

基于 Softmax 回归模型的协同过滤算法研究与应用

孟 佩,曹 菡,师 军
(陕西师范大学,陕西 西安 710119)

摘 要:针对传统的协同过滤推荐算法所存在的推荐精度不高的问题,提出了基于 Softmax 回归模型的协同过滤算法。根据用户的属性特征将用户分为不同的簇,再从目标用户所在的簇中实现协同过滤推荐,有效缩减了最近邻居的查找范围,提高了推荐效率。主要将此改进算法应用于饮食推荐中,根据用户的饮食记录对用户按口味偏好进行精准分类,将偏好相同的用户划分到同一个簇中,再从目标用户所在的用户簇中查找最近邻居,完成推荐。从两方面对此方法进行了实证分析:基于 Softmax 的用户口味偏好分类的准确率分析和基于 Softmax 的协同过滤推荐精准度分析,验证了该方法的有效性和可行性。

关键词:Softmax 回归;口味偏好;协同过滤;营养饮食

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2016)12-0153-03

doi:10.3969/j.issn.1673-629X.2016.12.033

Research and Application of Collaborative Filtering Algorithm Based on Softmax Regression Model

MENG Pei, CAO Han, SHI Jun
(Shaanxi Normal University, Xi'an 710119, China)

Abstract: In view of the low accuracy for traditional collaborative filtering recommendation algorithm, the collaborative filtering algorithm based on Softmax regression model is proposed. According to the user's attributes, the users can be divided into different clusters, and the collaborative filtering recommendation is realized in the cluster from its target users, reduction of the nearest neighbors search scope, improvement of the performance of the recommendation system. The improved algorithm is applied to dietary recommendations, depending on the user's diet by recording the user taste preferences for accurate classification, the same user preferences will be divided into the same cluster, and then the nearest neighbor is searched from the user cluster where there is the target user to complete the recommendation. An empirical analysis about this method from two aspects is made, including the accuracy analysis of the user's taste preference classification based on Softmax and precision analysis of collaborative filtering recommendation based on Softmax, and the effectiveness and feasibility is verified.

Key words: Softmax regression; taste preference; collaborative filtering; nutrition diet

0 引 言

随着社会经济的发展和人民生活条件的改善,人们的饮食消费观念已经由温饱型转向营养型,因此出现了许多营养膳食系统及饮食推荐方面的研究。基于本体的个性化营养推荐系统^[1-2],主要采用协同过滤技术和上下文相关技术进行饮食推荐,忽略了数据稀疏的问题;基于 Web 数据挖掘的健康餐饮分析推荐系统的设计^[3],基于模糊决策的体质学饮食推荐建模及其系统实现^[4],都过分注重营养却忽略了用户的饮食

喜好,推荐的食物虽然健康,用户却不喜欢。

大多数用户在饮食上都有自己偏好的口味,而且短期之内不会改变,所以文中采用 Softmax 多分类回归算法,根据用户近期一周的饮食记录,预测用户的口味偏好,在此基础上,对目标用户所在的类采用协同过滤算法^[5-9],完成 top-N 推荐。此方法能够缩小邻居查询范围,减少计算量,缩短计算时间并提高用户口味满意度,弥补已有的营养膳食系统的不足,很好地平衡用户的喜好和饮食营养之间的关系。

收稿日期:2016-01-21

修回日期:2016-05-06

网络出版时间:2016-11-21

基金项目:国家自然科学基金资助项目(41271387)

作者简介:孟 佩(1989-),女,硕士研究生,研究方向为云计算、数据挖掘;曹 菡,博士,教授,通讯作者,研究方向为大数据处理、空间数据挖掘以及智慧旅游。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20161121.1641.024.html>

1 Softmax 回归模型

Softmax 回归模型^[10-11]是解决多类回归问题的算法,是当前深度学习研究中广泛使用在深度网络有监督学习部分的分类器。设训练 Softmax 回归模型的样本来自 k 个类,共有 m 个,则由这些样本组成的训练集为 $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$ 。其中 $x^{(i)} \in R^n$, 标签 $y^{(i)} \in \{1, 2, \dots, k\}$ 。给定测试输入 x , 用假设函数针对每一个类别 j 估算出概率值 $p(y = j | x)$, 即估计 x 的每一种分类结果出现的概率,其中出现概率最大的类别即为输出值。因此,假设函数要输出一个 k 维向量(向量元素的和为 1)用来表示这 k 个估计的概率值。假设函数 $h_\theta(x^i)$ 形式如下:

$$h_\theta(x^i) = \begin{bmatrix} p(y^i = 1 | x^i; \theta) \\ p(y^i = 2 | x^i; \theta) \\ \vdots \\ p(y^i = k | x^i; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{\theta_j^T x^i}} \begin{bmatrix} e^{\theta_1^T x^i} \\ e^{\theta_2^T x^i} \\ \vdots \\ e^{\theta_k^T x^i} \end{bmatrix} \quad (1)$$

