应用并行化 K - means 优化算法探究中医治疗高血压用药规律

宋欣霞 金 卫

（山东中医药大学理工学院 济南 250355）

〔摘要〕 介绍并行化相关概念，包括 Mapreduce 编程框架以及 Hadoop 分布式文件系统，对传统 K - means 算法进行优化，建立并行化 K - means 聚类模型，探究中医治疗高血压用药规律，挖掘出中医辨证治疗高血压的 8 组用药组合，结合中医理论分析，验证结论真实性。

〔关键词〕 并行化；K - means 算法；高血压用药

〔中图分类号〕R - 056 〔文献标识码〕A 〔DOI〕10. 3969/j. issn. 1673 - 6036. 2017. 06. 015

Exploration on the Medication Rules of Traditional Chinese Medicine for Hypertension Treatment through Parallel K - means Optimization Algorithm

SONG Xin - xia, JIN Wei, College of Science and Engineering, Shandong University of Chinese Medicine, Jinan 250355, China

〔**Abstract**〕 The paper introduces relevant concepts of parallelization, including Mapreduce programming frame and distributed document system of Hadoop, optimizes the traditional K - means algorithm, establishes parallel K - means clustering model, explores the medication rules of Traditional Chinese Medicine (TCM) for hypertension treatment, digs into 8 groups of prescriptions for differentiation treatment of hypertension through TCM, verifies authenticity of the conclusion by combining the theories of TCM.

〔**Keywords**〕 Parallelization; K - means algorithm; Anti hypertension drugs

1 引言

高血压（Hyperension）是以体循环动脉血压（收缩压和或舒张压）增高为主要特征，可伴有心、脑、肾等器官功能或器质性损害的临床综合症。高血压是常见的慢性病，也是心脑血管病最主要的危险因素，属于中医学“痰饮”、“水肿”、“眩晕”、“头痛”、“耳鸣”、“心悸”、“胸痹”、“中风”等疾病范畴，其病因与“气、火、瘀、痰”密切相关，“疏气、泻火、化痰、祛痰”是高血压最基本的治则治法。中医药治疗高血压具有较好疗效，为使中医药治疗高血压更加科学化、规范化，本文阐述一种并行化 K - means 优化聚类算法并对高血压的用药规律进行挖掘，以期更好地指导临床用药，提高中医治疗高血压的疗效。

〔修回日期〕 2017 - 04 - 05

〔作者简介〕 宋欣霞，硕士研究生，发表论文 2 篇；通讯作者，金卫。

2 并行化相关概念

2.1 Mapreduce 编程框架

Mapreduce 编程框架是云计算的基本框架，其

程序是根据一定的规律将彼此互不相关但是要进行分析的所有数据区分开，然后再将有联系的集合进行合并，最终得到结论。Mapreduce 框架由单独的主节点 Master 和几个集群节点 Slave 组成[10]，主节点 Master 的任务是统筹调度，监控执行所有与任务相关的子任务，对执行失败的任务重新安排调度，分派执行；而子节点 Slave 的任务是接受主节点的调度，完成被指派的要求执行的任务。这样的任务分配使得编程易于实现，程序员不需要关注并行程序的细节问题，只需要关注自我程序的实现，在系统平台的帮助下程序开发效率得到了大幅度提高。另外在大数据的环境下，Mapreduce 编程框架能够扩大存储空间，子节点的数量可以随着需求量的变化改变，从而提高运算速率[11]。当多个任务同时运行时经常会因为个别任务出错导致整个程序无法运行或系统崩溃，而 Mapreduce 编程框架的这种任务分配方式极大程度地避免了此种情况的发生。由于 Hadoop 平台是一个开源项目，因此 Mapreduce 编程框架可以免费使用，在成本上节约了很大一部分，而且这种分布式框架可以运行在普通的硬件系统上。一个大规模集群的建立 Mapreduce 编程框架有着不可忽视的重要作用。

