基于 LDA 主题模型和关联规则的冠心病方剂组方规律分析

李四海[1]，陈建国[1]，吕晓云[2]

1. 甘肃中医药大学，甘肃 兰州 730000；2. 兰州大学基础医学院中西医结合研究所，甘肃 兰州 730000

［摘要］**目的**：运用数据挖掘技术，分析中医治疗冠心病方剂的组方规律。**方法**：收集整理中医文献及 CNKI 中有关冠心病的方剂，建立方剂数据库。运用关联规则及 LDA 主题模型分别从药物关联性和方剂潜语义分析两个角度对冠心病方剂进行挖掘，分析冠心病方剂的组方规律。**结果**：收集冠心病方剂 356 首，涉及中药 318 味，药物频次统计结果表明冠心病方剂中的药物以补气药、活血化瘀药和理气药为主。LDA 主题模型共得到 3 个方剂主题，关联规则挖掘得到 12 个提升度较大的药对组合和药物关联规则。**结论**：中医治疗冠心病以益气、活血化瘀为主，兼顾理气祛痰，注重标本兼治。运用数据挖掘方法分析冠心病方剂组成，可以发现其组方规律，为临床冠心病治疗提供参考依据。

［关键词］冠心病；关联规则；潜狄利克雷分配；中药方剂；组方规律

［中图分类号］R541.4 ［文献标志码］A ［文章编号］0256-7415（2018）06-0047-04

DOI：10.13457/j.cnki.jncm.2018.06.012

Analysis on Composing Prescriptions Rules of Prescriptions of Coronary Disease Based on LDA Topic Model and Association Rules

LI Sihai，CHEN Jianguo，LYU Xiaoyun

Abstract：**Objective**：To analyze the composing prescriptions rules of prescriptions of Chinese medicine for coronary disease by data mining technology. **Methods**：Collected and arranged prescriptions of coronary disease in CNKI and literature of Chinese medicine，and established database of prescriptions. Association rules and Latent Dirichlet Allocation LDA topic model were used to respectively explore the prescriptions of coronary disease from the aspects of medicine association and latent semantic analysis of prescriptions，and to analyze the composing prescriptions rules of prescriptions of coronary disease. **Results**：356 cases of prescriptions of coronary disease were collected，including 318 kinds of Chinese medicine. The statistical results of medicine frequency showed that the medicines in the prescriptions mainly are qi-invigorating medicine，blood-activating and stasis-resolving medicine and qi-regulating medicine. Three topics of prescriptions were concluded by LDA topic model and 12 combinations of couplet medicines with higher upgrade degree and association rules of medicine were explored by association rules. **Conclusion**：Chinese medicine therapy for coronary disease focuses on tonifying qi and activating blood and resolving stasis，takes regulating qi and dispelling phlegm into consideration，and emphasizes that the treatment of the disease should attach importance to both the root cause and the symptoms. Through the analysis of the composition of prescriptions of coronary by the data method，the composing prescriptions rules can be found，which provides a reference for clinical treatment of coronary disease.

Keywords：Coronary disease；Association rules；Latent dirichlet allocation；Prescriptions of Chinese medicine；Composing prescriptions rules

冠心病是冠状动脉性心脏病的简称，是由于冠状动脉血管发生动脉粥样硬化病变而引起血管腔狭窄或阻塞，造成心肌缺血、缺氧或坏死而导致的心脏病，属于中医学胸痹、心痛、真心痛等范畴，胸痹、心痛的基本病机为本虚标实，根本原因在于心气鼓动无力，至痰浊、瘀血互结，内阻心脉，不通则痛，以气虚为本，血瘀为标。中医治疗冠心病能改善心肌缺血缺氧、保护心肌结构，明显改善患者临床症状，减少冠心病心绞痛复发频率，具有显著疗效[1]。

潜狄利克雷分配(Latent dirichlet allocation，LDA)模型是由 Blei DM 等[2]提出的针对文本集建模的三层贝叶斯概率主题模

［收稿日期］2017-11-15

［基金项目］兰州市科技计划项目（2015-2-70）；甘肃省中医方药挖掘与创新转化重点实验室开放基金项目(ZYFYZH-KJ-2015-010)

