基于电子病历可视分析的临床诊断模型①

商金秋[1,2], 朱卫国[4], 樊银亭[3], 李伟亨[1,2], 马翠霞[1], 滕东兴[1]

[1](中国科学院软件研究所 人机交互技术与智能信息处理实验室, 北京 100190)
[2](中国科学院大学, 北京 100190)
[3](中原工学院 计算机学院, 郑州 450007)
[4](中国医学科学院 北京协和医学院 北京协和医院 普通内科&信息管理处, 北京 100730)

摘 要: 针对当前医生在临床诊疗过程中缺乏系统有效的手段, 以及隐藏在大量电子病历中的医学知识没有得到充分利用的现状, 研究了利用可视分析和数据挖掘相结合的方法, 辅助医生进行临床诊疗服务. 本文以不明原因发热疾病为例, 首先对电子病历进行数据预处理和结构化提取, 然后结合具体需求进行可视组织与分析, 再利用数据挖掘相关算法对患者大量症状和发热原因之间的关系进行学习, 帮助医生发现病历中潜在的医疗知识, 辅助医生进行诊断. 在上述工作的基础上, 构建了一个面向临床诊疗的可视分析与辅助诊断框架, 并给出了系统实例加以验证, 结果表明该系统可以有效的帮助医生分析不明原因发热电子病历内的知识, 有利于进一步的疾病诊断, 缩短了平均确诊时间.

关键词: 电子病历; 可视分析; 数据挖掘; 辅助诊断

Clinical Diagnosis Model Based on Visual Analysis for Electronic Medical Record

SHANG Jin-Qiu[1,2], ZHU Wei-Guo[4], FAN Yin-Ting[3], LI Wei-Heng[1,2], MA Cui-Xia[1], TENG Dong-Xing[1]

[1](Intelligence Engineering Lab, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100190, China)
[3](School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China)
[4](Division of General Internal, Medicine Department of IT Management, Peking Union Medical College Hospital, Beijing 100730, China)

Abstract: To help doctors better diagnose diseases, overcome the lack of systematic and effective means in the process of clinical diagnosis and treatment and make the best of the medical knowledge in electronic medical records, a diagnosis model is proposed based on visual analysis and data mining for electronic medical record. Firstly, electronic medical records of fever of unknown origin are preprocessed into structured data by extracting patients' symptoms. Secondly, the structured data is organized and visualized based on specific requirements. Finally, a diagnosis model is trained to discover the relationship between symptoms and causes,helping doctors find the potential medical knowledge in medical records and assisting doctors to diagnose. A visual analysis and auxiliary diagnosis framework for clinical diagnosis and treatment is designed based on the above analysis. Experiments show that the system could help doctors analyze the knowledge of electronic medical records of unknown cause, which could help doctors diagnose diseases in a shorter period of time.

Key words: electronic medical records; visual analysis; data mining; computer aided diagnosis

医疗信息化的快速发展不仅提升了医生的工作效率, 减轻了医生的工作负担, 同时也改善了用户对医院的满意度和信任度. 尽管我国许多大中型医院在医疗信息化平台建设方面已经有了较快进展, 但是大部分医院的医疗信息系统仅仅是一种管理工具, 是对医院日常业务的一种支撑, 未能给医生提供更为有效的诊疗服务.

一方面, 目前医生在临床诊疗过程中主要依靠自

① 基金项目:北京协和医院杰出青年基金项目(JQ201509);国家高技术研究发展计划(863)(2012AA02A608);国家自然科学基金(U1304611)
收稿时间:2016-03-21;收到修改稿时间:2016-04-24 [doi:10.15888/j.cnki.csa.005465]

身的专业知识及经验积累, 疾病的诊治缺乏系统有效的手段, 但由于医生自身的知识更新速度很难赶上医学知识发展速度, 在形成诊断结论、制定诊疗计划时, 仍然需要依赖计算机的辅助支持以提高工作的准确性和效率. 计算机辅助的临床诊疗服务通过准确的、有针对性的方式提供给医生诊疗建议, 可以提高医疗服务质量和效率. 然而, 当前智能化的临床诊疗服务并未得到有效开展, 主要表现在: 医学知识获取水平不高、更新滞后; 医学知识服务水平难以保证海量资源信息的充分利用; 医疗服务环境和用户人群的复杂多样性使得传统的临床诊疗服务流程与模式难以满足广大用户的个性化需求.

