

双向聚类方法在医学文本挖掘中的实现过程及应用意义探讨^{*}

杨 颖

李子政

曹 姝 崔 雷

(中国医科大学图书馆
沈阳 110122)(中国医科大学附属第四医院
沈阳 110032)(中国医科大学图书馆
沈阳 110122)

〔摘要〕 介绍双聚类概念、基本原理以及相关研究、应用情况, 阐述双聚类使用工具和方法, 通过案例分析双聚类在医学文本挖掘中的具体实现过程, 提出双聚类在数据挖掘方法中的应用前景, 为相关研究提供参考。

〔关键词〕 数据挖掘; 双聚类; gCLUTO; PubMR

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.05.007

Discussion on the Realization Process and Application Significance of Bi-clustering in Medical Text Mining YANG Ying, Library of China Medical University, Shenyang 110122, China; LI Zizheng, 4th Hospital of China Medical University, Shenyang 110032, China; CAO Shu, CUI Lei, Library of China Medical University, Shenyang 110122, China

〔Abstract〕 The paper introduces the concept, basic principle, related study and application situation of bi-clustering, expounds the application tools and methods of bi-clustering, analyzes the specific realization process of bi-clustering in medical text mining through cases, and puts forward the application prospect of bi-clustering in data mining method, so as to provide references for related study.

〔Keywords〕 data mining; bi-clustering; gCLUTO; PubMR

1 引言

大数据时代, 利用数据挖掘技术抓取有价值信息具有重要意义。书目文献数据库中可用于文本挖

掘共现分析的字段主要有论文作者、机构、引文、关键词或主题词, 其中引文的共现分析可以进一步分解为作者同被引、期刊同被引等。最早开展共现分析的是高被引论文的同被引聚类分析, 另一条研究主线则是对主题词或关键词的共现分析。根据这

〔收稿日期〕 2020-12-15

〔作者简介〕 杨颖, 副研究馆员, 硕士生导师, 发表论文 40 余篇。

〔基金项目〕 2020 年辽宁省教育厅青年科技人才“育苗”项目“高校双一流学科建设与提升途径研究——基于 WOS、ESI、InCites、NI 平台指标的规范化使用”(项目编号: QNRW2020006); 2020 年度中国医科大学图书馆、医学信息学院规划课题“高校图书馆助力双一流学科建设的知识情报服务与政策建议”(项目编号: CMUGY006); 辽宁省社会科学规划基金项目“大数据驱动下智慧化学科精准服务平台设计与构建”(项目编号: L20BTQ002); 2020 年中国图书馆学会阅读推广课题“医学院校图书馆阅读推广影响力系统评价研究”(项目编号: Y2020B02)。

些共现数据部分学者开展了对高被引论文、高被引作者和高频主题词的聚类分析并将分析对象分类。常用传统聚类方法包括系统聚类、快速聚类 (K-means), 都是单向聚类, 即只对行 (对象) 或者列 (属性) 进行聚类, 有一定局限性。例如有的类是扩散且互相渗透的, 不同算法得出结果不同, 传统聚类分析不适用于类群界限相互渗透的数据分析。另外单向聚类分析容易忽视系统多样性以及多元因素的非线性因素^[1]。因此需建立面向高维数据的挖掘方法, 产生双聚类算法。其基本思想是: 直观简单地确定数据矩阵的几个双聚类, 同一类别中的个体有较强相似性, 不同类别个体差异较大, 根据多个观测指标找出能够度量样本 (变量) 之间相似度的统计量, 以此为依据采用各种聚类算法将所有样本 (变量) 分别聚到不同类别中。其原理是: 简单确定数据矩阵的几个双聚类, 矩阵中的对象根据值标上相应颜色形成矩阵图像; 再重新排序行和列以聚集相似的行和列, 进而形成具有相似颜色块的图像。这些块具有相似表达值, 是行和列的子集, 因此是双聚类^[2]。双向聚类可以有效抓取高维数据中隐藏的局部特征信息。

