DOI:10.13203/j.whugis20130081　　文章编号:1671-8860(2015)02-0253-05

基于 MI 聚类的协同推荐算法

袁汉宁[1]　周　彤[2]　韩言妮[3]　陈媛媛[4]

1　北京理工大学软件学院,北京,100081
2　武汉大学国际软件学院,湖北 武汉,430079
3　中国科学院声学研究所高性能网络实验室,北京,100190
4　武汉理工大学经济学院,湖北 武汉,430070

摘　要:在个性化推荐系统中,项目的内容特征是影响推荐精度的重要因素。针对传统协同推荐不能有效考虑项目内容特征的问题,在考虑传统用户-项目评分信息的基础上,引入项目的内容特征属性,构建基于多示例(MI)的用户评分信息表达模型。根据多示例学习模式具有一定容错性的特点,设计了基于多示例聚类的协同推荐算法,通过多示例聚类计算用户的最近邻集合,根据最近邻集合对用户评分进行预测。实验结果表明,基于 MI 聚类的协同过滤推荐算法提高了预测评分的准确度,且有效缓解了数据稀疏性问题。

关键词:协同推荐;MI 聚类;内容特征
中图法分类号:TP391　　**文献标志码**:A

推荐系统可以根据用户的兴趣特点和购买行为向用户推荐用户感兴趣的信息和商品。推荐技术主要分为基于内容的推荐、协同推荐和混合推荐。基于内容的推荐技术是基于项目内容特征学习用户的兴趣,依据用户资料与待预测项目的匹配程度进行推荐。协同推荐技术基于用户及其评分数据,通过最近邻的搜寻来产生推荐结果。

基于内容的推荐仅考虑用户评价过的项目,推荐结果的多样性差,推荐质量较低。协同推荐中,人们往往只关注挖掘用户与项目之间的关联关系,而较少考虑项目的内容特征。但是,在许多应用场景下,仅仅依靠用户与项目之间的关联关系并不能生成有效的推荐,项目的内容特征是影响推荐精度的重要因素。在协同推荐算法中,引入项目内容特征属性的混合推荐算法可以有效提高推荐算法的精度,如李聪等[1]在协同推荐算法中引入项目的属性,较好地解决了数据的稀疏性问题与冷启动问题。目前的混合推荐[2]策略主要包括加权、集成和串联过滤等,虽然提高了推荐精度,但增加了计算的复杂度,需要引入组合参数,易出现参数过度拟合的情况。

多示例学习是一种新的学习框架,其独特的性质为样本提供了一种更灵活的多样化的表达方法。本文在考虑传统用户-项目评分信息的基础上引入项目内容特征,构建基于多示例的用户评分信息表达模型。用包表达用户对项目的评分信息,包内示例为用户评价过的项目,每个示例包含用户对项目的评分以及项目的内容特征属性。基于协同推荐的思想,采用 MI (multiple instance)聚类算法计算用户的最近邻居集合,在此基础上,加权平均最近邻对未知项目的评分,计算出用户对未知项目的评分。实验证明,引入项目的内容特征,并采用基于多示例聚类的协同推荐算法不仅有效缓解了数据的稀疏性问题,而且提高了推荐的有效性。

1　相关理论和算法

1.1　基于聚类的协同推荐算法

基于聚类的协同推荐对数据进行聚类,在聚类结果的基础上查找目标用户的最近邻集合。这种算法能够有效地缓解数据稀疏性的问题。

基于聚类的协同推荐算法(一般采用 *K*-means 算法[3]聚类)首先取定聚类簇数,利用聚类算法对用户-项目评分数据进行聚类;在得到聚类结果之后,需要找到离目标用户最近的簇(也可以是最近的若干个簇),将整个簇中用户作为目标用

收稿日期:2013-04-25
项目来源:国家自然科学基金资助项目(61173061,61472039,61303252,71201120);中央高校基本科研业务费专项资金资助项目(2012-IV-053)。

