DOI:10.13245/j.hust.2013.s2.034

一种基于循环回归的推荐算法

许逸格[a] 张 可[a] 柯 朦[a] 谢倩倩[b] 章 文[a]

(武汉大学 a 计算机学院; b 数学与统计学院, 湖北 武汉 430072)

摘要 提出了一种基于循环回归的推荐算法.首先,对原数据集中的评分数据及缺失值进行离散化处理,然后对离散化数据进行回归模型训练,此过程循环执行并最终建立推荐系统.在离散化阶段,对比不同的离散方法,并对它们的分类粒度开展研究.在模型训练阶段,讨论回归算法对于模型性能的影响.数值计算实验表明,本算法较之近年非常热门的 SVDFeaute 方法,能够产生更小的均方根误差,验证了算法的有效性.

关键词 推荐系统; 回归分析; 循环; 数据离散化; 数据挖掘; 缺失数据

中图分类号 TP319 **文献标志码** A **文章编号** 1671-4512(2013)S2-0188-04

A recommendation algorithm based on cyclic regression

Xu Yige[a] *Zhang Ke*[a] *Ke Meng*[a] *Xie Qianqian*[b] *Zhang Wen*[a]

(a School of Computer; b School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)

Abstract A novel cyclic regression-based algorithm was presented, in which the user's rates and missing values were discredited and the original problem was thus transformed as the regression problem. Different discretization methods were compared, and different regression methods were used to construct the models. In computational experiments, compared with the popular method SVDFeaute, the proposed algorithm can lead to the better performance in terms of the root mean squared error.

Key words recommender system; regression analysis; recycling; data discretization; data mining; missing data

随着计算机网络的发展,人们在互联网上购物和享受服务成为一种发展趋势.推荐系统是对信息过滤的系统,它通过建立用户行为和项目信息的海量数据集,对用户兴趣模式进行发掘和分析,从而为用户推荐符合其爱好的信息和实物(如:电影、图书、电视节目等).

在工业界,随着长尾理论[1]的提出,电子商务领域(Amazon, Youtube, Netflix)对推荐系统的重视程度日益提高. Amazon 前首席科学家 Andreas Weigend 曾指出,亚马逊有 20%~30%的销售额都来源于推荐系统[2].在学术界,大量的竞赛如 Netflix 百万美元大赛[3]等不仅提供了海量数据集,同时也使得推荐算法受到广泛关注,思路迥异的算法不断被提出.

推荐系统自 20 世纪 90 年代提出后不断发展.协同过滤推荐算法对用户兴趣、项目特征进行分析,发掘相似集合或特征,挖掘相似集合或特征的关联和内涵并对用户兴趣做出预测.此类方法有基于用户的协同过滤推荐算法[4],基于项目的协同过滤推荐算法[5],隐语义协同过滤推荐算法[6]等.除此之外,还有一些协同过滤推荐算法通过增加偏见值、时间变迁[7-8]等影响因素或将其与其他模型融合[9-10].基于内容的推荐算法源于信息检索领域,根据对用户过去喜欢的项目的模式进行分析,为用户推荐与其喜欢项目相似的项目.主要方法有 TF-IDF[11]和 LSA[12]等.协同过滤推荐算法无需领域知识,但是存在冷启动问题和数据稀疏性问题.基于内容的推荐算法简单有效,结

收稿日期 2013-07-25.

作者简介 许逸格(1991-),女,硕士研究生; 章 文(通信作者), 副教授, E-mail: zhangwen@whu.edu.cn.

基金项目 教育部博士学科点专项科研基金资助项目 (20100141120049); 湖北省自然科学基金资助项目 (2011CDB454); 国家自然科学基金资助项目 (61103126).

果易理解，但存在应用领域狭窄和用户冷启动问题[13-15].

在前人工作的基础上，本文提出了一种基于循环回归的推荐算法，通过离散化评分矩阵数据，并对离散数据进行回归模型训练，从而预测评分.