［作者简介］李四海（1972-），男，副教授，研究方向：数据挖掘，中医药信息化。

型，由词、主题、文本三层构成。由于主题数通常远小于文档维度，LDA 主题模型通过建立文本的概率主题分布，能够同时实现对高维文本的潜语义分析和低维表达，从而有效提高大型文本的聚类和分类效率。将主题模型中的文本、词、主题分别与中医药中的方剂、药物和方剂主治证型相对应，就能够方便地实现对中医药方剂的潜语义分析，得到方剂在证型空间的概率表示及所有药物被分配至不同证型的概率分布，通过对每个证型主题中概率取值较大的药物进行分析，能够发现不同证型在遣方用药方面的差异，进一步揭示方证之间的对应关系。近年来，LDA 主题模型被逐步引入中医药研究领域[3-5]，通过提取症状或方剂隐含主题，并将主题与证型相关联，为症 – 证及方 – 证相关性分析提供了新的研究思路。

基于数据挖掘方法分析冠心病方剂用药规律已有相关报道[6-8]，关联规则分析被广泛用于中医药领域，通过设置支持度和置信度可以发现药物和证型之间的知识和推理规则。关联规则分析的优点是方法简单、结果直观，缺点是仅仅通过统计药物出现频次分析药物之间的关联性，无法反映中医方剂的潜在语义。本研究收集冠心病方剂 356 首，涉及中药 318 味，建立冠心病方剂数据库，综合运用关联规则分析和 LDA 主题模型，分别从药物关联性及方剂主题隐含语义两个角度，对冠心病方剂进行挖掘，总结治疗冠心病方剂的用药规律，为中医临床治疗冠心病提供新的思路和依据。

1 资料与方法

1.1 处方来源与规范化 收集整理《金匮要略》、《太平惠民和剂局方》、《医林改错》、《广济方》、《冠心病良方大全》[9]中冠心病方剂及中国知网、万方数据库知识服务平台、维普中文期刊服务平台等在线数据库相关论文中所涉及的冠心病方剂，共得到有效方剂 356 首，建立冠心病方剂数据库，录入包括序号、方名、组成、用量、主治功效等方剂信息。为避免录入中药因炮制、别名不同对数据分析结果的影响，对药名进行了统一规范化处理。如天门冬录为天冬，麦门冬录为麦冬，瓜蒌子、瓜蒌皮、瓜蒌实、瓜蒌仁、全瓜蒌均录为瓜蒌，川郁金录为郁金，炒酸枣仁录为酸枣仁。由于甘草与炙甘草功效区别较大，将甘草、生甘草统一录为甘草，炙甘草单独录入。

1.2 数据挖掘方法与软件 使用关联规则算法和 LPA 主题模型对冠心病方剂进行挖掘。关联规则算法使用 WEKA3.8.1 中的 Apriori 算法[10]，LDA 主题模型则在基于 JAVA 语言编写的开源软件开发包 JGibbLDA 中实现[11]。

1.2.1 关联规则 关联规则挖掘方法是一种基于频次的统计分析方法，通过支持度、置信度及提升度等指标来度量规则的有效性及强弱程度。支持度的定义如下：$sup(A\Rightarrow B)=P(A\cup B)$。支持度是一个概率值，是同时出现药物 A 和药物 B 的方剂数占方剂总数的比例，支持度衡量了规则的有用性。置信度的定义为：$conf(A\Rightarrow B)=P(B\mid A)=\frac{P(A\cup B)}{P(A)}$。置信度是一个条件概率值，是在所有出现药物 A 的方剂中，同时出现药物 A 和药物 B 的方剂数所占的比例。置信度反映了规则的确定性。提升度的定义为：$lift=(A\Rightarrow B)=\frac{P(A\cup B)}{P(A)P(B)}=\frac{conf(A\Rightarrow B)}{P(B)}$。提升度用来度量药物之间的相关性，当 $lift>1$ 时，二者正相关；当 $lift<1$ 时，二者负相关；$lift=1$ 时，二者相互独立。

1.2.2 LDA 主题模型 LDA 主题模型是一种三层的贝叶斯产生式有向概率图模型，主要用于文本降维、聚类和分类。LDA 模型假设文档集包含多个主题，这些主题为所有文档所共享，每个文档由多个主题所构成，每个主题则由文档集中的词所表达。文档 – 主题和主题 – 词分布分别服从超参数为 α 和 β 的多项分布，α 和 β 均代表先验知识，α 反映了文档集中隐含主题之间的相对强弱，α 越大，主题越趋向于均匀分布；β 刻画了词在主题上的概率分布。α 和 β 的经验取值一般为 $\alpha=50/K$，$\beta=0.01$，其中 K 为设定的主题数。

