一种基于基因表达式编程的串行聚类算法并行化研究

蔡宏果[1,2],元昌安[1,*]
(1 广西师范学院 科学计算与智能信息处理广西高校重点实验室,南宁 530023;
2 广西教育学院 培训学院,南宁 530023)

摘 要 为进一步解决基于用户的协作过滤技术的扩展性问题,利用基因表达式编程(GEP)的并行性优势,与已有的串行聚类 DBSCAN 算法进行融合,使得串行程序并行化,提出了一种 GEP-DBSCAN 协作过滤聚类算法来寻找最近邻居,改进基于密度的协作过滤方法,实验证明了算法的有效性以及提高了时间效率.

关键词 聚类算法;基因表达式编程;协作过滤
中图分类号 TP181 **文献标识码** A **文章编号** 1672-4321(2017)04-0112-04

Research on GEP-based Cluster Algorithm for Serial Program to be Parallelized

Cai Hongguo[1,2], *Yuan Chanan*[1]
(1 Computer and information engineering college, Guangxi Teachers Education University, Nanning 530023, China;
2 Training School, Guangxi College of Education, Nanning 530023, China)

Abstract To address the expansibility problem of collaborative filtering technology based on users, the GEP-DBSCAN algorithm for collaborative filtering clustering was proposed. It is key to fuse the parallelism of gene expression programming and the advantages of the DBSCAN algorithm. The new algorithm makes serial programs parallelized and can be used to find the nearest neighbors. It improves collaborative filtering method based on the density. The experimental results show that the GEP-DBSCAN algorithm is effective and can increase time efficiency.

Keywords cluster algorithm; gene expression programming; collaborative filtering

协作过滤技术分为基于用户的协作过滤和基于资源的协作过滤. 对于基于用户的协作过滤推荐系统,当有新用户加入时,推荐的扩展性将成为非常严重的问题[1],文献[2]为了解决扩展性问题,在用户评分数据上再做一次用户聚类分析是比较有效的办法,由于在特定用户分类中,最近邻居群相对变化较小,研究表明,用户分成多少个类别的群集,推荐系统的时间效率就能够提高到几倍. 文献[3]将基于邻居聚类的协同过滤方法用于云计算环境,实验表明,在数据量剧增环境下,基于协同过滤的个性化推荐技术比传统的推荐技术具有更高的准确率. 文献[4]用具有明确数学模型的基于隐语义概率模型的协同过滤方法来预测用户偏好,具有更好的预测精度. 文献[5]使用基于密度的划分方法在用户评分数据上做"最近邻居"聚类分析,聚类计算往往可以离线,没有实时要求的基础,从而极大的缓解了推荐系统的实时计算压力,增加了推荐系统的速度. 但是,以上协同过滤方法中的聚类算法都需要精确的数学模型或大量的实验来调整一些参数的设置,同时还存在推荐准确率有待提高、没有考虑空间拓展性能等问题,这给实际工作带来应用上的复杂性. 为解决这些问题,近年来,利用进化计算来解决聚类技术存在的问题成为研究热点之一. Murthy 等在 1996 年提出的基于遗传算法的聚类技术,开创了进化计算用于聚类的研究,算法采用二进制编码,只适合解决数据集较小的聚类问题. 文献[6]使用改进粒子群算法来实现聚类,该算法能够处理较大的数据集. 文献[7]提出基因表达式编程(GEP)自动聚类算

收稿日期 2017-06-27 * 通讯作者 元昌安,研究方向:智能计算和数据挖掘,E-Mail: yca@gxtc.edu.cn
作者简介 蔡宏果(1978-),男,高级工程师,研究方向:智能计算和文本挖掘,E-mail: webminning@163.com
基金项目 国家自然科学基金资助项目(61262028)

法,解决了进化计算自动聚类的问题. GEP 是在遗传算法和其它进化算法的基础上发展的新技术,比遗传算法编码更灵活,具有更强的解决问题能力,GEP 目前已经应用在多个领域[8-10]. 进化计算具有并行性和智能性的优势,但是,进化计算具有求解问题只有近似解而难以得到最优解的弊端,与经典的基于密度的聚类算法相比,聚类的准确率随机性较大.

本文在以上研究的基础上,依据传统基于密度的聚类原理和 GEP 自动聚类的思想,提出了基于 DBSCAN 和 GEP 的聚类算法来解决协作过滤技术中的最近邻居聚类问题,先通过密度原理形成几个好的中心区域,再通过 GEP 并行性和智能性的处理能力,对用户进行自动聚类,减少空间开销,改善基于用户的协作过滤方法的性能拓展,提高预测的准确率.

