面向中医辨证规范的交互式数据挖掘框架*

□王 波 张 斌 魏伟杰 马玉慧 (东北大学信息科学与工程学院 沈阳 110004)
梁茂新** 王雪峰 董 丹 (辽宁中医学院 沈阳 110032)

摘 要：传统的中医辨证规范采用经验式的逻辑分析方法和数理统计方法，数学模型都是预先选定的，不是由具体病证的内在规律决定，因而规范结果的客观真实性受到一定的质疑。本文在"十五攻关"项目的基础上，设计了基于小儿肺炎中医辨证的交互式数据挖掘框架，采用三层模型，在挖掘信息的同时，通过人机交互，推动了数据挖掘技术和传统中医现代化的结合。

关键词：辨证规范 数据挖掘 数据预处理 挖掘算法 技术平台

一、引 言

随着数据库技术、人工智能和数理统计等技术的发展与融合，数据挖掘(Data Mining, DM)技术应运而生。数据挖掘是一门新兴的交叉学科，也是现代科学技术相互渗透的必然结果，其基本目标就是从大量的数据中提取隐藏的、潜在的和有用的知识和信息。这一技术自20世纪末提出以来，引起了许多专家学者的广泛关注，已应用到金融业、零售业、医疗保健和政府决策等多个领域。当前，该技术在中医药学研究领域展示了其广阔的应用前景。

中医的诊疗方法主要是辨证论治，证决定治法，治法决定方药，方药决定疗效，辨证诊断正确与否事关重大。然而，病证诊断不规范问题一直困扰着学术界，并被认为是制约中医药学发展的瓶颈之一。因此，辨证规范一直是普遍关注的攻关课题。

最初，辨证规范依托的是中医的疾病，是借助专家经验完成的综合性判断。由于中医的一种病证通常见于现代医学的多种疾病，多半是多种西医疾病的一个症状或多个症状，故而基于中医疾病的辨证规范未能得到学术界的普遍认可，还产生许多新问题。继之，人们尝试对现代医学的疾病进行辨证规范研究。此类研究兵分两路，一是针对中医某证，采用循证医学(DME)和流行病学方法，确认其在现代医学疾病的分布，以及现代医学各病关于该证的具体证候(即症状)的构成规律；一是针对现代医学某病，采用判别分析、回归分析、聚类分析、主成分分析或因子分析等方法确定该病的中医各证和各自所属证候的基本构成，进一步探讨各证间的相关性或该病的基本证等。此类研究无论以中医的证或西医的病为切入点，均未摆脱获取统计量时的经验性和主观性，研究之初便留下不客观的缺陷；而对证的信息量

收稿日期：2005-10-17
修回日期：2005-11-21
* 科学技术部国家"十五"科技攻关计划项目(2004BA721A05)：以小儿肺炎为示范建立辨证规范及中医疗效评价方法体系的研究，负责人：王雪峰。
** 联系人：梁茂新，研究员，博士生导师，从事中药临床药理、中医基础理论和中药新药开发研究，Tel:024-31207279，E-mail:lmxin@126.com

化处理时所用的数学模型均是依靠经验判断预先选定的，不是基于中医证和证候的内在规律建立的专门的数学模型，不可避免存在一定的局限性。

如果把中医的证候看作是疾病状态下人体释放的共时或历时信息，那么传统的证则是借助疾病状态下每一个体传递的所有证候信息归类分析后做出的粗略的本质抽象(或诊断)。由于个体差异的存在，同一疾病会有或同或异的证候表现，于是基于不同的证候类别则有不同的证的属性；同为一证，证候构成会有异同；隶属某一病证的证候，各自对病、证诊断的贡献程度(或贡献率)也不尽相同；疾病的向愈，以证候的共时或历时性减弱、消失为特征，以证的解体为标志。诸如此类，借以实现的辨证规范和疗效评价指标的确认，只能把立足点放在潜在规律的揭示上，并且在样本量足够大时相关研究方可得以实施。于是，数据挖掘技术找到了新的用武之地。