第一作者:袁汉宁,副教授,博士,现从事机器学习、商务智能研究。E-mail:yhn1979yhn@163.com

户的最近邻(或者计算目标用户与最近的一个或者若干个簇中用户的相似度,选取相似度最高的若干个用户作为目标用户的最近邻集合);最后,加权平均最近邻集合中用户对目标项目的评分,可以得到目标用户对目标项目的评分,从而可以预测目标用户对所有未知项目的评分。

1.2 MI聚类

多示例学习最先由Dietterich[4]等在药物活性预测中提出,是一类特殊的机器学习任务。在多示例学习中,训练样本是由多个示例组成的包,包是有概念标记的,但示例本身没有概念标记。不同于常见的聚类算法,MI聚类中用到的数据是多示例数据形式的数据。Zhang[5]使用平均Hausdorff距离度量包间距离,在此基础上根据K-中心点(K-medoids)聚类算法,提出了BAMIC(BAg-level multi-instance clustering)聚类算法。

该算法聚类的结果会因计算包之间距离公式的不同而有所不同。常用的计算包间距离的方法有最大Hausdorff距离[6]和最小Hausdorff距离[7]。由于最大和最小Hausdorff距离对孤立点很敏感,因而在MI聚类中计算包间距离采用平均Hausdorff距离。平均Hausdorff距离考虑了两个包中所有示例之间的几何关系,得出的聚类结果更理想。假如给出两个包$A=\{a_1,a_2,\cdots,a_m\}$,$B=\{b_1,b_2,\cdots,b_m\}$,平均Hausdorff距离如式(1)所示,其中$\|a-b\|$表示示例a、b间的欧氏距离。

$$\mathrm{ave}H(A,B)=\frac{\sum_{a\in A}\min_{b\in B}\|a-b\|+\sum_{b\in B}\min_{a\in A}\|b-a\|}{|A|+|B|} \quad (1)$$

2 基于MI聚类的协同推荐

传统的基于用户聚类的协同推荐算法仅挖掘用户与项目之间的关联关系,而较少考虑项目的内容特征。基于多示例学习的协同推荐算法采用多示例数据表达模型,不仅可以兼顾用户与项目之间的关联关系和项目的内容特征的表达,还可以有效地缓解数据的稀疏性问题,通过多示例聚类得到的最近邻集合更为准确。

2.1 基于多示例的用户评分信息表达模型

准确表达用户评分信息是实现有效推荐的关键。多示例学习的独特性质为用户评分信息的表达提供了一条新的途径。以电影推荐为例,传统的协同推荐算法用到的数据模型与基于多示例的用户评分信息表达模型如表1、图1所示。

表1 基于用户-项目的评分信息表达

Tab.1 Representation of User-item Based Ratings Information

	电影1	电影2	电影3	…	电影n
用户1	R11	R12	R13	…	R1n
用户2	R21	R22	R23	…	R2n
用户3	R31	R32	R33	…	R3n
⋮	⋮	⋮	⋮	…	⋮
用户m	Rm1	Rm2	Rm3	…	Rmm

