

· 研究方法学 ·

医学文献中双向聚类分析和三维可视化方法与实践

甘叶娜¹ 李多多¹ 苑 艺¹ 张 颖² 刘建平²

摘要 双向聚类分析是一种数据挖掘方法,借助软件图形聚类工具包(gCLUTO)与书目共现分析系统(BICOMB),可实现文本数据的提取和结果的三维可视化表达。文献计量学作为开展文献数据分析、了解研究热点的关键手段,其实施依赖统计分析方法的应用。面对目前数据获取和利用模式的转变,传统的统计分析方法的局限性逐渐显露,对此,数据挖掘方法的应用有望给数据的处理提供新的途径。在医学研究领域内,把双向聚类分析引入文献计量学研究,将为海量文献数据的挖掘提供一种可行的方法。

关键词 双向聚类分析;文献计量学;医学文献;图形聚类工具包;书目共现分析系统

Bicluster Analysis and 3D Visualization Method and Practice in Medical Literature GAN Ye-na¹, LI Duo-duo¹, YUAN Yi¹, ZHANG Ying², and LIU Jian-ping² 1 Department of Tuina and Pain, First Affiliated Hospital, Beijing University of Chinese Medicine, Beijing (100700); 2 Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing (100029)

ABSTRACT Bicluster analysis is an important data mining method. By combining software Graphic Clustering Toolkit and Bibliographic Items Co-occurrence Matrix Builder, text data extraction and three-dimensional visualization of cluster analysis results can be realized. Bibliometrics, as a key means for carrying out literature data analysis and understanding research hotspots, the basis and core of its implementation lies in the application of statistical analysis methods. The transformation of the mode of data acquisition and utilization in the information age puts forward new requirements for statistical analysis methods. Traditional statistical analysis methods have gradually revealed their limitations when dealing with high-dimensional data, while the application of data mining methods provides a new way for the processing of complex data. In the field of medical research, the introduction of bicluster analysis into the research of bibliometrics will provide a feasible approach for the mining of massive literature data.

KEYWORDS bicluster analysis; bibliometrics; medical literature; Graphic Clustering Toolkit; Bibliographic Items Co-occurrence Matrix Builder

双向聚类分析(bicluster analysis)是探索研究中无先验知识的样本与变量间一致性的多元统计分析方法,依据某些能够度量样本与变量间相似程度的统计量,对大规模的样本与变量数据进行由小到大的分类,是生物信息学领域内最常用的数据挖掘方法之一^[1]。基于传统聚类分析原理发展起来的双向聚类分析方法,不仅继承了聚类分析处理海量数据的优势,还弥补了传统聚类分析在局部信息提取方面的不足^[2]。对以统计分析为基石的文献计量学来说,双向聚类分析在提高其处理高通量文献数据作用的方面具有较大潜

力。目前医学文献中双向聚类分析的应用较少,尚无关于此方面研究的介绍,本文将结合具体实例探讨双向聚类分析在医学文献计量中应用的可行性及结果三维可视化呈现的方法。

1 概述

1.1 聚类分析 聚类分析(cluster analysis)是简化数据结构最常用的探索性统计分析方法,可通过定义变量间的相似系数或距离来寻找多元变量间的相关性,其简明直观的数据呈现方式,在结果表达方面较其他多元统计分析方法更具优越性^[2]。聚类分析可根据研究对象的不同,分为针对样本的Q型聚类和针对变量的R型聚类,两者均属于单向聚类分析。若从分类的原理出发,主要包括了针对小规模数据开展的层次聚类,及可快速处理大规模连续型变量的划分聚类。无论是层次聚类还是划分聚类,都属于传统的非参数类数据处理方法,其作用的发挥主要依赖于

基金项目:国家自然科学基金重点资助项目(No. 81830115)

作者单位:1. 北京中医药大学第一临床医学院推拿疼痛科(北京 100700); 2. 北京中医药大学循证医学中心(北京 100029)