2 双向聚类在数据挖掘中的应用综述

2.1 双聚类思想发展概述

双聚类思想最早由 Hartigan 于 1972 年提出: 将对象在不同属性下的取值看作一个矩阵, 根据数据矩阵的对象和属性分别从行和列两个维度同时聚类, 由此找出其中满足条件的各个小矩阵, 即由对象子集和属性子集组成的聚类。2000 年有研究^[3]应用双聚类算法模型挖掘基因表达数据, 之后 Mirkin^[4]提出“盒式聚类”、“双向聚类”等概念。

2.2 相关研究应用

通过调研中国知网数据库发现 400 余篇关于双聚类的中文文献。学科领域以计算机为主, 其次是医药领域 (包括临床医学、基础医学、公共卫生、口腔、护理、生物、中医、药学), 再次是图书情报领域及教育、商务领域。大部分研究关注双聚类

方法应用, 如在医学数据挖掘中, 有研究选择全球医学研究领域自然指数期刊为研究对象展开科学计量分析, 利用 gCLUTO 对其词篇矩阵进行双向聚类, 探索其研究前沿结构构成^[5]; 有研究利用共词分析挖掘国际护理信息学、乳腺癌相关酶、h 指数研究领域的研究热点^[6-9]。其在临床医学博士论文中也有所应用, 例如利用双聚类研究 SIRT6 的计量学特征与临床基础相结合^[10]。gCLUTO 双聚类除应用于图情领域的文献计量学外还应用于临床实践, 例如开展人类神经干细胞的主题趋势和知识结构、支架植入在胰腺疾病治疗中应用的研究趋势分析^[11-13], 在绝经后骨质疏松研究中定位主题趋势和热点识别等, 以上均是临床医生应用双聚类挖掘医学领域相关知识的实例。

3 双向聚类工具介绍

3.1 概述

经历几十年发展, 目前已有数 10 种双聚类算法和工具, 尤其在生物医学领域有着广泛应用, 以增强对高通量组学技术产生的大型数据集中分析能力。这些算法和工具应用于多种数据类型, 包括但不限于基因组、转录组、外显子组、表观基因组、表型和药理基因组。通用流行算法主要包括 SAMBA、ISA、BIMAX、QUBIC 和 FABIA 等。另外已经开发出几种工具, 如图形聚类工具包 (Graphical Clustering Toolkit, gCLUTO), R 包, Web 服务器等。GEMS 是基于 Gibbs 采样范例、用于基因表达挖掘的 Web 服务器, 而 biclust 和 QUBICR 是两个 R 包, 集成多种现有算法, 用以实现数据预处理功能以及结果的解释和可视化。本文主要阐述应用较多的 gCLUTO 和新开发的基于 R 包的 PubMR 两种资源获取和应用。

3.2 gCLUTO

聚类分析通常应用 SPSS、SAS 等统计分析软件。目前针对双聚类分析较常用的软件是 gCLUTO, 以用户友好的图形方式进行聚类并提供交互式聚类结果可视化方法。通过网址 <http://glaros.dtc.umn.edu/>

gkhome/cluto/gcluto/download 可下载 gCLUTO 并安装使用。实际的可执行文件必须在其所在文件夹内才可保证程序执行。

3.3 PubMR

中国医科大学崔雷和周晓北联合开发了新的数据挖掘共现系统 PubMR, 是 R 平台下的高效 PubMed 文本挖掘工具, 集合检索下载、解析抽取、基本统计、多维矩阵、论文相似、热点分析、概念识别和网络分析等功能。利用 R 语言嵌入双聚类程序包可以通过简单的 R 语言进行复杂的双聚类操作, 实现聚类矩阵可视化 (<https://github.com/xizhou/pubMR>)。