2.2 Hadoop 分布式并行化文件系统

Hadoop 平台包括 Mapreduce 编程框架和 Hadoop 分布式文件系统（Hadoop Distributed File System，HDFS）。HDFS 可能由很多存储着各种数据的服务器构成。通常在 HDFS 上运行的应用都是一个很大的数据集，文件大小往往为 G 字节或者 T 字节，因此强大的数据传输带宽是必不可少的。HDFS 属于可扩展、高效的分布式文件系统，具有良好的容错机制，可及时处理突发故障，同时其更新机制不需要改变已存在的数据，只需对新添加的数据进行更新[14]。此外作为存储系统，HDFS 能够实现文件的增加、删除、修改等命令，还能够对文件进行备份，对数据进行校验。数据以数据块的形式存储到数据节点上，系统通过复制文件块保证数据安全性。通过数据备份进行保存，同样一个数据块往往会在多个服务器上进行复制保存，HDFS 设计中采用的是 Master/Slave 架构，一个 NameNode 全节点和多个 DataNode 数据节点组成一个 HDFS 集群，其体系结构，见图 1。

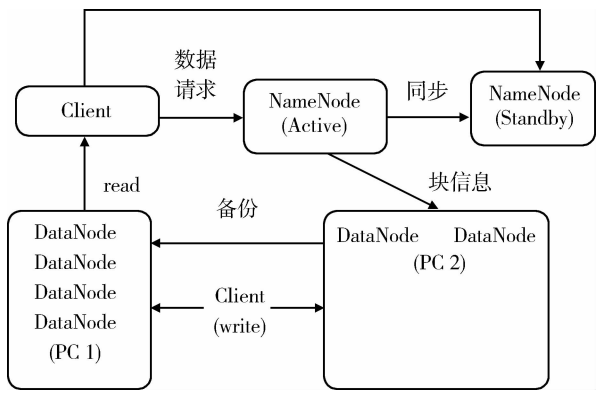

图 1 HDFS 体系结构

3 K－means 优化算法

在传统 K－means 算法的每次迭代中，都要重新计算给定数据点到所有集群中心之间的距离，对于庞大的数据量来说这种计算损耗是相当昂贵的。为了解决此问题，使得 K－means 算法更加高效，对存在的缺陷部分进行合理化分析，提出优化的 K－means 算法。对于每个数据点来说，可以保留其到与其最接近的簇之间的距离[5]，在下一次迭代中，只需要计算其到先前的集群中心的距离，如果新的距离小于或等于先前的距离，那么该点就保持在该集群中，这样也就没有必要计算其到其他群集中心的距离，节省用来计算到 $k-1$ 聚类中心距离的时间。优化算法的流程，见图 2。

