改进近邻传播聚类算法的抗生素处方行为①

叶 枫, 钟会玲

(浙江工业大学 经贸管理学院, 杭州 310023)

摘 要: 通过有效分析某医院抗生素处方行为, 提出了一种基于数据仓库改进的近邻传播聚类方法, 利用数据仓库透视出所需数据, 而利用改进的近邻传播聚类方法在数据仓库的基础上聚类出具有代表性医生的抗生素处方数据, 找到抗生素处方行为的影响因素以及规范医生处方行为的评价指标. 采用 2012 年浙江省某三甲医院信息系统中的抗生素处方数据, 对数据进行横断面研究, 首先对提取的数据建立数据仓库, 之后利用改进的近邻传播聚类算法对数据降维和分类, 得到医生的抗生素处方行为的训练集和测试集, 最后利用 SAS9.1 软件的多因素方差分析和配对非参数检验, 分析抗生素处方行为的影响因素以及评价指标. 结果表明, 不同科室、不同月份、不同抗生素种类对抗生素处方数据有显著影响, 而在该医院青霉素、头孢菌素随季节变化有显著差异, 可作为医生的处方评价指标.

关键词: 抗生素; 近邻传播聚类; 处方行为; 数据仓库; 评价指标

Antibiotic Prescribing Behavior of Physicians Based on Improved Affinity Propagation Clustering

YE Feng, ZHONG Hui-Ling

(College of Business Administration, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: A method based on improved affinity propagation clustering of data warehouse is proposed by analyzing physicians' antibiotic prescribing practices in this article. Loading pivottable by the data warehouse, we can select a representative physicians' antibiotic prescription data and find the internal factors of antibiotic prescribing behavior of physicians through improved affinity propagation. The data is from the antibiotic prescription of information system of a first-class hospital in Zhejiang province in 2012, which is studied through cross-sectional. Firstly, we build the data warehouse, then cluster and reduce the dimensionality of data set by improved affinity propagation clustering to get the training set and test set. Finally, the results show that different departments, months and types of antibiotics have significant impacts on the data of antibiotic prescriptions. There are significant differences of cephalosporin and penicillin with seasonal changes in this hospital which can be used as evaluation index of the doctor's prescription.

Key words: antibiotics; affinity propagation clustering; prescribing behavior; data warehouse; evaluating indicator

如今, 几乎所有重要病原菌都出现了多重抗生素耐药性, 如 ESKAPE 病原菌[1]; 并且新研发的抗生素药物十分有限, 大部分临床应用的抗生素是在 1941 至 1968 年发现的, 而过去近 40 年间仅发现了 3 种具有新型抗菌作用机制的药物, 即口恶唑烷酮类的利奈唑胺、脂糖肽类的达托霉素及 2006 年报道的由普拉特链霉菌产生的普拉特霉素(platensimycin), 这些药物主要是抗革兰阳性细菌[2]. 美国食品药品管理局(FDA)自 1983 年到目前的每 5 年新批准的新分子实体全身性抗生素(new molecular entity systemic antibiotics)的数量如表 1, 明确地反映了新抗生素逐年递减[3].

因此必须提高抗生素的利用率以遏制抗生素耐药性. 近年来国外研究者利用各种方法来研究抗生素处方行为的主要影响因素及其评价指标[4-14]. 国内有分析

① 收稿时间:2014-08-12;收到修改稿时间:2014-09-17

某一类抗生素处方的使用情况[15,16], 抗菌药物使用在某医院总体情况[17], 临床医生处方行为的影响因素[18], 患者抗菌药物期望对医生处方行为的影响[19], 实证研究处方点评制度[20]对处方行为能起到一定积极作用, 但是针对医生的抗生素处方行为分析较少. 国外 C. Pulcin 等人根据法国报销数据分析医生抗生素处方差异[21]. 由于抗生素使用情况在各国间有差异. 因此借鉴国内外研究方法, 找到影响国内医生抗生素处方行为的主要内部因素和评价指标, 从而合理使用抗生素.

