

基于分布特征的异常成绩检测方法

阳小华,李 萌

(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,提出了一种基于分布特征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 Grade – Outlier Detection Based on its Distribution Characteristics

YANG Xiao-hua, LI Meng

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

Abstract: Information Quality to Student grade has fundamental significance. Having combined the theory of pedagogy with data cleansing technique, we propose a detection approach of grade outliers based on distribution characteristics 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 have provided 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].

收稿日期:2008-09-30

基金项目:湖南省教育十五规划2003资助项目(XJK03CG021).

作者简介:阳小华(1963-),男,湖南衡阳人,南华大学计算机科学与技术学院教授,博士生导师. 主要研究方向:信息检索与系统管理、程序分析与建模、教育信息化工程等.

为了适用高等教育改革发展的需要,教育信息化已经取得了很大的成效,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,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.学生成绩若产生错误成绩,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.

文章尝试将数据质量提高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 数据质量与数据清洗

数据质量有不同的定义.从数据角度,数据质量是信息满足质量规范的程度^[2];从用户角度,数据质量是信息适合用户使用的程度^[3].尽管对数据质量有不同的理解,但一般认为数据质量是层次分类的概念,每个质量类都是由具体的数据质量维度构成的.

数据清洗(data cleansing, data scrubbing)是一种重要的数据质量提高技术^[1],研究如何检测并消除数据中的错误,主要包括重复对象检测、缺失数据处理、异常数据检测、逻辑错误检测、不一致数据处理等.重复对象检测就是要判断不同数据源中的表达是否代表现实世界的同一个实体,主要采用数据库和人工智能方法,包括关系数据库的重复记录检测和XML重复元素检测.缺失数据处理研究如何尽可能合理地填补缺失数据,主要集中在统计领域和数据库领域.逻辑错误是指数据中不符合特定应用领域业务逻辑的错误,通常根据应用依赖的领域知识建立规则体系来自动处理.多个数据源的数据集成时,各个数据源常常会提供相互重叠、不一致的数据内容,如何从不一致的数据中获得理想的数据答案是数据清洗中经常面临的问题.不一致数据通常采用排序、融合和根据规则进行处理.

数据异常有两种原因:一是数据固有的变异性,一是度量或者执行错误.异常数据检测主要采用数据审计方法(也称为数据质量挖掘, Data Quality Mining, DQM)^[4],其主要步骤有两个:首先是数据概化(Data Profiling),即采用数理统计方法对数据分布进行概化描述,以获得数据的总体分布特征;然后针对某个特定的数据质量问题

进行挖掘以发现异常.

2 异常成绩检测

设学生集合为 $S = \{s_1, s_2, \dots, s_k, \dots, s_m\}$, 课程集合为 $K = \{k_1, \dots, k_n, \dots, k_{(n+p)}\}$, 学生 s_k 当前全部成绩集合为 $C_{s_k}, C_{s_k} = \{c_{s_{k1}}, c_{s_{k2}}, \dots, c_{s_{kn}}, c_{s_{k(n+1)}}, \dots, c_{s_{k(n+p)}}\}$, 其中 $C_{s_{kn}}$ 是学生 s_k 所修课程 k_n 的成绩; C_{s_k} 由两个集合 $C_{s_{kold}}$ 和 $C_{s_{knew}}$ 组成, 其中 $C_{s_{kold}} = \{c_{s_{k1}}, c_{s_{k2}}, \dots, c_{s_{kn}}\}$ 为已经过复核的成绩集合, $C_{s_{knew}} = \{c_{s_{k(n+1)}}, \dots, c_{s_{k(n+i)}}, \dots, c_{s_{k(n+p)}}\}$ 是新录入待检验的成绩集合.

学生 s_k 的成绩集合 C_{s_k} 可划分为真实成绩和非真实成绩两个不相交的模糊子集 $C_{s_{ktrue}}$ 和 $C_{s_{kfalse}}$, 对于任意的 $x \in C_{s_k}$, $\mu_A(x)$ 和 $\mu_B(x)$ 分别表示 x 对 $C_{s_{ktrue}}$ 和 $C_{s_{kfalse}}$ 隶属度, $0 \leq \mu_A(x) \leq 1, 0 \leq \mu_B(x) \leq 1, \mu_A(x) + \mu_B(x) = 1$. 由于 $C_{s_{kold}}$ 中的成绩都已经过了成绩复核, 可以认定为真实成绩, 因此对于任意的 $x \in C_{s_{kold}}$, 有 $\mu_A(x) = 1$. 而对于任意的 $x \in C_{s_{knew}}$, 我们需要根据隶属度判断它是属于真实成绩集合 $C_{s_{ktrue}}$ 还是属于非真实成绩集合 $C_{s_{kfalse}}$. 当 $\mu_A(x)$ 大于设定的阈值时, 认为 x 是正常成绩; 否则, 认为是异常成绩.

