利用离群点算法预处理协同过滤推荐系统数据

周莹莹,王晓军
(南京邮电大学 信息网络技术研究所,江苏 南京 210003)

摘 要:由于电子商务系统的开放性和推荐系统用户的广泛参与性,推荐系统很容易受到攻击。出于某种目的的用户向系统中注入恶意信息,导致推荐质量受到威胁,因此过滤掉恶意信息成为迫切需要。离群点检测用于从数据集中找到明显偏离其他数据对象或不满足一般对象行为特征的对象。为了提高推荐系统的鲁棒性,保证推荐系统的高质量,文中利用局部离群点检测算法计算出每个用户的局部离群因子(LOF),过滤掉离群因子较高的用户,然后运用协同过滤算法为系统中剩下的用户做推荐。实验结果表明,与传统的协同过滤推荐算法相比,此方法在提高推荐质量上取得了一些好的效果。

关键词:推荐系统;协同过滤;离群点;离群因子
中图分类号:TP302.1 **文献标识码**:A **文章编号**:1673-629X(2015)09-0129-05
doi:10.3969/j.issn.1673-629X.2015.09.028

Pre-filtering Data of Collaborative Filtering Recommendation System by Outliers Algorithm

ZHOU Ying-ying, WANG Xiao-jun
(Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Due to the openness of the e-commerce system and extensive participation of recommended system users, recommendation system is vulnerable to attack. Some users who want to reach a particular purpose inject malicious information into the system, leading to under threat of the recommendation quality, and thus it is necessary to filter out malicious information. Outlier detection is to find the exceptional objects which do not satisfy the common patterns or deviate much from the rest objects of the dataset by some measure. In order to improve the robustness and guarantee the high quality of the system, compute user's Local Outlier Factor (LOF) and remove users who has a higher local outlier factor based on local outliers algorithm, and then use the collaborative filtering algorithm to recommend for the users. Compared with the traditional collaborative filtering algorithm, the experimental result shows some good results have been achieved on improving the quality of recommendation.

Key words: recommendation system; collaborative filtering; outliers; outlier factor

0 引 言

随着商业服务的迅速发展和大数据时代的到来,人们从大量涌现的信息数据中查找相关信息已显得力不从心,无从下手。推荐系统的出现给人们的生活带来了极大的便利,为用户提供个性化服务[1-2]。协同过滤推荐算法[3]是推荐系统中最常用的一种推荐算法。随着它的广泛普及应用,推荐质量已成为一个非常重要的研究问题。由于电子商务系统的开放性和推荐系统用户的广泛参与性,推荐系统很容易受到攻击。一些不良商家为了牟取利益,向推荐系统中注入许多虚假信息来提高自己产品的推荐频率或降低竞争对手产品的推荐频率,改变系统的推荐行为,可能导致用户对推荐系统的信任度降低。这些注入系统中的虚假信息相对于大量正常信息来说,将被看作是异常数据或离群数据,因此将离群点检测技术运用到推荐系统中十分必要。

离群点检测是数据挖掘[4-5]领域的重要研究方向之一,也称为离群点挖掘,其目的是在大量的、复杂的

收稿日期:2014-10-22 **修回日期**:2015-01-26 **网络出版时间**:2015-08-26
基金项目:国家自然科学基金资助项目(61003237)
作者简介:周莹莹(1989-),女,硕士研究生,研究方向为推荐系统应用;王晓军,研究生导师,副研究员,研究方向为分布计算技术与应用、云计算、数据库技术等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1535.012.html

数据集合中消除噪音或发现潜在的、有意义的知识。研究离群点异常行为有助于发现特别有价值的知识。离群点检测主要检测系统中具有攻击信息或异常的数据,换言之就是在系统中查找到与一般正常数据偏差比较大的信息数据,这些数据偏离了正常行为被看作是离群点或异常点。为了保证系统的安全性,必须对这些异常信息进行处理。在推荐系统中,用户概貌是其主要组成部分,攻击者很容易通过一定的手段向系统中注入有偏差的用户概貌信息来干预推荐结果,因此文中利用基于密度的局部离群点检测[6]找出推荐系统中的恶意攻击者,并且阻止这些攻击者对推荐系统的影响,从而提高推荐系统的推荐质量。