另一方面, 电子病历系统的广泛应用, 产生了大量的电子病历数据, 但当前基于电子病历来辅助医生进行临床诊疗的服务却比较匮乏, 电子病历仅仅是简单的记录了病人的基本情况、体格检查、用药情况和病程等, 仅仅是记录诊疗过程的作用, 隐藏在病历中大量的医学知识和诊疗经验没有得到充分的挖掘和利用.

本文针对以上问题, 以用户需求为中心, 以不明原因发热电子病历为实例, 提出了一种面向临床诊疗的可视分析与辅助诊断方法. 主要工作包括: 1)研究了大量临床诊疗信息的复杂关联关系, 特别是不明原因发热疾病临床诊治的现状, 给出了一种临床诊疗环境下人机协同认知特性; 2)构建了符合特定数据组织方式的可视形态集及自然的交互任务集; 3)利用病历文件训练了辅助医生进行临床诊断的数据挖掘模型; 4)构建了一个面向临床决策推理的可视分析与辅助诊疗框架; 5)给出了系统实例加以验证. 结果表明本文所述方法能够为医疗从业人员提供更加便捷的信息利用方式, 辅助其对病历数据的分析、归纳、整理活动, 减轻医生对不明原因发热病历数据进行分析的负担, 在医疗诊断过程中提供决策支持服务.

1 相关工作

1.1 临床决策支持系统

斯坦福大学的 Shortliffe 等人在 20 世纪 70 年代研发的 MYCIN[1]是世界上首个功能较全面的临床决策支持系统, 它能够辅助医生对细菌感染病进行诊断和治疗, 在 MYCIN 系统框架基础上建立的肺功能专家系统 PUFF 曾在旧金山太平洋医疗中心使用过一段时间, 也是医学专家系统首次在临床得到应用. 近年由 Archimedes Model 设计推出的 IndiGO[2]针对“个体化指导和决策”目标, 利用数学语言分析了临床、诊治和生理学资料, 基于分析结果设计出一个诊断模型, 以及诊治方案和人体生理标准. 为了真正实现“个体化指导”的目标, IndiGO 对于每位患者, 选取了 30 多种不同特征来分析, 这些特征包括既往病史、高危致病因素、过往治疗信息以及提示不同疾病的生物标记物, 个体化指南也会自动将其推荐的治疗费用和医保报销范围进行比较, 这样可以帮助患者掌握诊疗费用. Auminence[3]系统对患者既往史、已有症状和其他相关资料进行综合分析, 从所提供的患者资料中搜索与之对应的各个可能的疾病类型及发生的可能性大小.

1.2 病历数据可视分析

陈湖山针对电子病历数据研究了可动态配置的集成可视化视图, 提出了一种分层次的集成视图动态配置方法[4]; Bui 等人采用 TimeLine 可视化形态展现各类医疗文档和医学影像[5]; 曾志荣利用可视化的方式研究了电子病历中关系型数据, 通过一系列交互式可视化形态帮助医生分析病历[6]; 郑威琳用具体的可视化形态将病人的历史医疗信息表达出来, 使得医生不用读取、分析医疗报告就可直观地了解病人的历史以及健康状态[7].

1.3 基于病历数据的数据挖掘

近年来, 数据挖掘结合医疗领域的研究也越来越多, 刘立刚将数据挖掘中经典算法 Apriori 应用到电子病历数据中, 挖掘了具有诊断价值的关联规则, 以此来提高医生的诊断效率[8]. 张连育等人将不同的数据挖掘算法运用到中医领域不同问题上, 对比不同方法的结果, 达到了最大化利用某一种方法的效果[9].

利用可视分析和数据挖掘方法相结合的方式帮助医生整理、组织、分析电子病历内容的研究也越来越多. 徐天明将 LDA 主题模型和可视分析方法结合起来分析中文电子病历语义层面的关系, 利用主题这一语义层面的概念来表达大量文本内容的原始病历文件, 通过计算主题向量内积的方式, 帮助医生快速理解病历间的相似性及对病历进行分类[10].