根据教育学原理, 一个学生的不同课程成绩是独立同分布的^[5]. 因为不同课程的成绩来自不同的试卷, 所以彼此独立; 又因为都是对同一名学生的教育评价, 所以不同课程服从同一种分布, 因此学生的课程成绩是独立同分布随机变量. 由于成绩只在一个有限的范围取值, 因此具有有限的期望和方差, 根据中心极限定律, 其概率分布服从正态分布, 所以真实成绩集合的隶属函数 $\mu_A(x)$

$$= \exp\left[-\left(\frac{x - \bar{X}}{S}\right)^2\right], \bar{X} \text{ 是 } C_{s_{ktrue}} \text{ 中成绩的均值, } \bar{X} = \frac{\sum_{i=1}^n C_{s_{ki}}}{n}, S \text{ 为 } C_{s_{ktrue}} \text{ 中成绩的标准方差, } S = \sqrt{\frac{\sum_{i=1}^n (c_{s_{ki}} - \bar{X})^2}{n}}, \text{ 如图 1 所示.}$$

为了评价异常成绩检测的效果, 引入查全率、查准率和 F 值 3 项指标. 设检测出的异常成绩集合为 A , 实际非真实成绩集合为 B , 则 $A \cap B$ 表示检测异常成绩中非真实成绩的部分, 而 $A - A \cap B$ 则表明检测出来的异常成绩中含有误报的部分; 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 $R = \frac{N_{A \cap B}}{N_B}$, 它描述检测方法发现非真实成绩的范围; $R = 1$ 时, 说明能检测出系统中全部非真实成绩;

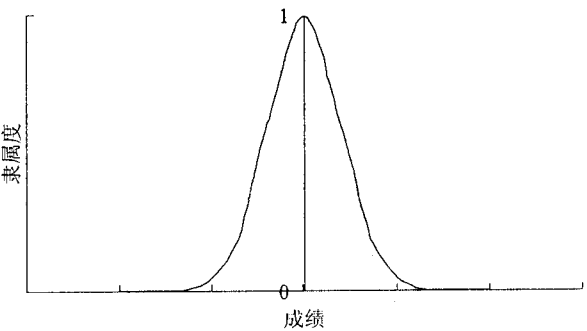


图 1 成绩隶属函数
Fig. 1 Membership function of scores

2) 查准率 P 为 $N_{A \cap B}$ 与 N_A 的比值, 即 $P = \frac{N_{A \cap B}}{N_A}$, 它描述提交的异常成绩中非真实成绩比率, $P = 1$ 时, 说明发现的异常成绩全部为非真实成绩;

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, 小于 1 则 R 更重要, 通常取 1, 表示两者同样重要.

3 实验

为了检验方法的实际效果, 我们选取了南华大学教务管理信息系统中某专业 603 名学生的 150 门专业课程成绩数据, 共 16 997 条记录; 这些成绩都经过了学生确认, 因此可以认为是真实成绩; 根据实际经验, 通常一门课程成绩中可能不存在不超过 10% 的非真实成绩. 所以, 我们选定一门课程, 随机从中选取 10% 的记录, 用 $(0, 100]$ 上均匀分布的随机数作为扰动, 与真实成绩合成构造非真实成绩.

不同课程试卷难度变化可能会造成课程成绩的波动, 此时原始分可能已无法准确描述学生实际水平, 所以实验中不但采用原始分, 而且使用 T 标准分, 后者反映成绩个体相对于团体的位置信息; 阈值的选取直接影响对成绩是否异常的判断, 从而影响查全率、查准率和 F 值. 实验结果如表 1 所示. 表 1 中各字段含义如下: $DataType$ 值 A 、 B 分别表示原始分、标准分; $Ratio$ 表示阈值; A 、 B 、 C 表示 N_A 、 N_B 、 $N_{A \cap B}$; $RecallRatio$ 表示查全率; $PrecisionRatio$ 表示查准率, F 表示综合评价函数 F .