1　基于循环回归的推荐算法

Yehuda Koren 在 Netflix 比赛的后续研究中，对推荐系统用户冷启动问题的解决方案做出了探讨[16]. 他指出，通过对电影数据集的已知评分矩阵的数据进行分析，可构建将不同电影作为节点的决策树. 其建立决策树的核心思想是，将所有用户评分映射到〈lovers，haters，unknowns〉的三维空间上，计算每个项目的总方差，选择最小的从根节点开始自顶向下建树. 当达到终止条件时，停止树的构造.

受到 Yehuda Koren 将未知评分映射到 unknowns 类的启发，本文提出了一种新的推荐算法—基于循环回归的推荐算法(cyclic regression based algorithm，下文简称 CR-Based 算法). 算法的主要思想有两点. a. 将推荐问题转化为回归问题. 对于评分矩阵，每个项目为一行，其中一列作为目标用户的决策属性，其他列作为条件属性. 离散化条件属性值时，条件属性的缺失值单独映射到一类(unknowns). 通过构建条件属性到决策属性已知的映射，将评分预测问题转化为回归问题，从而预测决策属性的未知属性值. b. 循环目标用户. CR-Based 算法循环地将其中一列作为决策属性，其他列作为条件属性，对所有的未知评分进行预测.

输入一个以 n 个项目为行、m 个用户为列的打分矩阵 $\boldsymbol{M}$(其中部分打分已知)，算法执行步骤描述如下. 对于 m 列，将其中一列作为决策属性，其他列作为条件属性. 按列，分别离散化 $m-1$ 个条件属性的值. 从矩阵的 n 行中，提取决策属性值已知的行作为训练数据集，采用回归方法训练预测模型后对其他行的数据预测得到评价值. 由于矩阵存在 m 个列，重复上述过程(过程彼此独立)，直到每个列都被作为决策属性使用过一次为止. 此时，矩阵中所有打分都被补充完全.

2　数值实验分析

2.1　实验数据及方法

采用两个数据集. 一个数据集是明尼苏达大学 Grouplens 小组的 Movielens 数据集[17]. 该数据集包含从其推荐系统项目 Movielens 上采集的来自 943 个用户对 1 682 部电影的100 000条评分. 另外一个是明尼苏达大学 Grouplens 研究组在 2008 年 2 月从其协同过滤推荐系统 Wikilens 上抽取的数据构成的 Wikilens 数据集[18]. 本文从数据集随机抽样 10% 用户，并进行三叠交叉验证. 现有推荐算法评测标准中，均方根误差(root mean squared error，记为 E_{RMS}) 最为重要，因此采用均方根误差指标评估模型性能.

在数据离散阶段，本文考虑了 4 种不同的离散化方法，如表 1 所示. 这里，数据集为评分式，范围为 1～5 分，从 1 到 5 表示喜欢程度递增.

表 1　数据离散化方法表

数值填补法	离散前	1	2	3	4	5	未知
(NumberFill)	离散后	1	2	3	4	5	0
简单填补法	离散前	1	2	3	4	5	未知
(SimpleFill)	离散后	A	B	C	D	E	U
混合填补法	离散前	1	2	3	4	5	未知
(MixFill)	离散后	H	H	O	L	E	U
分类填补法	离散前	1	2	3	4	5	未知
(ClassifyFill)	离散后	H	H	H	L	L	U

数值填补法：保持原有评分分数，将评分矩阵中的缺失值填补为 0. 简单填补法：保持原有评分分数，但将其映射为符号 A、B、C、D、E，缺失值表示为 U(Unknown). 混合填补法：根据评分分布，将原有评分映射为四类. 其中，5 分为 E(Excellent)类，4 分为 L(Like)类，3 分为 O(Ordinary)类，2 分和 1 分为 H(Hate)类，缺失值表示为 U. 分类填补法：根据评分含义，将原有评分映射为两类. 其中，大于 3 分映射为 L(Love)类，小于等于 3 分映射为 H(Hate)类，缺失值仍表示为 U.

算法模型训练阶段，本文对 WEKA 数据挖掘库提供的五种回归模型进行了探讨：GaussianProcesses，AdditiveRegression，RegressionByDiscretization，M5P，M5Rules.

