·药学服务·

# 基于文本分类技术的住院患者药源性变态反应自动监测模块研究

王啸宇，郭代红，徐元杰（解放军总医院药品保障中心，北京 100853）

**[摘要]** **目的：**利用医疗电子病历中的文本信息开展住院患者用药安全性评价，为住院患者ADR监测提供新方法。**方法：**在已有的主动监测系统基础上，设计、开发基于文本分类技术的住院患者药源性变态反应自动监测模块，利用优质文本进行分类算法的机器学习。**结果：**完成包括事件配置器、特征词集、自然语言处理器、文本分类器、结果展示器5部分的主动监测模块的开发；试用于头孢哌酮舒巴坦用药患者的主动监测，结果显示629例患者中出现变态反应的阳性预测值达到44.44%（4/9），其真实世界发生率0.64%，与说明书中所列的发生率0.68%相近。**结论：**本研究建立了文本信息主动监测方法，阳性预测值可通过特征词集、分类规则的深入研究加以改善。

**[关键词]** 药品不良反应；医疗电子病历；自然语言处理；文本分类技术

**[中图分类号]** R95 **[文献标识码]** A **[文章编号]** 1672－8157(2016)02－0117－04

## Study on automatic monitoring module of inpatient drug-induced allergy based on text categorization technology

WANG Xiao-yu, GUO Dai-hong, XU Yuan-jie(*Department of Pharmaceutical Care, PLA General Hospital, Beijing 100853, China*)

**[ABSTRACT]** **Objective:** To evaluate drug safety of inpatients by text information in electronic healthcare records, and provide the new method for inpatients ADR monitoring. **Methods:** On the basis of active monitoring system, automatic monitoring module of inpatient drug-induced allergy based on text categorization technology was designed and developed, and the machine learning of categorization algorithm was carried out through qualified text. **Results:** The active monitoring module consists of the event configurator, the representative feature set, the natural language processors, the text categorizer and the result display unit. The module had been tested, and the positive predictive value was 44.44% (4/9) in the 629 users of cefoperazone and sulbactam. The real world incidence rate was 0.64%, which was similar with the incidence rate (0.68%) in the drug directions. **Conclusion:** Text information active monitoring method had been built and the positive predictive value could be improved by further study of the representative feature set and categorization rules.

**[KEY WORDS]** Adverse drug reaction; Electronic healthcare record; Natural language processing; Text categorization technology

药品不良反应（adverse drug reaction，ADR）监测是药品安全性监测的主要手段，是药品上市后评价的重要组成部分。随着药品风险研究领域中ADR主动监测工作的日趋重要，计算机ADR自动监测系统的研究也日渐深入。我们采用触发器技术研发的“医疗机构ADE主动监测与智能评估警示系统”，能够监控患者检验指标的异常变化，对血小板减少、肝肾功能异常等多种药源性疾病实施主动监测，并已初步形成了国内首个药品安全信息化主动监测网络[1-2]。但医疗电子病历中包含有患者在院治疗期间的全部医疗数据，且其中大部分信息为采用医疗自然语言记录的文本，由于缺少高效的计算机自动监测系统，无法精准的筛选这些文本中包含的大量ADR相关信息。因此本研究采用文本分类技术开发医疗电子病历文本信息监测模块，并通过机器学习获取较高的阳性报警率，填补自动监测系统在该领域的空白，为临床药师更加精准的开展药学监护工作提供有效支撑工具，保障患者的用药安全。

## 1 模块设计思路

医疗电子病历文本信息监测模块的功能需求是能够通过自然语言处理手段挖掘医疗电子病历中的文本信息，以文本自动分类技术判定患者是否发生ADR，自动监测院内目标药物导致变态反应的发生情况。为实现这一目的，模块需具备划定监测范围、信息识别、信息收集、文本性质判定、结果呈现等功能。

**[基金项目]** 2014年全军后勤科研重点项目（BWS14R039）

**[通信作者]** 郭代红，女，主任药师，硕士生导师，主要从事临床药学及药物警戒研究。E-mail：guodh301@163.com

**[作者简介]** 王啸宇，男，药师，硕士研究生，主要从事临床药学研究。E-mail：metallica365@126.com

因此，应包括以下几个部分：

### 1.1　事件配置器

事件配置器是使用者的操作平台，其作用在于制定监测计划，包括设置监测方式、建立纳入排除标准、选定监测范围、调整监测指标、调取医疗数据等。事件配置器是监测模块的中心，主动监测的每一个步骤都在其中得以体现；通过限定监测规则，事件配置可以调节模块各部分的功能，影响监测结果。