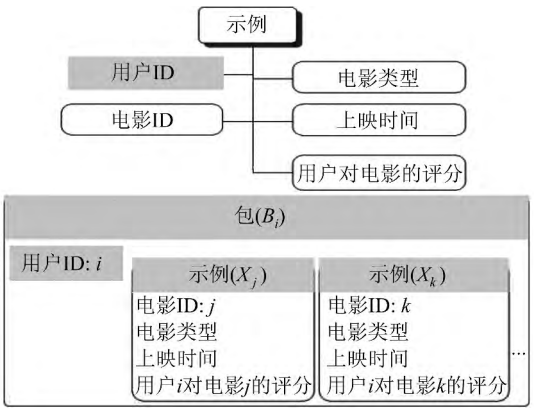

图1 基于多示例的表达模型

Fig.1 MI-based Representation Model

表1中,传统的协同推荐算法中用到的数据是用户-项目评分矩阵,矩阵中的数据为用户对电影的评分。图1所示的基于多示例的用户评分信息表达模型中,每个包B_i代表一个用户对所有看过电影的评分状况,$B_i=(X_1,X_2,\cdots,X_n)$,其中i表示第i个用户,n表示包中示例的个数,即该用户对n部电影进行了评分。包中的每个示例$X_i=(X_{i1},X_{i2},\cdots,X_{im-1},X_{im})$表示用户对某部电影的评分及该电影的内容特征,其中$(X_{i1},X_{i2},\cdots,X_{im-1})$是电影的内容特征向量,$(X_{im})$是用户对该部电影的评分。

与传统的用户-项目评分矩阵相比,基于多示例学习的用户评分表达模型具有以下特点:

1) 多示例学习中,包(用户评分信息)和示例(用户对某个项目,评分信息)是一对多的映射关系[8],不仅适合描述用户评分状况,而且很好地表达了用户评分信息与用户对每个项目评分信息的结构关系。

2) 可以兼顾表达用户对项目的评分和项目的内容特征。

3) 由于有包的限制(包是有标记的,而示例是无标记的),多示例学习中的概念知识具有一定的模糊性和不确定性,学习模式具有一定的容错性[9],对用户评分数据中的噪声、模糊性和缺省具有较强的容忍。

2.2　基于 MI 聚类的协同推荐算法

基于 MI 聚类的协同推荐算法在利用 BAMIC 算法的基础上,在整个用户集中找到与目标用户相似度最高的若干个用户作为其最近邻集合,在计算最近邻集合的基础上预测用户对未知项目的评分。该算法具体描述如下。

1) MI 聚类。① 任意选取 k 个包作为初始的簇中心 C_1、C_2、…、C_k,初始的簇中心组成一个集合 $G_j=\{C_j\}$,其中 $j=1,2,\cdots,k$;② 对于集合 B 中的所有包,逐个计算其到每个簇中心的包的距离,将其赋给最近的簇;③ 重新计算找到每个簇的簇中心的包;④ 重复步骤②和③,直到不再发生变化。

2) 最近邻查找。① 计算目标用户到 k 个簇的距离,找到最近的一个簇;② 依式(2)计算目标用户与簇中用户 U_1、U_2、…、U_n 的相似度,得到 sim(user,U_1)、sim(user, U_2)、…、sim(user,U_n)。

$$\mathrm{sim}(a,b)=\frac{\sum_{j\in I_{a,b}}(r_{aj}-\bar{r}_a)(r_{bj}-\bar{r}_b)}{\sqrt{\sum_{j\in I_{a,b}}(r_{aj}-\bar{r}_a)^2(r_{bj}-\bar{r}_b)^2}} \quad (2)$$

式中,$\mathrm{sim}(a,b)$ 表示用户 a 与用户 b 之间的相似度;r_{aj}、r_{bj} 分别表示用户 a 和用户 b 对项目 j 的评分;$\bar{r}_a$、$\bar{r}_b$ 分别表示用户 a 和用户 b 对项目的平均评分。③ 将 sim(user, U_1)、sim(user, U_2)、…、sim(user, U_n)排序,取相似度最高的 N 个用户放入最近邻居用户集合 N。④重复上述过程,直到找到所有用户的最近邻居集合。

3) 推荐集生成。① 找到最近邻居集合中对未知项目有评分的用户,根据式(3)预测目标用户对项目的评分,得到用户对未知项目的评分。

$$P_{aj}=\bar{r}_a+k\sum_{i=1}^{N}\mathrm{sim}(a,b)(r_{bj}-\bar{r}_b) \quad (3)$$