4 双聚类在医学数据挖掘中的实现过程

4.1 步骤

4.1.1 收集数据 选择数据库和研究方向、检索信息、下载数据, 对于共词分析选取 XML 格式; 对于引文分析下载引文记录。

4.1.2 数据预处理 对下载数据进行处理, 利用书目共现分析系统^[14] (Bibliographic Item Co-Occurrence Matrix Builder, BICOMB) 抽取高频词或高

频引文, 生成词篇矩阵 (引文-引证文献矩阵) 与共词矩阵 (高被引文献的共引文献矩阵)。

4.1.3 双聚类分析及可视化 在 gCLUTO 里输入数据矩阵, 对结果进行可视化分析, 结果更加直观、可读性强。选择合适类别进行双向聚类分析, 根据聚类报告中参数对照表、山峰图和可视化聚类结果选取效果最佳聚类。

4.1.4 结果解读 根据聚类参数对照表和可视化聚类结果以及山峰图进行结果分析。

4.2 双聚类在文献计量分析中的应用案例

4.2.1 主题词-高被引文献矩阵双聚类 以基本科学指标数据库 (Essential Science Indicators, ESI) 免疫学高被引论文为数据源, 在 PubMed 数据库下载文献 XML 数据, 利用 BICOMB 生成高频主题词-来源文献的词篇矩阵, 通过 gCLUTO 软件进行双聚类分析并制作山峰图, 挖掘该学科研究热点和研究前沿。在此基础上绘制战略坐标, 探索各个主题在整个领域所属地位和发展趋势。利用 UCINET 和 NetDraw 进行社会网络分析, 从微观角度探索主题内部之间联系程度。再将社会网络分析图谱和战略坐标融合, 展示出该领域宏观结构中微观主题的贡献度, 见图 1。

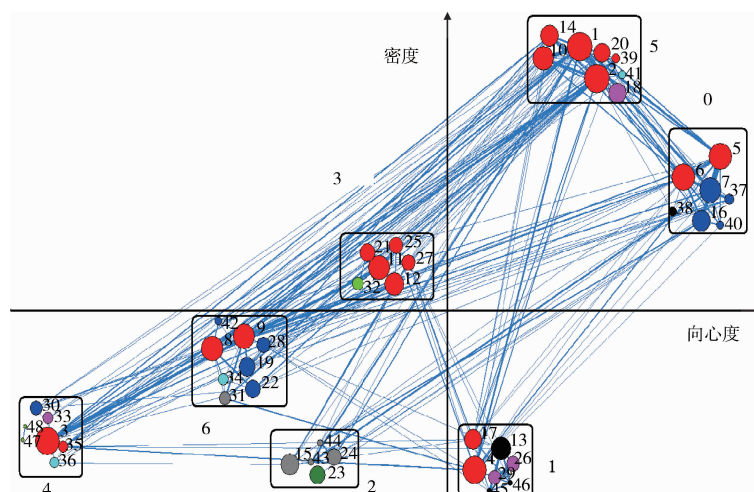


图 1 共词网络分析与战略坐标相互嵌入

4.2.2 以引文-被引矩阵为例 以 SCIE 数据库中某时间段的医学信息学领域论文作为数据源, 应用挖掘软件 BICOMB 统计得到共被引矩阵, 利用

gCLUTO 进行引文耦合分析和被引论文共被引分析, 得到知识基础结构和研究前沿。根据时间序列计算各类别新颖度和关注度, 绘制引文战略坐标图

谱^[15]。文献计量学中引文耦合分析和同被引分析两种统计方法常用于探索领域研究结构和前沿知识,其区别在于同被引方法关注被引用文献,展示本学科领域学科知识基础;引文耦合方法关注施引文献,代表本学科研究前沿。针对两种方法的研究都是单独进行而没有整合分析。双聚类恰好从同被引分析和引文耦合角度探索研究结构和前沿知识。双聚类用于引文分析与 Persson^[16]观点相符,同被引文献集的引证论文集表示研究前沿,同被引文献集表示知识基础,不同之处在于双向聚类引入同被引分析同时实现同被引聚类分析和基于高被引论文的引文耦合分析。