需要说明的是，数据挖掘技术绝不排斥逻辑分析方法，在中医辨证规范研究方面更是如此。为了克服先前辨证规范存在的问题，本文以小儿肺炎中医辨证规范为示范，把逻辑分析方法和数据挖掘技术紧密结合起来，在完善并建立小儿肺炎中医辨证规范的同时，系统建立辨证规范的技术平台。

二、面向中医辨证规范的数据挖掘基本内容

面向中医辨证规范的数据挖掘主要包括病证所属症状体征的规范、西医疾病所属各证的辨证规范，另外我们还提出了一种客观反映小儿肺炎中医药干预的临床疗效评价指标和方法。

1. 病证所属症状体征的规范[1]

(1)病证所属症状术语规范；

(2)症状间逻辑关系规范；

(3)症状体征分级规范(三级量表的建立)；

(4)症状体征测量方法规范；

(5)症状体征诊断规范、体征(舌、脉象)诊断客观化等。

2. 西医疾病所属各证的辨证规范

(1)西医疾病所属各证基本构成规范；

(2)西医疾病所属各证构成比确定；

(3)病证所属症状的基本构成规范；

(4)西医疾病中医各证临床诊断标准规范；

(5)西医疾病分期、分类、分型、疾病发展阶段、病情与所属中医各证的对应关系的确认等。

3. 验证并提出一种客观反映小儿肺炎中医药干预的临床疗效评价指标和方法

(1)症状数据在整个病程中作为时间序列数据考虑时，对疗效评价的一个重要贡献；

(2)西医疾病中医各证基本演变趋势确定；

(3)临床疗效评价指标的确认。

三、适合中医辨证规范的数据挖掘关键技术

1. 粗糙集

粗糙集(Rough Set)理论是波兰数学家Z. Pawlak在1982年提出的一种分析处理数据的数学理论，该理论提出了一种处理不确定、不完整、不精确问题的新的数学工具。粗糙集理论能有效地分析和处理各种不完备信息，并从中发现隐含的知识，提示潜在的规律[2]。粗糙集的最大优点在于不需要问题所需处理数据之外的任何先验信息，如统计中的先验概率和模糊集中的隶属度，算法也易于实现。

在数据挖掘前，可以利用粗糙集进行症状属性的约简，保留与挖掘有关的重要症状属性；同时可以把粗糙集理论与关联分析、聚类过程相结合，实现算法的优化。

2. 关联分析

在数据库的数据挖掘中，关联规则就是描述在一个事物中物品之间同时出现的规律的知识模式。更确切地说，关联规则通过量化的数字描述物品甲的出现对物品乙的出现有多大影响。设$I=\{i_1,i_2,\cdots,i_m\}$是二进制文字的集合，其中的元素称为项。记D为交易T(Transaction)的集合，这里交易T是项的集合，并且$T\subseteq I$。对应每一个交易有唯一的标识，如交易号，记作TID。设X是一个I中项的集合，如果$X\subseteq T$，那么称交易T包含X。一个关联规则是形如$X\Rightarrow Y$的蕴涵式，这里$X\subset I, Y\subset I$，并且$X\cap Y=\Phi$。规则$X\Rightarrow Y$在交易数据库D中的支持度是交易集中包含X和Y的交易数与所有交易数之比，记为$support(X\Rightarrow Y)$，即

$\text{support}(X\Rightarrow Y)=\frac{|\{T:X\cup Y\subseteq T,T\in D\}|}{|D|}$[3]。

规则 X⇒Y 在交易集中的可信度是指包含 X 和 Y 的交易数与包含 X 的交易数之比，记为 confidence(X⇒Y)，即 $\text{confidence}(X\Rightarrow Y)=\frac{|\{T:X\cup Y\subseteq T,T\in D\}|}{|\{T:X\subseteq T,T\in D\}|}$。给定一个交易集 D，则挖掘关联规则的问题就是产生支持度和可信度分别大于用户给定最小支持度(minsupp)和最小可信度(minconf)的关联规则。

使用关联规则方法主要是为了挖掘症状与症状、病机与症状之间的关联关系，以发现症状的相关规律。通过关联规则挖掘算法，挖掘出频繁同时出现的症状-症状等组合，对发现中医学规律有着重要的作用。