式中,N 表示目标用户 a 对应的最近邻居的集合。② 设定一个阈值,选取用户对其评分超过该值的项目形成用户的推荐集。

基于 MI 聚类的协同推荐算法是建立在基于多示例的用户评分信息表达模型基础上的,包和示例的一对多的映射关系能充分利用已评价的项目信息缓解数据的稀疏性问题;示例的信息表达兼顾了项目评分信息和项目的内容特征,可以有效提高推荐的有效性。

3　实　验

3.1　数据及环境

为了验证本文算法的有效性,实验选择 MovieLens 站点提供的数据集(http://movielens.umn.edu/)。数据集中有 943 名用户,1 682 部电影,用户对电影的评分记录有 100 000 条。数据集中用户的评分数据为 1～5 间的整数,每个用户至少对 20 部电影进行了评分。实验中,传统的协同过滤算法使用的数据是用户-项目评分矩阵,仅考虑用户对电影的评分,而基于多示例聚类的协同过滤算法使用的数据是包和示例,不仅考虑了用户对电影的评分,而且还考虑了电影的内容特征(电影类型、上映时间)。

实验采用五折交叉验证法,取数据集中 80% 的数据作为训练集,20% 的数据作为测试集,对预测评分达到 4 分及以上的项目做推荐。本文进行了十次五折交叉验证,取其均值作为对算法准确性的估计。实验通过 Java 平台和 Weka 软件来完成数据集的格式化以及推荐算法的实现。

3.2　评价标准

本文采用 Tsai 等人[10]的方法来衡量算法推荐的结果。对于一个未被用户评分的电影,可能出现的结果有 4 种,即推荐给用户且用户很喜欢,推荐给用户但是用户不喜欢,用户喜欢但是没有推荐,用户不喜欢且没有推荐。表 2 总结了这 4 种可能的情况[11],其中 TP (true positive)、FN (false negative)、FN 以及 TN (true negative)分别表示 4 种情况的数目。

表 2　待预测电影的可能的 4 种情况

Tab. 2　Four Possible Cases of Movies to be Predicted

用户偏好	算法推荐	算法不推荐
喜欢	TP	FN
不喜欢	FN	TN

评价推荐结果的标准有准确率、精确率与召回率。准确率表示正确推荐的电影占总电影数的比例,精确率表示所有推荐的电影中应该推荐的电影所占的比例,而召回率则表示应该推荐的电影中推荐的电影所占的比例。准确率、精确率和召回率的值越大,表明推荐结果的质量越高。

另外,本文选取常用的协同推荐质量的评价指标平均绝对误差(mean absolute error, MAE)[12],通过计算预测得出的评分与用户的实际评分间的偏差来衡量预测的准确性。MAE 越小,表明预测越准确,推荐效果越好。

3.3　实验结果及分析

为了评价本文提出的模型和方法的有效性,将传统的基于聚类的协同推荐算法和基于 MI 聚类的协同推荐算法进行了比较。计算准确率、精确率、召回率与 MAE 时,聚类个数从 10 增加到 30,间隔为 5。实验结果如图 2、图 3 与图 4 所示。