通讯作者:刘建平, Tel: 010-64286760, E-mail: liujp@bucm.edu.cn

DOI: 10. 7661/j. cjim. 20210315. 022

大样本理论,对空间样本量的需求常随数据维数呈指数变化。数据自身的空间样本量取决于数据分布的离散情况,通常数据离散度与数据维数成正相关,空间样本量与数据离散度成负相关,因而高维数据的空间样本量相对较少,这使得此类方法在处理高维数据时常难以保证统计分析的渐进性和稳定性。再者,此类方法更专注于全局数据间的相似性分析,常忽略了隐藏在整体数据背后的局部信息。这种整体性分析模式极易受到数据间无关属性的干扰,而直接影响聚类分析的结果。数据间的无关属性与数据的维数密切相关,通常数据的维数越高,数据间存在的无关属性就越多,因而,对高维数据来说,局部相关性的探讨会比整体的数据研究更具优势^[3]。

1.2 双向聚类分析 传统的聚类分析方法仅可实现基于样本或变量全局数据的单向聚类,无法获取样本与变量间的局部信息,在应对离散性较大的高维数据时效果不佳^[4]。因而,在以高维数据为特点的基因测序领域,有学者提出了同时针对样本和变量进行聚类,以探索两者间局部关系的双向聚类分析^[2]。该方法既保留了聚类分析的整体观,又补充了其局部观,可同时实现对数据的整体和局部把控,为聚类分析能有效应对由高维数据所带来的挑战提供了途径。聚类分析常用树状图的方式从聚类类别和类间相对距离这两个方面,对聚类的结果进行可视化呈现,即从研究的变量或样本出发,沿着相对距离由小到大的方向,将聚类类别的数量以横线的方式由多到少依次罗列,并以竖线表示类间的相关关系,最终将所有数据合并为两类,该图对聚类的全过程及类间关系的表现形式类似于谱系,因而亦被称为谱系图。此图极大提升了聚类分析结果的可解释性,根据研究需求选择合适的类间距离即可获得相应类别特征下的聚类结果。不同于仅可记录样本或变量单一方向频次分布情况的单向聚类分析,双向聚类分析还可获得两个方向组合后的频次数据,并将此数据以可视化的形式呈现,优化了分析的结果,提高了分析的效率,拓展了聚类分析在高维数据中的应用,现已成为基因分析、文本聚类、推荐系统等诸多领域内的关键技术^[2-4]。

1.3 医学文献计量 随着现代科学技术的发展,数据库的信息规模已发生了从量到质的转变^[3]。作为医学数据库的一个重要组成部分,医学文献的数量正急剧上升,并且还在以更高的数量级继续增长。在大数据背景下,要实现对医学文献的高效管理,数学、统计学等方法的引入就显得尤为重要。文献计量学的形成和应用,便为此趋势下文献数据挖掘和定量分析

工作的开展提供了思路^[5]。文献计量学借助其内在的统计与推理特点,可对文献及其计量特征进行系统描述、概括、评价和预测,极大提高了研究者对文献数据的提取和应用能力。该方法最早被运用于情报领域内的相关研究,随后被广泛应用于医学信息领域,为临床工作的开展提供了大量有价值的信息^[5,6]。其中,统计分析方法的应用是文献计量学发挥作用的基础和核心^[5],把善于进行数据挖掘的统计分析方法引入,将有助于其适应信息化时代下数据模式的转变,充分发挥量化知识体系的作用。

1.4 双向聚类分析与医学文献计量 医学文献作为一种文本数据,从中所获取的信息常以关键词的形式进行呈现,这些关键词正是进行文献分类和信息检索的关键所在,但同一关键词在不同文献中的分布情况存在一定差异性,同一文献中不同关键词的使用情况又具有多样性,这使得由文本数据聚类所获矩阵的维数高且离散度大,因此从全体关键词出发进行文献的整体相似性度量,常可能出现文本数据聚类效果不佳的情况^[6]。而双向聚类分析,通过探索文献与关键词的局部对应关系,可快速获取文本相关性的聚类分析结果,并可利用高频关键词对文本的主题进行预测,对包括医学文献在内的诸多文本数据挖掘具有较大的潜能。有学者对双向聚类分析的文本聚类情况进行研究,发现其效果明显优于单向聚类分析,且聚类结果的准确性高,尤其当应对维数越高的数据时其分析的效率也越高^[7]。对强调统计分析影响的医学文献计量学来说,双向聚类分析的应用,为其能更好的应对大规模文献数据提供了一种可行的方法。

