使用二分图网络提高协同推荐的准确性

冷亚军[1] 陆 青[1] 张俊岭[2]

(上海电力学院经济与管理学院 上海 201300)[1] (浙江师范大学经济与管理学院 金华 321004)[2]

摘 要 协同过滤是推荐系统中广泛使用的最成功的推荐技术,但却面临着严峻的稀疏性问题。评分数据稀疏性使得最近邻搜寻不够准确,导致推荐质量较差。使用二分图网络缓解协同过滤推荐系统中的稀疏性问题,即将用户和项目抽象为二分图网络中的节点,重新分配项目资源并计算项目间资源贴近度,据此填充用户未评分项目,将稀疏评分矩阵转化为完全矩阵。采用近邻传播聚类对评分矩阵进行聚类,提高算法的可扩展性。最后提出了两种不同的在线推荐策略:(1)通过加权目标用户所在类的邻居用户评分产生推荐(BNAPC1);(2)通过各个类的总体偏好产生推荐(BNAPC2)。在MovieLens和Netflix数据集上进行了实验,结果表明BNAPC1的预测精度优于BNAPC2,且与其他几种常用的推荐算法相比仍具有一定优势。

关键词 推荐系统,协同过滤,二分图网络,近邻传播聚类

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.053

Using Bipartite Network for Enhancement of Collaborative Filtering

LENG Ya-jun[1] LU Qing[1] ZHANG Jun-ling[2]

(College of Economics and Management, Shanghai University of Electric Power, Shanghai 201300, China)[1]

(School of Economics and Management, Zhejiang Normal University, Jinhua 321004, China)[2]

Abstract Collaborative filtering is one of the most successful and widely used techniques among recommender systems. However, it suffers from serious problem in sparsity. Sparsity in ratings makes the formation of neighborhood inaccurate, thereby resulting in poor recommendations. In this paper, bipartite network was used to alleviate the sparsity problem in collaborative filtering. Users and items are mapped to nodes in bipartite network, and resources on items are redistributed. Resource approach degree between items is computed, and the original rating matrix is converted to complete matrix based on the resource approach degree. Then affinity propagation clustering was applied to cluster the rating matrix to improve the scalability of our approach. Finally, two different recommendation methods were presented. One is generating recommendations according to neighbors in the cluster which active user belongs to (BNAPC1), and the other is generating recommendations according to clusters' preferences (BNAPC2). Experiments on MovieLens and Netflix datasets show that BNAPC1 is more accurate than BNAPC2, and is also superior to existing alternatives.

Keywords Recommender systems, Collaborative filtering, Bipartite network, Affinity propagation clustering

1 引言

随着互联网上信息的增长和用户个性化需求的提高,推荐系统(recommender system)的应用日益广泛,成为电子商务、社会网络、视频/音乐点播等主流Web 2.0服务的核心技术[1]。推荐系统根据用户偏好,向用户提供个性化的信息、商品和服务的推荐,帮助用户解决信息超载(information overload)问题带来的困扰。根据推荐过程中使用方法的不同,推荐系统可以分为[2]:基于内容的推荐系统(content-based)、协同过滤推荐系统(collaborative filtering)和混合推荐系统(hybrid approach)。基于内容的推荐系统对项目属性进行分析,为用户建立用户描述,对项目属性和用户描述进行比较确定被推荐项目[3]。协同过滤推荐系统根据其他用户的偏好向目标用户进行推荐。它首先找出一组与目标用户偏好一致的邻居用户,然后对邻居用户进行分析,把邻居用户喜欢的项目推荐给目标用户[4]。混合推荐系统结合了基于内容的方法和协同过滤方法,它试图避免以上两种推荐系统存在的不足[2]。

协同过滤推荐系统不需要考虑项目的内容,且易于实现,因此它是迄今为止应用最为成功的推荐技术[5]。许多大型网站都应用了协同过滤推荐系统,如Amazon.com、Yahoo.com、Netflix.com等。尽管协同过滤在个性化推荐方面取得了巨大成功,但却面临着严峻的稀疏性问题(sparsity problem)[6,7]。实际网站中用户和项目的数量庞大且在不断增长,使得评分矩阵成为高维矩阵;同时用户通常只对一小部分项目进行评分,导致矩阵中的数据极端稀疏、用户之间共同评分项过少。数据稀疏性问题由此产生,使得最近邻用户难以