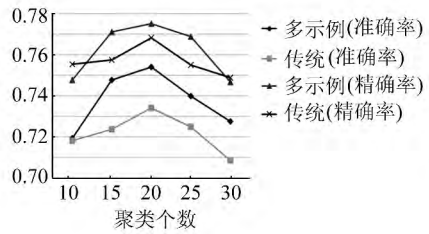

图2 准确率、精确率比较

Fig. 2 Comparison of Accuracy, Precision

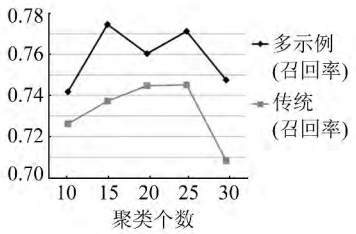

图3 召回率比较

Fig. 3 Comparison of Recall

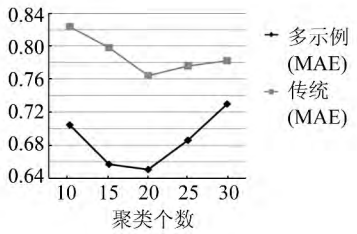

图4 MAE比较

Fig. 4 Comparison of MAE

通过图2、图3与图4可以看出，与传统的基于用户聚类的协同推荐算法相比，基于多示例聚类的协同推荐结果都有更高的准确率、精确率和召回率以及更低的MAE，推荐质量高于传统的基于用户聚类的协同推荐算法。随着最近邻居个数的增加，两类算法的推荐质量均先升后降，当邻居个数为20个时，具有较好的推荐质量。

综上所述，与传统的基于聚类的协同推荐算法相比，基于MI聚类的协同过滤算法由于考虑了项目的内容特征和充分利用了用户评分信息，推荐的结果更趋近于用户的真实评分，推荐精度更高。

4 结 语

传统的协同过滤推荐算法无法准确地表达用户的偏好，受数据稀疏性影响较大，导致推荐结果不够理想。针对上述问题，本文引入以多示例学习为理论基础建立的用户评分信息表达模型，从包的角度重新表示一个用户的评分状况，引入项目的内容特征属性，以每个示例表示用户对某个项目的评分，这种方法可以较为准确地表达用户偏好。实验结果表明，与传统的协同推荐算法相比，基于MI聚类的协同过滤推荐算法全面考虑了项目的内容特征和用户与项目的关联关系，不仅提高了推荐结果的准确性，也有效地缓解了数据的稀疏性问题。

MI聚类算法可能出现簇中包的数量过少，找不到目标用户的最近邻居，因而有必要探索更适合聚类多示例数据的聚类方法。另外，在基于多示例的协同推荐算法的基础上，合理地运用多示例学习方法，研究更有效的推荐算法，是未来要做的工作。

参 考 文 献

[1] Li Cong, Liang Changyong. A Collaborative Filtering Recommendation Algorithm Based on Attributes-Value Preference Matrix[J]. *Journal of the China Society for Scientific and Technical Information*, 2009, 27(6): 884-890(李聪，梁昌勇. 基于属性值偏好矩阵的协同过滤推荐算法[J]. 情报学报，2009，27(6)：884-890)

[2] Li Zhongjun, Zhou Qihai, Shuai Qinghong. System Model Based on Isomorphic Integrated to Content-based and Collaborative Filtering[J]. *Computer Science*, 2009, 36(12): 142-145(李忠俊，周启海，帅青红. 一种基于内容和协同过滤同构化整合的推荐系统模型[J]. 计算机科学，2009，36(12)：142-145)

[3] Zheng Dan, Wang Qianping. Selection Algorithm for *K*-means Initial Clustering Center[J]. *Journal of Computer Application*, 2012, 32(8): 2 186-2 188(郑丹，王潜平. *K*-means初始聚类中心的选择算法[J]. 计算机应用，2012，32(8)：2 186-2 188)

[4] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the Multiple Instance Problem with Axis-parallel Rectangles[J]. *Artificial Intelligence*, 1997, 89(1): 31-71

[5] Zhang M L, Zhou Z H. Multi-instance Clustering with Applications to Multi-instance Prediction[J]. *Applied Intelligence*, 2009, 31(1): 47-68

[6] Zhang Zhizheng, Xing Hancheng. A Concept Learning Method in Case-based Reasoning[J]. *Computer Engineering and Application*, 2006, 42(10): 87-90(张志政，邢汉承. 一种基于实例推理的概念学习方法[J]. 计算机工程与应用，2006，42(10)：87-90)

[7] Wang J, Zucker J D. Solving the Multiple-instance Problem: A Lazy Learning Approach[C]//Proceedings of the 17th ICML. San Francisco: Morgan Kaufmann Publishers, 2000