1 相关研究

推荐系统的用户参与性为用户提供参考平台,推动推荐系统的进一步发展,给用户带来极大的便利,但同时也暴露了自身的缺点。一些恶意用户利用这一特点向系统中注入恶意信息,使推荐结果朝着利于自己的方向发展,最终导致用户对推荐系统失去信任。近年来,如何提高推荐质量已成为一个炙手可热的研究课题,很多国内外学者在此方面进行了探索,提出了许多防御和检测推荐系统攻击的方法。

在不同的场景中寻找恶意用户或异常数据的算法是不同的。之前有学者利用信任度来避免恶意用户或异常数据的攻击。Patil V A 和 Ragha L 把信任度分成两部分:直接信任和推荐信任。利用相似度与信任度的权重找出该用户的近邻形成最终推荐[7]。秦继伟等也做过类似的研究[8]。为了提高推荐系统的质量,很多学者认为提前对数据做一些处理(比如过滤掉恶意用户或异常数据)很有必要,因此他们对查找这些特殊数据做了一些研究。Mehta 等提出基于主成分分析(PAC)的检测算法,对数据矩阵进行降维分析,过滤出彼此之间高度相关的攻击概貌。该算法需要预先知道恶意用户注入攻击概貌的数量,但在实际应用中该参数难以确定,参数选取的不合理直接影响到算法检测的精度[9]。Itaf 等将推荐集分成多个区间,通过差异函数计算每个区间的差异度,差异度最大的推荐区间即为恶意推荐用户[10]。Chung 等通过三组实验筛选恶意攻击者,当某个用户至少被其中的两组认为是不正常的用户时,此用户就最终确认为攻击者,该实验使用 beta 分布模型[11]。孙启林等结合聚类和相似度思想过滤异常数据。首先初始化每个对象为一个类,再运用相似度函数计算两个对象的相似度,找到相似度最大的两个类,当这两个类的相似度大于给定的阈值时合并两个类,重复执行直到相似度小于给定的阈值时计算各个类的密度,如果某个类的密度小于离群点阈值密度,则该类中的对象都是离群点[12]。

在大规模的数据集中恶意用户的行为偏离正常用户的程度较大,可以将其看成是离群点,因此可以从离群点的角度寻找异常用户。离群点算法有五类:基于统计的离群点检测、基于距离的离群点检测、基于深度的离群点检测、基于偏差的离群点检测、基于密度的离群点检测[13]。不同于其他四种离群点算法,基于局部密度离群点检测算法不是将离群点看作一种二元性质,即不简单用是或不是来断定一个点是不是离群点,而是用一个权值来评估它的离群度。它结合点之间的距离与点个数得到一个区域,然后针对此区域计算局部密度及其离群点因子[14]。局部离群程度依赖于对象相对于其领域的孤立情况。文中将局部密度离群点算法运用到协同过滤推荐系统[15-18]中,通过计算每个用户的离群因子后降序排序,过滤掉离群因子较大的用户,从而提高推荐系统的质量。

2 局部密度离群点算法

概念 1:对象 p 的第 k 距离(k_distance)。

对任意的自然数 k ,当对象 p 与某个对象 o 之间的距离满足以下条件:

(1)至少存在 k 个对象 $o' \in D\backslash\{p\}$ 满足 $d(p,o) \leqslant d(p,o')$;

(2)至多存在 $k-1$ 个对象 $o' \in D\backslash\{p\}$ 满足 $d(p,o) < d(p,o')$ 。

其中, $d(p,o)$ 为对象 p 和对象 o 之间的距离, D 为对象集合对象时, p 的第 k 距离 k_distance$(p) = d(p,o)$ 。

该定义通过计算每一个对象与其他对象之间的距离,然后升序排列查找出第 k 个不同数值的距离来确定该对象的局部空间区域范围。对于密度较大的区域,第 k 距离的数值一般情况下较小;对于密度较小的区域,第 k 距离的数值一般情况下较大。