其中, $\theta_1, \theta_2, \dots, \theta_k \in R^n$ 是模型的参数向量,可用矩阵表示为 $\theta = [\theta_1, \theta_2, \dots, \theta_k]^T$; $p(y^i = j | x^i; \theta)$ 表示样本 x^i 属于第 j 类的概率。

该模型的代价函数^[12-13]为:

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m \sum_{j=1}^k 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{i=1}^k e^{\theta_j^T x^{(i)}}} \right] + \frac{\lambda}{2} \sum_{i=1}^k \sum_{j=0}^n \theta_{ij}^2 \quad (2)$$

其中, $1\{\cdot\}$ 为指示函数,如果花括号中的表达式为真,则指示函数取值为 1,否则为 0。

原始的代价函数没有权重衰减项,即加号后面的部分,加入权重衰减可以解决 Softmax 回归的参数冗余所带来的数值问题,能够保证得到唯一的解。解决 $J(\theta)$ 的最小化问题有两种常用方法:梯度下降法和最小二乘法。文中采用第一种方法进行优化。具体思想是先对 θ 取一个随机初始值(其目的是使对称失效),然后不断迭代改变 θ 的值使 $J(\theta)$ 减小,直到最终收敛取得一个 θ 值使得 $J(\theta)$ 最小,最终确定出假设函数 $h_\theta(x)$, 以此对新输入的数据进行预测分类。

2 基于 Softmax 的口味偏好分类

(1) 数据归一化处理。

① Min-max 标准化,也称为离差标准化,是对原始数据的显性变换,使结果值映射到 $[0, 1]$ 之间,转换公式为:

$$X = (x - \min) / (\max - \min) \quad (3)$$

其中, \max 为样本数据的最大值; \min 为样本数据的最小值。

② Z-score 标准化,该方法对原始数据的均值

(mean) 和标准差(standard deviation)进行数据的标准化。经过处理的数据符合标准正态分布,即均值为 0, 标准差为 1,转换公式为:

$$X = \frac{x - \mu}{\delta} \quad (4)$$

以用户在最近一周内所食用的各种菜品的次数作为实验数据进行实验,由于每个人的饮食喜好不同,喜欢吃菜的程度不同,则实验数据内容可能会存在很大差异,进而影响数据分析结果,因此该实验采用线性函数法对数据进行归一化处理,使其归一化到 $[0, 1]$ 范围内,增强可比性。

(2) Softmax 回归模型的训练。

该实验训练数据 200 条,测试数据 150 条,采用梯度下降法对代价函数进行迭代优化,当结果达到收敛或经过指定步长数之后模型训练完毕。

3 基于 Softmax 的协同过滤推荐

区别于传统的基于用户的协同过滤算法,基于 Softmax 的协同过滤推荐算法首先对用户进行分类,将具有相同特征的用户划分到同一类中,然后在目标用户所在的用户簇中进行邻居查找及 top-N 推荐,避免了在整个用户空间上进行算法的实现,减少了一定的计算量和时间,并提高了推荐结果的精准度。

具体方法:使用 Softmax 对用户进行分类,寻求目标用户所在的用户簇,对每一类具有相同口味偏好的用户建立用户-饮食频次矩阵^[14]和食物-营养素矩阵,求取食物间的相似度,完成预测并进行最终推荐。其中,食物间的相似度由用户饮食相似度和食物营养素相似度共同决定,均采用 Pearson 相关系数进行计算,相关公式如下:

$$\text{sim}_{\text{food}}(i, j) = a * \text{sim}_{\text{diet}}(i, j) + (1 - a) * \text{sim}_{\text{nutrients}}(i, j) \quad (5)$$

$$\text{sim}_{\text{diet}}(i, j) = \frac{\sum_{c \in I_{i,j}} (R_{i,c} - \bar{R}_i) (R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_{i,j}} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_{i,j}} (R_{j,c} - \bar{R}_j)^2}} \quad (6)$$

$$\text{sim}_{\text{nutrients}}(i, j) = \frac{\sum_{c \in I_{i,j}} (R_{i,c} - \bar{R}_i) (R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_{i,j}} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_{i,j}} (R_{j,c} - \bar{R}_j)^2}} \quad (7)$$

式(6)中, $I_{i,j}$ 表示用户 i 和用户 j 共同吃过的食物的集合; $R_{i,c}$ 表示第 i 个用户吃过第 c 个食物的频次; \bar{R}_i 表示第 i 个用户一周之内吃过的食物的平均频次。

式(7)中, $I_{i,j}$ 表示 i 和 j 两种食物中共同含有的营养素; $R_{i,c}$ 表示 i 食物中营养素 c 的含量; \bar{R}_i 表示 i 食物

