基于免疫遗传算法的抗菌药物数据挖掘①

王一敏, 梁治钢

(甘肃省人民医院 网络中心, 兰州 730000)

摘 要: 本文主要研究基于免疫遗传算法的抗菌药物数据挖掘. 在数据挖掘的过程中, 传统挖掘方法的精确度较低, 因此, 将免疫遗传算法技术应用到抗菌药物数据挖掘中, 可以提高数据挖掘的准确性和及时性. 数据挖掘技术为有效地分析疾病间的关系以及其出现的规律提供了新思路, 以此来更好地治疗疾病, 提升治疗效果. 在 HIS 系统中对抗菌药物的数据进行分析和挖掘, 获得潜在的规律和趋势, 逐渐建立抗菌药物诊断知识库. 依据 HIS 系统的医嘱数据, 根据规则自主学习并更新知识库数据, 从而为医生治疗患者提供合理的辅助决策.

关键词: 免疫遗传算法; 抗菌药物; 知识库; 数据挖掘

Data Mining of Antimicrobial Drug Based on Immune Genetic Algorithm

WANG Yi-Min, LIANG Zhi-Gang

(Networks Center, Gansu Provincial Hospital, Lanzhou 730000, China)

Abstract: In this paper, we study data mining of the antimicrobial drug based on immune genetic algorithm. The accuracy of traditional approach for data mining is poor. The immune genetic algorithm technology, which is applied to the data mining of antimicrobial drug can improve the accuracy and timeliness of data mining. Data mining provides a new idea for the effective analysis of the relationship between the disease and its occurrence regularities, which helps better curing the disease and improving the treatment effect. By immune genetic algorithm, it analyzes and mines antibacterialdrug data in HIS, and obtains the rules and trends of potential, which gradually establishes the diagnosis knowledge database of antibacterial drug. According to the doctor's order data of HIS system and autonomous learning and updatding the knowledge database, the approach provides a reasonable assistant decision for doctors to treat patients.

Key words: immune genetic algorithm; antibacterial drug; knowledge database; data mining

1 前言

手术过程中的感染是目前医疗机构里手术患者常见的情况, 其发生率大约是 2%-20%[1,2], 它不仅会增加患者的住院时间及再住院率, 而且可导致致死率、平均住院日以及住院费用的增加. 抗菌药物预防应用的主要目的是预防手术部位感染的发生, 目前, 世界各地的卫生及医疗机构制定了抗菌药物应用指南, 我国医疗行政管理部门也颁布相应的法规推出了符合我国国情的应用指南, 相应法规条款已经被列入到医院等级评审中.

抗菌药物经过了多年的发展, 已经被广泛的应用到临床中, 它的使用虽然可以有效降低患者手术部位的感染发生率, 但是与此同时, 抗菌药物的使用也带来了一定的负作用, 抗菌药物不合规格使用的情况时有发生, 有效控制细菌耐药、加强医疗质量和医疗安全已经是政府和医疗机构共同面临的问题. 医生在抗菌药物的选择和使用上具有随从性和经验性, 忽略了抗菌药物的适应症和患者生理指征, 结果不仅没有达到预防感染的治疗效果, 反而导致患者细菌的耐药性明显增强以及药品不良反应的增加, 使用抗菌药物的患者的住院时间延长的同时, 治疗费用也增加了[3]. 因此, 适合患者的抗菌药物使用研究意义重大, 是当

① 基金项目:甘肃省青年科技基金(2014GS03498)

收稿时间:2016-06-20;收到修改稿时间:2016-08-08 [doi:10.15888/j.cnki.csa.005657]

前所有医疗机构都面临的难题, 虽然有医院已经使用了合理用药支持系统, 但是并没有真正的起到为患者服务的作用, 抗菌药物的合理性使用还有待进一步的提高.