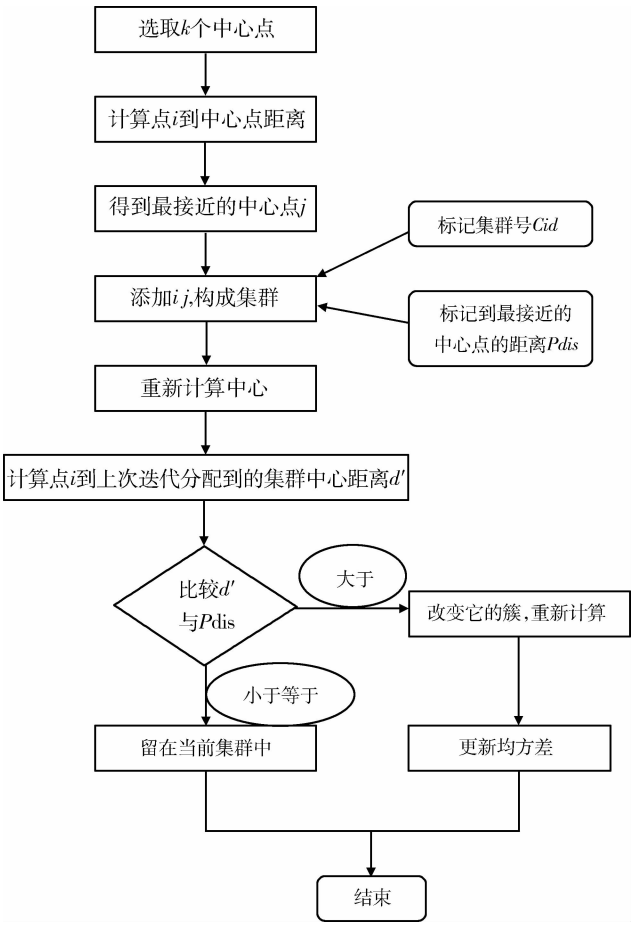

图 2 优化算法流程

4 并行化 K – means 算法模型

4.1 设计思路

基于 Mapreduce 并行化 K – means 算法的设计思路是：将 K – means 算法的每次迭代计算[4]都变成可以独立操作的情况下并行实现，将串行运算转化为一次 Mapreduce 计算，算法的迭代运算包括计算样本到聚类中心的距离，还包括新的聚类中心的局部运算。并行化算法的实现过程主要包括 Map 数据处理过程、Combine 数据合并过程和 Reduce 再处理过程。Mapreduce 编程框架是对输入的数据集进行并行操作，首先要将数据集分割成多个子集(这里关心输入数据的内部逻辑结构)，具体分割方法要根据实际情况去指定，或者用已经定义完成的几个算法简单分割[13]，最终文件分割体对应一个新的任务。

4.2 并行化函数设计

4.2.1 Map 函数设计 Map 函数将文件数据以 key/value 键值对的形式表示，每一行作为一个样本，通过计算每个样本到各集群中心的距离，挑选出距离最小的中心，将样本保留在该集群中。同时对样本进行标记，归属到新的集群类别[12]，再以 key/value 键值对的形式输出。形式 < 行 – ID，属性 > 表示 < key，value > 对，Map 函数输入的 key 是行 – ID，相当于起始数据输入点的偏移量[11]，value 是当前记录的属性值；输出中间结果以 < 局部集群中心 – ID，记录属性向量 > 形式表示 < key′，value′ > 对；输出数据同样为 < key′，value′ > 对，key′ 表示局部集群中心 – ID，value′表示记录属性向量。

4.2.2 Combine 函数设计 作用是将 Map 过程产生的中间结果进行 Reduce 本地化处理，减少数据在节点之间传输的时间消耗。其任务是预处理所处节点内的映射结果，同时处理具有相同 key 值的 value 值，得到局部聚类结果，再将值传给集群中的 Reduce 函数进行 Reduce 操作。

4.2.3 Reduce 函数设计 作用是整合 Combine 函数处理得到的局部聚类结果，计算出新的聚类中心，将新的聚类中心用于下一次迭代运算。Reduce 函数首先计算从每个节点得到的输出局部聚类结果的样本个数，分析每个样本的各维属性值，然后叠加各维度累加值，得到的数据除以总样本个数，计算结果就是新的聚类中心坐标。

4.3 算法时间复杂度分析

根据上述对算法的描述可知，需要进行计算的步骤包括计算给定数据点到聚类中心的距离以及迭代之后重新计算新的聚类中心[6]。进行一次迭代的时间复杂度是 $O(nkd)$，总的时间复杂度为 $O(nkdt)$，其中 t 为迭代次数，时间复杂度通常可用 $O(n)$ 表示。K – means 聚类 Mapreduce 并行化的计算过程中，将 m 设为集群中节点的数目，并行计算在 m 个节点上同时执行，每个样本数据对象在每次迭代中，计算距离的时间复杂度是 $O(kd)$，因此所有数据对象用于计算距离的时间复杂度是 O

(nkd)。在较为理想的情况下所有数据对象的距离计算方式均匀地分布到 m 个节点上同时并行执行，所以时间复杂度被缩小为 O（$nkdt/m$）。根据算术式可以看出，集群中节点数目增加，每个节点距离计算所消耗的时间反而会大幅度减少。根据 Hodoop 设定分块大小，将原始文件分割成多个由 Hadoop 设定的数据分块，Map 任务每次分配 1 个数据分块，因为每个节点运行多个 Map 任务，所以每个节点拥有的 Map 任务数是 P 个。这样原来按节点处理的 $nkdt$ 次距离计算，分配给多节点并行处理，时间复杂度则降为 O（$nkdt/mp$）。