2 临床诊疗环境下人机协同认知特性分析

认知心理学将认知过程定义为由信息的获取、分析、归纳、编码、储存、概念形成、提取和使用等一

系列阶段组成的, 按一定程序进行信息加工的系统. 其中, 信息提取指依据一定的线索从记忆中寻找并获取已经储存的信息, 信息使用指利用提取的信息对信息进行认知加工[11]. 临床诊疗过程中, 临床医生经常会面临诊断、决策任务, 临床诊疗过程中往往需要反复阅读分析病人病历. 如何为决策者提供自然的、辅助医生进行思维决策的可视化形态和交互方式, 对于诊断来说具有重大意义.

人机交互的过程就是人与计算机借助各种符号和动作进行信息双向交换的过程, 人和计算机系统是交互主体[12]. 作为可视分析的认知主体, 分析决策者一方面需要通过不断“动手”与机器交互, 完成信息获取、加工等决策支撑活动, 另一方面还需要不断“动脑”来分析判断信息进而决策, 如图 1 所示.

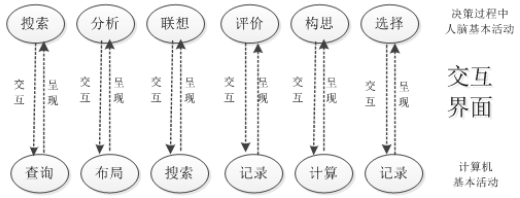

图 1　决策分析活动中的人机协同工作模式

从人的认知角度出发, 电子病历中海量的、具有复杂关联关系的医学信息给临床医生带来了信息过载的问题. 计算机在处理速度、存贮容量、数值计算、逻辑推理等方面具有非常明显的优势, 但却无法超越人的大脑在学习能力、创造能力、环境适应能力以及经验总结和知识归纳等方面的能力. 人机交互过程中, 当信息的表现形式与人的认知能力越接近, 人感到认知负担越小、交互越自然, 相应地对计算机信息加工处理的能力要求也越高. 反之, 当信息的表现形式越接近计算机可处理的计算模型, 信息加工处理过程越简单, 但却会增加人的认知负担.

目前可视分析系统在决策支撑方面的应用注意到了可视分析活动的渐进特点, 提供了渐 进式的使能技术来支持用户展开深入分析活动, 但仍然较多关注系统的建设成本, 而对人类的阅读和认知成本关注不够[13]. 一方面由于缺乏对信息利用过程中大脑思维过程的深入研究, 另一方面缺乏对人在信息利用与分析过程中的交互习惯和方式的研究, 没有充分利用人们长期生活中已经习得的交互方式和手段, 导致人在信息利用的过程中的思维活动经常被信息搜索和整理等基本活动所打断, 干扰了思维活动的有效进行. 可视分析的过程可以看作是用户与数据不断会话的过程, 因此, 自然高效的、对用户思维活动干扰少的交互方式是降低分析决策者的交互负担、有效提高可视分析效率的重要手段, 它需要满足:

① 能真实有效地记录医疗行为和医疗知识;

② 能很好的组织专家经验, 支持信息的复用和重组;

③ 能利用相关历史行为数据和病人诊疗历史为分析决策过程提供必要的支持;

④ 能将分析过程与可视化界面关联起来.

3　基于不明原因发热电子病历的可视分析

本文在分析了上述人机协同认知特性的基础上, 针对不明原因发热疾病实际诊疗过程医生的认知负担, 并结合计算机智能化处理数据的高速计算能力, 研究了利用可视分析和数据挖掘方法辅助医生进行诊疗活动, 减轻医生的认知负担.

3.1 不明原因发热电子病历的收集与整理

本文研究过程中所用的电子病历来自于北京协和医院普通内科科室, 共 288 份发热待查患者病历, 时间从 2012 年 3 月到 2015 年 12 月, 其中在出院时确诊的有 260 例, 未确诊的有 28 例, 将 260 例出院确诊患者的病历作为数据挖掘诊断模型的训练数据. 根据临床医生的医学知识, 将出院诊断中共 29 个病因划分为共四个大类, 分别是: 感染、免疫、肿瘤和其他. 每个患者的病历包括入院记录, 出院记录, 病程记录三部分. 病程文件记录了医生每次对患者的治疗过程, 包括做了哪些检查及用了什么药; 出院记录中记录了患者出院时的诊断结论和医嘱.