[8] Zhou Z H, Yu Y. Ensembling Local Learners Through Multimodal Perturbation[J]. *IEEE Trans SMC—Part B: Cybernetics*, 2005, 35(4): 725-735

[9] Zafra A, Romero C, Ventura S. Multiple Instance Learning for Classifying Students in Learning Management Systems[J]. *Expert Systems with Applications*, 2011, 38(12): 15 020-15 031

[10] Tsai C F, Hung C. Cluster Ensembles in Collaborative Filtering Recommendation[J]. *Applied Soft Computing*, 2012, 12(4): 1 417-1 425

[11] Zhu Yuxiao, Lv Linyuan. Evaluation Metrics for

Recommender Systems[J]. *Journal of University of Electronic Science and Technology of China*, 2012, 41(2): 163-175(朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163-175)

[12] Li Chun, Zhu Zhenmin. Collaborative Filtering Recommendation Algorithm Based on Neighbor Decision-making[J]. *Computer Engineering*, 2010, 36(13): 34-36(李春, 朱珍民. 基于邻居决策的协同过滤推荐算法[J]. 计算机工程, 2010, 36(13): 34-36)

Collaborative Recommendation Algorithm Based on MI Clustering

YUAN Hanning[1]　*ZHOU Tong*[2]　*HAN Yanni*[3]　*CHEN Yuanyuan*[4]

1 International School of Software, Beijing Institute of Technology, Beijing 100081, China

2 International School of Software, Wuhan University, Wuhan 430079, China

3 High Performance Network Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

4 School of Economics, Wuhan University of Technology, Wuhan 430070, China

Abstract: In a personalized recommendation system, the context feature of item is an important factor affecting recommendation accuracy, but traditional collaborative recommendation algorithms cannot take context feature of item into account effectively. To solve this problem we constructed a user-item ratings information representation model with multiple instance learning(MIL) based on traditional user-item ratings information, considering the context features of an item. Exploiting a characteristic trait of MIL, its strong tolerance to fault, a collaborative recommendation algorithm based on MI clustering was designed which computes users' nearest neighbors by MI clustering and predicates user ratings according to the neighbors. Experimental results confirmed that collaborative recommendation algorithm based on MI clustering improved accuracy of predictions and alleviated the problem of sparse data effectively.

Key words: collaborative recommendation; MI clustering; context feature

First author: YUAN Hanning, associate professor, PhD, specializes in machine learning and business intelligence. E-mail: yhn1979yhn@163.com

Foundation support: The National Natural Science Foundation of China, Nos. 61173061, 61472039, 61303252, 71201120; the Fundamental Research Funds for the Central Universities, No. 2012-IV-053.

(上接第 252 页)

Selection of Antarctic Research Stations Based on GIS and Fuzzy AHP

LIU Haiyan[1,2]　*PANG Xiaoping*[1,2,3]

1 School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China

2 Key Laboratory of Polar Surveying and Mapping, SBSM, Wuhan 430079, China

3 Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan 430079, China

Abstract: Site selection for the Antarctic research stations affects safety, functionality, and operational efficiency of the stations. This study aims to build a criteria system and conduct the site selection aided by Geographical Information Systems and a Fuzzy Analytical Hierarchy Process. Fifteen criteria were proposed as multiple evaluation sub-criteria and were grouped into four main criteria: scientific interest, environment, accessibility and topography. Comparisons were made between the predicted suitable areas and locations of existing stations in Antarctica to show the fitness-for-use of allocation results. This work offers a comprehensive allocation methodology for decision-makers in the assessment of Antarctic station allocation problems.

Key words: GIS; fuzzy analytical hierarchy process; criteria system; Antarctic research station

First author: LIU Haiyan, PhD candidate, specializes in cartography and GIS application, data model analysis and uncertainty. E-mail: liuhaiyan@whu.edu.cn

Foundation support: The State Oceanic Administration Polar Exploration Office Polar Science Strategy Research Fund Project, No. 20120202.