2 结果

2.1 药物频次分析 纳入分析的 356 首治疗冠心病的方剂中共涉及中药 318 味，药物出现总频数 3 557 次，其中使用频数最高的是丹参，共使用 203 次，频数超过 30 次的药物如表 1 所示。结果显示冠心病方剂常用药物主要以补益药、活血化瘀药、祛痰药为主，兼顾理气药、化痰药、温里药的使用。既有益气养血之党参、黄芪、炙甘草、白术、人参，又有养阴之麦冬、五味子、白芍，安神之酸枣仁，活血化瘀之丹参、当归、川芎、赤芍、桃仁、红花、延胡索，化痰之瓜蒌、陈皮，温里之制附子、桂枝。

表 1 使用频次超过 30 次的药物

药物	频次	药物	频次
丹参	203	人参	62
黄芪	167	郁金	59
川芎	148	三七	55
炙甘草	145	枳壳	55
当归	116	薤白	52
麦冬	98	白术	51
桂枝	94	延胡索	50
茯苓	89	白芍	48
赤芍	80	柴胡	45
瓜蒌	78	酸枣仁	42
党参	75	陈皮	40
法半夏	74	细辛	40
红花	74	桃仁	39
五味子	70	葛根	33
制附子	63	淫羊藿	31

2.2 关联规则分析 在 WEKA3.8.1 中调用 Apriori 算法，设定支持度阈值为 5%，即只统计频次超过 18 次的药物之间的关

联性，Delta 为 0.05，置信度阈值为 0.9，共得到 23 条药物之间的关联规则。一般认为，提升度越大，药物之间的关联性越强，得到的关联规则越有价值[12]。对以上规则按照提升度的大小进行筛选，表 2 列出了提升度值较大的前 12 条药物关联规则。

表 2 冠心病方剂药物关联规则

关联规则	支持度(%)	置信度	提升度
川芎、当归、桃仁→红花	5.62	0.95	4.58
当归、桃仁→红花	6.18	0.92	4.41
丹参、法半夏、薤白→瓜蒌	5.06	0.95	4.32
黄芪、川芎、五味子→麦冬	5.06	1.00	3.63
丹参、川芎、五味子－麦冬	6.74	0.96	3.49
党参、五味子→麦冬	5.34	0.95	3.45
丹参、五味子→麦冬	11.80	0.91	3.32
川芎、白芍→当归	6.18	0.92	2.81
当归、赤芍、红花→川芎	5.34	1.00	2.41
黄芪、当归、红花→川芎	6.18	0.92	2.20
当归、红花、桃仁→川芎	5.62	0.91	2.19
桂枝、红花→黄芪	5.62	0.91	1.94

2.3 *方剂潜语义分析* 《金匮要略·胸痹心痛短气病脉证治》曰：“夫脉当取太过不及，阳微阴弦，即胸痹而痛。所以然者，责其极虚也。今阳虚知在上焦，所以胸痹、心痛者，以其阴弦故也。”阳微为本虚，阴弦为标实。虚性证候以血虚、气虚、阴虚、阳虚为主，实性证候以血瘀、痰浊、寒凝较为常见[13]。气虚血瘀证为冠心病最常见的证型。将冠心病常见证型与方剂隐含主题相对应，建立冠心病方剂的 LDA 主题模型，对方剂进行潜在语义分析。