概念 2:对象 p 的第 k 距离邻域($N_{k\_distance}(p)$)。

已知对象 p 的 k_distance ,则 p 的第 k 距离邻域定义为与 p 之间的距离不超过 k_distance(p) 的数据对象的集合,即

$$N_{k\_distance}(p) = \{o \in D\backslash\{p\} \mid d(p,o) \leqslant k\_distance(p)\}$$

对象 o 即为对象 p 的 k 最近邻,其邻域实际上是以对象 p 为圆心、p 的 k 距离为半径的区域内所有对象的集合(对象 p 除外)。由于数据集中可能同时存在多个第 k 距离的数据对象,因此该集合至少包含了 k 个数据对象。每个数据对象的该集合所涵盖的区域可能存在较大的差异性,偏离程度较高的数据对象该集合涵盖的区域比较大,而偏离程度较低的数据对象该集合

涵盖的区域较小,对于处于同一群组中的对象来说,涵盖的区域面积则相当。

概念 3:对象 p 对于对象 o 的可达距离(reach_distance(p,o))。

对象 p 相对于对象 o 的可达距离定义为:

$$\text{reach\_distance}_k(p,o) = \max\{d(p,o),\ k\_\text{distance}(o)\}$$

如图 1 所示,当 p_1 在对象 o 的邻域内,p_1 与 o 的可达距离取 o 的可达距离 k_distance(o) ,当 p_2 不在对象 o 的邻域内,p_2 与 o 的可达距离取两个对象间的实际距离 $d(p_2,o)$ 。

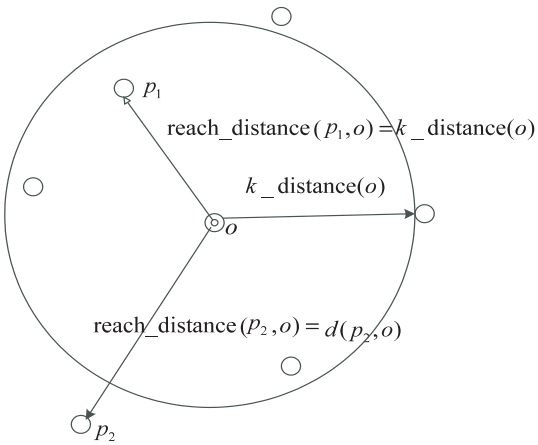

图 1　p_1 和 p_1 可达距离示意图(k = 4)

概念 4:对象 p 的局部可达密度 $\text{lrd}_k(p)$ 。

对象 p 的局部可达密度定义为:

$$\text{lrd}_k(p) = 1/\left(\frac{\sum_{o\in N_{k\_\text{distance}}(p)} \text{reach\_distance}_k(p,o)}{|N_{k\_\text{distance}}(p)|}\right) \quad (1)$$

由局部可达密度定义公式可知,如果对象 p 的周围对象分布稀疏,即对象 p 远离自己的 k - 近邻,则 p 与邻域中各对象的可达距离取值为两个对象之间的实际距离,从而导致 p 与邻域对象的可达距离之和较大,因此其局部可达密度会相应较小。$\text{lrd}_k(p)$ 数值表明了对象所处于的局部空间区域的密度情况。

概念 5:对象 p 离群因子 LOF。

$$\text{LOF}_k(p) = \frac{\sum_{o\in N_{k\_\text{distance}}(p)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(p)}}{N_{k\_\text{distance}}(p)} \quad (2)$$