3.2 不明原因发热电子病历预处理

医院存储患者电子病历的方法通常是采用 XML 格式存储在数据库中, 本文首先采用 java 工具包 dom4j 对 XML 格式的原始病历文件进行解析, 提取入院记录中的现病史部分, 对于 word、excel 等格式的数据先转换成纯文本格式, 再利用 NLPIR 系统中的中文分词器进行去停用词处理, 再以特定符号为分隔符进行断句处理; 根据医生提供的症状词典, 采用正向最大匹配算法从文本中提取出<key,value>结构, 其中

key 的取值为医生提供的症状词典中的症状, value 的取值为症状key的患者临床表现, 例如当key的取值为“呕吐”这一症状时, value 的值为“有”或“无”; 对于体格检查中的生理指标, 直接提取其中的具体数值, 最终将原始的病历文本处理成结构化格式, 如图 2 所示, 医生参照原始病历内容, 利用数据预处理页面中的工具核对提取结果, 最后将核对过的结构化数据提交到服务器. 这样的数据格式不仅是本文进行可视分析和数据挖掘的基础, 也是帮助临床医生进行病历整理和病情讨论的参考.

病症	症状描述	病症	症状描述
干咳	有	血尿	无
气短	无	胸闷	无
咳痰	无	眼干	无
咯血	无	畏寒	无
胸痛	无	寒战	无
咳嗽	有	关节肿痛	无
踩棉花感	无	黄痰	无
脱发	无	腹泻	无
尿量减少	无	尿急	无

图 2 病历提取结果截图

在对病历中现病史和体格检查进行结构化提取之后, 利用 NLPIR 系统中的分词工具对病历所有的文本内容进行分词处理, 该工具在用户指定了自定义词典之后, 在完成中文分词的基础上, 还能够对分词结果进行词性标注, 如名词、动词、形容词. 在不明原因发热电子病历中, 对病人症状的描述大多采用形容词, 分析分词结果中形容词的使用情况可以帮助实习医生快速掌握常用医学术语, 熟悉医务流程.

3.3 面向不明原因发热电子病历的可视化形态

本文基于上述得到的病历处理结果, 结合不明原因发热临床诊治的特点及临床诊疗环境下人机协同认知特性, 构建了一系列可视化形态.

3.3.1 患者治疗过程可视化形态

不明原因发热患者的住院时间通常是数周到数月的时间不等, 往往要经过多个病程的治疗才会出院, 在病程记录文件中, 每次的病程都记录了当时患者的身体状况和医生的治疗方法, 如针对哪些症状用了哪些药等记录, 所以每位住院患者的病程病历文件中完整的记录了从入院到出院的治疗过程, 但从以文本段落的方式记录显示的病程中发现诊治规律往往很难, 通过可视化形态可以简单直观的展现患者的整个诊治过程.

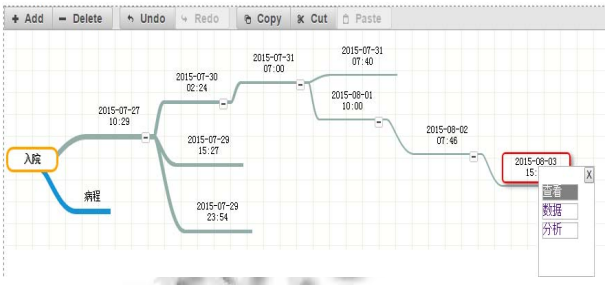

图 3 病程数据可视化

如图 3 所示, 为某一患者的病程病历的可视化结果: 图中每个节点代表一次病程, 右键后可以查看本次病程的具体治疗细节.

3.3.2 患者病症描述词分布比例可视化形态

在一份发热待查患者的病历中, 有大量的文本用来描述患者的相关症状临床表现, 标签云可视化形态将文本中每个词的出现频率作为权重, 用特定的布局算法, 在一定空间内用不同的颜色和大小编码每个标签, 很直观的表示出了哪些词是出现频率比较高的. 如图 4 所示, 出现频率比较高的所占的空间比较大. 图 5 用柱状图的方式显示了在一份病历中, 形容词出现的最多的 10 个词的分布情况, 从中可以看出哪些形容词最多的用在病历中用来描述病人的身体状况.

图 4 病历文本内容可视化

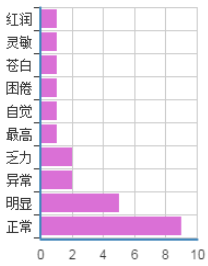

图 5 一份病历中使用频率最高的 10 个形容词

3.3.3 患者诊治过程中体温变化情况可视化

医生在患者的每次病程中的都会记录患者当天的最高体温，下图用折线图和柱状图可视化形态展示了患者从入院到出院的每天最高体温变化情况，图中底部的滑块可以用来进行筛选日期帮助医生查看感兴趣时间段内的体温变化.