3. 聚类分析

聚类(clustering)就是将数据对象分组成为多个类或簇(cluster)，在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象间差别较大。也就是说把整个数据分成不同的组，并使组与组之间的差距尽可能大，组内数据的差距尽可能小[4]。

在机器学习领域，聚类是无指导学习(unsupervised learning)。与分类不同，聚类不信赖预先定义的类和带类标号的训练。由于这个原因，聚类是观察式学习，而不是示例式学习。而分类是用户知道数据可分为几类，将要处理的数据按照分类分入不同类别也称为有监督学习，聚类分析和分类通常是一个互逆的过程。

基本的聚类方法：划分方法，层次方法，基于密度的方法，基于网格的方法和基于模型的方法。

传统的中医辨证分类是借助专家经验完成的综合性判断，因此这种分类法带有一定的主观性。通过聚类算法，挖掘出某些症状之间的相似度，形成几个簇，即多个证，从而确立一种客观的辨证规范。

4. 孤立点的处理

孤立点是样本集中一些远离其他数据点的数据，孤立点可能是表明一些特殊的情况信息，对这种孤立点要重点地研究和分析。孤立点问题的处理，应当从医学和统计学专业两方面去判断，尤其应当从医学专业知识判断。离群值的处理应在自态检查时进行，如果试验方案未预先指定处理方法，则应在实际资料分析时，进行包括和不包括离群值的两种结果比较，研究它们对结果是否不一致以及不一致的直接原因。在收集数据的过程中造成错误从而导致了孤立点的出现，对于这种孤立点应予剔除[5]。由于孤立点在整个数据集中只占很小的一部分，所以以前在处理数据的时候，往往会把这些孤立点忽略，从而导致一些重要信息的丢失。目前主要有两类发现孤立点的方法：主观发现法和客观发现法[6]。在主观方法中，直接把自己的知识融入定义参数的过程中，如“非常远”、“低密度”等，从这一点来说，主观发现方法对于不同用户可能会导致不同的结果，并且其可扩展性也比较差，而且这种主观发现的方法只适用于一些小数据集，对于大数据集来说，将是一个非常耗时的过程。在客观发现的方法中，模糊聚类算法将根据数据点的分布情况，自动地发现一些孤立点，取得了很好的结果。

通过聚类算法，可以发现症状的特异情况。这种特异情况表现为与正常取值有很大的差异，或者说远离了正常取值。比如一个人的病程期间，症状是否有特异情况；或者是某个症状在所有病例中是否有特异情况，这将有助于推动中医辨证向深度方面扩展。

四、面向中医辨证规范的数据挖掘系统(CMDM)的总体框架

针对中医辨证的特点，我们建立了面向中医辨证规范的智能平台，该平台以小儿肺炎为背景，建立中医的辨证规范和疗效评价的体系结构。平台为用户提供了一个良好的数据操作和挖掘环境，主要包括数据库工具，Brio 分析工具及数据挖掘工具。

整个框架兼容现有的数据挖掘任务及过程，增加了用户与数据挖掘过程及算法的交互性，构建了一个用户友好的交互式的数据挖掘环境。本框架将数据挖掘划分为 3 个层次：数据预处理层、数据挖掘层、模式表示层，支持简单类型和复杂类型(时间序列)的挖掘/分析，提供了与外部程序的开放式接口。系统框架见图 1。

五、面向中医辨证规范的数据挖掘系统的设计

下面我们以小儿肺炎为例，分析数据挖掘的实现过程，见图2。

1. 数据采集

数据采集是整个系统工作流程的第一步，也是最为重要的部分。因为只有采集的数据准确、完全、符合标准，下面的其他工作才可以展开。数据采集指从各采集点采集数据，消除噪声数据，最终放入数据库中。

(1)CRF表的数据录入。

主要功能是针对某位具体的受试者填写相应的CRF表数据。其中还包括受试者CRF表数据的提交，保存及修改等功能。CRF表录入共两次：由两位录入员分别录入CRF表数据。在CRF表录入过程中会有一些分析处理工作，如时间窗、服药依从性等，并可以根据合并用药及不良事件等判定受试人是否终止试验。