数据挖掘 DM (Data Mining)是一个新兴的人工智能与机器学习技术的应用研究领域, 有着广阔的应用前景, 它是从大量的、不完全的、有噪声的、模糊的、随机的应用数据中, 发现隐含在其中的并且人们事先未知的、但又是潜在有用的信息和知识的非平凡过程. 数据挖掘算法中常用的有机器学习型算法和统计型算法两类, 机器学习型遗传算法 GA (Genetic Algorithm)被普遍运用. 遗传算法是一种借鉴生物界自然选择和自然遗传机制, 模拟自然进化过程搜索最优解的方法. 免疫算法 IA (Immune Algorithm)是模拟免疫系统对病菌的多样性识别能力而设计出来的多峰值搜索算法, 它旨在抽取生物免疫系统中独特的信息处理机制, 研究和设计相应的模型和算法, 进而解决各种复杂问题. 免疫遗传算法 IGA (Immune Genetic Algorithm)是将遗传算法和免疫算法的优点结合起来的算法, 它即具有遗传算法的全局性和并行性, 也具有免疫算法的记忆功能, 从而加快了搜索速度, 提高了传统遗传算法的总体搜索能力, 最终找到最优解.

本文研究重点是利用某三甲医院患者一个月的抗菌药物数据, 结合患者的诊断、生理指症、是否手术等可能影响抗菌药物使用的因素, 将免疫遗传算法应用到抗菌药物数据挖掘中, 对抗菌药物的预防使用和治疗使用情况进行分析, 建立分类模型, 利用免疫遗传算法中先验知识的引入能力, 较好地处理污染数据和缺失数据, 依靠该算法为医生的抗菌药物使用提供最适合个体患者的决策和依据, 有效提高医生对患者的治疗质量和效果.

2 基于IGA的抗菌药物数据挖掘

2.1 问题的设定

目前有不少的合理用药系统已经嵌入到 HIS 系统中, 医生下医嘱的过程包括事前提醒、事中干预、事后分析等过程中, 但目前医院所用的系统不具备自主学习功能, 不能有效的利用已经存在的知识, 只是简单的分析和发现, 针对医生不合理的用药医嘱给予提示并做出相应的调整, 有些数据需要经过人工对照患者病历和医嘱才能发现不合理, 所有的抗菌药物的使用表面看上去似乎都很合理, 但是实际上对医嘱过程行为进行分析审计, 患者诊断、生理特征、用药时机、疗程等指标与国家规定存在着差距, 用药不仅没有达到预防感染的治疗效果, 反而导致细菌耐药性的增强和药品不良反应的增加[4].

手术期间经常存在给手术患者随意时间段内用药, 患者的实际情况并没有被完全考虑, 医生大多数情况下靠自己的临床经验来下医嘱, 由于每个患者的个体情况差异, 导致增加抗菌药物剂量的情况时常发生. 例如, 某些抗菌药物的使用要求是术前 2h 内才能使用, 但有些临床医师则在手术前几天就已经给手术患者使用, 甚至部分手术医师则在手术前几天就已经给患者使用了抗生素[5]. 导致患者的耐药性升高, 增加了患者术后感染的风险, 预防用药的疗程无形中也增加了. 因此将免疫遗传算法的最优解搜索能力应用到 HIS 系统中, 从而给医生提供最适合不同患者的医嘱方案中是病房医生站和电子病历信息系统需要解决的问题. HIS 系统抗菌药物的数据主要集中在病区医嘱表中, 具体字段名称为医嘱名称、药品序号、开医嘱时间、停医嘱时间、用药方式、医嘱类型等.

2.2 编码方式

本文使用免疫遗传算法在 HIS 数据库中进行数据挖掘, 算法首先要解决的就是编码问题, 编码要根据实际数据的特点, 它不仅决定了个体染色体排列形式, 而且决定 k 个个体从搜索空间的基因型变换求解. 编码方法还影响到遗传算子、交叉算子、免疫算子的运算操作. 因此, 编码方法在很大程序上决定了如何进行群体的遗传进化运算以及遗传进化运算的效率. 本文中在数据库中挖掘蕴含在其中的有效数据, 需要建立一个规则来进行数据的挖掘, 具体用下面的伪代码实现:

if (rule1 && rule2 && ….&& rule*n*) {Result} (1)