5 中医治疗高血压用药规律探究

5.1 数据来源

试验数据采用山东中医药大学附属医院心血管门诊高血压医案 2 000 例（其中单纯高血压 1 571 例，合并其他疾病 429 例），患者共 1 378 例，男患者 516 例，女患者 862 例，年龄 2 ~ 86 周岁，就诊次数 1 ~ 14 次。由于数据量有限，因此在进行并行化试验时，将预处理完的数据进行 n 倍处理，以达到足够量的数据需求。分别使用单机和 Mapreduce 并行处理，进行算法运行效率的对比。医案信息包括患者年龄、性别等基本信息，此外还有疾病的症状、证候、疾病的中医诊断或者西医诊断等，将数据规范化（高血压、高血压病、血压高、原发性高血压、一级高血压、早期高血压等统一为高血压），对数据进行预处理，转化为计算机处理的二进制模式数据单元，使之规范、准确。

5.2 试验环境

试验使用 3 台计算机，其中 1 台作为 Master 和 Job Tracker 服务节点，另两台作为 Slave 和 Task Tracker 服务节点，3 台计算机使用 1 台普通百兆交换机进行连接，具体的硬件配置和基本设置，见表 1、表 2。

表 1 各个节点硬件配置情况

CPU	内存	硬盘	网卡
Intel Pentium4	2G	500G	百兆以太网卡

表 2 各个节点基本设置情况

用户名	主机名	IP	网关
Hadoop	Mater	172. 20. 0. 11	172. 20. 0. 1
Hadoop	Mater	172. 20. 0. 11	172. 20. 0. 1
Hadoop	Mater	172. 20. 0. 11	172. 20. 0. 1

5.3 结果与讨论

5.3.1 并行化 K-means 算法执行效率结果分析 程序并行化的性能和效果通常采用加速比来衡量。加速比是指同一个任务在单处理器系统（串行）和并行处理器系统（并行）中运行消耗时间的比率，其计算公式为 $S_p = T_1/T_p$，式中：S_p 表示加速比，T_1 表示串行的运行时间，T_p 表示并行的运行时间，p 表示处理器的个数。单机处理速率与并行化 K - means 算法处理速率通过加速比来表示。K - means 算法与并行化算法的时间对比，见表 3。

表 3 基于 K - means 的 3 种算法运行时间对比

算法类型	读取文件	通信	数据计算	总时间（毫秒）
传统 K - means 算法	1. 03e - 3	0	50	50
优化 K - means 算法	1. 03e - 3	0	30	30
并行化 K - means 算法	2. 89e - 7	2. 46	18	20. 46

3 种算法在聚类肝阳上亢型高血压的用药组合的正确聚类数目（即聚类质量）分析，见表 4。

表 4 聚类质量对比

算法名称	正确聚类数目（例）
传统 K - means 算法	1 975
优化 K - means 算法	1 985
并行化 K - means 算法	1 990

通过以上结果可知，当处理较小数据量时，Hadoop 集群系统处理效率优势不是很大，但是从算法的时间复杂度分析来看，当采用较大数据时，在保证聚类质量的前提下，聚类时间会成倍缩短，明显发挥了并行化的优势。

5.3.2 探究高血压辨证用药规律结果分析 高血压病症用药规律，见表 5。

表 5 高血压辨证用药规律

症型	例数	用药组合
肝阳上亢或热毒型	621	钩藤、黄连、黄答、黄柏、丹皮、栀子、泽泻、荻荃、稀签草、野葛根
肝肾阴虚型	237	杜仲、牛膝、制首乌、女贞子、枸杞子、五味子、女贞子、金樱子
高血压兼有血癖	119	当归、川芎、元胡、丹参、生地
高血压伴有气阴两虚型冠心病	489	黄芪、麦冬、五味子、三七粉、冰片
高血压伴有失眠	518	炒枣仁、夜交藤、紫石英
高血压伴有高血脂病	355	钩藤、丹皮、草决明
高血压伴有快速性心律失常	23	青蒿
高血压伴有腹痛、腹胀	35	木香