### 1.2　特征词集

特征词集是在药源性变态反应的诊断、治疗过程中医疗电子病历记录内可能出现的专业词语的集合。特征词集是文本分类的依据，应能够全面地描述ADR，涵盖目标ADR相关的专业词、同义词、特殊词；为兼顾系统运行效率，特征词集不宜过大，要求每个特征词都具有代表性和特异性。因此，特征词集的建立是本课题的研究重点之一。

### 1.3　自然语言处理器

自然语言处理是基于文本信息的自动监测的基础，也是研究的难点。合格的处理器应具备3种基本能力：①目标药物使用人群的查询；②利用病人特征编码如住院号、病历号等，通过电子病历系统接口取得患者病历资料；③结合特征词集提取患者与目标ADR有关的文本信息及完整的药物治疗记录，并能按照标准化格式储存、传递上述数据，便于数据的进一步利用。

### 1.4　文本分类器

文本信息自动监测的实质，是利用病历资料中的文本信息，判断病历资料的性质，将其分为发生ADR与未发生ADR两类[3]。性能良好的文本分类器是这一过程顺利进行的保证，也是自动监测模块的核心组成部分。特征词集内的特征词与文本性质均存在一定的逻辑关系，分类器利用统计学方法，计算出现在同一份病历资料内的特征词所代表的逻辑关系的总和，对文本进行分类，得到自动监测所需要的结果。

### 1.5　结果展示器

监测的结果交由结果展示器呈现，包括报警、正常、排除三类。报警结果为自动监测系统判断为阳性的病例，正常为阴性病例，排除结果为按照监测计划的纳入排除标准排除的相关病例。在结果展示器中可查看由患者电子病历中提取的文本数据、特征词集记录、药物治疗记录等信息，由专业人员分析、评估自动监测的结果。

在这种模块构架下的自动监测一般流程为：在事件配置器中设定自动监测计划，启动监测任务，而后配置器在HIS系统中调取计划所需数据，传递给自然语言处理器；自然语言处理器结合特征词表，提取有意义的数据，用于文本分类；在文本分类器中，处理上一阶段产生的数据，判断文本性质，完成分类，提交到结果展示器，供药师人工甄别。如图 1 。

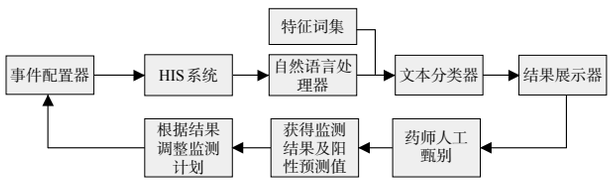


**图1　自动监测模块运行流程图**

**Fig 1　Program flow chart of the automatic monitoring module**

## 2　关键性技术问题

### 2.1　建立特征词集

在以往的ADR主动监测或自动报告系统中，亦有使用关键词法作为触发器技术的报道出现，但其使用的词库通常来自药品说明书，通常对某一ADR事件仅用1～2个词语作为关键词，词语简单、笼统、特异性差，与电子病历中医疗文书的真实情况相去甚远，不能满足要求[4]。

在本次特征词集建立过程中，我们深入研究了变态反应这一ADR事件在临床医疗文书中的表达。首先，研读《过敏性疾病诊疗指南》等5本药源性变态反应或皮肤疾病专著，划定了一个包括43个词语的待选范围，入选的均为药源性变态反应临床表现相关的专用术语；而后利用院内已确诊的药源性变态反应患者医疗电子病历，对照备选特征词进行病案研究，统计特征词在阳性病例中出现的词频，分析词语与ADR事件的关联性，确定特征词集。同时收集病案中出现的特殊词、非标准词，用以扩充词量，提高特征词集的特异性。最后确定的药源性变态反应特征词集包括临床表现、解救措施、相关排除等多个维度，共44个特征词。

### 2.2　自然语言处理方法探索

自然语言处理是利用特征词集，采用计算机程序解读文本的语义，提取有效信息的过程，是人工智能和语言学领域的分支学科[5]。中文的自然语言处理，经过多年的发展，已经取得了一定的成绩，但本项研究所涉及的文本主要为病程等医学资料，较一般文本来说，具有一定的特殊性。