中营养素的平均含量。

4 实验及结果分析

(1) 基于 Softmax 的口味偏好分类准确率验证。

在参考校园一卡通数据的基础上,以 350 条生成数据为例验证改进算法的可用性。其中 200 条为训练数据,共含有清淡味、甜味、麻辣味和酸辣味 4 类口味偏好的用户,分别用 1、2、3、4 表示。通过训练数据训练出模型参数 θ ,得到预测函数 $h_{\theta}(x)$,再通过 150 条测试数据进行验证,检测假设函数的正确性。该实验最终正确率为 99.333%。用户饮食记录数据结构如表 1 所示。

表 1 饮食记录数据结构表

字段名	数据类型
学号	字符串型
消费日期	日期型
消费时间	时间型
餐次	字符串型
就餐地点	字符串型
食物	字符串型
价格	浮点型
数量	整数型
单位重量	浮点型
口味偏好	字符串型

在训练数据和测试数据中,各类口味偏好的用户数据分布情况如表 2 所示。

表 2 实验样本分布表

数据	口味			
	清淡味	甜味	麻辣味	酸辣味
训练数据	44	33	66	57
测试数据	52	22	27	49

Matlab 下的实验结果如图 1 所示。

```
ans =  
  
4      15  
  
ans =  
  
150    15  
  
Accuracy: 99.333%
```

图 1 Softmax 实验结果图

(2) 基于 Softmax 的协同过滤精准度验证。

该实验将分两组进行对比:一组实验是直接进行协同过滤推荐,另一组是基于 Softmax 的协同过滤,在上述分类的测试数据

对误差(MAE)对算法进行度量,主要是通过计算目标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,因而 MAE 的值越小,推荐的质量越高。

实验结果如图 2~4 所示。

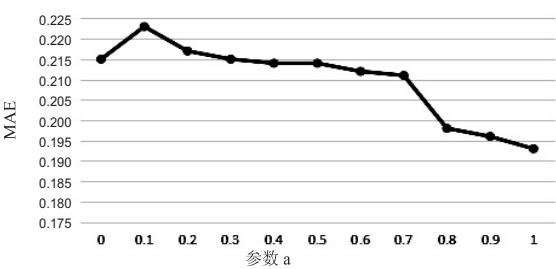


图 2 传统的协同过滤

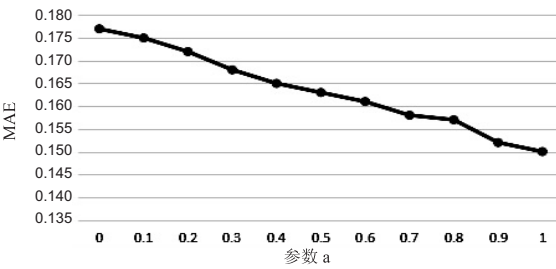


图 3 改进的协同过滤

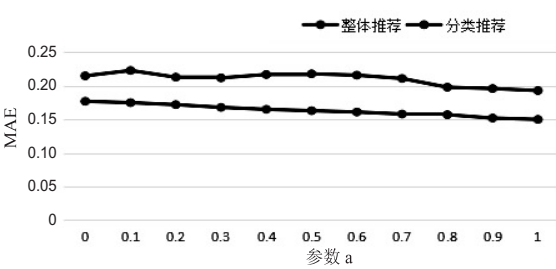


图 4 实验对比

从实验结果看,基于 Softmax 的协同过滤的推荐质量高于传统的协同过滤,所以该方法具有可行性。

5 结束语

文中主要提出了一种基于 Softmax 的协同过滤改进算法,通过用户的某种特征属性对用户进行分组,根据目标用户所在的用户群,采用协同过滤算法实现推荐。模拟饮食数据,采用 Softmax 多分类回归算法,对用户按口味进行分类,在此基础上采用协同过滤算法进行饮食推荐研究。实验结果证明了该方法的可行性,能够为用户提供满意度更高的服务。

参考文献:

[1] 刘 浩. 基于本体的个性化营养推荐系统[D]. 天津:天津大学,2007.

[2] 唐建华,张秀南. 营养食疗个性化推荐系统设计与开发[J]. 扬州大学烹饪学报,2014,31(2):23-26.