(2)CRF表数据打印。

根据用户所选择的受试者将其CRF数据打印出来。

(3)二次录入校验。

提供对两次录入数据进行对比，如果没有差别，可认为所输入数据是准确的，如果两次录入数据有差别，则需要核对手工填写的CRF表数据并进行适当的修改。

(4)CRF表的内容查询。

可以针对某个受试者进行CRF表数据内容浏览，也可以用报表的形式对所有受试者的数据进行浏览。

2. 数据质量监控

为了防止数据的错误以及在操作中的失误，对数据采取了多种有效性质量监控方案：

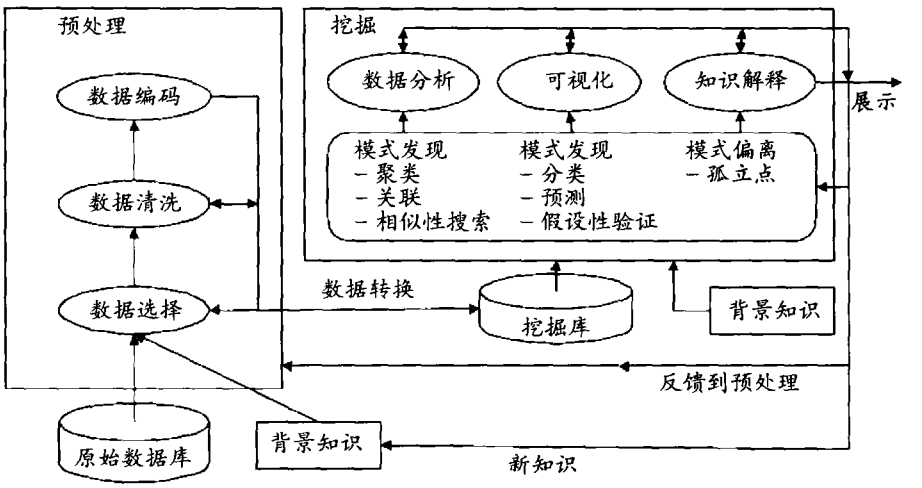

图1 CMDM系统数据挖掘框架

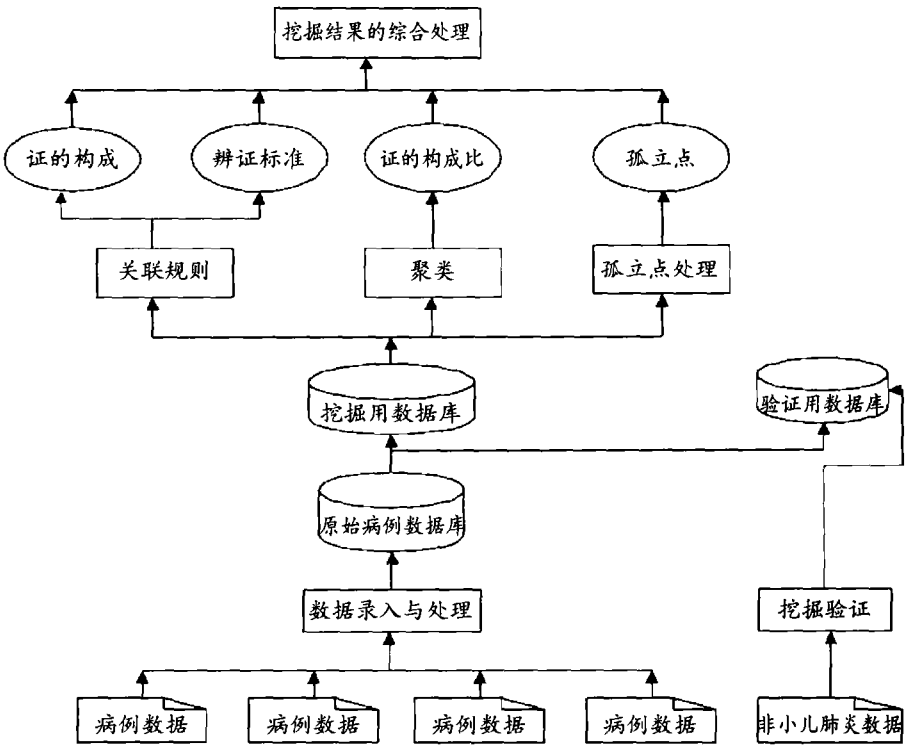

图2 小儿肺炎辨证规范的实现

(1)数据录入时的完整性检查,包括检查内容的完全性、一致性,制定统一检查标准。

(2)连续数据关联性检查,保证连续性特征。

(3)对孤立点数据进行特殊处理,通过主观干预或客观筛选方式实现。

(4)双人双盲。

(5)建立原始病例的扫描查询系统,保证数据真实性,减少人为因素的干扰。

3. 数据预处理

清洗数据,比如描述不完备的数据等。这里采用粗糙集的数据过滤方法、构造差分矩阵,筛掉多词一义、词义模糊、词义涵盖或交叉特征的数据记录。

4. 数据挖掘

(1)关联挖掘算法。

使用关联规则方法主要是为了挖掘证的分布,证的构成比,从而得到静态关联度。采用 FP-growth 或加权的 Apriori 算法找出频繁项目集,确定关联规则。