上式中 rule*i* 采用三元组<Type,Operator,Value>, 其中 Type 代表数据库中某个字段值, 本文中该变量值取以下字段: 药品名称、药品适应症、开医嘱时间、用药时间、生理指标等. Value 代表某个字段值得一个取值, 药品名称用药品代码表示, 药品适应症用相应疾病编码表示, Operator 是连接 Type 和 Value 的一个关系运算符, 一般情况下该运算符取值为“与”或者“或”, Result 则是考虑上述字段值综合取值. 因此在这里采用结构体数组编码的方法进行编码, 数组的元素个数

与事务数据库中的字段个数相对应, 这里的 Y*i* 表示 rule*i* 的取值, 具体编码方式如图 1 所示.

Y_1	Y_2	…	Y_n

图 1　编码方式

2.3 适应度计算

适应值是评价个体好坏的唯一标准, 适应值高的个体将被保留, 所有的算法都是基于适应度函数来进行数据挖掘, 因此建立合适的适应度函数对整个算法的执行很重要, 算法的要求需要满足如下条件: 在规定的约束条件内搜索达到时间复杂度最小, 能够动态分析数据以及得到最优解. 置信度表示结论成立的可相信程度. 覆盖度表示结论包含于条件的正确程度, 次数越大, 说明该规则越完备. 在确定适应度函数时必须考虑以上条件, 以下是取适应度函数:

$$\mathrm{Fit(r)} = a*\mathrm{Conf}(r)/\mathrm{Confmin} + b*\mathrm{Supp(r)}/\mathrm{Suppmin} + c*\mathrm{Cover}(r)/\mathrm{Covermin} \quad (2)$$

在上式中, a、b、c 分别代表置信度、支持度、覆盖度的权值, $a + b + c = 1$(a>=0, b>=0, c>=0). Confmin 是最小置信度阀值, Suppmin 是最小支持度阀值, Covermin 是最小覆盖度阀值, 默认的三个最小阀值都为 1. 适应度函数反映了支持度、置信度和覆盖度这三者综合作用的结果. 在进化过程中, 只有这三者都高的规则才能在竞争中生存下来. 在本文中, 适应度函数定义为从某个时间段内某个病种使用抗菌药物数量中去掉不适合该病种的抗菌药物数量.

2.4 免疫算子

免疫算子的选择是用来判断抗体的多样性及等位基因概率的变化过程. 设免疫系统有 N 个抗体组成, 每个抗体有M位基因. 每个基因位可供选择的字符(等位基因)共有 S 个. 根据信息论原理, N 个抗体第 j 位基因的信息熵可表示为:

$$M_j(x) = \sum_{i=1}^{n_j}\left(H_{ij}\log H_{ij}\right) \quad H_{ij} \neq 0 \quad (3)$$

式中, n 为第 j 基因位上基因总数. H_{ij} 为第 j 位基因取 i 个等位基因的概率, 在该文中 H_{ij} 为 x 个抗菌药物医嘱中, 第 j 位为基因的概率. 当第 j 位基因的所有等位基因都相同时, H_{ij}= 1, 则 $M_j(N)$=0, 信息熵可看作免疫系统中表示抗体多样性的一种度量. N 个抗体所有基因位的平均信息熵为:

$$H(Y) = \frac{1}{N}\sum_{i=1}^{N}\left(H_j(Y)\right) \quad (4)$$

为了从抗体中找到适应度较高的抗体, 需要比较抗体之间、抗体和抗原间的亲和度, 任意两个抗体 u 与 v 间的亲和度表示为:

$$\left(A_t\right)_{uv} = \frac{1}{1+H(2)} \quad (5)$$

上式中为两个抗体的平均信息熵. $H(2)$可以由下面的公式得到:

$$H(2) = \frac{1}{N}\sum_{i=1}^{N}\left(H_i\right) \quad (6)$$

在式(6)中, $H(2)$表示同一病种和同一生理指征下两条抗菌药物医嘱之间的相似度.