2.2　实验结果分析

主要在 Movielens 数据集上分析离散化方法和分类方法的作用. 不同的数据离散化方法的均方根误差 E_{RNS} 如图 1 所示. ClassifyFill 方法的均方根误差较小，其平均均方根误差仅为 0.768，SimpleFill 方法的均方根误差较大；数值化离散方法比符号化离散方法均方根误差更小；对符号化离散方法，随着分类粒度变大其均方根误差减小. 一般说来，数值化离散方法可比符号化离散方法取得更好的结果. 然而，符号化离散方法随着分类粒度的增大能取得更好的实验结果.

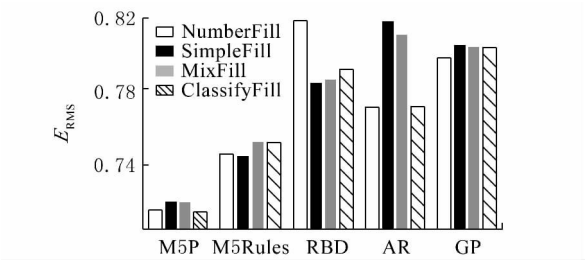

图 1 数据离散化方法之均方根误差分析图

不同的回归模型的均方根误差如图 2 所示. M5P 和 M5Rules 方法表现最好,其中 M5P 更为精确,其均方根误差值均在 0.72 左右,比 M5Rules 平均分低 0.03 左右. 其余三种方法总体均方根误差较大的. 结合其他指标如相关系数和相对平方根误差(结果略),M5P 的方法无论在哪种评判方式下都为最优. 由于推荐算法的主要评判方式为均方根误差,因此 M5Rules 方法次之.

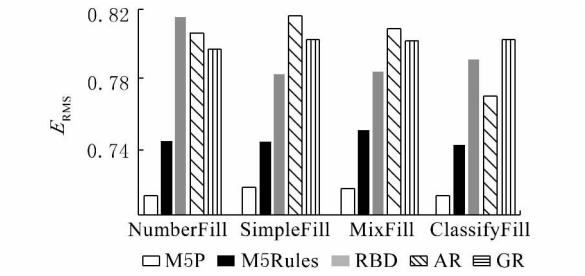

图 2 模型训练方法之均方根误差分析图

根据以上的研究,CR-Based 算法模型的最终执行方案为:采用分类填补法进行数据离散化,使用 M5P 方法训练回归模型.

2.3 算法对比

为了验证算法的有效性,本文选取上海交通大学 SVDFeature Toolkit[19] 中的默认算法作为对比,CR-Based 采用 2.2 节中讨论的最优方案.

SVDFeature 在 Movielens 数据集上的平均结果如表 2 所示. 经过默认 10 轮训练后,SVDFeature 的均方根误差为 1.02,而 CR-Based 算法的均方根误差仅为 0.72,CR-Based 的预测结果有了提升.

表 2 SVDFeature 在 Movielens 数据上的运行结果

轮数	结果	轮数	结果
0	3.24	6	1.05
1	1.62	7	1.04
2	1.28	8	1.03
3	1.16	9	1.02
4	1.10	10	1.02
5	1.07		

SVDFeature 在 Wikilens 数据集上的平均结果如表 3 所示. 经过默认 10 轮训练后,SVDFeature 的均方根误差为 1.18,而 CR-Based 算法的均方根误差仅为 0.69,也取得了较好的结果.

表 3 SVDFeature 在 Wikilens 数据集上的运行结果

轮数	结果	轮数	结果
0	3.30	6	1.24
1	1.79	7	1.22
2	1.48	8	1.21
3	1.36	9	1.19
4	1.30	10	1.18
5	1.26		

3 结束语

基于循环回归的推荐算法采用数据离散化方法填补评分矩阵中的缺失值,然后建立回归模型对评分矩阵缺失值进行了细化的预测. 该算法也存在一些缺点:a. 数据稀疏性的问题并没有得到完全的解决,对未知数据的处理实际上仍旧在于统一为其赋予初值;b. 实际问题中存在超大规模矩阵,因此评分矩阵的空间需求会很大. 而在模型训练阶段,需要分别处理评分矩阵的每一列,也会使得计算复杂度较高.

参 考 文 献

[1] Anderson C. The long tail[J]. Wired, 2004, 12(10): 1-5.

[2] Francesco Ricci, Lior Rokach, Bracha Shapira, et al. Recommender systems handbook [M]. Berlin: Springer, 2011.