自然语言处理的难点之一在于词语的正确提取，如“皮疹”和“疹”都是模块中的特征词，但在词语提取时，将“皮疹”中的“疹”作为独立的词语提取出来，则会产生词语重复提取的情况；医疗文书的语义分析也存在难度，例如“皮疹”与“无皮疹”、“未见皮疹”语义完全相反，如果不能正确识别，则会产生极

大的歧义。对于上述问题，我们将中文分词技术和文本关键词检索与语言逻辑处理相结合，分词得到的结果需通过逻辑检验才予以采信、保留，解决了以上问题。

另外，自然语言处理对计算机性能有很高的要求，在大规模处理文本时，会对服务器带来极大的压力；我们以时间为标志增量提取电子病历，避免了重复读取病历全文，有效地缓解了这一压力。

### 2.3　文本分类方法的选择

中文文本分类研究中比较成熟的技术主要分为两类[6]，一类是根据特征词在各类文本中分布概率来进行文本分类的，比较具有代表性的方法为朴素贝叶斯法；另一类是根据特征词间存在的具有方向性的联系对文本分类，如：决策树法。本研究采用朴素贝叶斯法，在机器学习后能够得到每个特征词与文本性质间的逻辑关系，即在是否发生了ADR两类文档中的分布概率，在分类过程中，计算每个病历的总分布概率，以达到分类的目的。采用朴素贝叶斯法的优势是：①分类结果准确，系统调试难度小；②特征词间不存在关联性，便于调整特征词集，适用于系统初期开发；③算法易行，系统运算压力小；④能够进行自主学习。使用决策树法亦能达到系统所需要的分类性能，但决策树法依赖于特征词间的相互关系，不利于特征词集的调整，而且其自主学习机制搭建复杂，应用难度大，并不适用于系统初期开发，在特征词集调整完毕后，可添加为比较算法，提升系统效率。

### 2.4　分类算法的机器学习

在基于朴素贝叶斯法的文本分类器构建完成后，仍需经过机器学习方能使获得特征词分布概率形成文本分类的能力[6]。机器学习是指针对算法，通过统计分析分类确定的优质文本中特征词的分布情况，计算特征词在各分类中的概率，用以开展真实样本分析。在机器学习的过程中，我们发挥医疗电子病历资源丰富的优势，通过人工筛选的方式在HIS系统中取得阳性样本132例，阴性样本68例，完成了系统建设阶段的机器学习；并通过程序设定，将每次自动监测任务的结果返回系统，进一步调整特征词的分布概率，从而完成运行后的机器学习，在使用过程中自动优化软件性能。

## 3　结果

### 3.1　模块开发

本系统模块开发工具为Microsoft Visual Studio 2012，其中服务器端配置为：CPU 2.0 GHz，内存4 G，硬盘200 G；操作系统兼容Win2003 Server/Win2008 Server；数据库采用Microsoft SQL Server 2008。客户端配置：CPU 2.0 GHz，内存2 G，硬盘空间20 G；操作系统兼容Win2003/XP/Win7/Win10等，IE6.0及以上环境。

### 3.2　模块内容

按照模块设计思路，完成了文本监测模块的初步开发工作。模块共包括事件配置器、特征词集、自然语言处理器、文本分类器、结果展示器5部分，其中事件配置器、特征词集、结果展示器有可操作界面，自然语言处理器和文本分类器为后台运行。见图2～4。


**图2　事件配置器**

**Fig 2　The event configurator**


**图3　事件配置器及特征词库**

**Fig 3　The event configurator and the representative feature set**

### 3.3　实际应用效果

利用本系统对院内HIS电子病历数据库中正在使用头孢哌酮舒巴坦的住院患者进行自动监测，共纳入监测病人629例，系统报警9例，经人工甄别确定的阳性病例为4例，阳性率44.44%，头孢哌酮舒巴坦致变态反应的真实世界发生率0.64%，与说明书中所列的0.68%相近。