[3] 康钟荣. 基于项目特征分类与填充的协同过滤算法研究(下转第 159 页)

行为 $K=47$, 得到相似度矩阵。随机抽取一位用户, 列出与其相似度最高的十位用户, 如表 1 所示。可根据

用户之间的相似关系提供食品推荐服务、群体饮食行为研究等。

表 2 与用户 1 相似度最高的十位用户

s_{ij}	user25	user7	user28	user29	user16	user5	user23	user4	user24	user14
user1	1.42	1.29	1.11	1.05	1	0.85	0.84	0.78	0.78	0.77

4 结束语

针对美食互动社区中的 UCG 数据, 结合 LDA 模型的文档-主题-词分层模型的特点, 用 UCG 数据来代表用户, 进而提出了用户-饮食行为-词的用户饮食行为模型, 为数据挖掘在饮食行为方面的研究提供了一个新思路。今后的研究工作可结合更多的社交网络特征, 通过数据挖掘, 为解决饮食行为干预、疾病预防和控制、食品推荐等问题提供更大的帮助。

参考文献:

[1] 毛 茅,王 洋,赵好婕,等. 基于社交网络的美食互动网站设计与评估[C]//第七届和谐人机环境联合学术会议(HHME2011)论文集. 出版地不详;出版者不详,2011.

[2] 杨正雄,赵文华,陈君石. 饮食行为干预的研究进展[J]. 中国学校卫生,2008,29(6):573-576.

[3] 贡浩凌,戴莉敏,刘 媛,等. 医院-社区-家庭护理干预模式对 2 型糖尿病患者饮食控制的效果[J]. 中华护理杂志,2014,49(4):399-403.

[4] 张雅楠,丁 虹,杜玉萍. 回顾性膳食调查辅助工具的应用现状与评价方法[J]. 职业与健康,2015(9):1294-1296.

[5] 安宜沛. 慢性心衰患者膳食现况调查及中医药膳调养研究[D]. 广州:广州中医药大学,2015.

[6] Blei D M,Ng A Y,Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.

[7] 孙昌年,郑 诚,夏青松. 基于 LDA 的中文文本相似度计算[J]. 计算机技术与发展,2013,23(1):217-220.

[8] 曾利明. 中国民众存在五大饮食“误区”[N]. 光明日报,2004-11-26.

[9] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning,2001,42(1-2):177-196.

[10] 张 斌,张 引,高克宁,等. 融合关系与内容分析的社会标签推荐[J]. 软件学报,2012,23(3):476-488.

[11] Griffiths T,Steyvers M. Probabilistic topic models[M]//Latent semantic analysis. Hillsdale,NJ:Laurence Erlbaum,2006.

[12] Asuncion A,Welling M,Smyth P,et al. On smoothing and inference for topic models[C]//Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. [s. l.]:AUAI Press,2009:27-34.

[13] 石 晶,胡 明,石 鑫,等. 基于 LDA 模型的文本分割[J]. 计算机学报,2008,31(10):1865-1873.

[14] 刘振鹿,王大玲,冯 时,等. 一种基于 LDA 的潜在语义区划分及 Web 文档聚类算法[J]. 中文信息学报,2011,25(1):60-65.

[15] 李文峰. 基于主题模型的用户建模研究[D]. 北京:北京邮电大学,2013.

(上接第 155 页)

[J]. 河南科技,2013(6):3-5.

[4] 侯翠琴,焦李成,张文革. 一种压缩稀疏用户评分矩阵的协同过滤算法[J]. 西安电子科技大学学报,2009,36(4):614-618.

[5] Ar Y,Bostanci E. A genetic algorithm solution to the collaborative filtering problem[J]. Expert Systems with Applications,2016,61:122-128.

[6] 李晓城,张增杰,夏勇明,等. 基于 web 数据挖掘的健康餐饮分析推荐系统的设计[J]. 微型电脑应用,2011,27(1):44-46.

[7] 付德坤. 基于模糊决策的中医饮食推荐建模及嵌入式系统实现[D]. 成都:电子科技大学,2013.

[8] 黄 洋. 基于聚类和项目类别偏好的协同过滤推荐算法研究[D]. 杭州:浙江理工大学,2014.

[9] Abdelwahab A, Sekiya H, Matsuba I. Collaborative filtering based on aniterative prediction method to alleviate the sparsity

problem[C]//Proceedings of the 11th international conference on information integration and web-based applications & services. [s. l.]:ACM,2009.

[10] Tino P. Bifurcation structure of equilibria of iterated Softmax[J]. Chaos,Solitons & Fractals,2009,41(4):1804-1816.

[11] 付 鹏,姚建刚,龚 磊. 利用红外特征和 Softmax 回归识别绝缘子污秽等级[J]. 计算机工程与应用,2015,51(13):181-185.

[12] 汪海波,陈雁翔,李艳秋. 基于主成分分析和 Softmax 回归模型的人脸识别方法[J]. 合肥工业大学学报:自然科学版,2015,38(6):759-763.

[13] 王 晟. 基于 Softmax 回归的电力仪表分类[J]. 研究与开发,2014(6):25-28.

[14] Majid A,Chen Ling,Chen Gencai. A context-aware personalized travel recommendation system based on geotagged social media data mining[J]. International Journal of Geographical Information Science,2013,27(4):662-663.