到稿日期:2014-05-01 返修日期:2014-07-01 本文受国家自然科学基金项目(71201145),教育部人文社会科学研究基金项目(11YJC630283),上海高校选拔培养优秀青年教师科研专项基金项目(sdl10021),上海市教育委员会科研创新项目(15ZS064)资助。

冷亚军(1985-),男,博士,讲师,主要研究领域为电子商务、数据挖掘,E-mail:huayi2001@163.com;陆 青(1982-),男,博士,讲师,主要研究领域为进化计算、数据挖掘;张俊岭(1981-),男,博士,副教授,主要研究领域为智能决策支持系统、进化算法。

搜寻或搜寻准确度不高，从而严重影响系统的推荐质量。

本文使用二分图网络来缓解协同过滤推荐系统中的稀疏性问题。本文提出的算法分为离线和在线两个阶段。在离线阶段，采用基于二分图网络的评分预测方法来填补评分矩阵中的缺失值，降低评分数据的稀疏性，并采用近邻传播聚类对评分矩阵进行聚类，减小最近邻搜寻范围，提高算法的可扩展性。在在线阶段，采用两种不同方法预测目标用户对未评分项目的评分，并完成推荐。MovieLens 和 Netflix 数据集的实验结果表明本文算法具有较高的预测精度。

本文第 2 节介绍新算法的离线处理阶段；第 3 节介绍新算法的在线推荐过程；第 4 节给出实验结果和分析；最后，总结全文并指出未来的工作。

2 离线阶段

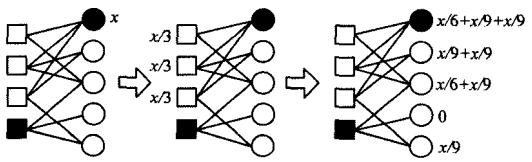
2.1 稀疏性处理

二分图(bipartite network，BN)是图论中的一种特殊模型，其在复杂网络的理论研究和实际应用中具有非常重要的意义。现实生活中很多系统可以用二分图来表示，比如包含学生和课程信息的选课网络、包含化学物质和化学反应的代谢网络、包含论文和作者的论文作者网络等[8,9]。一个二分图模型通常可以表示为 $G(X,Y,E)$，X 为顶层节点集合，Y 为底层节点集合，E 为网络中连接(边)的集合，X 或 Y 中任意两个在同一集合中的节点都不直接相连。

假设推荐系统中有 m 个用户和 n 个项目，我们用二分图 $G(U,I,E)$来表示该推荐系统。构造一个 $m\times n$ 阶权重矩阵 $W=\{w_{pq}\}(1\leqslant p\leqslant m,1\leqslant q\leqslant n)$，如果用户 u_p 对项目 i_q 的评分不为空，则 $w_{pq}=1$；否则 $w_{pq}=0$。对于用户 u_p 未评分的任一项目 i_t，为其分配 x 单位的资源。在二分图中，该资源首先传递给与 i_t 相连接的用户，再通过这些用户传递给所有项目，如图 1 所示。则任一项目 i_s 从项目 i_t 处获得的资源可以通过式(1)计算得到[9]：

$$d_{st}=\frac{x}{k(i_t)}\sum_{p=1}^{m}\frac{w_{ps}w_{pt}}{k(u_p)} \tag{1}$$

其中，$k(i_t)$表示对项目 i_t 进行过评分的用户数，$k(u_p)$表示用户 u_p 评价过的项目数。



方块和圆分别表示用户和项目，黑色方块表示目标用户，黑色圆表示待预测项目

图 1 二分图网络中的资源分配过程

对式(1)重复计算 n 次，可以得到所有项目的最终资源向量 $F=(d_{1t},d_{2t},\cdots,d_{nt})$。则项目 $i_s(1\leqslant s\leqslant n)$与未评分项目 i_t 的资源贴近度为：

$$c_{ts}=d_{st}/d_{tt} \tag{2}$$

其中，c_{ts}表示项目 i_t 愿意把自身资源分配给项目 i_s 的程度。c_{ts}的取值范围为[0,1]，越接近于 1，表示项目 i_t 越青睐于把资源分配给项目 i_s。

