构成 $C_{old_{wk}} = \{c_{old_1}, c_{old_2}, \cdots, c_{old_n}, c_{old_{n+1}}\}$, $c_{old_{n+1}} = c_{new_i}$, $C_{old_{wk}}$中成绩点 i的前 k个最近邻距离之和称为 i的权,记为 $wk(i)$, $wk(n+1)$是 c_{new_i}的权,如果权最大的前 m 个成绩点中包含了 c_{new_i},那么 c_{new_i}是一个 *WK* 异常成绩,即 if $\exists wk(n+1) \in \mathrm{Max}[wk(i)]_m$, c_{new_i}是 *WK* 异常成绩.

WK 异常成绩定义与 *WK* 孤立点定义有所不同,是因为:根据 *WK* 原始定义计算出的 m 个异常成绩中可能包含了 C_{old} 中的异常成绩,但由于之前已经认定了 $C_{old} \subseteq C_{true}$,并且我们仅对 C_{new} 中的异常成绩感兴趣,因此仅当 c_{new_1} 的权 $wk(n+1)$ 落入 $\mathrm{Max}[wk(i)]_m$ 集合时,才认定其为 *WK* 异常成绩.

2.2 参数选取

设 d_{old}是 C_{old}中任意两点距离,对于 *DB*,其最小距离 d存在以下 3种自然的取值:

1)乐观值,认为 C_{new} 中的成绩可信程度很高,如:教师认真负责,很少出现非真实成绩,此时 d取历史成绩间的最大距离, $d = \mathrm{Max}(d_{old})$;

2)悲观值,认为 C_{new} 中的成绩可信程度很低,如:某教师工作不负责,经常引起学生复核成绩,此时 d取历史成绩间的最小距离, $d = \mathrm{Min}(d_{old})$;

3)适中值,认为 C_{new} 中的成绩可信程度适中,如:尽管教师很负责,但难免视觉疲劳,偶尔出现成绩录入出错,此时 d取历史成绩间的平均距离, $d = \mathrm{Avg}(d_{old})$.

对于 *WK* 中最近邻位置 k和异常成绩数量 m,当取全部距离之和,即 $k = n$时, $w_i^k = w_i^n$,可避免用户设置最近邻位置 k[6],此时用户只需要设定合适的异常成绩数量 m.

3 实验

为了检验方法的有效性,选取了南华大学教务管理信息系统中某专业 2004级 244名学生的 32门课程共 8 052条成绩记录. 这些成绩都经过了学生确认,因此认定为真实成绩. 每名学生的成绩集合 C,从中取出一门课程成绩作为 C_{new},其余成绩构成 C_{old}. 根据实际经验,通常一门课程成绩中可能存在不超过 10%的异常成绩. 因此从这 244个 C_{new} 中随机选取 10%,用 $[-100, 100]$上服从均匀分布的随机数作为扰动,与真实成绩叠加,模拟人员操作失误造成的不真实成绩,使用第 2节中的异常成绩定义对其进行检测.

为了综合评价检测效果,引入查全率(Recall)、查准率(Precision)和函数 F 共 3项指标. 设提交的异常成绩集合为 A,不真实成绩集合为 B,则 $A \cap B$ 表示被检测出的不真实成绩;各集合中记录数分别为 N_A、N_B、$N_{A\cap B}$,定义如下:

1)查全率 R 为 $N_{A\cap B}$ 与 N_B 的比值,即 $R = \frac{N_{A\cap B}}{N_B}$,它描述识别出的不真实成绩占全部不真实成绩的比例, $R = 1$时,说明能检测出全部不真实成绩;

2)查准率 P 为 $N_{A\cap B}$ 与 N_A 的比值,即 $P = \frac{N_{A\cap B}}{N_A}$,它说明提交的异常成绩中,不真实成绩所占比例, $P = 1$时,表明检测结果全部为不真实成绩;

3)函数 F 是查全率、查准率的加权几何平均

值, $F = \frac{(b^2+1)PR}{b^2P+R}$, 这里 b表示 P和 R的相对权重, b大于 1表示 P更重要, 反之则 R更重要, 通常取 1, 表示两者同样重要.