表 1 新分子实体全身性抗生素的数量

年份	新发现的抗生素数量
1983-1987	16
1988-1992	14
1993-1997	10
1998-2002	7
2003-2007	4
2008-2011	2

基于相似性度量的数据聚类是在数据分析和工程系统中关键的一步. 常用典型的聚类方法 k-means 算法对选择初始的聚类中心比较敏感, 通常需要用不同的初始化运行很多次, 只有当数据量很小并且至少有一个合适的初始化时才适用[22], 使用 K-means 算法聚类医生的抗生素数据会在一定程度上影响分类结果的精确度. 利用将所有数据点作为聚类中心的方法, Brendan J. Frey 等人发明了近邻传播聚类算法[23].

1 方法与数据

1.1 方法

1.1.1 数据仓库技术

数据仓库(Data Warehouse)是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合, 用于支持管理决策. 数据仓库中的数据包含基本数据、历史数据、综合数据和元数据, 这些数据不是大量数据的简单堆积, 而是将数据通过清洗、筛选、稽核和入库加载, 形成 DWD, 进而根据所需形成数据应用[24].

CUBE 是数据仓库中进行 OLAP 分析的重要操作, 根据主题和需求定义分析模型建立 CUBE 并根据用户需求实时生成虚拟立方体. 本文建立医生、抗维生素药品、日期等维表, 构造数据仓库如图 1, 得出包含医生抗生素日期等信息的事实表.

1.1.2 改进的近邻传播(Improved Affinity Propagation)聚类算法

近邻传播(AP)聚类算法把 n 个样本的数据集都作为候选的聚类中心, 算法基于所有样本相似度矩阵 s, 反复迭代找到一组类代表点, 然后把所有的样本点分配到最近的类代表点中, 适合处理大规模数据并且与 K-means 聚类算法相比, AP 聚类算法本身克服了随机性, 因此比较稳定[23].

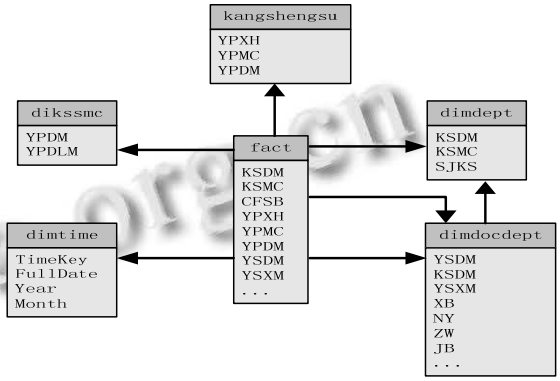


图 1 数据仓库结构

AP 聚类算法聚类过程如下:

$r(i,k)$ 表示从点 i 发送到候选类代表点 k 的消息, 反映 k 点是否适合作为 i 点的类代表点; $a(i,k)$ 则表示从候选类代表点 k 发送到点 i 的消息, 反映 i 点是否选择 k 作为其类代表点. 在聚类之前, 初始假定所有的样本被选为类代表的可能性相同, 即设定所有的 $s(k,k)$ 为相同的值 p. p 值称之为偏向参数(preference).

输入: 任意两个节点 i 和 k 之间的负欧几里得距离 $\mathbf{s(i,k) = -\|x_i - x_k\|^2}$

输出: 数据聚类结果集

算法过程:

初始化: $a(i,k) = 0$ (1)

更新节点之间的 r 和 a 值的积累过程:

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k} \left\{ a(i,k') + s(i,k') \right\} \tag{2}$$

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\} \tag{3}$$

AP 聚类算法传递吸引度 R(responsibility)和归属度 A(availability)两种类型的消息. 对于数据点 i, 使得式(4)最大的数据点 k 可以作为数据点 i 的类中心点.

$$\max\{a(i,k) + r(i,k)\} \tag{4}$$

当迭代次数超过最大值或者当类代表点在连续多次迭代中都不发生改变时终止计算.