例如采用加权的 Apriori 算法:以权重累加值 a_weight 替代原算法中的交易数累积值。因此,在改进的算法中,支持度为 a_weight/total_weight。通过这种算法,可以挖掘出中医数据库中隐含的关联规则,并能充分利用中医学专家的专业知识,挖掘出更有意义的规则。

通过计算支持度确定证的分布和构成比情况(见表 1)。

(2)聚类挖掘。

在关联分析的基础上进行相似聚类,通过给定阈值,发现症状的构成以及症状的贡献率,得到症状所属证的聚类。通过遗传算法搜索和 K-means 局部优化相结合,按照最近基因匹配的交叉算子,在交叉过程中不断产生新个体,保证群体的多样性,减少了 K-means 算法的早熟现象,很好地解决了全局最优的问题[7]。K-means 局部聚类可以发现孤立点,特殊处理。

例:给定阈值:>=0.75,特异症状;在 0.5 和 0.75 之间,主症状;<0.5,次症状。

所有症状聚类后,会得到 N 个聚类。因为数据挖掘的结果不带有主观因素,所以这些聚类需要专家根据挖掘结果重新命名。

例如:XXXX 证(需人工设定名称)

特异症状:高热、咳嗽、气急、肺部啰音

主要症状:烦躁、微喘、面色红赤、口渴、咽红

次要症状:便干、尿黄、指纹紫滞。穷举出所有症状对证的贡献度,挖掘出症状量化值,放入各种证的量化表中,对逻辑确定的症状量化值进行修正。

(3)孤立点。

进行聚类的同时,可以发现特异症状,根据情况单独处理。

孤立点处理算法:

① 初始化算法参数 c,m,ε,t,其中 c 为聚类数目,m 为权重指数,ε 为一个很小的正数,设算法迭代记数 t=1,JH 为目标函数,$JH(t)=\mu_{ik}\cdot w_i^m$。