根据其他项目与项目 i_t 的资源贴近度，可以计算用户 u_p 对项目 i_t 的预测评分：

$$B_{pt}=\frac{\sum_{R_{ps}\neq\varnothing}c_{ts}\times R_{ps}}{\sum_{R_{ps}\neq\varnothing}c_{ts}} \tag{3}$$

其中，R_{ps}为用户 u_p 对项目 i_s 的真实评分。

对原始评分矩阵进行填补，得到较为密集的评分矩阵 A：

$$A_{pj}=\begin{cases}R_{pj}, & R_{pj}\neq\varnothing\\ B_{pj}, & R_{pj}=\varnothing\end{cases} \tag{4}$$

2.2 聚类

为了提高在线推荐的实时性，本文对填补后的矩阵进行聚类。常用的 k-means 聚类[10]和 k-medodis 聚类[11]随机选择一些数据点作为初始类代表点，致使聚类结果对初始类代表点的选择非常敏感，聚类准确性难以保证。本文采用近邻传播聚类(affinity propagation clustering，APC)[12]完成聚类工作。

APC 在数据点形成的相似度矩阵的基础上进行聚类，它将所有数据点作为候选类代表点。对于任意两数据点 x_i 和 x_k，它们之间的相似度可以通过欧氏距离来测度，$sim(i,k)=-\|x_i-x_k\|^2$。给定一组数据点和它们之间的相似度，APC 搜寻每一类别所包含的数据点和该类别对应的类代表点。APC 在各数据点之间传播两种信息：吸引度(responsibility)和归属度(availability)。算法的迭代过程就是这两个信息量交替更新的过程。吸引度 $res(i,k)$从点 x_i 指向点 x_k，代表点 x_k 积累的证据，用来表示 x_k 适合作 x_i 类代表点的程度；归属度 $av(i,k)$从点 x_k 指向点 x_i，代表点 x_i 积累的证据，用来表示 x_i 选择 x_k 作为类代表点的合适程度。对于任意数据点 x_i，计算所有数据点的吸引度 $res(i,k)$和归属度 $av(i,k)$之和，则 x_i 的类代表点为 x_k：$\arg\max_k\{av(i,k)+res(i,k)\}$。

设定相似度矩阵对角线元素 $sim(k,k)$为一相同值 $p=\delta\times sim_{ave}$[1]，初始化 $res^{(0)}(i,k)=av^{(0)}(i,k)=0$，则吸引度和归属度根据式(5)－式(7)进行更新[12]。

$$res(i,k)\leftarrow sim(i,k)-\max_{k'\text{ s.t. }k'\neq k}\{av(i,k')+sim(i,k')\} \tag{5}$$

$$\text{If } i\neq k, av(i,k)\leftarrow\min\{0,res(k,k)+\sum_{i'\text{ s.t. }i'\notin\{i,k\}}\max\{0,res(i',k)\}\} \tag{6}$$

$$av(k,k)\leftarrow\sum_{i'\text{ s.t. }i'\neq k}\max\{0,res(i',k)\} \tag{7}$$

对所有数据点求和信息量 responsibility 和 availability，找到每个点的类中心点。当满足以下条件，算法终止：(1)超过某一迭代最大数目；(2)信息改变量低于某一固定阈值；(3)选择的类中心在连续几步迭代过程中保持稳定。

判断得到的聚类数是否满足要求，如果不满足，则改变 p 值，重复进行上述算法直至聚类数满足要求为止。

3 在线推荐

3.1 算法描述

在在线推荐阶段，预测目标用户对未评分项目的评分，将评分较高的项目推荐给目标用户。本文采用两种方式计算预测评分：(1)通过加权目标用户所在类的邻居用户评分产生预测(BNAPC1)；(2)通过各个类的总体偏好产生预测(BNAPC2)。假设通过 APC 将用户集 U 划分为 t 类，得到用户类集合 $C=\{c_1,c_2,\cdots,c_t\}$，其中 $c_1\cup c_2\cup\cdots\cup c_t=U,c_i\cap c_j=$

[1] 称 p 为偏向参数，sim_{ave}是相似度矩阵中所有元素的平均值，δ 为任意实数。

$\emptyset(1\leqslant i\leqslant t, 1\leqslant j\leqslant t, i\neq j)$；相应的聚类中心为 $L=\{l_1, l_2, \cdots, l_t\}$。