本文在原来维度 AP 聚类算法代码基础上增加一个维度, 在原聚类结果的基础上使聚类结果更加精确, 能够直观分辨其立体空间相对位置, 如图 3(b). 从医院信息系统中提取、转换后加载到数据仓库的数据, 其数据量是较大的, 应用改进的 AP 聚类分析的主要目的是降维和分类, 得到最具代表性医生的抗生素处方.

1.1.3 方差分析及配对非参数检验

方差分析是 1923 年由英国统计学家 R.A.Fisher 首先提出的, 其基本思想是对方差进行变异分解, 将总体方差分解到不同因素的不同水平上, 考察不同因素不同水平是否对总体变异具有显著的影响, 进而考察多个总体之间的均值是否有明显的不同[25].

本文通过运用 SAS9.1 软件利用多因素方差分析了科室与月份、抗生素种类与月份之间有无交互性, 以所选取的处方数分别为对应因素的水平, 进行显著性检验. 并且利用配对非参数检验分析了总体抗生素、青霉素、头孢菌素、合成抗菌药、抗分支杆菌、抗真菌药和抗病毒药物处方数随季节有无显著变化, 以选取处方数分别为对应因素的水平进行分析.

1.2 数据

为了验证所提方法的有效性, 采用 ETL 技术抽取 2012 年浙江省某三甲医院的医院信息系统有关医生, 药品, 病人等信息数据, 经过数据预处理后进入数据仓库后形成所需的抗生素处方数据.

1.3 实验流程

原始医院信息系统中数据经过 ETL 处理后进入数据仓库, 分析数据具体实验流程如图 2 所示, 包括四个步骤.