② 初始化样本集中样本隶属度参数。

③ 计算向量样本到类中心 v_i 距离 d_{ik},i=1,2,…,n,k=1…,c。

④ 重新计算每一样本向量的隶属度 μ_{ik},i=1,2,…,n,k=1…,c。

⑤ 计算得到每个样本向量权重系数 w_i,i=1,2,…,n。

⑥ 如果 | JH(t) -JH(t-1) | >ε,t ← t +1,转步骤 c

⑦ 否则,算法终止。

孤立点数据处理的结果见图 4。

从图 4 可以看出,当算法结束时,根据最终得到的权值,可以做出等权重线,在一定的等权重线之外

表 1 关联挖掘结果

id	item	supp	conf
1	高热,气急	34	90.1%
2	气促,发热	28	87.5%

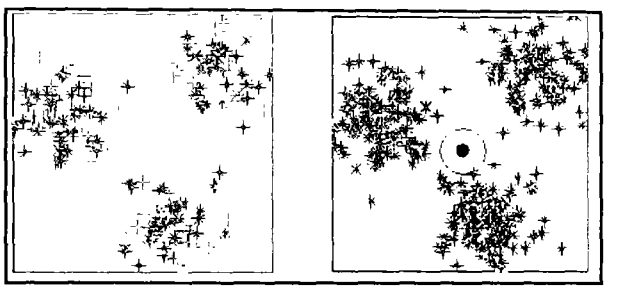

图 3 原始数据集和聚类结果集

(1)数据录入时的完整性检查,包括检查内容的完全性、一致性,制定统一检查标准。

(2)连续数据关联性检查,保证连续性特征。

(3)对孤立点数据进行特殊处理,通过主观干预或客观筛选方式实现。

(4)双人双盲。

(5)建立原始病例的扫描查询系统,保证数据真实性,减少人为因素的干扰。

3. 数据预处理

清洗数据,比如描述不完备的数据等。这里采用粗糙集的数据过滤方法,构造差分矩阵,筛掉多词一义、词义模糊、词义涵盖或交叉特征的数据记录。

4. 数据挖掘

(1)关联挖掘算法。

使用关联规则方法主要是为了挖掘证的分布,证的构成比,从而得到静态关联度。采用 FP-growth 或加权的 Apriori 算法找出频繁项目集,确定关联规则。

例如采用加权的 Apriori 算法:以权重累加值 a_weight 替代原算法中的交易数累积值。因此,在改进的算法中,支持度为 a_weight/total_weight。通过这种算法,可以挖掘出中医数据库中隐含的关联规则,并能充分利用中医学专家的专业知识,挖掘出更有意义的规则。

通过计算支持度确定证的分布和构成比情况(见表 1)。

(2)聚类挖掘。

在关联分析的基础上进行相似聚类,通过给定阈值,发现症状的构成以及症状的贡献率,得到症状所属证的聚类。通过遗传算法搜索和 K-means 局部优化相结合,按照最近基因匹配的交叉算子,在交叉过程中不断产生新个体,保证群体的多样性,减少了 K-means 算法的早熟现象,很好地解决了全局最优的问题[7]。K-means 局部聚类可以发现孤立点,特殊处理。

例:给定阈值:>=0.75,特异症状;在 0.5 和 0.75 之间,主症状;<0.5,次症状。

所有症状聚类后,会得到 N 个聚类。因为数据挖掘的结果不带有主观因素,所以这些聚类需要专家根据挖掘结果重新命名。

例如:XXXX 证(需人工设定名称)

特异症状:高热、咳嗽、气急、肺部啰音

主要症状:烦躁、微喘、面色红赤、口渴、咽红

次要症状:便干、尿黄、指纹紫滞。穷举出所有症状对证的贡献度,挖掘出症状量化值,放入各种证的量化表中,对逻辑确定的症状量化值进行修正。

(3)孤立点。

进行聚类的同时,可以发现特异症状,根据情况单独处理。

孤立点处理算法:

① 初始化算法参数 c,m,ε,t,其中 c 为聚类数目,m 为权重指数,ε 为一个很小的正数,设算法迭代记数 t=1,JH 为目标函数,$JH(t)=\mu_{ik}\cdot w_i^m$。