如果对象 p 的离群程度较大,则它的邻域中大多数是离对象 p 较远且处于一个聚类中的数据对象,那么这些数据对象的局部可达密度 lrd 应该偏大,而对象 p 本身的可达密度 lrd 偏小,所以最后得到的 LOF 值也偏大。反之,如果对象 p 处于某一个群组之中,则其第 k 距离邻域内的数据对象与其属于同一个群组的可能性较大,所以得出邻域中所有对象 o 的平均分布密度 lrd 和对象 p 的 lrd 相似,最后所得的 LOF 值应该接近 1。

利用离群点预过滤改进协同过滤推荐算法的步骤为:

(1)输入用户-项目评分二维矩阵 $\boldsymbol{R}$,参数 k,离群点个数 n。

(2)将每个用户 u_i 对所有项目的评分看作是一个向量,即 $u_i = \{r_{i,j} \mid i \in U, j \in I\}$ 。其中,U 是所有评分用户的集合,I 是所有项目的集合。该算法采用欧几里得距离公式计算两个用户之间的距离,此距离是基于两个用户的共同评分项目来计算的。

$$\text{distance}(u_a,u_b) = \sqrt{\sum_{j\in(I_a\cap I_b)} (r_{a,j} - r_{b,j})^2}$$

其中,I_a 是用户 u_a 的评分项目集合;I_b 是用户 u_b 的评分项目集合;$r_{a,j}$ 与 $r_{b,j}$ 分别是用户 u_a 、u_b 对项目 j 的评分。

(3)根据用户间的距离,计算每个用户的第 k 距离邻域。针对每个用户 u_i ,降序排列用户 u_i 与其他用户之间的距离,选取前 k 个不同距离,第 k 个距离为用户 u_i 的可达距离,同时用户 u_i 的邻域是与 u_i 之间的距离等于或小于第 k -距离的所有用户的集合,集合中的用户个数至少 k 个。

(4)对于用户 u_i 来讲,u_i 与它的邻域组成了一个局部群,然后利用式(1)计算用户 u_i 相对于它的局部群的可达密度,依次类推,计算用户集合 U 中各用户的局部可达密度,此密度代表局部群中用户分布疏密的情况。

(5)利用式(2)计算每个用户的 LOF,即用户在它的局部群中的离群程度。若离群因子的数值接近 1,则说明此用户与其邻域对象所属一个局域中;若离群因子数值较大,则说明此用户距离自己局部群中的其他用户比较远,可视为离群用户。

(6)从大到小排列用户集合 U 中各个用户的离群因子,选择前 n 个离群因子最大的用户,并从用户-项目评分矩阵 $\boldsymbol{R}$ 中删除这些用户所对应的各个项目的评分。

(7)对剩下的用户利用协同过滤算法对用户进行推荐。协同过滤推荐算法的过程主要分为三步实现:建立用户信息模型、寻找用户的近邻和推荐结果评估。

文中算法流程如图 2 所示。

3　实验与结果分析

为了验证上述理论研究是否能在推荐系统邻域取得较好的作用,文中进行实验仿真。实验采用 GroupLens 研究小组通过 MovieLens 网站收集的 MovieLens 电影评分公开数据集。它包括来自 943 个用户对 1 682部电影的 10 万条评分记录,其中每个用户至少对 20 部电影进行评分,评分值表示用户对电影的喜欢

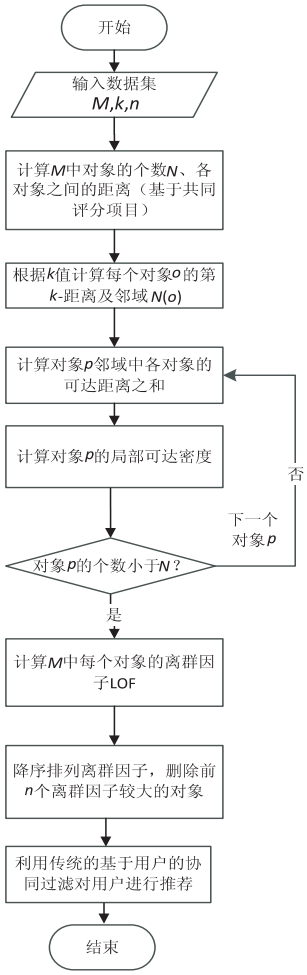

图 2　局部密度离群点算法流程

程度,用 1 ~5 之间的整数表示,1 表示最不喜欢,5 表示最喜欢。实验采用 Matlab 来实现离群点预处理算法。为了较好的进行实验,首先对 MovieLens 数据进行适当的修改,即将实验数据按照一定的比例进行填充。本实验分别在 1% 、2% 填充规模和 1% 的攻击规模下进行实验。通过实验得出以下结论:

(1)通过图 3 中 MAE 曲线比较表明局部离群密度算法取得了一定的效果。图(a)按 1% 的填充规模进行实验,填充后的 MAE 与除去离群点后的 MAE 之间保持一个稳定的变化趋势,且除去离群点后的 MAE 小,说明通过离群点算法预处理后推荐质量有一定的提高。图(b)按 2% 的填充规模进行实验,观察填充前后 MAE 变化,邻居数目较小时 MAE 之间的差异不大,随着邻居数目的增加,差异也逐渐明显。通过比较图中 MAE 的变化得出在进行推荐之前对数据进行一定的预处理,筛除异常数据会提高系统的推荐效果。

(2)针对填充后的数据进行实验检测离群点时,发现检测出的离群点中包含一些原始数据中的用户,于是提出这样一个假设:原始数据可能存在一些异常数据。为了验证假设是否成立,利用文中算法在原始数据集进行实验。图(c)表明在邻居数目不是很大的情况下,除去离群点计算的 MAE 值不是稳定的;图(d)表明随着邻居数目的增大,除去离群点后的 MAE 值趋于稳定,并且比原始的 MAE 值小,同时也证明了原始数据集中确实存在一些异常数据。

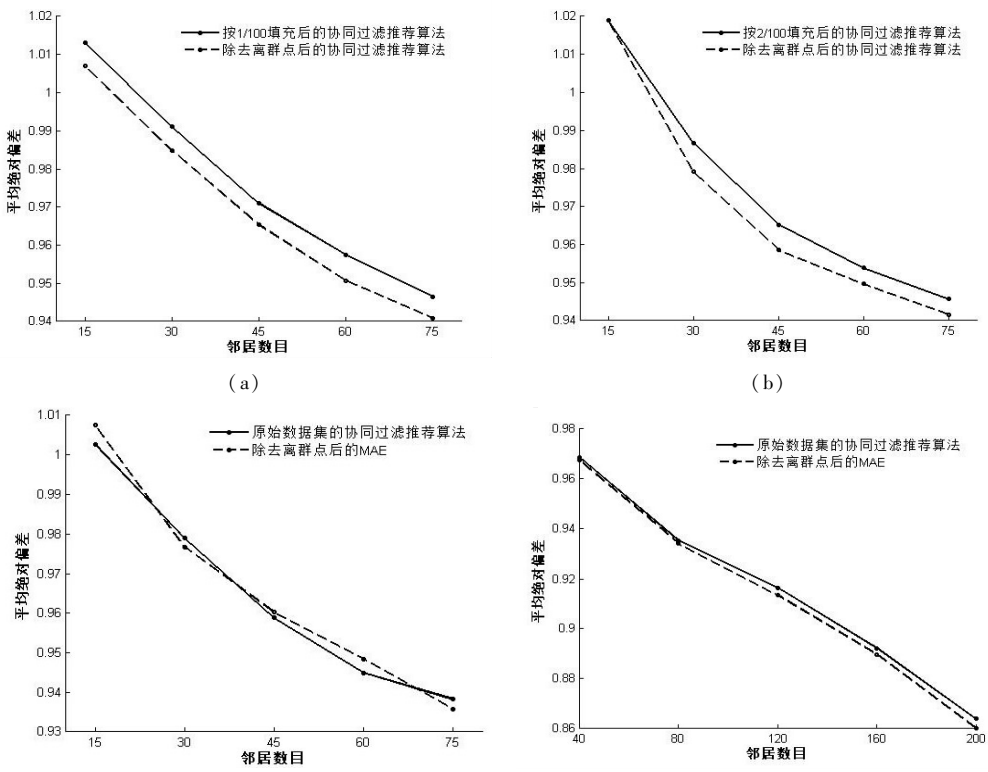

图 3　填充规模为 1% 与 2% 的协同过滤 MAE 比较

4　结束语

文中研究了局部密度离群点算法,并将此算法运用到协同过滤推荐系统中。为了提高推荐系统的质量,先利用离群点算法预过滤数据集,通过离群因子衡量用户距离自己局部群的程度来除去一部分离群用户,然后再利用传统的协同过滤对用户进行推荐。实验结果表明,此算法在预处理协同过滤数据上取得了一定的效果。

参考文献:

[1] 曾　春,邢春晓,周立柱. 个性化服务技术综述[J]. 软件学报,2002,13(10):1952-1961.

[2] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.

[3] Resnick P, Varian H R. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-58.

[4] Han Jiawei, Micheline K. Data mining: concepts and techniques[M]. 2nd ed. San Francisco: Mogran Kaufmann Publishers, 2006.

[5] 薛安荣,鞠时光,何伟华,等. 局部离群点挖掘算法研究[J]. 计算机学报,2007,30(8):1455-1463.

[6] Guido B F, Flavio M. Outlier detection in large data sets[J]. Computers and Chemical Engineering, 2011, 35: 388-390.

[7] Patil V A, Ragha L. Comparing performance of collaborative filtering algorithms[C]//Proc of 2012 international conference on communication, information & computing technology. Mumbai, India: [s. n.], 2012: 1-6.