2.5 抗体浓度

抗体的浓度反映群体中相似抗体所占的比例,

$$C_i - \frac{1}{N}\sum_{i=1}^{N}\left(S_{i,s}\right) \quad (7)$$

式中, N 为抗体总数; $S_{i,s} = \begin{cases}1, S_{i,s} > T\\ 0, S_{i,s} < T\end{cases}$, T=0.75 为预先设定的一个相似度值. 浓度是由种群中和抗体具有很大相似度的抗体(抗菌药物医嘱)个数表示.

2.6 免疫遗传算法的记忆库更新

免疫遗传算法在每次更新记忆库时, 采用精英保留策略, 先将适应度较高的若干个抗体存入记忆库, 然后按照繁殖概率在剩余群体中选择优秀抗体存入记忆库, 这样可以避免适应度高的抗体因其浓度高而受到抑制. 父代抗体群的形成与记忆库更新策略类似, 首先, 将适应度排序较高的若干个父代抗体直接加入到子代抗体群, 然后随机从剩余父代抗体中进行选择操作, 选择优秀抗体加入到子代抗体群, 父代抗体被选择的概率即为式(7)计算出的抗体的繁殖概率.

2.7 算法设计

免疫遗传算法将待求解的问题作为抗原(Antigen), 在抗菌药物数据挖掘的系统中, 对应就是针对不同病症、症状、体征的病人生成一套最适合患者的医嘱抗菌药物组合, 即治疗方案; 将问题的解作为抗体(Antibody), 对所求问题进行合理分析和计算, 产生出多种数据的组合, 最终形成最适合患者治疗方案的数据, 即疫苗(Vaccine); 免疫系统(Immune system)确认抗原入侵, 然后根据疫苗信息产生相应的抗体来解决问题[6,7]. IGA 具体算法如下:

① 参数初始化: 设置种群规模 N、记忆库容量 C、

变异概率 P_m 等参数;

② 产生初始抗体群: 抗体通常是随机产生的, 如果识别的抗原是已经有的记忆抗原, 则从记忆库中取出相应的抗体组成初始种群, 否则就随机产生, 抗体采用图 1 的编码;

③ 计算抗体适应度: 根据适应度函数计算公式, 计算群体中每个抗体的适应度, 按照适应度大小降序排列, 选择其中适应度较高的 K 个抗体组成群体;

④ 抗体选择操作: 对抗体群中的各个抗体进行评价. 在 IGA 中对个体的评价是以个体的繁殖概率为标准, 保留全局最优抗体;

⑤ 更新记忆库: 将抗体群分别按适应度和繁殖概率排序, 并分别取按适应度排序的前三分之一的个体和按繁殖概率排序的前三分之二的个体存入记忆库中;

⑥ 依次执行选择操作、交叉操作、变异操作得到下一代群体;

⑦ 子代群体与记忆库的群体合并, 构成新一代抗体群;

⑧ 终止条件: 重复执行步骤③至步骤⑦, 判断是否满足结束条件, 是则结束[5].

本文利用 IGA 实现从抗菌药物知识库中进行不同病种抗菌药物医嘱自动组合, 按照一些约束条件(如: 病症、生理特征、年龄、性别、地区, 既往史等)从知识库中根据算法来生成一套抗菌药物医嘱, 其中每个病种的抗菌药物使用方案即为一个抗体, 病种中不同生理指征、年龄、性别、职业为抗体中一个基因, 这样反复选择一个病种不同指征来组成初始种群, 然后按照上述算法的流程进行免疫遗传操作, 最终得出与患者病症最适合的一套抗菌药物医嘱方案.

3 数据挖掘应用

3.1 基本数据

本文以抗菌药物辅助决策的数据挖掘为实例, 对提出的基于机器学习抗菌药物数据挖掘模型进行研究. 本实例的优化目标就是判定患者使用抗菌药物的合理性、建立抗菌药物知识库以及数据挖掘的分类技术在该实例应用. 对 2015 年医院某一个月住院病人抗菌药物的数据进行分析, 其中住院病人 5356 人, 使用抗菌药物的病人 2479 人, 根据合理用药使用规范及标准, 符合标准的 1946 人(达到 78.51%). 采集到的生产数据一般都是比较复杂的, 必须进行数据清洗和规范化, 使其既能反应出生产的需要, 也能适合数据挖掘. 预处理的功能就是利用各种统计规律对数据进行分析, 去掉无用数据, 从而达到数据挖掘的目标. 经过预处理的数据各个指标变量见表 1.