表 1 实验结果
Table 1 Experimental result

DataType	Ratio	A	B	C	RecallRatio	PrecisionRatio	F
A	0.95	103	20	20	1	0.194 174 752	0.325 203
A	0.9	94	20	20	1	0.212 765 962	0.350 877
A	0.85	86	20	20	1	0.232 558 146	0.377 358
A	0.8	83	20	19	0.949 999 988	0.228 915 662	0.368 932
B	0.95	86	20	19	0.949 999 988	0.220 930 234	0.358 491
B	0.9	75	20	19	0.949 999 988	0.253 333 33	0.4
B	0.85	64	20	19	0.949 999 988	0.296 875	0.452 381
B	0.8	54	20	17	0.850 000 024	0.314 814 806	0.459 459
B	0.75	45	20	16	0.800 000 012	0.355 555 564	0.492 308
B	0.7	38	20	16	0.800 000 012	0.421 052 635	0.551 724
B	0.65	30	20	16	0.800 000 012	0.533 333 361	0.64
B	0.6	26	20	16	0.800 000 012	0.615 384 638	0.695 652
B	0.55	23	20	15	0.75	0.652 173 936	0.697 674
B	0.5	21	20	15	0.75	0.714 285 731	0.731 707
B	0.45	19	20	14	0.699 999 988	0.736 842 096	0.717 949
B	0.4	19	20	14	0.699 999 988	0.736 842 096	0.717 949
B	0.35	16	20	11	0.550 000 012	0.687 5	0.611 111
B	0.3	14	20	9	0.449 999 988	0.642 857 134	0.529 412

技术条件,选取两种典型回采方案进行回采过程模拟,分别对比分析了两种方案回采过程中的应力、位移及塑性区分布规律,优选方案二为矿山最终回采顺序。

参考文献:

- [1] 施建俊,孟海利. 采场结构参数与回采顺序的数值模拟优化研究[J]. 有色金属(矿山部分),2005,57(2): 9-12.
- [2] 杨承祥. 深井金属矿床高效开采及地压监控技术研究[D]. 长沙:中南大学,2007.
- [3] 曹明,吕广忠,王志国. 金厂峪地下采场回采顺序弹塑性数值模拟分析[J]. 金属矿山,2004(12):23-26.
- [4] 杨典森,陈卫忠,杨为民,等. 龙滩地下洞室群围岩稳定性分析[J]. 岩土力学,2004,259(3):391-395.
- [5] LUO Zhouquan, LIU Xiaoming, SU Jiahong, et al. Deposit 3D modeling and application[J]. Journal of Central South University of Technology, 2007, 14(2): 225-229.
- [6] Gong Jian - ya, Cheng Peng - gen, Wang Dong - yang. Three dimensional modeling and application in geological exploration engineering[J]. Computers&Geosciences, 2004,30(4):391-404.
- [7] Houlding S W. Practical geostatistics, modeling and spa-

tial analysis[M]. New York and Heidelberg:Springer - Verlag,2000.

- [8] 罗周全,刘晓明,吴亚斌,等. 地质统计学在多金属矿床储量计算中的应用研究[J]. 地质与勘探,2007,43(3):83-87.
- [9] Lemon A M, Jones N L. Building solid models from boreholes and user - defined cross - sections[J]. Computers & Geosciences, 2003, 29(3): 547-555.
- [10] 罗周全,刘晓明,吴亚斌,等. 基于 Surpac 和 Phase - 2 耦合的采空区稳定性模拟分析[J]. 辽宁工程技术大学学报(自然科学版),2008,27(4):485-488.
- [11] 王纯祥,白世伟. 三维地层信息系统与有限元方法集成研究[J]. 岩石力学与工程学报,2004,23(21): 3695-3699.
- [12] 罗周全,吴亚斌,刘晓明,等. 基于 Surpac 的复杂地质体 FLAC^{3D}模型生成技术[J]. 岩土力学,2008,29(5):1334-1338.
- [13] Hoek, Brown E T. Practical Estimates of Rock Mass Strength[J]. International Journal of Rock Mechanics of Mining Science,1997,34(8):1165-1186.
- [14] 朱维申,李晓静,郭彦双,等. 地下大型洞室群稳定性的系统性研究[J]. 岩石力学与工程学报,2004,23(10):1689-1693.

(上接第9页)

从实验结果可以看出,当采用标准分、隶属度阈值为0.5时,异常成绩检测方法表现出了优异的性能,查全率达到0.75,查准率达到0.71,综合评价函数 F 为0.73达到最高分。

4 小结

本文将教育学原理与数据清洗技术相结合,提出了一种基于分布特征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。我们下一步的工作包括两个方面:一是进一步改进异常成绩检测方法,并将它集成到教务管理信息系统中;而是探索数据质量评估与提高技术在教育信息化中的其它应用。

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] Kahn B, Strong D, Wang R Y. Information Quality Benchmarks: Product and Service Performance[J]. Communications of the ACM,2002,45(4):184-192.
- [3] Richard Y Wang, Diane M Strong. Beyond accuracy: What data quality means to data consumers[J]. Journal of Management Information System,1996,12(4):5-34.
- [4] Hipp J, Guntzer U, Grimmer U. Data quality mining: making a virtue of necessity[C]//Workshop on Research Issues in Data Mining and Knowledge Discovery, Santa Barbara:2001,52-57.
- [5] 黄光扬. 教育测量与评价[M]. 上海:华东师范大学出版社,2002.