1　相关工作

1.1　基于用户的协作过滤和密度聚类概述

基于用户的协作过滤的基本思想是根据用户兴趣的相似性来推荐资源,通过研究不同用户的兴趣,主动为当前用户推荐相似的其它用户的已访问资源,将 N-top 资源提供给当前用户[11]. 基于密度的聚类方法与其它方法的一个根本区别是,它不是基于各种各样的距离的,而是基于密度的,文献[12]和文献[13]研究了基于密度原理的新的聚类方法. 基于密度的方法指导思想是,对给定簇中的每个数据点,在给定半径的邻域内至少必须包含规定的阈值个点. 代表算法有: DBSCAN 算法、DENCLUE 算法、OPTICS 算法等. DBSCAN 算法有几个重要的定义如半径 eps、E 领域、核心对象、阈值 $minpts$ 等[11]. DBSCAN 算法描述参见文献[11].

1.2　基于 GEP 和 DBSCAN 协作过滤算法

基于 GEP 和 DBSCAN 的协作过滤聚类算法基本思路是: 将日志文件以用户作为个体,用户访问序列中的访问页(项目、属性等)作为对象相似性度量,预处理成数据库 D,从数据库 D 中选取用户,通过半径和密度阈值确定是否为核心对象或者噪声,并创建核心对象的簇,结果标记为数组 $arr[k]$,利用基因表达式编程算法的深度递归搜索在 $arr[k]$ 中寻找聚类族.

基于基因表达式编程的 DBSCAN(GEP-DBSCAN)算法:

输入: 数据库 D; GEP 参数; Eps(邻域或称为半径); $MinPts$(密度阈值).

输出: 聚类簇结果.

步骤:

(1) 扫描原始数据,数据清洗,预处理成数据库 D;

(2) 读取数据库 D 中的任意一个还没有分类的用户 U,检索出与 U 的距离不大于 Eps 的所有对象 $Neps(U)$;

(3) 如果($Neps(U) < MinPts$),则将 U 作为一个噪声标记,并执行步骤 2,否则给这组对象标记一个 $classN$ 标签后,记入一个数组,执行步骤 4;

(4) 初始化种群,生成一定规模的染色体个体,计算 $Neps(U)$ 的每个用户与核心对象的相似性,公式 2 作为染色体个体的适应性函数;

(5) 依照轮盘规则实施 GEP 遗传操作;

(6) 按照 GEP 的交叉算子对个体进行遗传交叉;

(7) 按照 GEP 的变异算子对个体进行遗传变异;

(8) 若达到最大迭代次数则执行步骤 9,否则执行步骤 4;

(9) 选择出最优个体;

(10) 输出聚类簇结果.

GEP-DBSCAN 算法空间复杂度分析: 通常情况下,并行算法对空间-内存需求非常大. 经常会出现内存不能一次装载所需的数据,需要备用存储设备的情况,此时如果处理不好就会严重降低算法的性能. 开发并行聚类方法一个好的途径就是,对已有的串行算法进行改进,挖掘其中的并行性质并加以利用,使得串行程序并行化. 一般地,并行聚类算法比传统的串行聚类算法在空间复杂度上可以高几倍的效率,DBSCAN 算法的空间复杂度为 $O(N^2)$,GEP-DBSCAN 算法如果种群规模是 m,则 GEP-DBSCAN 算法空间复杂度为 $\frac{1}{m}O(N^2)$,GEP-DBSCAN 算法通过改进原串行聚类算法-基于密度的聚类 DBSCAN 方法,加入 GEP 并行性优点,结合起来,在降低空间复杂度上提高了效率.

1.3　GEP 适应度计算

用户之间的相似计算主要是通过用户对目标项目(属性)评分等决定,研究中,项目评分的次数、项目内容的相近度都可以提高相似性计算,但是,一项协作过滤推荐算法比较研究表明: 从所有用户评分

矩阵中抽离出维度更小的评分矩阵不仅提高了预测效率,而且在聚类的基础上有时可以提高算法的预测精度,相似度计算方法可采用皮尔逊相关系数计算.

$$sim(i,j) = \frac{\sum u \in v_{i,j}(V_{ui} - \overline{v.i})(V_{uj} - \overline{v.j})}{\sqrt{(\sum u \in v_{ij}(V_{ui} - \overline{v.i})^2)}\sqrt{\sum (V_{uj} - \overline{v.j})^2}}, \quad (1)$$

(1)式中$sim(i,j)$是用户i和j的相似度;v_{ij}表示被用户i和j共同评分过的项目的用户数;V_{ui}表示被用户i评分过的编号为u的项目用户数;V_{uj}表示被用户j评分过的编号为u的项目用户数;$\overline{v.i}$、$\overline{v.j}$表示对用户评分过的项目的总用户数和项目数的平均值.由于在GEP中一般按照适应度由大到小选择个体,每个个体代表一个最近邻居聚类簇,因此,算法中适应度函数f的度量公式为:

$$f = \frac{1}{n}\sum_{x=1}^{n} sim(i,j). \quad (2)$$

2 实验与性能分析

实验数据来源于movielens数据集,本文实验使用了数据集提供的用户评分数据ml数据集用于算法测试,实验环境,Windows XP; Microsoft Visual studio(c#)及SPSS Clementine 10.0.