图 6　患者最高体温变化情况

3.3.4 患者体格检查中各项生理指标可视化

为了确诊患者发热的原因，医生要对患者进行各项生理指标的辅助检查，常见的包括 CRP、铁蛋白，血压等十几项化验和检查，通过对比病人各项体格检查结果和正常值的高低，找出可能引起病人发热的原因，所以各项辅助检查结果是一种维数比较高的数据，单纯的从病历文本中分析这些数据的关系显然对医生是一个很大的挑战. 平行坐标可视化形态通过将每一维的数据映射到一个坐标轴上，使得可以在有限的空间内展示大量数值型数据之间的大小和变化关系，如图 7 所示，图中表示的是一位发热待查患者住院期间辅助检查结果，每一条不同颜色的折线代表一天中各个生理指标，通过观察同一个坐标轴上数据的波动情况，可以分析哪些检查结果对于发热待查疾病的诊治有意义.

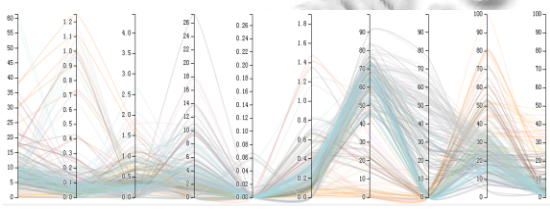

图 7　一位患者生理指标变化情况

4　基于数据挖掘算法的辅助诊断

针对当前医生在临床诊疗过程中缺乏系统有效手段的现状，结合当前计算机在快速运算和智能学习方面的优势，充分发挥人机结合的优势，以不明原因发热电子病历为例，利用数据挖掘的相关模型从大量的患者病历中学习众多症状和诊治结论之间的内在关系，辅助医生对新的不明原因发热患者病因进行诊断.

4.1 训练数据处理和模型特征选择

将 3.2 节中预处理后的病历文件作为学习模型输入的原始数据，根据临床医生的经验和模型的复杂度，结合病历文件中各个症状的分布情况，从 87 个症状中选择了 26 个症状作为特征，如对于发热这一症状，几乎在每位患者病历中都出现了，对于诊断病因没有区别性，不作为模型的特征；根据症状的临床表现阴阳性，数值化症状的临床表现，有设为 1，无设为 0. 这样，将原始的每个文本病历结构化为一个带有分类标签的多维向量.

4.2 诊断模型的选择

不明原因发热患者的出院诊断中往往有多个病因，各个病因的诊断可能性不同，在数据挖掘领域中这是一个多标签分类问题. 根据已有病历数据的特点和规模，选择神经网络算法和决策树算法对数据进行建模，这样可以对比分析两种模型在处理不明原因发热诊断问题上的效果，找到适合该疾病诊断的最佳模型，最后选择学习效果比较好的模型作为辅助医生进行诊断的模型.

神经网络算法的实现选用了基于 java 的开源框架 Neuroph，Neuroph 是轻量级的 java 神经网络的框架，可以用来模拟常见的神经网络架构，用户可以模块化定义需要的网络结构，在将训练数据处理成特定的格式后，模型会自动进行训练学习，直到收敛或迭代指定的次数. 本文选用框架中提供的分类能力较好、学习过程收敛速度较快的多层感知机(Multi-Layer Perceptron)神经网络模型，训练算法采用后向传播(Back Propagation 简称 BP)算法，通过输出后的误差来估计前一层的误差，一层一层的反传下去，在输入样本不断的刺激下，改变网络连接的权重，以使网络的输出逐步接近期望的输出. BP 神经网络以其非线性映射能力和自适应能力在分类问题上得到了广泛的应用. 决策树的学习采用 ID3 算法，通过计算每个特征的信息增益选择分裂节点，每一次的决策是向树的底部深度遍历的过程，直到遇到叶子节点，给出数据的分类标签.

4.3 模型训练及结果评估

由于出院诊断明确的病历数量较少，为了提高模型训练结果的精度，模型的训练采用交叉验证的方式，将 260 份出院诊断明确的病例随机均分成 10 组，每轮

选择 9 组作为训练数据, 其余的 1 组病历数据作为验证, 进行 10 轮的训练和验证, 最后计算 10 轮训练的平均错误率. 下图为最终 2 个模型的训练结果, 决策树和神经网络的平均错误率分别是 0.33 和 0.38, 从图中可以看出, 整体上决策树模型的学习效果优于神经网络模型.