[8] 秦继伟,郑庆华,郑德立,等. 结合评分和信任的协同推荐算法[J]. 西安交通大学学报,2013,47(4):100-104.

[9] Mehta B, Hofmann T, Fankhauser P. Lies and propaganda: detecting spam users in collaborative filtering[C]//Proceedings of the 12th international conference on intelligent user interfaces. Honolulu, Hawaii: ACM, 2007: 14-21.

[10] Itaf N, Ghafoor A, Zia U. An attack resistant method for detecting dishonest recommendations in pervasive computing environment[C]//Proc of 18th IEEE international conference on network. Singapore: IEEE, 2012: 173-178.

[11] Chung Chen-Yao, Hsu Ping-Yu, Huang Shih-Hsiang. A novel approach to filter out malicious rating profiles from recommender systems[J]. Decision Support Systems, 2013, 55(1): 314-325.

[12] 孙启林,方宏彬,张　健,等. 一种基于相似度量的离群点检测方法[J]. 重庆工商大学学报:自然科学版,2013,29(10):96-100.

[13] 揭财明. 基于密度的局部离群点检测算法分析与研究[D]. 重庆:重庆大学,2012.

[14] Breuning M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[C]//Proc of ACM SIGMOD conference. New York, USA: ACM Press, 2000: 427-438.

[15] Schafer J B, Frankowski D, Herlocker J, et al. Collaborative filtering recommender systems[M]//The adaptive web. Berlin: Springer, 2007: 291-324.

[16] 张　瑶,陈维斌,傅顺开. 协同过滤推荐研究综述[J]. 微型机与应用,2013,32(6):4-6.

[17] 肖来元,殷　明. 基于身份和上下文的个性化服务研究[J]. 计算机工程与科学,2012, 34(11) 28-33.

[18] 夏建勋,吴　非,谢长生. 应用数据填充缓解稀疏问题实现个性化推荐[J]. 计算机工程与科学,2013, 35(5) 15-19.

利用离群点算法预处理协同过滤推荐系统数据

作者： 周莹莹， 王晓军， ZHOU Ying-ying， WANG Xiao-jun
作者单位： 南京邮电大学 信息网络技术研究所,江苏 南京,210003
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(9)