算法 1 BNAPC1

输入：填补后的评分矩阵 A、用户类集合 C、聚类中心集合 L、目标用户 u 的评分向量 R_u、最近邻用户数 k、被推荐项目数 N

输出：目标用户 u 的 top-N 推荐集 I_r

过程：

Step1 采用 Pearson 相关系数计算目标用户 u 与 $L=\{l_1, l_2, \cdots, l_t\}$ 中每一聚类中心的相似性，取相似性最大的聚类中心对应的类 c_{max} 作为 u 所在的类

$$sim(u,l)=\frac{\sum_{i\in I_{ul}}(R_{u,i}-\overline{R_u})(R_{l,i}-\overline{R_l})}{\sqrt{\sum_{i\in I_{ul}}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{i\in I_{ul}}(R_{l,i}-\overline{R_l})^2}} \tag{8}$$

Step2 根据式(8)计算 u 与 c_{max} 中每位用户的相似性，取相似性从大到小排列的前 k 个用户作为 u 的最近邻集合 $U_n=\{u_1, u_2, \cdots, u_k\}$。

Step3 采用式(9)预测 u 对未评分项目 i 的评分 $P_{u,i}$。

$$p_{u,i}=\overline{R_u}+\frac{\sum_{u_k\in U_n} sim(u,u_k)\times(R_{u_k,i}-\overline{R_{u_k}})}{\sum_{u_k\in U_n}|sim(u,u_k)|} \tag{9}$$

Step4 按 $P_{u,i}$ 值从大到小取前 N 个项目组成 top-N 推荐集 $I_r=\{i_1, i_2, \cdots, i_N\}$ 并输出。

算法 2 BNAPC2

输入：用户类集合 C、聚类中心集合 L、目标用户 u 的评分向量 R_u、被推荐项目数 N

输出：目标用户 u 的 top-N 推荐集 I_r

过程：

Step1 采用式(8)计算目标用户 u 与 $L=\{l_1, l_2, \cdots, l_t\}$ 中每一聚类中心的相似性 $sim(u, l_j)$。

Step2 计算 $C=\{c_1, c_2, \cdots, c_t\}$ 中每一用户类的密集度。

$$\gamma(c_j)=|c_j|/\sum_{i=1}^{t}|c_i| \tag{10}$$

Step3 计算 L 中每一聚类中心对 u 进行项目推荐的可能度。

$$\rho(u,l_j)=sim(u,l_j)\times\gamma(c_j) \tag{11}$$

Step4 预测 u 对未评分项目 i 的评分 $P_{u,i}$。

$$p_{u,i}=\overline{R_u}+\frac{\sum_{j=1}^{t}\rho(u,l_j)\times(R_{j,i}-\overline{R_{l_j}})}{\sum_{j=1}^{t}|\rho(u,l_j)|} \tag{12}$$

Step5 按 $P_{u,i}$ 值从大到小取前 N 个项目组成 top-N 推荐集 $I_r=\{i_1, i_2, \cdots, i_N\}$ 并输出。

3.2 计算复杂度分析

传统搜寻最近邻的方式基于整个用户空间搜寻目标用户 u 的最近邻，计算复杂度为 $O(m\times n)$，因为 m 和 n 为同一数量级，所以 $O(m\times n)\approx O(n^2)$。BNAPC1 中 Step1 计算目标用户 u 与每一聚类中心的相似性，计算复杂度为 $O(t\times n)$；Step2 在 u 所属用户类中搜寻最近邻，计算复杂度为 $O(m_a\times n)$（m_a 表示 u 所属用户类的用户总数）；所以 Step1－Step2 整体计算复杂度为 $O(t\times n)+O(m_a\times n)$，因为 t 和 m_a 都远小于 n，所以整体计算复杂度为 $O(n)$。BNAPC2 中 Step1－Step3 计算每一聚类中心对目标用户 u 进行项目推荐的可能度，计算复杂度为 $O(t\times n)\approx O(n)$。BNAPC1 和 BNAPC2 由于采用了近邻传播聚类，因此计算复杂度显著降低。

4 实验结果及分析

4.1 数据集

本文使用 MovieLens 数据集[7]和 Netflix 数据集[13]对算法进行评估。MovieLens 数据集包含 943 位用户对 1682 部电影的 100000 条评分记录（评分值为 1－5 的整数）。Netflix 数据集包含 480189 位用户对 17770 部电影的 100480507 条评分记录（评分值为 1－5 的整数）。由于 Netflix 数据集规模过大，我们从中随机抽取 25413 条评分记录作为实验数据集。表 1 给出了实验用数据集的基本特征。我们将每一数据集划分为 80%的训练集用户和 20%的测试集用户，对于测试集用户采用 Given K 方案进行验证。