实验一选取数据集中的1000条评价数据.根据文献[8]和文献[9]对GEP遗传算子参数的实验结论和讨论,选择GEP参数.设置不同的*Eps*、*MinPts*值,聚类结果如图1和图2所示.

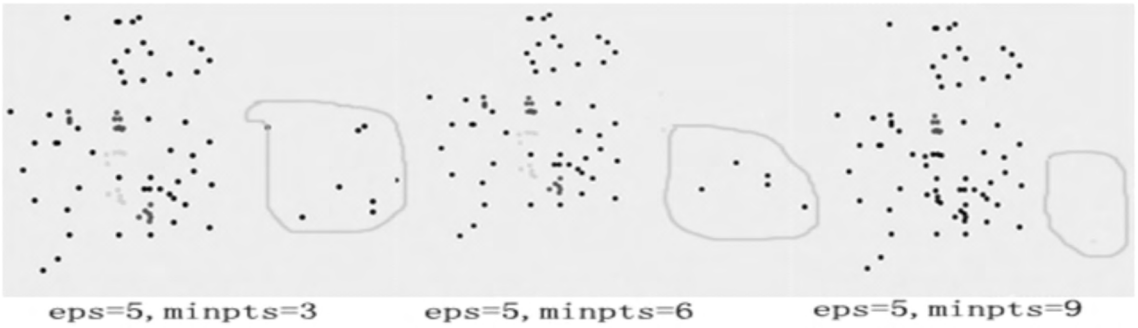

图1 聚类结果的比较(相同的邻域,不同阈值条件下)

Fig. 1 Comparison results on clustering (same neighborhood, different threshold conditions)

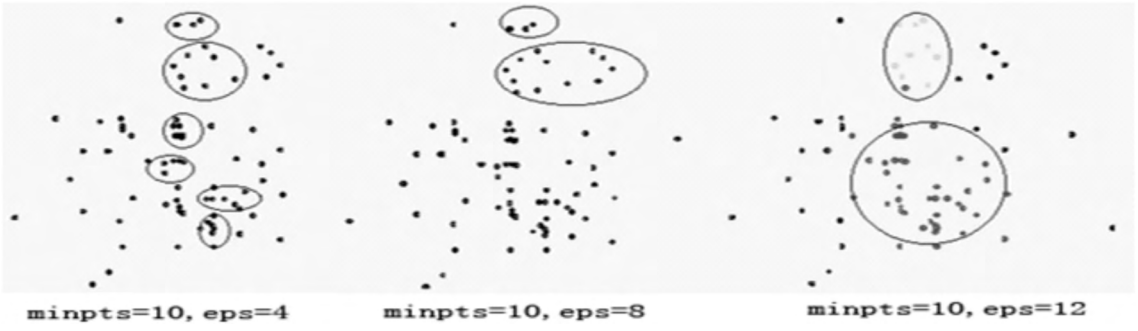

图2 聚类结果的比较(不同的阈值,相同邻域条件下)

Fig. 2 Comparison results on clustering (different threshold, same neighborhood conditions)

根据图1所示,*eps*不变,随着*minpts*的值不断增加,圆圈所示噪声数据不出现在结果集合中,GEP-DBSCAN算法能过滤噪声,类与类之间的间隔也比较明显;根据图2所示,*minpts*不变,随着*eps*的值不断增加,圆圈所示中的聚类个数减少.实验表明,GEP-DBSCAN算法能够发现数据集中的聚类模式,较好的解决最近邻居的聚类,形成聚类中心.

实验二对u1~u5这5个数据集进行实验,将每个数据集其中80%作为训练集,20%作为测试集,通过GEP-DBSCAN算法进行一次聚类后,依据基于用户的协作过滤方法,获得项目推荐集,抽取测试集其中的5个用户的访问情况与推荐集比较.与传统基于用户的协作过滤方法和文献[7]自动聚类后的协作过滤方法进行实验比较和分析,通过推荐的准确率来度量推荐质量,得到实验结果见表1,通过最常用的平均绝对偏差MAE值来度量三种算法推荐质量如图3所示,通过三种算法对数据集的运行时间和平均运行时间来度量时间效率,得到实验结果

见表 2.