2 图形聚类工具包联用书目共现分析系统实现医学文献双向聚类分析及三维可视化 双向聚类分析需先依据样本与变量的相关信息建立数据矩阵,再从矩阵中提取具有相似特征的子矩阵,以进一步深入挖掘数据中潜在的局部信息,高效数据处理软件的应用可大幅降低该过程中数据计算的复杂程度,并实现结果的三维可视化呈现^[1]。本文将结合具体实例,介绍图形聚类工具包(Graphic Clustering Toolkit, gCLUTO)与书目共现分析系统(Bibliographic Items Co-occurrence Matrix Builder, BICOMB)联用实现医学文献双向聚类分析及三维可视化的方法。BICOMB是一款由中国医科大学信息管理与信息系统(医学)系开发,基于文献数据库中书目信息共现关系进行文本挖掘的免费文献管理软件,具有基本的文献计量学分析功能,可对文献数据进行共现分析及可视化表达,并为进一步的聚类分析、关联分析和科学计量分析奠定

基础^[6, 7]。gCLUTO 是目前进行双向聚类分析, 并实现结果三维可视化最简便易行的工具, 其可通过消除所纳数据本身的不确定因素, 以提高聚类结果的稳定性及科学性^[8]。本文先利用 BICOMB 对文献数据进行提取, 再借助 gCLUTO 开展数据分析, 并以棋盘图和山峰图的形式导出结果。

文中所用软件为 BICOMB-2.0 和 gCLUTO-1.0, 可分别通过下方的网址获取 <https://202.118.40.32/dmi/nr.jsp?urltype=news.NewsContentUrl&wbtreeid=2161&wbnewsid=4589>; <http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview>。由于篇幅的限制, 本文不再赘述软件的具体操作流程。

本文所用数据及参数均只作为举例说明, 读者可根据研究需求选择相应的数据或参数。此处所选数据均来源于中国知网 (CNKI) 数据库, 以 “SU= ('HIV' + '艾滋病' + 'AIDS' + '获得性免疫缺陷综合征' + '爱滋病') * '中西医结合' ” 为检索式, 检索 CNKI 数据库中所有相关文献 (检索时间为 2020 年 11 月 22 日), 共获得 127 篇文献, 将所有文献以 “.xml” 格式导出, 可获得名为 “CNKI_result.xml” 的文件, 此文件是一个以 “树” 状结构存储数据的纯文本文件, 对文件中数据的提取操作, 可看作是对 “树” 上指定 “节点” 部位 “叶子” 的收集过程, 本文借助 BICOMB-2.0 来完成此过程。设定 “关键词” 为关键字段, 获取当 “109>频次阈值>2” 时的 “词篇矩阵” 和 “共现矩阵”。本文中, “共现矩阵” 展现的所有数据均表示关键词间两两组合出现的频次, 根据统计结果中各关键词出现频次的大小, 按 “从左到右” “从上到下” 的顺序分别在 “行” 和 “列” 上排布, 归属于同一关键词下的 “行” “列” 数据分布相同, “行” “列” 数据的交汇点代表了该关键词出现的最大频次, 所有关键词出现的最大频次数据在矩阵中呈现出 “左上 - 右下” 的分布规律, 如表 1 所示 (此处仅展示局部数据分布特点)。

“词篇矩阵” 亦可称为 “词频矩阵”, 其以 “行” 代表某关键词在各篇文献中的使用情况, 以 “列” 代

表 1 “共现矩阵” 局部数据分布特点

关键词	中西医结合	艾滋病	中医药	HAART	免疫重建
中西医结合	109	97	14	24	13
艾滋病	97	103	33	19	2
中医药	14	33	45	9	1
HAART	24	19	9	39	7
免疫重建	13	2	1	7	14

注: 此处所列 “中西医结合” “艾滋病” “中医药” “HAART” “机会性感染” 均为关键词

表该篇文献中各关键词的使用情况, 不同于 “共现矩阵” 的是, “词篇矩阵” 仅以 “1” 和 “0” 两种形式, 分别记录关键词的 “出现” 与 “未出现” 情况, 如表 2 所示 (此处仅展示局部数据分布特点)。

无论是 “共现矩阵” 还是 “词篇矩阵”, 均展现了关键词出现频次的分布情况, 但两者各有侧重, “共现矩阵” 表示了关键词间两两组合出现的频次数据, “词篇矩阵” 则表示了关键词在文献中出现的频次数据。