图 8 模型训练结果对比

在模型训练结束后, 系统保留错误率最低的模型参数, 医生利用诊断模型页面中提供的操作方法, 上传新的病历后, 模型输出该病人发热原因, 供医生参考.

5 基于不明原因发热电子病历可视分析与辅助诊断框架

在以上工作基础上, 本文开发了基于不明原因发热电子病历可视分析与辅助诊断系统, 系统架构图如图 9 所示, 系统分为 5 层.

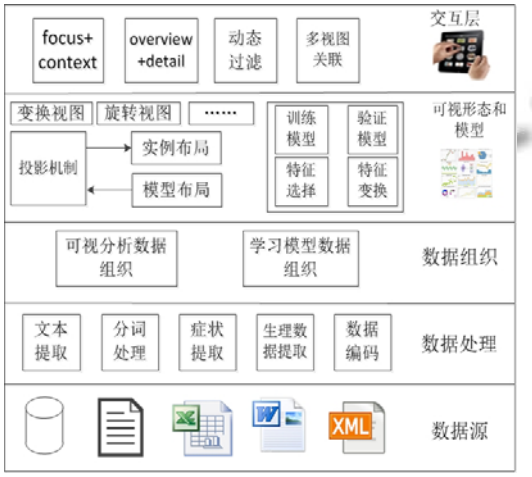

图 9 系统架构图

① 数据源层: 主要负责从不同的数据源以不同的格式读取原始的电子病历, 现在各大医院的电子病历系统存储病历的方式还未统一, 常见的格式如数据库、word 格式、excel 格式、XML 格式, 纯文本格式等, 数据源层通过不同的数据接口, 从系统外读取电子病历, 将其转换成纯文本的格式, 供数据处理层使用.

② 数据处理层: 对上述处理得到的纯文本数据进行基本的去停用词和标点符号处理, 对相关文本进行分词处理, 同时基于医生提供的症状词典, 运用自然语言处理的相关方法从病历文本中提取症状和症状描述以及某些生理指标, 如体温数据, 为后面的可视化和模型训练做数据准备.

③ 数据组织层: 根据上层可视化形态和交互任务的要求, 将相关数据组织成特定的格式供上层使用, 将提取出来的症状数据编码成数值型数据.

④ 可视形态层和模型层: 根据医生的具体需要, 从可视化组件库中选择可视化形态, 利用相关布局算法和投影映射机制, 将数据组织层提供的结构化数据, 用可视化形态展现病历内的规律和病历间的关系; 在完成了特征选择和特征变化之后, 通过数据挖掘中的算法模型学习病历内不明原因发热相关症状和诊治结论之间的关系, 再用学习完成的模型帮助医生对新病人的诊断.

⑤ 交互层: 提供基本交互任务集, 如平移、旋转、过滤等, 医生通过和视图之间不断的交互操作, 渐进式的完成可视分析的任务.

6 相关技术

6.1 症状提取算法-正向最大匹配算法

症状提取的目标是从病历中的现病史和体格检查部分中提取医生对患者临床症状的具体描述, 根据医生提供的不明原因发热疾病症状词典, 在病历文本中搜索症状词典中的词, 从中找到最长匹配的症状以及症状的具体表现. 正向最大匹配算法在待搜索中文字符串中, 从左到右扫描中文字符串, 当有与词典中的词匹配的字符串时, 暂存该词及词的长度, 当整个中文字符串搜索结束后, 取长度最大的词为最终提取结果.

6.2 决策树构造算法-ID3

决策树构造的关键是从众多特征中, 怎样选择合适的特征作为根节点以及分裂节点, ID3 算法基于信息论中熵的概念[14], 通过计算每个特征的信息增益的方法选择分裂节点. 下面具体介绍 ID3 算法的计算过程.