5.4 中医理论分析

中医治疗高血压通常采用辨证论治方法，包括主证、治法和方药组合[3]。从 2 000 例数据中可归纳出 8 种不同症状的证候表现。由于证候不同需要采用不同的治法，因此得到多种不同的用药规律组合。试验中聚类最终产生 8 种辨证用药规律组合如表 4。药组一[2]，这一组合的药均有降压功效，其中黄连可清热燥湿、泻火解毒；丹皮、泽泻均可清热凉血、活血散瘀，因此该药组可治疗肝阳上亢或热毒型高血压。药组二，其中杜仲有补益肝肾、强筋壮骨、调理冲任、固经安胎的功效；牛膝入药有逐瘀通经、补肝肾、强筋骨、利尿通淋等效用；制首乌可补益精血、养肝安神、强筋骨；女贞子是一味补肾滋阴、养肝明目的中药，可治肝肾不足、头晕耳鸣、头发早白及两目昏糊等病症，因此该药组可治疗肝肾阴虚型高血压[9]。药组三，这一组合药可治疗血虚诸证、月经不调、经闭、痛经、症瘕结聚等，因此该药组治疗高血压兼有血癖。药组四，这一组合的药可直接扩张外周血管，降低外周阻力，从而降低血压，同时可养阴生津、润肺清心，用于治疗内热消渴、心烦失眠、肠燥便秘，因此该药组可治疗气阴两虚型高血压。药组五，具有镇心、安神、降逆气的功效，可用于缓解失眠。单味药改善并发症也可通过试验结果看出来。药组六，这类药物可润肠通便，降脂明目，有缓泻作用，因此可用于治疗高血压高血脂症。药组七，青蒿具有清热解暑、除蒸、截疟的功效，用于暑邪发热、阴虚发热等，因此可用于治疗高血压伴又快速性心律失常的症状。药组八，木香具有行气止痛、调中导滞的功效，用于胞胁胀满足、脘腹胀痛，因此可用于治疗高血压伴有腹痛腹胀的症状。

6 结语

随着现代医学对高血压的广泛关注，探究高血压的用药规律显得尤为重要。本文通过聚类算法对高血压用药规律进行挖掘，在挖掘过程中发现传统的 K－means 算法存在迭代性高、运行速度慢等缺点，在处理数据量巨大的医学类数据方面存在缺陷，因此对算法进行优化同时引入并行化的概念，大大提高了算法的运行速率。通过开展试验，对传统 K－means 算法、优化 K－means 算法和并行化 K－means 算法进行比较。在个体数相同的前提下，比较 3 种算法的运行时间，通过比较均方误差的不同值来判断聚类结果的质量，得出优化 K－means 算法具有更高的优势。对并行化 K－means 算法，分析其算法耗时以及不同节点数量对算法的影响，任务分块在节点不同时有何变化，并行算法的总体运行时间，对并行算法进行性能测试。以高血压已有相关理论为依据，分析挖掘结果，得到不同证候对应所用药方之间的联系，从而获得中医治疗高血压辨证用药的一般规律。

参考文献

1 翟瑶瑶. 基于中医传承辅助平台系统挖掘治疗原发性高血压的组方规律研究［D］. 北京：北京中医药大学，2016.

2 马宁，侯雅竹，王贤良，等. 基于文献的中医治疗高血压病阴虚阳亢证用药规律探析［J］. 中国中西医结合杂志，2016，36（4）：403－410.

3 张琳，卢笑晖. 基于中医传承辅助平台治疗肝阳上亢高血压组组方规律系统综述［J］. 实用中医内科杂志，2016，30（3）：3－6.