[3] Töscher A, Jahrer M, Bell R M. The bigchaos solution to the netflix grand prize[R]. [s. l]. Netflix prize documentation, 2009.

[4] Herlocker J L, Konstan J A, Borchers A L, 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 Retrieval, New York: ACM, 1999: 230-237.

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

[6] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[7] Bhaskar Mehta, Thomas Hofmann, Wolfgang Nejdl [C]// RecSys'07 Proceedings of the 2007 ACM Conference on Recommender Systems, New York:

ACM，2007：49-56.

[8] Koren Y. Collaborative filtering with temporal dynamics[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York：ACM，2009：447-456.

[9] Koren Y. Factorization meets the neighborhood：a multifaceted collaborative filtering model[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008：426-434.

[10] Koren Y. Factor in the neighbors：scalable and accurate collaborative filtering[J]. ACM Transactions on Knowledge Discovery from Data (TKDD)，2010，4(1)：1-24.

[11] Salton G，Wong A，Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM，1975，18(11)：613-620.

[12] Deerwester S，Dumais S T，Furnas G W，et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science，1990，41(6)：391-407.

[13] Aizenberg N，Koren Y，Somekh O. Build your own music recommender by modeling internet radio streams[C]// Proceedings of the 21st International Conference on World Wide Web，2012：1-10.

[14] 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.

[15] Schafer J B，Frankowski D，Herlocker J. Collaborative filtering recommender systems[M]. Heidelberg：Springer 2007：291-324.

[16] Golbandi N，Koren Y，Lempel R. Adaptive bootstrapping of recommender systems using decision trees[C]// Proceedings of the fourth ACM International Conference on Web Search and Data Mining，New York：ACM，2011：595-604.

[17] Universityo f Minnesota：Grouplens Research Project. Wikilens Dataset[DB][2013-04-01]，MovieLens Datasets. http://www.grouplens.org/datasets/movielens.

[18] Wikilens Data Set. http:// grouplens.org/datasets.wikilens.

[19] Chen T，Zheng Z，Lu Q，et al. Feature-based matrix factorization[R]. Shanghai：Shanghai Jao Tong Universitt Aoex Data and Knowledge Management Lab，2011.

（上接第187页）**索引的方式，利用谓词之间的覆盖关系来加快匹配速度；并且能够及时发现不存在与事件匹配的订阅的情形，从而可以终止匹配过程，减少不必要的时间消耗，提高了匹配效率.**

参　考　文　献

[1] TRAN D A，PHAM C. PUB\SUB：a content-based publish/subscribe framework for cooperative P2P networks[C]// 8th International IFIP TC 6 Network Conference. Heidelberg：Springer-Verlag，2009：770-781.

[2] **陈继明，鞠时光，潘金贵，等. 基于内容的快速匹配算法**[J]. **通信学报**，2011，32(6)：78-85.

[3] **马建刚，黄涛. 面向大规模分布式计算发布订阅系统核心技术**[J]. **软件学报**，2006，17(1)：134-137.

[4] Ashayer G，Leung H K Y，Jacobsen H A. Predicate matching and subscription matching in publish/subscribe systems[C]// Proc of Workshop on Distributed Event-Based Systems (DEBS). Vienna：IEEE，2002：539-546.

[5] Carzaniga A，Wolf A L. Forwarding in a Content-Based Network[C]// Proceedings of ACM SIGCOMM 2003. New York：[s.n.]，2003：163-174.

[6] **薛涛，冯博琴，李波，等. 基于内容的发布订阅系统中快速匹配算法的研究**[J]. **小型微型计算机系统**，2006(3)：529-533.

[7] Yan T W，Garc H. Index structures for selective dissemination of information under the Boolean model[J]. ACM Trans Database System，1994，19(2)：332-334.

[8] Aguilera M K，Strom R E，Sturman D C，et al. Matching events in a content-based subscription system[C]// Proc of the 18th ACM Symp on Principles of Distributed Computing. [s.l.]，1999：53-61.

[9] **张晓丰，张凤鸣，郭建胜. 发布/订阅系统中基于属性划分的并行搜索树**[J]. **计算机工程**，2007，33(3)：45-47.