LDA 主题模型的参数设置如下：主题数 K=3，α=50/K=16.7，β=0.05，迭代次数为 2 000。LDA 主题模型通过蒙特卡洛吉布斯采样算法，为每一味中药采集主题，当马尔可夫链收敛时，就得到了所有药味的主题分布，利用得到的药物主题分布可方便地计算出方剂主题和主题药物两个概率分布，最终将方剂映射到主题空间，3 个主题中概率值较大的前 10 味中药分别为：主题 1：川芎、黄芪、赤芍、瓜蒌、法半夏、红花、枳壳、三七、薤白、陈皮；主题 2：当归、麦冬、茯苓、党参、五味子、人参、白术、柴胡、酸枣仁、黄连；主题 3：丹参、炙甘草、桂枝、黄芪、郁金、制附子、延胡索、白芍、细辛、泽泻。将每一个主题与方剂主治证型相对应，则主题中概率值较大的药物即为该证型对应的核心药物。通过对主题中核心药物的分析，可以发现不同证型冠心病在用药方面的差异。图 1 显示了所有 318 味中药在主题 1 上的概率分布。结果显示，主题 1 上不同药物的概率值具有非常大的差异，说明 LDA 主题模型较好地提取了药物的语义。在概率值较大的前 10 味药中，黄芪为补气圣药，具有补气升阳、益卫固表的功效。赤芍养血和血，破血行瘀，《别录》言赤芍能“通顺血脉，缓中，散恶血，逐贼血”，川芎为血中之气药，辛香行散，温通血脉，上走下达，活血行气，行血之中以和血，行气之中以散郁。二者相伍，则行血而不破血，补血而不滞血，《本草汇言》谓“川芎，上行头目，下调经水，中开郁结，血中气药”，配伍组合，可除瘀血心痛。红花、三七活血化瘀止痛，瓜蒌涤痰散结，宽胸理气，调畅血脉，通达阳气，与薤白相伍，辛散能助阳气以行，苦降能涤痰散瘀，二药相用，涤痰之中能通阳，散瘀之中能通脉，走心窍而除痹，兼疗痰中有瘀、瘀中有痰之胸痹。枳壳、陈皮开提气结，理气宽胸。以主治证型作为方剂隐含主题，结合以上分析，可以推断主题 1 所对应的冠心病证候极有可能为气虚血瘀证[14~15]。

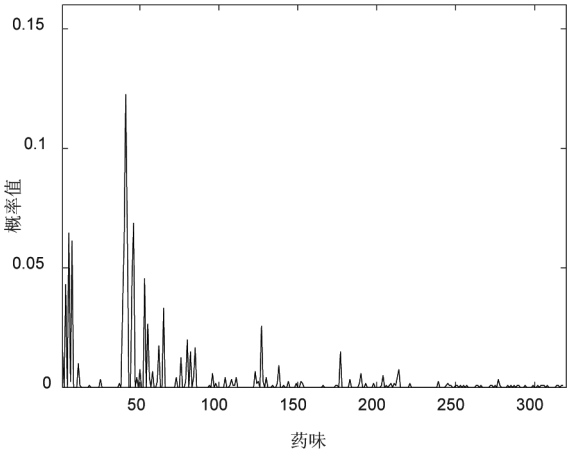

图 1 所有药物在主题 1 上的概率分布

3 讨论

本研究建立了冠心病方剂数据库，基于 WEKA3.8.1 数据挖掘平台及 JGibbLDA 主题模型软件对冠心病方剂数据库进行挖掘。结果显示使用频率最高的前 10 味药分别为：丹参、黄芪、川芎、炙甘草、当归、麦冬、桂枝、茯苓、赤芍、瓜蒌。其中丹参、川芎、当归、赤芍为活血化瘀药，黄芪、炙甘草为补气药，麦冬为养阴药，桂枝温阳通阳，茯苓健脾渗湿。以上药物以益气药、活血化瘀药为主，兼用健脾补气、祛痰理气及养阴药物，这与目前临床上治疗冠心病的基本方法十分吻合。其中使用频次最高的丹参具有活血祛瘀、通经止痛、清心除烦、凉血消痈之功效。使用频次排在第 2 位的黄芪为补气圣药。中医学认为，气血同源，用黄芪可补气养血，起到祛瘀散结之效。现代药理研究表明，黄芪有扩张血管、降压、强心、提高心肌耐缺氧能力、抗心肌缺血、镇痛等作用，是治疗冠心病心绞痛最常用的药物之一。关联规则挖掘方法共得到 12 个相关性较强的药物组合，这些药物大多为冠心病经典方剂的核心药物，如生脉散、瓜蒌薤白半夏汤、桃红四物汤、血府逐瘀汤等。