4 王辉，张望，范明. 基于集群环境的 K－Means 聚类算法的并行化［J］. 河南科技大学学报：自然科学版，2008，29（4）：42－45，116－117.

5 蒋利顺，刘定生. 遥感图像 K－Means 并行算法研究［J］. 遥感信息，2008，（1）：27－30，115.

6 Fahim A. M，Salem A. M，Torkey F. A，et al. An Efficient Enhanced K－means Clustering Algorithm［J］. Journal of Zhejiang University Science A（Science in Engineering），2006，（10）：1626－1633.

7 陈守强. 丁书文教授用药规律的计算机辅助分析［D］. 济南：山东中医药大学，2005.

8 王海霞. 肝阳上亢证中医文献研究［D］. 济南：山东中医药大学，2005.

9 周超凡，陈京莉. 中医治疗高血压病的用药思路与方法［J］. 中国中医药信息杂志，2003，10（4）：72－73.

10 冯波，郝文宁，陈刚，等. K－means 算法初始聚类中心选择的优化［J］. 计算机工程与应用，2013，49（14）：182－185，192.

11 张依杨，向阳，蒋锐权，等. 朴素贝叶斯算法的 MapReduce 并行化分析与实现［J］. 计算机技术与发展，2013，23（3）：23－26.

12 李晓飞. 云计算环境下 Apriori 算法的 MapReduce 并行化［J］. 长春工业大学学报：自然科学版，2013，34（6）：736－740.

13 幸莉仙，黄慧连. MapReduce 框架下的朴素贝叶斯算法并行化研究［J］. 计算机系统应用，2013，22（2）：108－111.

14 张磊，张公让，张金广. 一种网格化聚类算法的 MapReduce 并行化研究［J］. 计算机技术与发展，2013，23（2）：60－64.

（上接第 70 页）

3 糖尿病－实况报道［EB/OL］.［2016－11－20］. http://www.who.int/mediacentre/factsheets/fs312/zh/.

4 韩玲革，周东花. 我国抗高血压药物专利的现状分析［J］. 医学信息学杂志，2008，（12）：26－29.

5 李瑞丰，欧阳雪宇，吕飞，等. DPP－IV 抑制剂列汀类抗糖尿病药物专利分析和专利布局［J］. 中国新药杂志，2015，（1）：8－17.

6 宋聪雨. 聚氨酯橡胶领域的重要专利申请人分析［J］. 专利代理，2015，（3）：76－80.

7 德国默克与百时美施贵宝拟携手进军中国糖尿病药物市场［EB/OL］.［2016－03－28］. http://endo.dxy.cn/article/50075.

8 默克糖尿病新药一周吃一次，试验效果略胜旧药佳糖维［EB/OL］.［2016－09－21］. http://mt.sohu.com/20150921/n421657085.shtml.

9 中国慢病管理网. 赛诺菲 PK 诺和诺德：糖尿病市场谁革谁的命？［EB/OL］.［2016－03－03］. http://www.ncd.org.cn/Article/index/id/5341.

10 赛诺菲新一代基础胰岛素 Toujeo 获 FDA 批准［J］. 中国新药杂志，2015，（5）：508.

11 丁香园. 罗氏糖尿病检测业务预计 2017 年可恢复增长［EB/OL］.［2016－05－12］. http://yao.dxy.cn/article/492186.

12 美国 CDC 报告：近十分之一美国人患糖尿病［EB/OL］.［2016－11－24］. http://news.medlive.cn/all/info－progress/show－70822_46.html.

13 倪娜. 药品专利政策对生物制药业的影响研究［D］. 北京：中国人民解放军军事医学科学院，2010.

14 姚雪芳，丁锦希，邵蓉，等. 我国医药专利申请审批现状分析与政策建议［J］. 中国发明与专利，2011，（5）：47－51.

15 赵霞，王玉光，胡瑞法. 基于专利分析的转基因农作物技术布局态势研究［J］. 情报杂志，2014，（9）：51－55，92.