· 中医药信息学 ·

基于中药功效的聚类分析

何前锋[1], 周雪忠[1], 周忠眉[1], 崔 蒙[2], 吴朝晖[1]

(浙江大学人工智能研究所, 浙江 杭州 310027; 中国中医研究院中医药信息研究所, 北京 100700)

关键词: 数据挖掘; 单味药; 功效; 聚类; 配伍规律

中图分类号: R289.1　　**文献标识码**: B　　**文章编号**: 1005-5304(2004)06-0561-02

笔者使用数理统计的方法, 采用较新的数据挖掘技术, 对中药库中的大量数据进行分析, 研究的焦点转向组成方剂的单味药, 从对单味药按照功效进行分类入手, 逐步由简到繁探索方剂的配伍规律。

1 数据挖掘方法

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的以及最终可理解的模式的非平凡过程。它是一门涉及面很广的交叉学科, 包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术。数据挖掘有以下 6 种不同的分析方法: ①分类(Classification); ②估值(Estimation); ③预言(Prediction); ④相关性分组或关联规则(Affinity grouping or association rules); ⑤聚类(Clustering); ⑥描述和可视化(Description and Visualization)。

2 聚类(Clustering)方法研究单味药与功效

2.1 聚类方法简介

聚类方法是对数据分组的一种方法, 它把相似的数据归在一个类里。聚类不依赖于预先定义好的类, 不需要训练集, 这是和普通分类的区别。

2.2 单味药与功效聚类分析建模

我们的目标是把具有相同或者相似功效的单味药归类在一起。因此首先需要建立中药数据库, 存贮单味药以及功效有关的信息。数据库的结构见表 1～表 3。然后是建模单味药之间在功效上的相似程度(见表 4)。

表 1　单味药表

字段名称	字段含义	字段类型	其它说明
DID	单味药编号	INTEGER	主关键字
MINCHEN	单味药名称	VARCHAR(100)	

表 2　功效表

字段名称	字段含义	字段类型	其它说明
GID	功效编号	INTEGER	主关键字
MINCHEN	功效名称	VARCHAR(100)	

列表示所有的不同功效, 行表示所有的单味药。药行与功效列相交的格表示药是否具有该功效。如药 1 与功效 1 相交

基金项目: 国家科学技术部科技基础性工作专项资金项目(2002DEA30042)

表 3　单味药与功效对应表

字段名称	字段含义	字段类型	其它说明
DID	单味药编号	INTEGER	主关键字(DID, GID)
GID	功效编号	INTEGER	

表 4　单味药与功效对应表

单味药	功效 1	功效 2	功效 3	功效 4	功效…
药 1	0	1	1	0	…
药 2	1	1	0	0	…
药 3	0	0	1	1	…
…	…	…	…	…	…

的格为 0, 表示药 1 不具有功效 1, 反之, 如果是 1, 则表示具有该功效。如药 1 与功效 3。

设 y 表示药, g 表示功效, 那么变量 y 可以表示成(g1, g2, g3, …gi…, gn), 其中 gi=0, 或者 1。显然, 变量 y 是 0, 1 值所组成的数据集合。

对于这样的数据集合, 可以用如下的方法来计算两变量之间的相似度(见表 5)。

表 5　两变量之间的相似度

		Yj		
		1	0	总和
yi	1	q	r	q+r
	0	s	t	s+t
	总和	q+s	r+t	p

相似度 d(yi, yj)=(r+s)/(q+r+s+t)。根据该表达式, 可知, 两单味药之间越相似, d 就越小, d 的取值在[0, 1]之间。

由于中药数据库中, 功效表的数据有 1 460 个不同的功效, 而对于每一单味药, 其功效大约在 4 个左右, 其中 3 个左右的功效最多。因此如果采用上式计算, 那么两单味药都不具有的功效的统计值, 即上表中的 t 值将变得很大, 而 r+s 基本上都很小, 所得的相似度就不能在[0, 1]之间分散开来。为了改变这种状况, 我们使用如下公式: d(yi, yj)=(r+s)/(q+r+s)。