因为原始分无法准确反映成绩个体在整体中的位置, 所以实验中采用了标准分. 标准分的计算采取正态化转换方案, 该方案同常见的标准分计算方案 $Z = \frac{x-X}{S}$ 相比, 避免了不同课程间最高分悬殊过大的问题.

根据第 2节异常成绩定义可知, 参数 d、pct、k、m 的设定将直接影响异常成绩的判断, 进而影响查全率、查准率、F 值. 因此实验中采用在某个范围内依次取值的方式进行参数值设定. 对于 DB 中 d, 除上述 3种取值外, 还采用 $[1, 100]$ 范围内步长为 1的步进值, 同样 pct取 $[0.1, 1]$ 范围内步长为 0.1的步进值; 因为 C_{old} 只有 32条成绩, 所以 k在 $[1, 31]$ 范围内以步长 1递增取值; 同样的, 由于 $C_{old_{wk}}$ 只有 32条成绩, 因此 m 在 $[1, 32]$ 范围内以步长 1递增取值.

需要注意的是, 实验数据来自现行系统, 肯定存在一定程度的数据质量问题, 因此在数据准备阶段, 对其进行数据清洗是很有必要的, 如: 重复记录消除、域约束冲突消除等. 实验结果如表 1所示, 其中各字段含义如下: Recall是查全率; Precision为查准率; Ratio在分布特征中为隶属度阈值, 在 DB 中指百分比 pct, WK 中表示最近邻位置 k; MinDistance在 DB 中为最小距离 d, WK 中表示异常成绩个数 m; A、B、C 表示 N_A、N_B、$N_{A\cap B}$.

从表 1可知, 当采用 DB 异常成绩定义, 并且最小距离取 7, 百分比取 0.9时, 查全率达到 0.8, 查准率达到 1, 综合评价 F为 0.88, 效果最好.

采用基于分布特征的方法[8]进行实验, 结果表明隶属度阈值取 0.1时, 最佳 F 为 0.80, 该方法检测效果略低于基于距离方法.

表 1　异常成绩检测结果

Table 1　The result of grade-outlier detection

model	F	Recall	Precision	A	B	C	Ratio	MinDistance
分布特征	0.807 0	0.920 0	0.718 7	32	25	23	0.100 0	
DB (乐观值)	0.529 4	0.359 9	1	9	25	9	0.100 0	
DB (悲观值)	0.185 8	1	0.102 4	244	25	25	0.300 0	
DB (平均值)	0.851 0	0.800 0	0.909 0	22	25	20	0.800 0	
DB (步进值)	0.888 8	0.800 0	1	20	25	20	0.900 0	7
WK	0.863 6	0.760 0	1	19	25	19	7	2

DB (平均值)的 F 值为 0.85, 与最佳 F 很接近, 这是因为注入的不真实成绩比例为 10%, 使得 C_{new}中的成绩可信程度适中, 因而 DB (平均值) 表现出了较好的检测效果.

实验结果中的查准率均获得了较高数值, 根据其定义可知, 异常成绩检测能够有效的发现不真实成绩. 检测出的异常成绩均与其历史成绩均存在较大出入, 这说明学生成绩是稳定的、渐变的. 也就是说, 上述检测方法所发现的是新录入成绩集合中那些相对历史成绩存在突变的成绩. 这也很好的回答了异常数据检测的第 3个问题.

基于分布特征的方法要求 C_{true}样本量不能太少. 实验中, 当 WK 在最近邻位置为 7时, 取得最佳检测结果. 换句话说, 只要历史成绩集合中有不小于 7个成绩对象时, 使用 WK 就能够有效的进行检测. 因此, 在实际应用中, 当 C_{true}样本量较少时, 应优先考虑基于距离的方法.

4　总结

异常成绩不一定是不真实成绩, 如: 成绩异常可能是由于作弊所取得的高分, 但其并非为不真实成绩; 同样, 不真实成绩也不一定就是异常成绩, 如: 成绩录入时, 教师将实际成绩 68误写为 69, 此时 69是不真实成绩, 但由于偏差很小并符合异常成绩定义.