假设训练数据集为 D, 共有 m 个不同的类别 $Ci(i=1,\ldots,m)$,$|Ci,D|$表示数据 D 中类别为 i 的样本个数, $|D|$表示训练集的大小, 则 D 的熵定义为:

$$Info_A(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

其中 pi 为 D 中任何一个样本属于类别 Ci 的概率，用 |Ci,D|/|D|来估计；假设我们在特征 A 上对整个数据集进行划分，即将 A 特征作为根结点，A 在数据集 D 中有 v 个不同的取值{a1,a2,…,av},则特征 A 可以根据其取值将数据集 D 划分为 v 个不同子集{D1,D2,…,Dv}，即对于子集 Dj 的所有样本在特征 A 上的取值都为 aj ,则特征 A 的熵定义为:

$$Info_A(D) = \sum_{j=1}^{v}\left(\frac{|D_j|}{|D|} \times Info(D_j)\right)$$

则特征 A 的信息增益定义为:

$$Gain(A)=Info(D)- Info_A(D)$$

同理可以计算数据集 D 中其他特征的信息增益，最后选择信息增益最大特征作为根结点；在 A 划分的 v 个子集上继续使用上述方法选择信息增益大的特征作为分裂结点，直到某个子集中的类别全部一样，将类别标签作为叶子结点.

6.3 标签云中关键词布局算法

标签云(word cloud)可视化形态常常用来表达大量文本中出现频率较高的关键有哪些，通过词的大小和颜色帮助用户快速浏览文本中的关键信息，常常用来做网站的导航和个人主页特点展示. 其中，用来布局的重要参数就是每个词的权重，本文中所用的布局算法将每个词在病历中出现的次数作为权重，在完成了词的权重和标签云中字体大小、颜色的映射之后，将词表和词表中词的权重作为算法的输入数据，首先随机的将权重最大的词放置在某个起点位置，通常是靠近中间或中央水平线某处，如果该词与任何先前放置的词相交，移动它，沿着螺旋上升一步. 重复，直到没有交叉点. 下面的伪代码简单的描述了算法的流程:

算法输入：按词频降序排列的关键词列表和对应词频列表

```
while(关键词列表不空){
  取关键词表中下一个词;
  随机将该词放置在中间位置;
  设置字体大小;
  while(该词与之前放置的词相交){
  沿着某个向量方向螺旋移动该词一步;
  }
  将该词放置在当前位置;
}
```

7 应用实例

基于上述研究，开发了基于不明原因发热电子病历可视分析与辅助诊断系统，如图 10 所示，针对已有的电子病历数据进行了实例验证. 系统采用经典的 MVC 设计模式，服务器端由 java 编写，前端页面由 html+css+javascript 完成，其中可视化工具选择了成熟的 d3.js 和 ECharts.js.

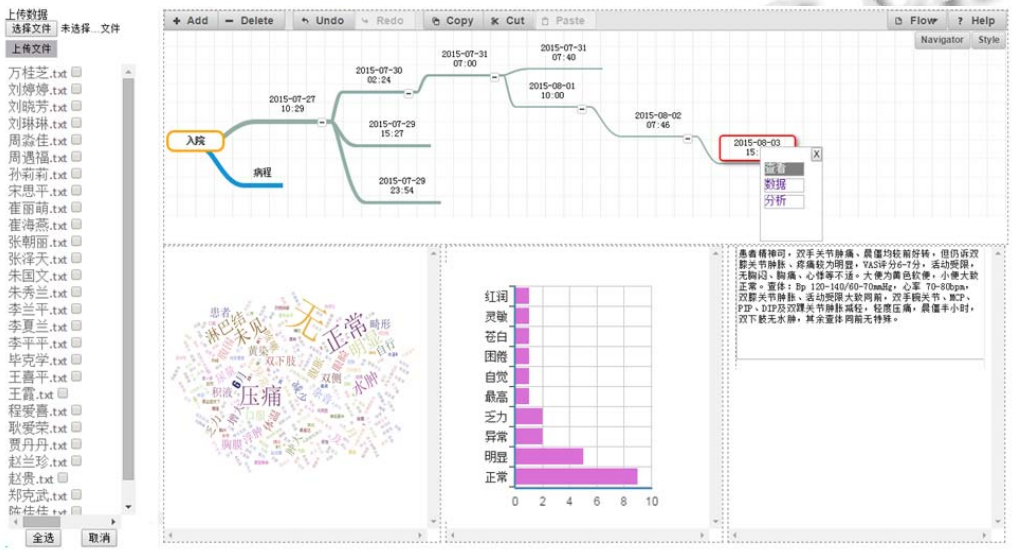

图 10　系统可视分析页面

系统可以直接读取服务器端后台中存储的从 HIS 系统中导出的 XML 格式电子病历，利用 java 工具包解析 XML，从中提取中病历中的现病史、体格检查等病历内容，再利用自然语言处理方法中的正向最大匹配算法从病历文本中提取患者和发热相关的症状以及该症状的具体表现，最后在浏览器中以表格的形式呈现

给医生, 医生参照原始病历内容, 核对提取结果, 最后将核对过的提取结果存到服务器端.