表 1 实验数据集基本特征

	用户数目	项目数目	评分数目	稀疏等级
MovieLens	943	1682	100000	0.9370
Netflix	500	1333	25413	0.9619

注：稀疏等级＝1－数据集中的评分总数/(用户数×项目数)

4.2 评价标准

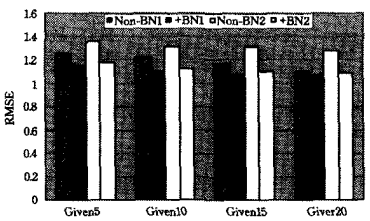
实验采用均方根误差（Root Mean Squared Error，RMSE）[14,15]作为度量算法优劣的标准。RMSE 通过计算用户的预测评分与实际评分之间的偏差来度量预测的准确性，RMSE 值越小，预测准确性越高。假设测试集用户共有 H 条评分数据被隐藏，分别为 $\{q_1, q_2, \cdots, q_H\}$，算法对这些评分数据的预测值为 $\{p_1, p_2, \cdots, p_H\}$，则算法的 RMSE 为：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{H}(p_i-q_i)^2}{H}} \tag{13}$$

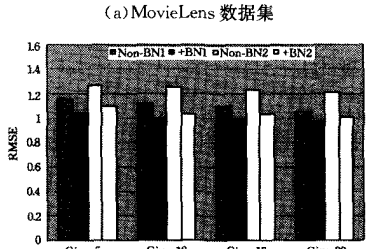
4.3 实验结果

4.3.1 BN 的作用

首先考查 BN 对算法准确性的影响，比较本文算法采用 BN(＋BN1、＋BN2)和不采用 BN(Non-BN1、Non-BN2)情况下的 RMSE。从图 2 的实验结果可以看出：(1)随着测试集用户可见评分的不断增多，各算法的预测准确性越来越高；(2)采用了 BN 的推荐算法，其准确性优于没有采用 BN 的推荐算法；(3)本文提出的两种推荐算法中，BNAPC1 的预测准确性优于 BNAPC2；(4)两个数据集上的实验结果基本相同，Netflix 数据集具有更低的 RMSE。实验结果证明了 BN 可以提高推荐算法的准确性。这是由于 BN 有效地缓解了评分数据的稀疏性，使最近邻搜寻更加准确，从而提高了算法的预测精度。



(a)MovieLens 数据集

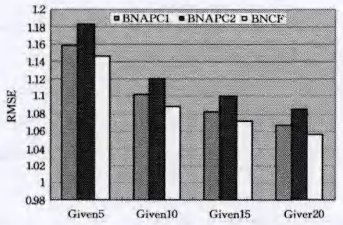


(b)Netflix 数据集

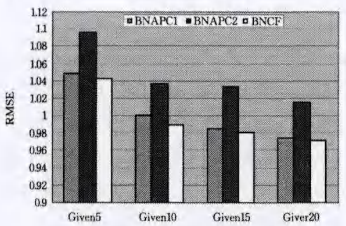
图 2 BN 的作用($\delta=2.5$，最近邻数 $k=30$)

4.3.2 APC的影响

本文算法采用APC对用户进行聚类，以提高在线推荐速度。APC会对算法的准确性产生影响，本节实验考查采用APC和不采用APC情况下算法的RMSE。在填补后的矩阵的整个用户空间搜寻最近邻，为传统的推荐算法，记为BNCF。实验结果如图3所示：(1)随着测试集用户可见评分的不断增多，各算法的预测准确性逐渐提高；(2)BNAPC2的准确性最低，与BNCF和BNAPC1有较大差距；(3)BNAPC1的准确性略低于BNCF，但差别不大。