表 1 变量指标及名称

变量名	值范围	指标值	变量解释
Age	<20岁	1	年龄
	20岁<=age<30岁	2	
	30岁<=age<40岁	3	
	>=40	4	
Ismale	1 男	1	性别
	2 女	2	
InputStatus	1 危急	1	入院情况
	2 急诊	2	
	3 一般	3	
Ismedical	1 无过敏	1	过敏药物
	2 青霉素	2	
	3 其它药物	3	
Inputdiagnosis	1 单病种	1	入院诊断
	2 并发症	2	
InHospital	<7	1	住院日
	7天<= and <14天	2	

	14天<= and <21天	3	
	>21天	4	
Mzvalue	1 全麻	1	麻醉方式
	2 局麻	2	
	3 腰麻	3	
	4 硬膜外	4	
Qkdj	1 I类切口	4	切口等级
	2 II类切口	3	
	3 III类切口	2	
	4 无手术	1	
Operatetime	<=60分钟	1	手术时长
	60分<and<=120分	2	
	121分<and<=180分	3	
	>180分钟	4	
Antidrug	预防用药	2	抗菌药物
	非预防用药	1	
Overtwo	单药品	1	抗菌联合用药
	2种联合	2	
	3种联合	3	
	大于4种联合	4	
Yysj	术前2小时到术后24小时之间	1	用药时间
	大于术前2小时或者术后24小时后	2	

3.2 抗菌药物辅助决策指标参数

以抗菌药物辅助决策作为数据挖掘的设定参数, 对于不同的抗菌药物、生理指症及病种进行比较和判别, 根据抗菌药物使用数据得知, 围手术期疾病诊断、用药品种和给药时机三项符合标准则定义该病例抗菌药物符合标准, 其中某一项不符合, 则判断该病例用药不符合标准, 需要进行相应的改造, 符合置标志为“1”, 否则为“0”.

另外, 在后面的最优解挖掘算法里相关参数设置如下: a=1(支持度权值), b=1(置信度权值), c=1(涵盖度权值), P_m=0.6(变异率), P_x=0.8(交叉率).

3.3 最优解模式挖掘

本文中利用免疫遗传算法来求解患者使用抗菌药物的合理性及知识库的建应立与自学习能力. 数据挖掘的目标是求解某医院一个月抗菌药物的合理性使用情况, 希望能发现最优解, 由于每个患者的生理情况不同, 抗菌用药的使用也没有一个具体的标准, 只是根据住院病人的相关药物信息进行分析探索性研究, 从各个医疗数据中获取最适合病人的有用知识. 在求解的过程中, 将抗菌药物数据按照表 1 中的数据变量指标值进行相应的判断, 表 2 是具体判定标准.

表 2　最优解模式挖掘判定标准

变量累计值	判定情况
>30	抗菌药物辅助决策第1级, 加入知识库;
25-30	抗菌药物辅助决策第2级, 加入知识库;
20-25	抗菌药物辅助决策第3级, 不加入知识库;
<20	抗菌药物辅助决策第4级, 不加入知识库.

3.4 验证结果

在实验进程中, 为了验证本文算法的合理性, 针对抗菌药物辅助决策系统, 分别采用遗传算法(Genetic Algorithm)、蚁群算法(Ant Colony Algorithm)、神经网络算法(Neural Networks Algorithm)、免疫遗传算法(Immune Genetic Algorithm)等 4 种算法进行仿真求解, 优化结果主要是判断抗菌用药的辅助规则使用及加入知识库情况, 求解时间是计算上述优化结果所使用的时间, 最终求解的输出结果如表 3 所示, 根据实验结果, 无论是求解质量和速度, 本文的方法优于其它 3 种[5].