## 4　讨论

模块的监测目标是医疗过程中产生的患者医疗文

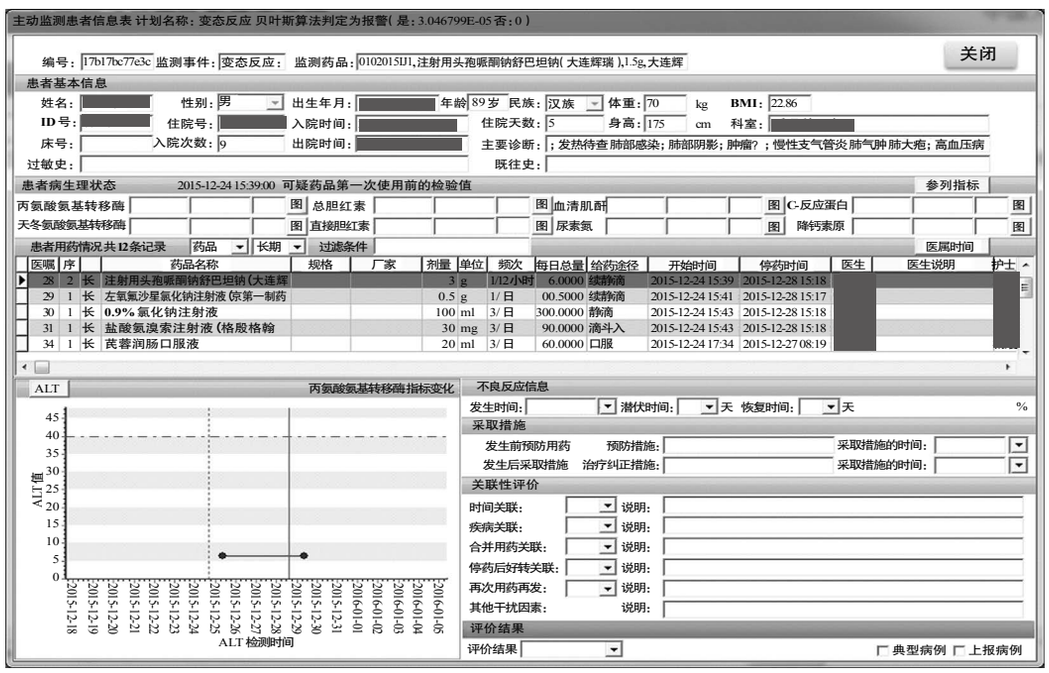


**图4　结果展示器**

**Fig 4　The result display unit**

本数据，因此其报警提示的ADR信号已经被临床医生发现，并做出客观描述，具有一定的滞后性，但对于及时补充漏报ADR并分析其发生原因具有重要的实际意义，能够减轻临床药师的工作负担，提高临床药学干预的效率和准确性，增加患者用药安全性。

经过测试，本次研究开发的自动监测模块阳性率为44.44%，低于已有模块的平均水平[7]，除了医生病案书写的规范性外，还可能与以下3个因素相关：①测试覆盖的患者范围窄，ADR发生率低，阳性病例数少，监测结果受误差影响较大；②特征词集中，存在1个或多个特征词的特异性不强，对主动监测结果造成干扰，致使结果阳性率较低；③特征词集所包含的词语间存在一定的联系，而朴素贝叶斯法未能将这些联系纳入考察范围，使文本中蕴含的部分信息被忽略，影响自动监测模块性能未能达到预期的目标。

针对上述可能的影响因素，我们将继续展开以下研究：①进一步扩大使用范围，增加样本量，减少随机误差对实验结果的影响；②深化特征词集研究，对再次特征词与ADR之间的联系，筛选可能存在干扰的词语，同时在扩大试用范围的基础上，加大机器学习的强度，得到更加准确的特征词分布概率；③文本分类算法由朴素贝叶斯调整为贝叶斯网络，将特征词间的相互联系纳入到文本性质考察中来，更全面地分析文本中的信息。

**[参考文献]**

[1] 郭代红，陈超，马亮，等.5所医院住院患者ADE警示系统主动监测数据分析与评价[J].中国药物应用与监测，2014，11(6)：368-371.

[2] 陈超，郭代红，薛万国，等.住院患者药品不良事件主动监测与评估警示系统的研发[J].中国药物警戒，2013，10(7)：411-414，418.

[3] 吕婷，姜友好.文本挖掘在生物医学领域中的应用及其系统工具[J].中华医学图书情报杂志，2010，19(4)：56-64.

[4] 耿魁魁，刘圣，沈爱宗，等.医院信息系统中药品不良反应主动监测系统的构建[J].中国医院药学杂志，2012，32(14)：1147-1149.

[5] 余凯，贾磊，陈雨强，等.深度学习的昨天、今天和明天[J].计算机研究与发展，2013，50(9)：1799-1804.

[6] 苏金树，张博锋，徐昕.基于机器学习的文本分类技术研究进展[J].软件学报，2006，17(9)：1848-1859.

[7] 裴斐，陈超，郭代红.阿托伐他汀致转氨酶异常升高的主动监测研究[J].中国药物应用与监测，2014，11(1)：31-33.

（收稿日期：2016-01-20　修回日期：2016-02-25）