同时, 医生对处理后的病历, 可以通过可视分析页面左侧的文件目录选择要分析的病历文件, 再选择相关的可视化形态查看分析病历内容, 如可以通过病程可视化形态查看每次治疗的具体用药和检查结果等, 通过折线图和热力图分析病历体温和症状在治疗前后的变化情况; 通过标签云和柱状图分析一份病历文件中哪些词出现的频率最高, 以及哪些形容词在描述病人体征情况时用的较多. 系统支持在线上传新的病历文件, 系统解析后, 返回提取结果给医生.

图 11 系统辅助诊断页面

在辅助诊断页面, 医生可以利用已经训练好的模型对新的不明原因发热患者的病历进行分析, 系统给出可能的发热原因; 此外, 医生也可以从已经处理完成的患者病历中选择一部分作为训练数据, 在线训练诊断模型, 根据返回的训练结果, 下一步可以进行诊断或继续训练优化模型.

8 结论和展望

本文针对当前医生在临床诊断中缺少系统有效手段的问题, 提出了一种以可视分析和数据挖掘方法相结合的方式, 辅助医生进行临床诊疗, 并以不明原因发热电子病历为实例进行了实验验证. 通过和临床医生沟通后确定了关注的症状集合, 运用自然语言处理的相关算法对电子病历进行了预处理和症状的结构化提取, 根据不明原因发热疾病的特点和病历数据内的规律, 设计了一系列相关的可视形态帮助医生分析病历内潜在的规律; 利用数据挖掘相关模型挖掘大量患者症状和出院诊断之间的关系, 用计算机的智能对新的不明原因发热患者的病因给出参考, 在一定程度上减轻了医生的诊疗负担.

同时, 本文的研究内容还存在以下不足: 设计的可视化形态不够丰富, 交互性还有待进一步的提高; 特征选择和模型选择还可以采用更加科学有效的方法进行验证, 模型还可以尝试更多, 比如随机森林, 朴素贝叶斯等, 在后续工作中, 收集到更多的病历后, 完善以上的不足, 更好的辅助医生进行诊断.

参考文献

1 Shortliffe EH, Axline SG, Buchanan BG, et al. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Computers and Biomedical Research, 1973, 6(6): 544–560.

2 Bellows J, Patel S, Young SS. Use of IndiGO individualized clinical guidelines in primary care. Journal of the American Medical Informatics Association, 2013.

3Bagchi S, Barborak MA, Daniels S D, et al. User interface for an evidence-based, hypothesis-generating decision support system. U.S. Patent Application 13/448,607. [2012-4-17].

4 陈湖山.可动态配置的电子病历数据集成视图研究与开发[学硕士位论文].杭州:浙江大学,2012.

5 Bui AAT, Taira RK, Churchill B, et al. Integrated visualization of problemcentric urologic patient records. Annals of the New York Academy of Sciences, 2002, 980(1): 267–277.

6 曾志荣.电子病历中关系型数据的质量分析可视化技术[学位论文].北京:中国科学院研究生院,2012.

7 郑威琳.病人医疗信息多维可视化表达方法与实现技术研究[博士学位论文].上海:中国科学院研究生院上海技术物理研究所,2014.

8 刘立刚,钟锐,杨娟.基于兴趣度的 Apriori 算法在电子病历数据分析中的应用.江西理工大学学报,2013,34(5):72–76.

9 张连育,吕立.基于策略模式的中医数据挖掘平台.计算机系统应用,2010,19(11):5–9.

10 徐天明,樊银亭,马翠霞,等.面向电子病历中文医学信息的可视组织方法.计算机系统应用,2015,24(11):44–51.

11 陈为,沈则潜,陶煜波.数据可视化.北京:电子工业出版社,2013.12

12 滕东兴,王子璐,杨海燕,等.基于交互式可视组件的分析决策环境研究.计算机学报,2011,34(3):555–565.

13 董士海,王衡.人机交互.北京:北京大学出版社,2004.

14 Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Elsevier, 2011.