4.2.3 以作者-高频词矩阵为例 假设根据发表论文的主题对某领域学者进行分类,要计算每两位作者在所有主题上的相似性,而实际上如果两位作者在某个局部主题上一致即可证明二者比较相似,不需要在所有主题上都完全一致。以某学科领域中部分高产作者作为聚类对象,以其高频主题词作为分类属性,运用双向聚类方法对高产作者和论文关键词进行聚类并进行可视化表达,同时对各类别作者及对应关键词类别进行详细分析并总结其在该研究领域的特点,可以探索该领域高产作者及其研究特色。

4.2.4 以疾病-物质名共现矩阵为例 利用 PubMR 进行双聚类展示,在 PubMR 环境下输入以下语句,可以展示异质共现矩阵疾病-物质名的共现关系。

```
library (data.table)
library (tidyr)
options (datatable.prettyprint.char = 10L)
obj1 = data.table (PMID = obj@PMID, MH = obj@MH)
obj1 = obj1 %>% unnest (MH)
obj1 [, n: =. N, by =. (MH)]
obj1 <- obj1 [n >= 5,]
V <- crossprod (table (obj1 [, 1: 2]))
V <- crossprod (table (obj3 [, 1: 2]))
V [1: 2, 1: 2]
idr <- which (rownames (V) %in% meshtree [class = " D", mesh])
idc <- which (rownames (V) %in% meshtree [class
```

```
= " C", mesh])
V1 <- V [idr, idc]
x <- V1 [1: 40, 1: 40] x <- sweep (x, 1L,
rowMeans (x, na.rm = TRUE), check.margin = FALSE)
sx <- apply (x, 1L, sd, na.rm = TRUE) x <-
sweep (x, 1L, sx, "/" , check.margin = FALSE)
x [is.na (x)] <- 0
library (dendextend)
kr <- 6; kc <- 4
Rowv <- x %>% dist %>% hclust %>%
as.dendrogram %>%
set (" branches_ k_ color", k = kr) %>%
set (" branches_ lwd", 1.2)
Colv <- t (x) %>% dist %>%
hclust %>% as.dendrogram %>%
set (" branches_ k_ color", k = kc, value = c (" orange", " blue", " green", " red")) %>%
set (" branches_ lwd", 1.2)
cluster <- cutree (as.hclust (Rowv), k = kr)
clustab <- table (cluster) [unique (cluster [as.hclust
(Rowv) MYMorder])]
m <- cumsum (clustab)
m <- m [-length (m)]
heatmap (x, Rowv = owv, Colv = Colv, scale = "
none", add.expr = abline (h = m + 0.5, lwd = 3))
```

5 双向聚类在数据挖掘中的应用展望

5.1 双向聚类在科研中的作用

对于数据挖掘中的异质矩阵:高被引-关键词、同被引-引文、机构-关键词、作者-关键词、疾病-药物、疾病-基因、药物-靶向位点之间的二维关系无法用传统聚类方法来实现,需要用双向聚类挖掘科研中更多的信息与知识。

5.2 应用展望

双聚类应用于引文分析中是同被引分析和引文耦合分析的共现过程,可同时挖掘学科基础结构及学术前沿;应用于共词分析中可以探索双聚类在词篇分析中的应用,挖掘主题词关联性并在文献对比中挖掘新知识;在关键词-作者分析中的应用可以

解释学科研究者主要分布与研究方向；应用于机构-关键词分析可以揭示不同机构、地区的科研活动及研究方向；应用于疾病-药物分析中可以发现治疗疾病的有效药物；应用于疾病-基因分析中可以解释疾病与基因的关系；应用于药物-靶向位点可以挖掘药物作用的靶向位点；了解基因组信息与特定疾病之间的关联可用于发现新药。因此双聚类可以从不同角度阐释科研活动进展与趋势。