利用 gCLUTO-1.0 可对上述 “词篇矩阵” 和 “共现矩阵” 进行双向聚类分析, 此处以重复二分法为例对聚类结果进行解读。从棋盘图着手, 通过观察图中与行和列对应颜色的深浅分布可了解矩阵中频次的大小情况。在 “词篇矩阵” 中, 为避免出现频次较低的关键词被忽略, 常用较深的颜色进行表示, 而在 “共现矩阵” 中, 颜色较深者常表示出现频次较高的两个关键词组合。图 1 和图 2 分别展示了由 “词篇矩阵” 与 “共现矩阵” 所获的棋盘图, 在 “词篇矩阵” 的棋盘图中, 以横向谱系图表示文献的随机数字编号, 以纵向谱系图表示关键词; 在 “共现矩阵” 的棋盘图中, 横向和纵向谱系图均表示关键词。

文中对聚类结果的另一种表达方式是山峰图, 此图主要以山丘的形式进行数据呈现, 每个山丘的形状均为一条高斯曲线。通过观察山丘的高度、体积和颜色, 可对每个类内的数据分布进行粗略估计, 其中山丘的高度和体积分别与类内相似性及类群所含对象数量成正性相关, 峰顶的颜色代表了类内标准差的情况^[1]。

表 2 “词篇矩阵” 局部数据分布特点

关键词	0000000001	0000000002	0000000003	0000000004	0000000005
中西医结合	1	1	1	1	0
艾滋病	0	1	1	1	1
中医药	0	0	0	0	0
HAART	0	0	1	0	0
免疫重建	0	0	0	0	0

注: “0000000001” “0000000002” “0000000003” “0000000004” “0000000005” 均为文献的随机数字编号; HAART 为高效抗逆转录病毒疗法 (highly active antiretroviral therapy)

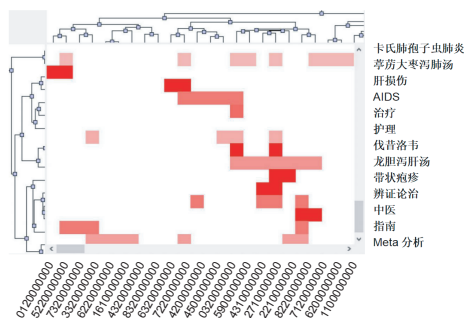
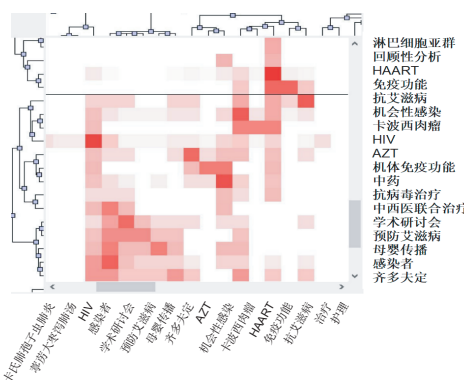


图 1 “词篇矩阵”对应的棋盘图



注: HIV 为人类免疫缺陷病毒 (human immunodeficiency virus);

AZT 为艾滋病防护药 (azidothymidine)

图 2 “共现矩阵”对应的棋盘图

通常,类内标准差越小,则类内相似性越高,类内的数据分布越集中,所呈现的山丘性状就越瘦高,峰顶的颜色也越靠近红色,反之,则山丘的性状趋于平坦,峰顶的颜色更靠近蓝色。需注意的是,山峰图中可视化的山丘地形是各条高斯曲线叠加的结果,因而,山丘所展现的高度往往会超出其类内数据的实际对应高度,对类内数据间相似情况的判断仅需依据峰顶的颜色,其他区域的颜色混合只是一个平滑过渡的效果^[4]。可视化山峰图的形状常取决于所选聚类方法的种类,在 gCLUTO 软件中共提供了“Repeated Bisection (重复二分法)”“Direct (直接聚类)”“Agglomerative (凝聚聚类)”和“Graph (图形聚类)”4 种聚类方法^[8]。结合谱系图分布特点,上述 4 种聚类方法又可分别对应理解为平均连接法、最短距离法、最长距离法和重心法。通常,由重复二分法所得聚类结果,其类间对象分布较均匀,而图形聚类的分布则会有明显的重心偏向。在直接聚类中,类间相对距离越小则类间分布越均匀,凝聚聚类则相反,类间相对距离越大时类间分布越均匀。虽然各种聚类方法均可得到相应的山峰图,但具体方法的选择还应遵循山峰图分布均匀的原则。图 3 和图 4 分别展示了与“词篇矩阵”和“共现矩阵”相对应的山峰图。