(a)MovieLens数据集



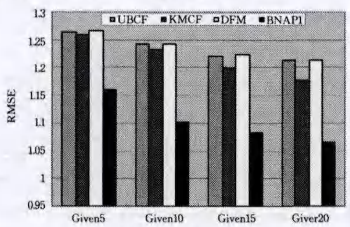
(b)Netflix数据集

图3 APC的作用($\delta=2.5$，最近邻数$k=30$)；

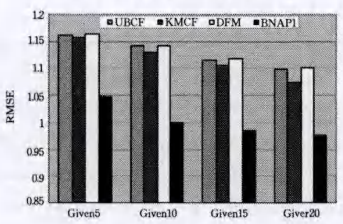
虽然BNAPC1的准确性稍逊于BNCF，但其在缩小了的聚类中搜寻最近邻，在线推荐速度较快；而BNCF在整个用户空间上搜寻最近邻，实时性较差。从准确性和推荐速度两方面综合考虑，BNAPC1的性能更高。

4.3.3 与其他算法比较

最后我们将本文准确性稍高的算法BNAPC1与基于用户的算法UBCF[16]、基于k-means聚类的算法KMCF[17]和基于均值偏差的算法DFM[18]进行比较。



(a)MovieLens数据集



(b)Netflix数据集

图4 与其他算法RMSE比较($\delta=2.5$，最近邻数$k=30$)

实验结果如图4所示，在MovieLens和Netflix数据集上的实验图形基本相同。随着Given K方案中K值的不断增加，即测试集用户可见评分的不断增多，各算法的RMSE越来越低。UBCF、KMCF和DFM的预测准确性大致相当。本文算法BNAPC1具有最低的RMSE，其预测准确性明显优于其他3种算法，这再次证明了BN在缓解数据稀疏性方面所起的作用。另外，BNAPC1离线阶段采用了近邻传播算法对用户进行聚类，在线阶段只需在目标用户所属聚类中搜寻最近邻即可，缩小了最近邻搜寻空间，其可扩展性优于传统的协同过滤算法。

结束语 推荐系统在向网络用户提供个性化服务方面发挥着至关重要的作用，目前许多大型网站都应用了推荐系统。协同过滤是推荐系统中广泛使用的最成功的推荐技术，但却面临着严峻的稀疏性和可扩展性问题。本文使用二分图网络缓解了协同过滤推荐系统中的稀疏性问题，根据项目资源在二分图网络中的分配来计算项目资源贴近度并填补用户未评分项，将稀疏评分矩阵转化为完全矩阵。采用近邻传播聚类对用户进行聚类，提高算法的可扩展性。最后提出了两种不同的在线推荐策略：(1)通过加权目标用户所在类的邻居用户评分产生推荐(BNAPC1)；(2)通过各个类的总体偏好产生推荐(BNAPC2)。在MovieLens和Netflix数据集上进行了实验，结果表明BNAPC1的预测精度优于BNAPC2，且与其他几种常用的推荐算法相比仍具有一定优势。未来的工作将尝试更加准确的稀疏性处理方法；另外，还将考虑引入性能更高的聚类方法。

参考文献

[1] 吴湖，王永吉，王哲，等. 两阶段联合聚类协同过滤算法[J]. 软件学报，2010，21(5)：1042-1054

[2] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749

[3] Barragáns-Martínez A B, Costa-Montenegro E, Burguillo J C, et al. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition[J]. Information Sciences, 2010, 180(22): 4290-4311

[4] Jeong B, Lee J, Cho H. An iterative semi-explicit rating method for building collaborative recommender systems[J]. Expert Systems with Applications, 2009, 36(3): 6181-6186

[5] Leung C W K, Chan S C F, Chung F L. A collaborative filtering framework based on fuzzy association rules and multiple-level similarity[J]. Knowledge and Information Systems, 2006, 10(3): 357-381

[6] Kim H N, Ji A T, Ha I, et al. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation[J]. Electronic Commerce Research and Applications, 2010, 9(1): 73-83

[7] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C] // Proceedings of the 10th International Conference on World Wide Web. 2001: 285-295

[8] 张译，靳雪翔，张毅，等. 基于二分图的城市公交网络拓扑性质研究[J]. 系统工程理论与实践，2007，7：149-155

[9] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4):

046115

[10] MacQueen J. Some methods for classification and analysis of multivariate observations[C] //Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297

[11] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341

[12] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976

[13] Ahn H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1): 37-51

[14] Jeong B, Lee J, Cho H. Improving memory-based collaborative filtering via similarity updating and prediction modulation[J]. Information Sciences, 2010, 180(5): 602-612

[15] Bogdanova G, Georgieva T. Using error-correcting dependencies for collaborative filtering[J]. Data & Knowledge Engineering, 2008, 66(3): 402-413