表 1　不同数据集中三种推荐方法的准确率

Tab. 1　The accuracy rate of three recommended methods in different data sets

不同的数据集	传统基于用户的协作过滤方法 /%	文献 [7] 自动聚类 GEP-Cluster 方法 /%	GEP-DBSCAN 算法的协作过滤方法/%
U1	35.0	44.6	60.6
C	32.5	44.0	44.6
U3	33.5	42.0	51.0
U4	35.0	46.0	48.6
U5	22.5	37.8	56.5
平均准确率	31.7	42.8	52.2

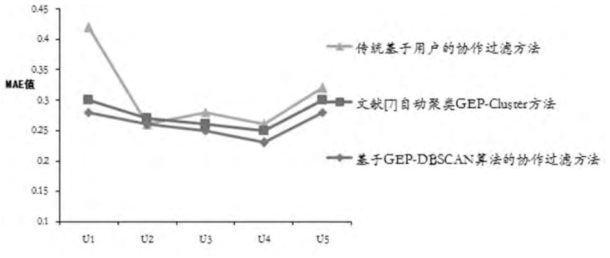

图 3　三种方法的 MAE 值

Fig. 3　MAE value of three methods

表 2　同数据集中三种推荐方法的运行时间

Tab. 2　The running time of three recommended methods in different data sets

不同的数据集	传统基于用户的协作过滤方法/s	文献 [7] 自动聚类 GEP-Cluster 方法/s	基于 GEP-DBSCAN 算法的协作过滤方法/s
U1	60.73	40.60	31.00
U2	50.70	40.00	28.50
U3	57.22	38.00	29.50
U4	74.16	42.00	31.00
U5	76.24	33.80	18.50
平均运行时间	63.81	38.88	27.70

根据表 1 实验结果表明，基于 GEP-DBSCAN 算法的协作过滤方法平均推荐准确度则为 52.2 %，比基于用户的协作过滤推荐算法的推荐准确度提高了 39.7%，比文献 [7] 中改进的协作过滤方法的推荐准确度提高了 18%，能显著提高协作过滤的推荐质量. 图 3 表明，基于 GEP-DBSCAN 算法的协作过滤方法平均绝对误差最小. 表 2 表明，基于 GEP-DBSCAN 算法的协作过滤方法平均运行时间最少，时间效率最高.

4　结语

协作过滤推荐系统已经被广泛应用于 Web 网站、电子商务、电子图书馆等众多领域，随着用户和产品数量的不断增加，传统算法的许多缺点逐渐暴露了出来. 本文通过对协作过滤存在的计算瓶颈问题，提出了一种 GEP-DBSCAN 的协作过滤聚类算法，来解决用户的“最近邻居”问题，通过算法的实验，提供了可靠的支撑依据.

参　考　文　献

[1] 雷建云，何　顺，李白杨. 基于标签和云模型的协同过滤算法 [J]. 中南民族大学学报(自然科学版)，2016，35(3)：117-122.

[2] 陈天昊，帅建梅. 一种基于协作过滤的电影推荐方法 [J]. 计算机工程，2014，40(1)：55-62.

[3] 朱　夏，宋爱波. 云计算环境下基于协同过滤的个性化推荐机制 [J]. 计算机研究与发展，2014，51(10)：2255-2269.

[4] 胡　堰，潘启民. 一种基于隐语义概率模型的个性化 Web 服务推荐方法 [J]. 计算机研究与发展，2014，51(8)：1781-1793.

[5] SabeurAridhi，Laurentd'Orazio etc. Density-based data partitioning strategy to approximate large-scale subgraph mining [J]. Information Systems，2013，48(2015)：213-223.

[6] 陈小全，张继红. 基于改进粒子群算法的聚类算法 [J]. 计算机研究与发展，2012，48(S1)：287-291.

[7] 陈　瑜，唐常杰. 基于基因表达式编程的自动聚类方法 [J]. 四川大学学报(工程科学版)，2007，39(6)：108-112.

[8] 元昌安，彭昱忠，覃　晓，等编著. 基因表达式编程算法原理与应用 [M]. 北京：科学出版社，2010：38-108.

[9] Ferreira. C. Gene Expression Programming：A New Adaptive Algorithm for Solving Problems [J]. Complex Systems，2001，13(2)：87-129.

[10] Qin Xiao，Yuan Chang-an. A GEP-Based Text Classification Algorithm [J]. Journal of Information & Computational Science，2009，6(3)：1303-1309.

[11] Han Jiawei，Micheline Kamber，Jian Pei，etc. Data mining：concepts and techniques [M]. Waltham：Morgan Kaufmann Press，2007：246-290.

[12] 黄创光，印　鉴. 不确定近邻的协同过滤推荐算法 [J]. 计算机学报，2010，33(8)：1369-1377.