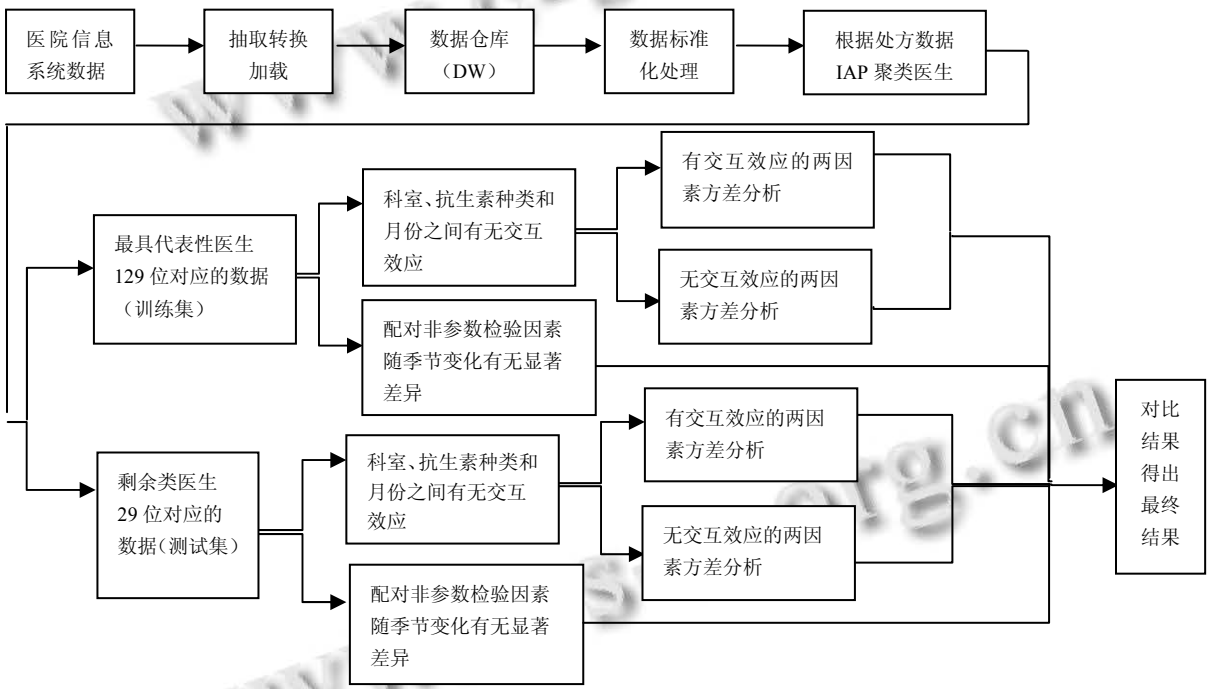


图 2　基于数据仓库改进的近邻传播聚类算法的实验流程

步骤 1: 在数据预处理后基于改进的近邻传播聚类(IAP)算法聚类数据. 对于标准化后的实验数据, 偏向参数 p 取相似矩阵 s 的中位值时, 聚类的数据结果最优, 根据抗生素处方行为聚类医生后的 25 簇数据, 挑选簇数最大的 3 簇作为最具代表性的医生数据训练集, 剩余 22 簇医生数据为测试集, 分成两组数据: 其中训练集 129 位医生对应的抗生素处方数据, 测试集 29 位医生对应的抗生素处方数据.

步骤 2: 由于尚未清楚科室与月份、抗生素种类与月份之间有无相互作用, 因此先用多因素方差分析, 得出其有无交互效应, 再用有交互效应或者无交互效应的两因素方差分析每一个抗生素处方影响因素.

步骤 3: 由于数据总体分布不了解, 使用配对非参数检验方法分析总体抗生素、青霉素、头孢菌素、合成抗菌药、抗分支杆菌、抗真菌药和抗病毒药物处方数随季节有无显著变化, 使其作为医生的处方评价指标.

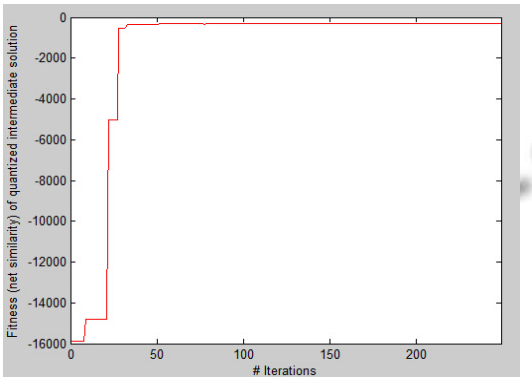
步骤 4: 经过方差分析后的训练集与测试集结果

对比, 得出具有显著相关的影响因素结果.

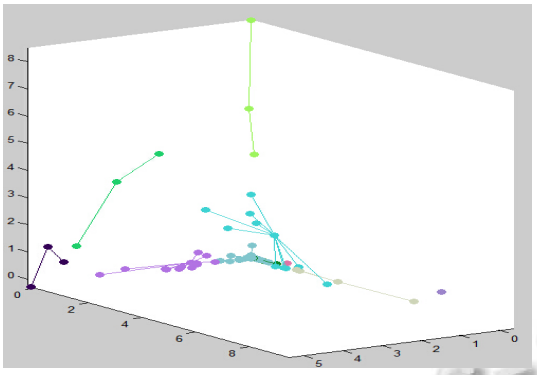
2 结果

2.1 聚类结果

在医生抗生素处方数据上的实验显示, IAP 聚类算法迭代 232 次后的 25 簇数据如图 3(b)所示, 将异常点 22 簇作为一组, 剩下紧簇的 3 簇作为一组. Brendan J. Frey 等人已证明, 利用 AP 聚类算法得到的分类结果较 K-means 聚类结果更为精确.