以药 1 与药 3 之间的相似度计算为例(见表 6)。

表 6　药 1 与药 3 之间的相似度

		y3		
		1	0	总和
y1	1	2	2	4
	0	2		
	总和	4		

d(y1,y3)=(2+2)/(2+2+2)=0.66667

按照上述方法求得数据库中任意两单味药之间的 d(yi,yj)就可以得到相似度矩阵 D,D 是 n×n 的二维矩阵。

上述建模方法的假设是基于单味药在功效上是完全的,也就是说单味药要么具有该功效,要么不具有该功效,忽略了单味药在疗效程度上的差异。

需要分析的单味药比较多时,即 n 的取值比较大时,所得的相似度矩阵 D 将非常大,但是由于很多单味药之间的相似度为 1,也就是说两单味药之间不存在相同的功效,因此为了节约存贮空间以及加快计算的整个过程,可以考虑使用稀疏矩阵来存贮相似度矩阵 D。

3 分层聚类方法的应用

分层聚类按照聚类的顺序可分为两种,一种是自底向上的合并,一种是自上向下的分裂。合并聚类的方法是:根据相似度矩阵 D,找出离得最近的两个类合并为一类,于是只剩了 n-1 个类。相似度矩阵 D 由 n×n 维变成了(n-1)×(n-)维的矩阵 D1,接着再从 D1 中找到离得最近的两个类将其合并,……直到合并的类满足一定的条件为止,如合并成指定的 k 个类,或者合并的两个类之间的距离要小于某个指定的值。而自上向下的分裂聚类方法则与合并方法相反,最开始是一个类,一直不断分裂,直到满足一定的条件。

在本研究中,我们采用合并的聚类方法。算法详细描述如下。

输入:相似度矩阵 D,指定的两类之间最大合并距离(相似度)maxDistance;输出:类的合并状态 state,数据结构为:第 i 次合并的两个类(m,n),合并类的距离 distance,第 i 次合并后所有的分类状况。

步骤:①始化有关的数据结构,如将每个点初始化为各自成一类;②找到相似度矩阵 Di 中的最小值,以及所在的位置;③合并最小值所在位置的行下标与列下标所指示的两类;④更新第 i 次合并的有关数据,如 distance、类的状态、被合并类与其它类之间的距离,以及在相似度矩阵中删除被合并的类等;⑤判断合并的距离是否满足大于指定的 maxDistance,如果小于 maxDistance,则回到第 2 步。

在第3步合并类更新相似度矩阵D数据时,计算合并类与其它类之间的距离有多种方式,较为常用的方法是:①合并的两个类中与其它类的较小距离,作为合并后新类与其它类之间的距离;该种方法,聚类的结果比较集中,容易形成一个类就聚集了很多其它相关类的情况;②取两合并类距离间的平均值;③找出合并的两个类中与其它类的较大距离,作为合并后新类与其它类之间的距离;该种方法,聚类的结果比较分散,使得各个类都有机会选择与该类最近的类来合并。经实验表明,该种方法简单,而且比较适合本实验的数据,因此得到的结论相对比较有效。

4 结论

现中药数据库中收集单味药数据 3968 条,功效数据有 1460 条,单味药与功效对应数据 8871 条,全部数据的分析比较复杂。因此,在初步的研究中,首先抽样基础库中具有 4 个(包括 4 个)以上功效的单味药。含有 4 个以上功效的单味药共有 639 种,单味药对应功效数据共有 2822 条,在聚类算法中,指定的两类之间最大合并距离为 0.4,因此所得到的结论是聚类分析的部分结论。

5 探讨

对单味药的功效描述,现阶段的数据还很不完备,且不十分准确。如清热解毒、清热、清热解暑等功效之间的独立性如何?在现有的数据库中,这些功效是分开描述的,因此与其相对应的单味药之间的相似度的计算就需要进一步探讨。

分层聚类方法本身的缺点。分层聚类方法一旦把类归并,在接下来的聚类过程中,所属类就不再发生变化,因此不能做动态优化改变。在聚类合并后,合并类与其它类之间的距离的计算还可以采用其它方法。

对于具有功效数目比较少的单味药的分类,存在着一定的困难。如:如果某一单味药 1 的功效完全包含了另一单味药 2,那么归类时,应该把这两单味药归在一起,而经过相似度的计算,如果药 1 与药 2 相同功效的数量比较少,那么得到的相似度可能会比较大,在指定比较小的最大合并距离进行聚类后,两单味药并不能归属到同一类中。