② 初始化样本集中样本隶属度参数。

③ 计算向量样本到类中心 v_i 距离 d_{ik},i=1,2, …,n,k=1…,c。

④ 重新计算每一样本向量的隶属度 μ_{ik},i=1,2,…,n,k=1…,c。

⑤ 计算得到每个样本向量权重系数 w_i,i=1,2,…,n。

⑥ 如果 | JH(t) -JH(t-1) | >ε,t ← t +1,转步骤 c。

⑦ 否则,算法终止。

孤立点数据处理的结果见图 4。

从图 4 可以看出,当算法结束时,根据最终得到的权值,可以做出等权重线,在一定的等权重线之外

表 1 关联挖掘结果

id	item	supp	conf
1	高热,气急	34	90.1%
2	气促,发热	28	87.5%

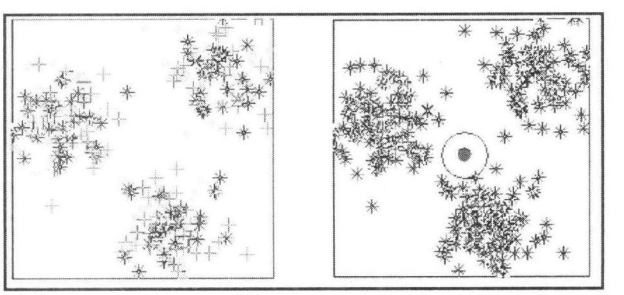

图 3 原始数据集和聚类结果集

法运行的时间以及占用的空间。当用户自行开发算法时,用户按照系统嵌入算法的要求,完成相关的接口,则算法的评价系统将能够对算法的耗用时间,占用空间做出评价,并且与系统中其它算法相互比较。

六、结 论

利用数据挖掘技术研究中医辨证,从病例中提取特征,寻找规律性信息,形成用数字描述和表达的中医症状信息,推动了中医辨证的规范化进程,为中医药信息的深度挖掘及现代化研究提供了一个平台。

当然,任何一种数据挖掘方法都需要先对数据进行评估、处理,形成数据模式的特征空间后,针对不同的问题,选择合适的数据挖掘算法和工具。模型的优劣在很大程度上依赖于数据的质量。为了确保数据反映物质的客观性和可靠性,在建模过程中数据的处理和识别必须融入专业知识,以便对可能产生的错误进行及时修正,提高数据挖掘的质量。中医药数据挖掘的对象是中医药领域中积累的海量数据,挖掘过程需要人机交互、多次反复,选择合适的数据挖掘方法,才能使数据挖掘技术在中医药规范化研究进程中开创新局面。

参考文献

1 梁茂新, 王雪峰, 董丹. 中医辨证规范所要解决的基本问题. 世界科学技术-中医药现代化,2005,7(3):18.

2 Padhraic Smyth..Breaking Out of the Black-Box: Research Challenges in Data Mining .Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD. CA, 2001.

3 Micheline Kamber.数据挖掘概念与技术. Jiawei Han.机械工业出版社,2001.

4 Ankerst. Human Involvement and Interactivity of the Next Generation's Data Mining Tools. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. CA, 2001.

5 R., Anand T. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. Advances in Knowledge Discovery and Data Mining . AAAI Press, Menlo Park, CA, 1~30.

6 Ronny Kohavi, Mehran Sahami. KDD-99 Panel Reports: Data Mining into Vertical Solutions. SIGKDD Explorations, 2000,1 (2), 55~57.

7 Fayyad U. M., Piatetsky-Shapiro G., Uthurusamy R. Summary from the KDD-03 Panel Data Mining: The Next 10 Years. SIGKDD Explorations, 2003, 5(2): 191~196.

Interactive Data-mining Framework Based on Normalization of Syndrome Differentiation in Traditional Chinese Medicine

Wang Bo, Zhang Bin, Wei Weijie and Ma Yuhui

(College of Information Science and Engineering, Northeast University, Shenyang 110004)

Liang Maoxin, Wang Xuefeng and Dong Dan (Liaoning Institute of Traditional Chinese Medicine, Shenyang 110032)

In the traditional normalization of syndrome differentiation in traditional Chinese Medicine the methods of empirical logic analysis and mathematic statistics are usually used and mathematic models are previously chosen instead of being determined by the inherent laws of respective syndromes, and as a result the objective truth of the normalization is in doubt. This article presents the study that on the basis of the "projects tackling key Scientific and technical problems in the 10th Five-Year Plan of China" an interactive data-mining framework with models of three tiers for the differentiation of syndromes of pneumonia in children is designed in TCM, which is able to promote the combination of data-mining technology with the modernization of traditional Chinese medicine via man-machine interaction while mining data involved.

Key Words: normalization of syndrome differentiation, data mining, pre-processing of data, mining algorithm, technical platform

(责任编辑:付建华 周应群,英文译审:秦光道)