表 3 4 种算法求解 10 个实例的计算结果

序号	优化结果				求解时间(S)			
	GA	ACA	NNA	IGA	GA	ACA	NNA	IGA
1	2029	2026	2033	2037	35	32	30	31
2	2193	2203	2195	2203	34	35	32	29
3	2098	2107	2109	2114	34	34	32	30
4	2113	2109	2116	2117	36	35	35	33
5	2103	2111	2112	2115	35	35	33	33
6	2121	2126	2129	2133	35	34	32	31
7	2118	2115	2113	2118	37	38	37	34
8	2111	2116	2114	2119	36	34	33	32
9	2109	2106	2112	2117	33	32	31	29
10	2112	2116	2120	2125	38	36	33	34

4 结语

影响疾病的因素具有不确定性, 确定一个正确的治疗方案有时非常困难, 随着抗菌药品种类不断发展, 新的药品被推出, 医疗抗菌用药的不合理现象与不良反应也随之增加. 本文中采取机器学习-免疫遗传算法对抗菌药物进行数据挖掘, 进行的是探索性研究, 对于患者的抗菌药物数据利用信息技术进行尝试和创新, 为药物利用深入研究提供新的思路, 有助于建立医疗数据仓库并进行知识发现的使用.

随着医院信息化的发展, HIS 系统中抗菌药物的使用分析已经由原来以“收费”为中心的信息系统向以“电子病历”为中心的合理性研究转变. 由于患者个体生理各项指标的不确定性以及医疗环境的特殊性, 抗菌药物智能辅助判断需要慎重, 专业知识、用药习惯和临床经验以及合理的算法起着至关重要的作用, 因此本研究在对某类疾病的抗菌药物使用合理情况与否进行智能判断时, 没有直接判断是否合理, 而是以既定指标的“符合”或“不符合”标准进行分类.

测试结果表明该挖掘方法在一定程度上能够帮助医生对抗菌药物的辅助使用及诊断准确性问题, 对抗菌药物辅助知识库的数据可以进行有效更新, 对其中的干扰数据进行了加权修正, 为医生的辅助决策提供了良好的数据基础[10]. 总之, 抗菌药物的合理使用是一个复杂过程, 为了增强模型的说服力, 需要采用更多的样本数据进行模型的完善.

参考文献

1 Afzal KAK, Mirshad PV, Rashed MR, Banu G. A study on the usage pattern of antimicrobial agents for the prevention of surgical site infections (SSIs) in a tertiary care teaching hospital. J. Clin. Diagn. Res., 2013, 7(4): 671–674.

2 Alp E, Elmali F, Ersoy S, et al. Incidence and risk factorsof surgical site infection in general surgery in adeveloping country. Surg. Today, 2014, 44(4): 685–689.

3 史占军,张亚莉,景宗森.规范化与长期应用抗生素预防术后伤口感染的效果对比.中华医院感染学杂志,2003,13(1): 57–59.

4 杜建强,聂斌.数据挖掘在中医药领域应用研究进展.中国中医药信息杂,2013,20(6):109–112.

5 肖伟平,何宏.基于遗传算法的数据挖掘方法及应用.湖南科技大学学报,2009,24(3):82–86.

6 於时才,梁治钢.基于免疫遗传算法的移动机器人路径规划.微计算机信息,2008,24(2):264–266.

7 Jiao LC, Wang L. A novel genetic algorithm based on immunity. IEEE Trans. on System, Man and Cybernetics-Part A: Systems and Humans, 2000, 30(5): 552–561.

8 苏娅,刘杰,黄亚楼.在线医疗文本中的实体识别研究.北京大学学报(自然科学版),2016,52(1):1–9.

9 谭文明,甘琴,龚世菊,赖小红,覃海坤.三甲医院处方点评软件系统的开发和应用.北方药学,2015,(2):142–144.

10 翟晓波,何志高,方芳,鲍思蔚,徐婷,文传民. “围手术期抗菌药物监控系统”的临床应用.药学与临床,2012,10: 1458–1460.