(a) AP 聚类算法迭代次数



(b) 医生 AP 聚类算法结果

图 3 158 位医生根据抗生素处方行为 AP 聚类算法结果

2.2 显著性分析

根据样本观测值计算检验统计量 F 及其对应的 P 值, 当 F 统计量较大时, 说明组间差异比组内差异大很多, Pr 值与显著性水平 α 比较, 通常取 α=0.05, 如果 Pr 值小于 α, 则拒绝原假设, 认为因素的各水平差异显著; 反之差异不显著. 假设每一个因素对抗生素处方数量无显著影响.

利用方差分析科室、抗生素种类和月份间有无交互效应, 考虑交互效应的两因素方差模型, 结果显示, 科室、抗生素种类和月份的训练集、测试集对应的 Pr 值均大于 α, 即因素间无交互效应, 即可用无交互效应的两因素方差分析得出因素对抗生素处方行为有无显著影响.

利用两因素方差分析科室、抗生素种类和月份之间无交互效应, 在上文结果的基础上可继续可用无交互效应的两因素方差分析得出显著性结果, 计算 F 统计量的 Pr 值与显著性水平 α 比较, 结果显示, 科室、抗生素种类、月份的训练集和测试集对应的 Pr 值均小于 α, 即三个因素水平均差异显著, 对抗生素处方数量有显著影响.

利用配对非参数检验分析因素随季节变化, 由于药物数据总体分布不清楚, 因此使用配对非参数检验. 在显著性水平 α=0.05 下, 计算 Wilcoxon 的 S 统计量及其对应的 P 值, 与显著性水平 α 比较, 结果显示, 青霉素、头孢菌素随季节变化对应的训练集和测试集的 P 值均小于 α, 即该因素水平差异显著, 对抗生素处方数量有显著影响.

3 讨论

首先本文为合理使用抗生素找出影响抗生素处方行为的因素, 对抗生素处方行为的三种因素以及六种评价指标是根据医院现有的数据提出的, 主要是用于医院内部研究之用. 不同科室、不同抗生素种类以及不同月份对抗生素处方数据均有影响, 本文在分析科室和抗生素种类时候去掉部分抗生素使用量较少的科室和种类, 能够分析出有显著影响的科室、抗生素种类和月份.

其次从研究结果看, 儿科专家、普外科、普通妇科专家对抗生素处方数据影响最为显著, 由于外科、妇产科属于手术科室, 术中、术后需使用抗生素, 儿科的呼吸系统感染病例较多, 是抗生素使用的重点科室, 因而使得抗生素的使用率偏高. 十二个月中一月二月三月对抗生素处方数据影响最为显著. 头孢菌素类、大环内酯类抗生素以及喹诺酮类药物对抗生素处方数据影响较为显著, 该医院数据显示头孢菌素类抗感染药物远高出其他类别的品种, 并且第三代头孢菌素类药物 66%的使用率远高于第一、二代头孢菌素类. 喹诺酮类药物是临床上最常用的抗生素之一, 但是随着喹诺酮药物的广泛应用, 细菌对其耐药越来越严重[26], 在美国, 喹咯酮类药物中的大肠埃希菌耐药率已经达到 25%; 而在欧洲, 有些国家的耐药率位于 20%~40%

之间[27]. 因此, 无论是国内还是国外, 细菌对喹诺酮类抗生素的耐药率都比较高, 因而选择喹诺酮类药物随季节变化情况作为医生评价指标, 但本文由于喹诺酮类药物种类数据缺失尚未研究. 针对以上结果说明该院抗生素使用存在滥用现象. 在研究青霉素、头孢菌素、合成抗菌药、抗分支杆菌、抗真菌药和抗病毒药物随季节变化有无显著差异, 结果得出青霉素与头孢菌素随季节变化在该医院可以作为医生处方行为的评价指标, 而合成抗菌药、抗分支杆菌、抗真菌药和抗病毒药物随季节变化没有显著性差异, 因此不能作为医生处方行为的评价指标.