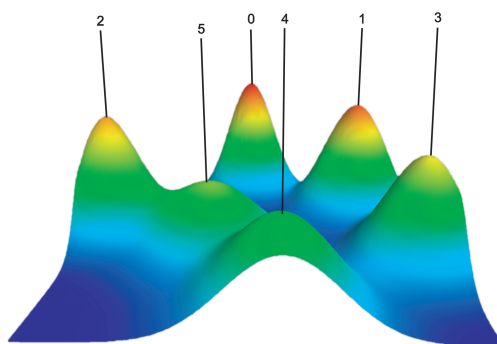


图 3 “词篇矩阵”对应的山峰图

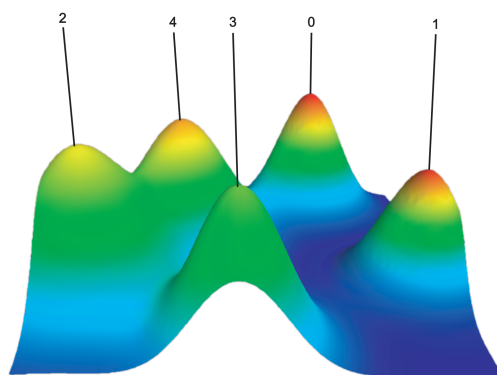


图 4 “共现矩阵”对应的山峰图

从“词篇矩阵”的山峰图来看,每个山丘均代表聚类结果中的一类,从“0”开始以自然数顺序标注于山丘的上方,通过对本研究所纳数据进行多次试验,最终选定聚类数为 6,得到如图所示的山丘分布,读者可根据需求自行调整聚类的数量。图中多个山丘联合在一起便组成了类群,此处可直观显示出 4 个类群,其中以 2、4、5 三个类别的类间聚类相似程度较为一致,构成了总体区域内最主要的一个类群。从山丘的高度来看,0、1、2 类相对较高,说明此三类的类内相似性相对较好。从体积上来看,1、4、5 类相对较大,说明类内所含对象数量相对较多。从峰顶颜色来看,0、1 类峰顶呈红色,说明类内偏差较小。综合这些特点可知,在 6 个聚类类别中,以 1 类的类内相似性高、所含对象多、分布集中,提示其是目前研究相对热门的内容,若要进一步提炼 1 类内受关注的研究方向,还可结合聚类结果中各类内对象的主要贡献情况进行判断。同理,从“共现矩阵”的山峰图中可看出,区域总体内共包含了 5 个山丘,4 个类群,以 2、4 类所含对象数量最多,0 类的偏差最小,类内相似最好。

3 讨论 数据化时代加快了大型权威医学数据库建设的步伐,使医学文献的管理朝着更全面、规范的方向发展,为医学文献计量工作的开展奠定了基

基础,大幅提升了分析结果的科学性。医学文献计量的研究对象主要以关键词的形式呈现,关键词是对核心研究内容的直观概括,常作为文献检索的指示性标志,以关键词为桥梁开展医学文献计量分析,能使分析结果更具针对性、代表性和可比性^[9]。出现频次是反应关键词使用情况的重要数据,其高低可直接展现研究的现状,通过对不同属性关键词的出现频次进行统计,即可从整体上把握研究的进展,并为后续研究工作的开展提供有价值的决策信息^[7]。通常,出现频次较高的关键词可反应出研究的热点,出现频次较低者则指明了研究的薄弱点,提示此方面的研究工作有待加强。双向聚类分析的聚类目标也是关键词,通过联用软件 gCLUTO 和 BICOMB,可获得反应关键词频次数据的共现矩阵和词频矩阵,依据矩阵信息,可将聚类分析的结果以棋盘图和山峰图的形式展现,在两个棋盘图中,分别以深色提示研究的热点及研究中需加强的内容。在两个山峰图中,研究的热点常隐藏在高度较高、体积较大且峰顶颜色偏红的山丘中,通过观察山丘的位置分布可对研究内容间可能的合作方向进行预测^[8]。通过分析本文所纳数据可知,采取当下临床公认针对 HIV 感染最有效的 HAART 疗法和近年来日渐被关注的中药干预手段来调节患者机体免疫功能的相关研究是目前中西医结合在艾滋病领域应用所关注的重点内容,由于本文仅例举了部分数据,尚不能展现该领域应用的全貌,若要深入探索还需借助更多的研究数据。