聚类算法中应该满足的最大距离的确定,到底应该取多少才能使得所得聚类结果比较理想?在本文中取值为 0.4,对于结果的相似性要求比较严格。

所得到的结论的有用性需要经过多方的认证,也需要经过较长时间的确认。因此,通过算法的改进以得到更好的结论就仍假以时日。

6 展望

上述的研究只是配伍规律研究的初步探索,大量的工作需要进一步开展,比如对层聚类算法的优化与改进,以满足更大量数据的快速、准确分析;运用其它聚类算法对问题领域的研究;不单把单味药的功效属性作为聚类分析的特征,还要纳入更多维数的特征来进行聚类分析,如四气五味、归经、升降沉浮等;优化中药数据库,包括性能、数据的完整性、准确性、完备性等;对方剂进行聚类分析;建立结论的评价体系以加快研究进程等等。

在聚类基础上可以发现新的药对、药队,为旧药新用以及设计新方开辟道路;为运用计算机辅助的手段研究中医药方剂配伍规律奠定基础;聚类的结果可以作为中医药配伍规律系统运行的知识库;聚类药所具有的相似性,意味着它们之间存在着一定的关系,为中医药学理论研究提供客观的分析材料;为中医其它领域的研究提供参考,如药的有效化学成分的分析研究等。

(收稿日期:2003-09-25)

基于中药功效的聚类分析

作者： 何前锋， 周雪忠， 周忠眉， 崔蒙， 吴朝晖
作者单位： 何前锋,周雪忠,周忠眉,吴朝晖(浙江大学人工智能研究所,浙江,杭州,310027)， 崔蒙(中国中医研究院中医药信息研究所,北京,100700)

刊名： 中国中医药信息杂志ISTIC
英文刊名： CHINESE JOURNAL OF INFORMATION ON TRADITIONAL CHINESE MEDICINE
年，卷(期)： 2004,11(6)
被引用次数： 16次

引证文献(16条)

1. 申明金. 贾飞云. 柴震 化学计量学在中药研究中的应用[期刊论文]-山西医药杂志 2010(9)

2. 申明金. 柴震. 贾飞云 化学计量学在中药研究中的应用[期刊论文]-北方药学 2010(2)

3. 张林. 梁茂新. 宫俊. 董俊龙. 唐加福 基于数据挖掘技术的方剂配伍规律研究述评[期刊论文]-现代生物医学进展 2010(20)

4. 张媛媛. 莫芳芳. 张国霞 数据库技术在中医药领域的应用概述[期刊论文]-江西中医学院学报 2009(6)

5. 刘熙. 王崇骏. 叶亮. 范欣生 基于最大频繁项集的层次聚类方法[期刊论文]-广西师范大学学报（自然科学版） 2009(3)

6. 刘小溪 中药复方治疗抑郁症用药及组方规律探究[期刊论文]-吉林中医药 2009(5)

7. 顾向晨. 王怡. 张小鹿 慢性肾衰竭现代文献中药复方用药规律初探[期刊论文]-中国中西医结合肾病杂志 2008(7)

8. 张俊美. 王娜娜 数据挖掘技术在方剂文献研究中的应用现状[期刊论文]-甘肃中医 2008(1)

9. 尹爱宁 中医药数据共享平台的技术实现[期刊论文]-中国中医药信息杂志 2007(11)

10. 谢琪. 崔蒙. 潘艳丽. 范为宇. 何巍. 胡艳敏. 李凤玲. 苏大明. 童元元. 徐俊. 赵英凯. 张早华. 宿荣秦 基于海量信息分析的中医药文献评价方法与技术的思考[期刊论文]-中国中西医结合杂志 2007(8)

11. 李力恒 浅谈KDD技术在中医药领域的应用[期刊论文]-黑龙江科技信息 2007(14)

12. 张万水. 陈利国. 黄运坤. 陈咏梅. 王凤珍 数据挖掘技术及其在中医遣方用药规律中的应用[期刊论文]-辽宁中医药大学学报 2006(4)

13. 刘颖 活血化瘀类中药数据库知识发现研究——功效与药理关联关系的探讨[学位论文]硕士 2006

14. 李川 中医药数据挖掘系统TCMiner设计、实现与核心技术研究[学位论文]博士 2006

15. 吴朝晖. 封毅 数据库中知识发现在中医药领域的若干探索(Ⅱ)[期刊论文]-中国中医药信息杂志 2005(11)

16. 司国民 气郁证中医文献与证治研究[学位论文]博士 2005