再次本文研究提出的聚类算法不仅选择出分类结果, 并促使找到与抗生素处方行为可能相关的影响因素, 对抗生素研究有一定的指导意义. 利用统计方法研究医疗处方的有很多, 但是用具体聚类算法结合统计方法去分析处方数据的很少, 虽然聚类数据的算法有很多, 但是本文的 IAP 聚类算法分析是其中较为有效的一种方法, AP 聚类算法分析不仅降低特征维数和运算的复杂性, 而且使得聚类结果准确性高, 尤其是在聚类算法增加一个维度后聚类结果更为精确直观, 而 IAP 聚类算法结合多因素方差和配对非参数检验方法分析能更有效得出分类结果、影响因素和评价指标.

最后从本文实验结果可以对抗生素的影响因素进行干涉来改善医生抗生素处方行为, 通过评价指标来合理控制医生的抗生素处方行为, 但是与医生抗生素处方行为有关的影响因素在实践中还有其他因素, 本文未研究患者方面因素比如患者年龄、医疗总费用、患者对抗抗菌药物期望以及医生用药偏好等因素与抗生素处方行为的关系, 可以继续进一步探讨, 虽然现阶段我国抗生素处方行为研究较少, 但随着技术的成熟与研究的深入, 以及借助国外已有的研究成果, 相信在不远的将来抗生素处方行为研究在我国会更加成熟. 医疗制度更为合理, 延长抗生素使用周期, 降低细菌对药物的耐药性.

4 结论

本文通过有效分析医生抗生素处方行为, 提出了一种基于改进的近邻传播聚类方法, 利用该方法选择出具有代表性医生的抗生素处方数据, 找到抗生素处方行为的影响因素以及规范医生处方行为的评价指标. 该方法以最终获得医生数据的分类结果和降低特征维数为目的, 在 2012 年浙江省某三甲医院的医院信息系统数据的基础上进行实验, 从原系统数据库中抽取、转换并加载到数据仓库中, 得出包含医生、抗生素药品、日期等维度信息的事实表, 利用 IAP 聚类算法对数据降维和分类, 利用 SAS9.1 软件对抗生素处方行为数据进行多因素方差分析以及配对非参数检验分析, 实验结果得出不同科室、不同月份、不同抗生素种类对抗生素处方数据有显著影响, 而在该医院青霉素、头孢菌素随季节变化有显著差异, 可以作为医生的处方评价指标.

以上结果能在一定程度上辅助规范医生处方行为, 降低抗生素药物对细菌的敏感性, 具有一定的实际意义. 本文所用数据只是一家医院的数据, 是否能反映整个三甲医院抗生素处方行为还有待研究.

参考文献

1 Rice LB. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: No ESKAPE. Infectious Diseases, 2008, 197(8): 1079–1081.

2 李显志,凌保东.2006 年细菌对抗菌药物耐药机制研究进展回顾.中国抗生素杂志,2007,32(4):193–202,224.

3 Spellberg B, Guidos R, Gilbert D, Bradley J, Boucher HW, Scheld WM, Bartlett JG, Edwards J. The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. Clinical Infectious Diseases, 2008, 46(2): 155–164.

4 McGregor A, Dovey S, Tilyard M. Antibiotic use in upper respiratory tract infections in New Zealand. Family Practice, 1995, 12(2): 166–170.

5 Grigoryan L, Burgerhof JG, Degener JE, Deschepper R, Lundborg CS, Monnet DL, Scicluna EA, Birkin J, Haaijer-Ruskamp FM. Determinants of self-medication with antibiotics in Europe: The impact of beliefs, country wealth and the healthcare system. Antimicrobial Chemotherapy, 2008, 61(5): 1172–1179.

6 Ferech M, Coenen S, Malhotra-Kumar S, Dvorakova K, Hendrickx E, Suetens C, Goossens H. ESAC Project Group. European sSurveillance of antimicrobial consumption (ESAC): Outpatient antibiotic use in Europe. Antimicrobial Chemotheray, 2006, 58(2): 401–407.