本文将教育学原理与异常数据检测技术相结合, 提出了一种基于距离的异常成绩检测方法. 在理论上论证了方法的合理性, 并通过实验验证了方法的有效性. 本文的工作不仅对于提高成绩管理系统的运行质量有直接作用, 而且为将数据质量研究应用于教育信息化领域提供了很好的开端. 尽管如此, 但仍然有以下几个方面的工作值得进一步展开:

(下转第 78页)

10, 15, 20), 总注射容量毫升数为 L, 剩余注射容量毫升数为 l, 已滴注时间为 T_0, 则剩余滴注时间为 $T = \frac{\alpha l}{N} = \frac{\alpha L}{N} - T_0$

3)设前后两滴的时间间隔为 Δt_1, Δt_2, 又设标定的输液流速为每分钟 M 滴, 测得的实时输液流速为每分钟 N 滴, 如果 $N \geqslant M$ 且 $\Delta t_1 < \Delta t_2$ 或者 $N \leqslant M$ 且 $\Delta t_1 > \Delta t_2$ 则不予调节, 而是当 $N < M$ 且 $\Delta t_1 \geqslant \Delta t_2$ 时则加大输液流速, $N > M$ 且 $\Delta t_1 \leqslant \Delta t_2$ 时则减小输液流速, 实验表明, 这种差分智能调节算法能稳定快速地将速度调节到标定的输液流速范围内.

4 结束语

在终端控制器与主控微机相距 500 m 的环境, 挂带多个终端的情况下, 进行了临床实验, 实验结果表明: 在本地和远程监控两种情况下, 输液速度调节精度误差均为 ±1滴 /min, 系统性能稳定、安全可靠、操控直观方便. 该系统适合各种类型的医院临床使用, 具有较好的市场前景.

参考文献:

[1] 周海军, 赵 阳. 基于 RS-485的局域控制网络的构建[J]. 电子技术, 2001(7): 8-11.

[2] 张爱华, 张克平, 朱 亮. 电容式传感器用于检测医疗输液液位的研究[J]. 传感技术, 2004, 23(8): 9-11.

[3] 彭端云, 程自峰, 陈红波, 等. 基于超声波采集的无线数据传输的输液监控系统研制[J]. 医疗卫生装备, 2005, 26(8): 7-8

[4] 李云胜. 基于 VC 的液体点滴实时监控系统的设计[J]. 计算机应用, 2003, 23(12): 853-855.

[5] 刘 辉, 王新辉. 智能输液系统的研究与开发[J]. 湘潭师范学院学报(自然科学版), 2005, 27(2): 65-67.

[6] 梁远娣, 张晓霞, 赵引棉, 等. 静脉输液溶液毫升数与滴数换算关系再探[J]. 护理学杂志, 2004, 19(3): 46-47.

(上接第 73页)

1)基于分布方法要求 C_{old} 中成绩的数量 $N_{C_{old}}$ 不能太小, 实验中 *WK* 只需计算前 7个最近邻距离和就能较准确的发现异常成绩, 是否在 $N_{C_{old}}$ 较小时, *WK* 仍然比 *DB* 或基于分布方法更有效? 如何确定上述方法的 $N_{C_{old}}$ 临界阈值?

2)课程间是存在关联的, 如: 外语 1, 外语 2 成绩优秀的学生, 很少出现外语 3成绩不及格的情况, 但有可能数学成绩不佳. 因此实验中所使用的都是专业课程. 实际应用中, 如何确定课程间的关系?

3)当历史成绩集合为空的时候, 如何检测异常成绩?

参考文献:

[1] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述[J]. 计算机科学, 2008, 35(2): 1-5.

[2] TAN Pang-ning STEINBACHM, KUMAR V. Introduction to data-mining[M]. Boston: Pearson Addison-Wesley Education Inc, 2006.

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

[4] YufeiTao, Xiaokui Xiao, Shuigeng Zhou. Mining distance-based outliers from large databases in any metric space[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA: ACM, 2006: 394-403.

[5] Victoria Hodge, Jim Austin. A survey of outlier detection methodologies[J]. Artificial Intelligence Review, 2004, 22(2): 85-106.

[6] 李 强, 李振东. 数据挖掘中孤立点的分析研究在实践中应用[J]. 微计算机应用, 2006, 27(3): 323-327.

[7] 丁晓燕. 标准分制度在教学实践中的运用[J]. 考试周刊, 2008(38): 43.

[8] 黄光扬. 教育测量与评价[M]. 上海: 华东师范大学出版社, 2002: 141-164.