[16] Resnick P, Iacovou N, Suchak M, et al. Grouplens: an open architecture for collaborative filtering of netnews[C] // Proceedings of the 1994 ACM on Computer Supported Cooperative Work. 1994: 175-186

[17] Sarwar B, Karypis G, Konstan J, et al. Recommender systems for large-scale e-commerce: scalable neighborhood formation using clustering[C] //Proceedings of the 5th International Conference on Computer and Information Technology. 2002

[18] Herlocker J, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering[C] // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieve. 1999: 230-237

(上接第 244 页)

选择特征构建特征子空间，APBEFS 方法先采用排序聚合技术过滤剔除样本分类不相关特征，后采用基于 bicor 的特征关联距离度量的近邻传播聚类算法对特征进行分组后，从不同组随机选择一个特征构建特征子空间，二者提高了子空间特征的质量，降低特征之间关联性，从而获得了较好的分类性能；(2)特征排序聚合方法，尤其是基于数据扰动的方法，能够获得更为稳定的、准确的特征排序，而且选择前 500 个与样本分类关联紧密的特征用于后续分析是合适的。

结束语 本文提出了一种基于特征排序聚合和聚类的集成特征选择方法。该方法能够与一般的特征过滤方法相结合，具有更广阔的运用前景，同时采用排序聚合技术，降低了不同特征选择方法或数据扰动带来的影响，以 bicor 相关系数为关联准则利用近邻传播聚类对特征进行分组，并在此基础上采用随机方法构建特征子空间，从而获得了更优的、存在差异性的基分类器，最后采用多数投票法进行基分类器融合，获得了比单分类器更好更稳定的分类性能。对 7 个常用的基因表达数据的实验结果表明，本文提出的方法能够获得准确度较高的分类结果，同时分类性能稳定，具有较好的可扩展性。

参 考 文 献

[1] Opitz D W. Feature selection for ensembles[C]//Proceedings of 16th National Conference on Artificial Intelligence (AAAI-99). Orlando, FL, USA, 1999: 379-384

[2] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844

[3] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32

[4] De Bock K W, Coussement K, Van den Poel D. Ensemble classification based on generalized additive models[J]. Computational Statistics & Data Analysis, 2010, 54(6): 1535-1546

[5] 姚旭，王晓丹，张玉玺，等. 基于正则化互信息和差异度的集成特征选择[J]. 计算机科学，2013，40(6)：225-228

[6] Moon H, Ahn H, Kodell RL, et al. Ensemble methods for classification of patients for personalized medicine with high-dimensional data[J]. Artificial Intelligence in Medicine, 2007, 41(3): 197-207

[7] Liu Hua-wen, Liu Lei, Zhang Hui-jie. Ensemble gene selection by grouping for microarray data classification[J]. Journal of Biomedical Informatics, 2010, 43(1): 81-87

[8] Wald R, Khoshgoftaar T M, Dittman D. Mean aggregation versus robust rank aggregation for ensemble gene selection[C]// 2012 11th International Conference on Machine Learning and Applications (ICMLA). Boca Raton, FL, USA, 2012: 63-69

[9] Lin Song, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices[J]. BMC Bioinformatics, 2012, 13(1): 328

[10] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976

[11] Boulesteix A L, Slawski M. Stability and aggregation of ranked gene lists[J]. Briefings in Bioinformatics, 2009, 10(5): 556-568

[12] Wald R, Khoshgoftaar T M, Dittman D, et al. An extensive comparison of feature ranking aggregation techniques in bioinformatics[C]//2012 IEEE 13th International Conference on Information Reuse and Integration (IRI). Las Vegas, NV, USA, 2012: 377-384

[13] Wang Chang-dong, Lai Jian-huang, Suen C, et al. Multi-Exemplar Affinity Propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2223-2237

[14] Wang Yu-hang, Makedon F S, Ford J C, et al. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data[J]. Bioinformatics, 2005, 21(8): 1530-1537

[15] Sakellariou A, Sanoudou D, Spyrou G. Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data[J]. BMC Bioinforma-tics, 2012, 13(1): 270

[16] Hardin J, Mitani A, Hicks L, et al. A robust measure of correlation between two genes on a microarray[J]. BMC Bioinformatics, 2007, 8(1): 220

[17] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27