6 结语

双聚类方法应用于文献计量分析与评价可揭示科学研究者及科研活动的主要分布与研究方向，提出观测科学活动的新角度。双聚类方法渗透于各学科领域，深入挖掘高维数据中的有效信息，在数据挖掘中具有重要的理论意义及实践价值。随着大数据生成技术在生物医学和卫生保健信息学中的应用日益增加，其成本逐渐降低，在公共领域已使用大量生物学和临床数据集。一方面，双聚类在医学及生物学领域为确定新的治疗靶标、药物指示和药物反应生物标志物提供材料；另一方面，为数据挖掘方法带来更多挑战，包括检测疾病亚型、识别疾病生物标志物和基因信号性质以及基因药物关联。

参考文献

- 杨颖, 崔雷. 双聚类方法在同被引分析中应用的研究 [J]. 图书馆杂志, 2018, 37 (5): 67-73.
- Hartigan J A. Direct Clustering of a Data Matrix [J]. I JASA, 1972, 67 (337): 123-129.
- 唐小丽, 胡孔法, 陈峻, 等. 一种改进的双向聚类算法及其在销售分析中的应用 [J]. 计算机应用研究, 2007, 24 (1): 251-254.
- Bleuler S, Prelic A, Zitzler E. An EA Framework for Biclustering of Gene Expression Data [J]. IEEE Congress on Evolutionary Computation, 2004, 1 (161): 166-173.
- 杨颖, 许丹, 陈斯斯, 等. 基于自然指数刊文数据对全球医学研究领域热点的探析 [J]. 情报学报, 2019, 38 (11): 1129-1137.
- 李范, 李敏, 王丽, 等. 利用共词分析挖掘国际护理信息学研究热点 [J]. 医学信息学杂志, 2014, 35 (9): 48-53.
- 刘星吾, 王哲, 解夕黎, 等. 全膝关节置换术后血栓形成的数据挖掘及聚类分析 [J]. 中华医学图书情报杂志, 2018, 27 (10): 40-48.
- 宫小翠, 赵迎光, 安新颖. 基于双聚类方法的乳腺癌相关酶研究前沿 [J]. 中华医学图书情报杂志, 2016, 25 (2): 69-74.
- 方丽, 崔雷. 利用双聚类和突发检测算法探测学科前沿及知识基础的比较分析——以 h 指数研究领域为例 [J]. 情报杂志, 2015, 34 (2): 79-83, 88.
- 路铠宁. SIRT6 医学信息分析及其在肾癌中的作用 [D]. 沈阳: 中国医科大学, 2019.
- Wei W J, Shi B, Guan X, et al. Mapping Theme Trends and Knowledge Structures for Human Neural Stem Cells: a quantitative and co-word biclustering analysis for the 2013-2018 period [J]. Neural Regen Res, 2019, 14 (10): 1823-1832.
- Zhu X, Niu X, Li T, et al. Identification of Research Trends Concerning Application of Stent Implantation in the Treatment of Pancreatic Diseases by Quantitative and Biclustering Analysis: a bibliometric analysis [J]. Peer J, 2019, 24 (7): e7674.
- Zhou S, Tao Z, Zhu Y, et al. Mapping Theme Trends and Recognizing Hot Spots in Postmenopausal Osteoporosis Research: a bibliometric analysis [J]. Peer J, 2019, 25 (7): e8145.
- 崔雷. 生物医学数据挖掘: 步骤与工具 [J]. 中华医学图书情报杂志, 2017, 26 (3): 1-5.
- 杨颖, 许丹, 陈斯斯, 等. 高校图书馆推进 ESI 学科建设的情报服务 [J]. 中华医学图书情报杂志, 2019, 28 (3): 46-49.
- Persson O. The Intellectual Base and Research Fronts of JASIS 1986-1990 [J]. Journal of the American Society for Information Science, 1994, 45 (1): 31-38.