运用 LDA 主题模型共提取到 3 个方剂主题，对主题 1 中的核心药物进行了分析，318 味中药在主题 1 上的概率值差异

很大，说明主题模型对方剂语义的提取是有效的。以方剂主治冠心病证型为方剂主题的隐含语义，推断主题 1 对应的冠心病证候可能为气虚血瘀证，主题 1 中的前 10 味中药与临床对该证型治疗所用的药物十分吻合，说明 LDA 主题模型较好地提取了冠心病方剂的潜在语义。

LDA 主题模型与关联规则分析都能够挖掘得到不同证型冠心病的核心药物组合，两种方法的不同之处在于：①关联规则分析是通过统计不同药物组合出现的频次，得到满足指定支持度和置信度的关联规则。LDA 主题模型则通过吉布斯采样，为方剂数据库中的每一味中药分配一个概率值，在马尔科夫链收敛前每一味中药所属的证型是动态变化的，当马氏链收敛时每一味中药所属证型则是明确的，最终得到每个证型所包含的核心药物组合。②LDA 主题模型更适合于大型方剂数据库的聚类分析，每个主题中的药物都具有较为明确的隐含语义。关联规则分析适合于中小型方剂数据库的挖掘，虽然无法明确每一味中药的语义，但能够得到不同药物组合之间以及药物组合和证型之间的相关性，显得更为灵活。

与大型文本的主题建模相比，中医药方剂的 LDA 主题建模有自身的特殊性。主要表现在：①方剂中无重复药物且方剂文本较短，而短文本主题建模容易产生特征稀疏问题[16]。②方剂最优主题数及主题隐含语义的确定较为困难。进一步研究解决短文本主题建模时出现的特征稀疏问题，利用困惑度指标确定方剂最优主题数，更好地对中医药方剂进行语义挖掘，将是下一步的研究方向。

[参考文献]

[1] 赵爱梅，任钧国，刘建勋. 益气活血方治疗冠心病气虚血瘀证作用机制研究进展[J]. 中国实验方剂学杂志，2017，23(7)：215-220.

[2] Blei DM，Ng AY，Jordan MI. Latent dirichlet allocation[J]. Journal of machine Learning research，2003，3(3)：993-1022.

[3] 许珠香，江弋. 基于潜在狄利克雷分配模型的医疗数据研究[J]. 厦门大学学报：自然科学版，2013，52(3)：356-359.

[4] 张小平，周雪忠，黄厚宽，等. 一种改进的 LDA 主题模型[J]. 北京交通大学学报，2010，34(2)：111-114.

[5] 霍蕊莉，刘保延，何丽云，等. 基于主题模型的消渴病痹痿症药关系研究[J]. 北京中医药，2014，33(3)：163-166.

[6] 任毅，陈志强，张敏州，等. 当代名老中医治疗冠心病用药规律的聚类分析[J]. 中国中西医结合杂志，2016，36(4)：411-414.

[7] 吴焕林，于俏，林基伟，等. 基于数据挖掘的冠心病心绞痛方剂组方规律分析[J]. 中医药导报，2016，22(19)：38-40.

[8] 田松，何茜. 基于现代文献的冠心病中医证候特征数据挖掘[J]. 中国中医药信息杂志，2013，20(3)：29-30.

[9] 毛以林，吴彬才. 冠心病良方大全[M]. 太原：山西科学技术出版社，2016.

[10] Witten IH，Frank E，Hall MA，et al. Data Mining：Practical machine learning tools and techniques[M]. San Francisco：Margan Kaufmann，2016.

[11] Phan XH，Nguyen LM，Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections [C]. Proceedings of the 17th international conference on World Wide Web，ACM，2008：91-100.

[12] 关鹏，王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术，2016，32(9)：42-50.

[13] 李贵华，姜红岩，谢雁鸣，等. 基于大数据 84697 例冠心病中医证候及其中西药使用分析[J]. 中国中药杂志，2014，39(18)：3462-3468.

[14] 袁建，梁庭栋. 中医药治疗冠心病心绞痛用药规律探索[J]. 四川中医，2013，31(10)：37-38.

[15] 庄逸洋，郑升鹏，陈文嘉，等. 国医大师邓铁涛治疗冠心病用药规律的数据挖掘研究[J]. 时珍国医国药，2016，27(12)：3025-3027.

[16] 张志飞，苗夺谦，高灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用，2013，33(6)：1587-1590.

（责任编辑：冯天保）