gCLUTO 的应用是开展双向聚类分析、实现三维可视化的关键环节,由其所呈现的棋盘图,不同于单向聚类分析中所使用的谱系图,除可反应样本间或变量间的聚类分析结果外,其还可通过构建两个方向间的数学模型,展现出样本与变量间、变量与变量间的双向聚类分析结果。山峰图是 gCLUTO 所独有的可视化表达,根据目标聚类数引入合适的山峰图,即可进一步提炼出棋盘图中的聚类特征,使得聚类分析结果的呈现更加直观,有利于研究者快速提取聚类结果中的有用信息^[7, 8]。软件中多种交互式聚类结果可视化方法的运用,不但丰富了聚类分析结果的呈现形式,还显著提高了结果的可读性,为双向聚类分析的推广提供了有力的技术支持。

双向聚类分析与医学文献计量分析在研究特点上具有相通性,将双向聚类分析用于医学文献计量,既可满足文献计量应对海量医学数据的统计需求,又可充分发挥文献计量对热点的预测优势。简便高效的数据处理软件应用及直观可视化的数据呈现方式,大幅

提升了分析工作的效率和研究结果的可推广性。然而,双向聚类分析的应用也存在一定的局限性^[7-10],其对数据的处理主要基于样本或变量间的相关性,虽然在进行局部数据挖掘时能发挥较好的作用,但从相关性着手的数据分析,会使研究的范围局限于具有相同属性的数据类型,而使不同属性数据类型间的关联性被忽略。由棋盘图和山峰图所展现的聚类分析结果,仅能粗略反应类间的相对数量关系或类内成分的相对比例,而数据的具体分布情况还需结合矩阵中的相关信息才能获得。

双向聚类分析的应用,改变了传统的整体分析观念,通过寻找数据间的相关性来缩小研究的范围,使得研究范围内的信息具有某种相似性,依据这种相似性进行数据挖掘和信息提炼,极大提高了医学文献计量的数据处理与信息利用能力,将成为大数据时代下对医学文献进行热点研究的一项重要工具。尤其对以“辨证论治”为核心指导思想,强调个性化诊疗的中医临床研究来说,双向聚类分析的引入将为该领域内文献计量研究工作的开展提供一种可行的途径。

利益冲突:无。

参 考 文 献

- [1] 方匡南,陈远星,张庆昭,等.双向聚类方法综述[J].数理统计与管理,2020,39(1):22-34.
- [2] 冯艳霞,陈艳楠.基于基因表达数据的双向聚类算法的综述[J].科技风,2017,(26):56.
- [3] 章蓉,陈谊,张梦录,等.高维数据聚类可视分析方法综述[J].图学学报,2020,41(1):44-56.
- [4] Kung SY, Mak MW, Tagkopoulos I. Symmetric and asymmetric multi-modality biclustering analysis for microarray data matrix[J]. J Bioinform Comput Biol, 2006, 4(2): 275-298.
- [5] 谢琪,崔蒙,潘艳丽,等.基于海量信息分析的中医药文献评价方法与技术的思考[J].中国中西医结合杂志,2007,27(8):753-756.
- [6] 姜涛,赵俊玲.医学文献计量分析的研究现状探讨[J].河北医学,2014,20(12):2141-2143.
- [7] 姚强,张研,张士靖.双向聚类在文献计量学中的应用初探——以医院绩效评价为例[J].情报杂志,2012,31(3):54-59.
- [8] 程则续,柯海涛.基于 BICOMB 与 HCA 的我国空间数据结构的文献计量研究[J].造纸装备及材料,2020,49(2):242,256.
- [9] 李睿.基于文献计量和共词分析的国内开放获取研究热点探析[J].图书馆学刊,2020,42(1):99-105.
- [10] 任艳玲,董平,郑洪新.多元统计分析在中医药研究中的应用与思考[J].中国中西医结合杂志,2002,22(11):860-862.

收稿:2020-12-07 在线:2021-04-12

责任编辑:赵芳芳