7 Genevieve C, Michal A, Dale D, Robyn T. Are physicians

with better clinical skills on licensing examinations less likely to prescribe antibiotics for viral respiratory infections in ambulatory care settings? Medical Care, 2011, 49(2): 156–165.

8 Pan Y, Henderson J, Britt H. Antibiotic prescribing in Australian general practice: How has it changed from 1990–91 to 2002-03? Respiratory medicine, 2006, 100(11): 2004–2011.

9 Ineke W, Marijke K, Arno H, Theo V. Antibiotics for acute respiratory tract symptoms: Patients' expectations, GPs' management and patient satisfaction. Family Practice, 2004, 21(3): 234–237.

10 Roumie CL, Halasa NB, Edwards KM. Differences in antibiotic prescribing among physicians, residents, and nonphysician clinicians. The American journal of medicine, 2005, 118(6): 641–648.

11 Linder JA, Singer DE, Stafford RS. Association between antibiotic prescribing and visit duration in adults with upper respiratory tract infections. Clinical Therapeutics, 2003, 25(9): 2419–2430.

12 Thorpe JM, Smith SR, Trygstad TK. Trends in emergency department antibiotic prescribing for acute respiratory tract infections. The Annals of Pharmacotherapy, 2004, 38(6): 928–935.

13 Linder JA, Bates DW, Platt R. Antivirals and antibiotics for influenza in the United States, 1995-2002. Pharmaco-Epidemiology and Drug Safety, 2005, 14(8): 531–536.

14 Ashworth M, Charlton J, Ballard K, Latinovic R, Gulliford M. Variations in antibiotic prescribing and consultation rates for acute respiratory infection in UK general practices 1995-2000. The British Journal of General Practice, 2005, 55(517): 603–608.

15 温悦,吴寒寅,冯玲玲.某院 2008 年门诊头孢菌素类抗生素处方分析.中国药业,2010,19(9): 55–57.

16 李荣华.我院 2010 年门诊不合理使用抗生素处方分析.中外医疗,2012,31(3): 59–59.

17 陆妙.我院门诊抗生素处方分析.华夏医学,2007,20(2): 304–305.

18 柴佳鹏.临床医生处方行为的影响因素分析与实证研究[硕士学位论文].上海:复旦大学,2009.

19 周艳玲,阳昊,张新平.武汉市患者抗菌药物期望对医生处方行为的影响.医学与社会,2008,21(8): 24–26.

20 沈云峰,许百虹,梁丽梅,赖伟华,黄惠燕.处方点评系统对提高处方质量效果分析.广东药学院学报,2013,29(5):533–535.

21 Pulcini C, Lions C, Ventelou B, Verger P. Approaching the quality of antibiotic prescriptions in primary care using reimbursement data. European Journal of Clinical Microbiology & Infectious Diseases, 2013, 32(3): 325–332.

22 Kojima K. Proceedings of the fifth berkeley symposium on mathematical statistics and probability. American Journal of Human Genetics, 1969, 21(4): 407–408.

23 Brendan FJ, Delbert D. Clustering by passing messages between data points. Science (New York, N.Y.), 2007, 315(5814): 972–976.

24 Inmon WH. Building the Data Warehouse. 4th ed. United States, Wiley computer Publishing, 2005: 78–80.

25 刘桂芬,吕桦,刘玉秀.卫生统计学.北京:中国协和医科大学出版社,2003.

26 吴清芳,许丽,管红云,张玉华,张韵,李明珍.氟喹诺酮类药物在综合医院中应用现状调查分析.临床肺科杂志,2010,15(1):120–121.

27 Lautenbach E, Weiner MG, Nachamkin I, Bilker WB, Sheridan A, Fishman NO. Imipenem resistance among pseudomonas aeruginosa isolates: risk factors for infection and impact of resistance on clinical and economic outcomes. Infection Control and Hospital Epidemiology, 2006, 27(9): 893–900.