基于 KNN-SVM 的混合协同过滤推荐算法*

吕成戍[1a]，王维国[1b]，丁永健[2]

(1. 东北财经大学 a. 管理科学与工程学院；b. 数学与数量经济学院，辽宁 大连 116025；2. 大连理工大学 管理与经济学部，辽宁 大连 116024)

摘　要：数据稀疏性问题对协同过滤推荐系统的推荐精度有很大影响，为此，融合缺失数据平衡方法，提出了一个基于 KNN-SVM 的混合协同过滤推荐算法。利用 K-最近邻法对训练集中的缺失数据进行填补，然后通过支持向量机交叉验证进行分类，综合两者优点，从而克服数据质量对推荐算法的影响。在标杆数据集上进行了仿真实验，数值结果证明了方法的有效性。

关键词：数据稀疏性；支持向量机；K-最近邻；协同过滤

中图分类号：TP311　　**文献标志码**：A　　**文章编号**：1001-3695(2012)05-1707-03

doi:10.3969/j.issn.1001-3695.2012.05.027

Hybrid collaborative filtering algorithm based on KNN-SVM

LV Cheng-shu[1a], WANG Wei-guo[1b], DING Yong-jian[2]

(1. *a. School of Management Science & Engineering, b. School of Mathematics & Quantitative Economics, Dongbei University of Finance & Economics, Dalian Liaoning* 116025, *China*; 2. *Faculty of Management & Economics, Dalian University of Technology, Dalian Liaoning* 116024, *China*)

Abstract: The problem of data sparseness has great influence on collaborative filtering recommendation system's accuracy, balance for this missing data fusion method, this paper proposed a hybrid collaborative filtering algorithms based on KNN-SVM. K-nearest neighbor method used the training set to fill the missing data, and then cross-validated by SVM classification. Comprehend advantages both KNN and SVM in order to overcome impact of data quality on the recommended algorithm. The proposed approach was applied to benchmark problems, and the simulation results show it is valid.

Key words: data sparsity; support vector machine(SVM); K-nearest neighbor; collaborative filtering

基于模型的协同过滤算法是目前个性化推荐系统的研究热点，其核心思想是利用数据挖掘与人工智能技术来改进传统的协同过滤算法[1~3]。分类算法是基于模型的协同过滤算法常采用的一种建模方法[4]。其中，Vapnik[5,6]根据统计学习理论提出的支持向量机方法具有诸多的优良特性，近年来引起了广泛的关注，已在文本分类、图像分类、人脸识别等领域取得了良好的应用效果。推荐系统和文本分类具有很多共同特征[7]，支持向量机在文本分类领域取得的成功促使其被运用到推荐系统中。Zhang 等人[8]将标准的 SVM 分类器直接应用到推荐系统中，但由于推荐系统中数据的极端稀疏性，SVM 分类器的分类精度受到很大影响，不能取得令人满意的的分类效果。Grcar 等人[9]将 KNN 与 SVM 在协同过滤框架下的性能进行了对比，实验结果表明，SVM 处理高维度、高稀疏数据集的性能优于传统的协同过滤算法，但其推荐精度与数据质量高度相关。

针对存在的问题，本文结合 KNN 和 SVM 算法的功能特点，提出了一个基于 KNN-SVM 的协同过滤推荐算法。算法首先采用 K-最近邻法找出具有缺失值样本点的 K 个最近邻，用 K 个最近邻的相关属性值来对用户—项目矩阵中的缺失数据进行预测和填补，使用户之间拥有更多的共同评分项目，从而降低用户—项目矩阵的稀疏性；然后在得到的稠密矩阵上建立模型，以 SVM 算法进行推荐。

1　理论基础

1.1　K-最近邻法

K-最近邻法[10]是最近邻的一个推广。其工作原理是利用评分相似度来构造用户或者项目的 K 个最近邻，再使用 K-最近邻集合进行推荐。K-最近邻法具备较高的灵活性，易与其他模型整合，推荐结果直观，容易解释[11]。因此，在实际应用中可以将 K-最近邻法与其他推荐技术相结合以取得更好的实际效果。

1.2　支持向量机

支持向量机[5,6]基于结构风险最小化原理，通过核映射解决高维空间的学习问题，并具有更好的推广能力，能够克服局部极小、维数灾难和过学习等问题。支持向量机算法的原始形式为

$$\min \frac{1}{2}\| \boldsymbol{w} \|^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t.}\quad y_i(\boldsymbol{w}\times x_i + b) \geqslant 1-\xi_i \quad i=1,2,\cdots,l \qquad (1)$$

其中：$\boldsymbol{w}\times x_i + b = 0$ 是所要求解的超平面，$\boldsymbol{w}$ 是超平面的法向

收稿日期：2011-10-07；**修回日期**：2011-11-14　　**基金项目**：辽宁省社会科学规划基金资助项目(L10BJL035)；中央高校专项科研基金资助项目(DUT10RW302)

作者简介：吕成戍(1979-)，女，硕士，主要研究方向为机器学习、电子商务(lvcs@163.com)；王维国(1963-)，男，教授，博导，主要研究方向为数据挖掘；丁永健(1977-)，男，副教授，博士研究生，主要研究方向为个性化推荐系统.

量,b 是超平面的偏移量;ξ_i 是在近似线性可分的情况下引入的松弛变量; C 是惩罚参数,用于对错分样本进行惩罚。原问题的对偶问题为

$$\max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x})$$

$$\text{s.t.} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \quad C \geqslant \alpha_i \geqslant 0; i = 1, 2, \cdots, l \tag{2}$$

其中:$k(\boldsymbol{x}_i, \boldsymbol{x})$ 是核函数;α_i 是与每个样本对应的 Lagrange 乘子。通过求解式(2),可得最优解 $\alpha^* = (\alpha_1^*, \cdots, \alpha_1^*)^{\mathrm{T}}$,计算 $\boldsymbol{w}^* = \sum_{i=1}^{l} \alpha_i^* y_i \boldsymbol{x}_i$,选择 α^* 的一个正分量 α_j^k,据此计算 $b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$,由此求得分类函数:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \right) \tag{3}$$

2 基于 KNN-SVM 的混合协同过滤推荐算法

2.1 问题描述

假设推荐系统中有 m 个用户的集合 $U = \{U_1, U_2, \cdots, U_m\}$ 和 n 个产品的集合 $I = \{I_1, I_2, \cdots, I_n\}$。用户评分数据集可用一个 $m \times n$ 阶矩阵 $\boldsymbol{R}$ 表示。某一用户 U_i 对产品 $I_j (U_i \in U, I_j \in I)$ 的评分为 $R_{i,j}$,这一评分体现了用户 U_i 对产品 I_j 的兴趣和偏好。

随着电子商务系统规模的进一步扩大,商品数目急剧增加,每个用户购买或评价的只是其中很小的一部分,因此矩阵 $\boldsymbol{R}$ 中存在大量未评数据,这导致了用户—项目矩阵的极端稀疏性。如何选择合适的方法对稀疏的数据集进行分析,是目前协同过滤算法面临的一个瓶颈问题。

本文针对这种情况提出了 KNN-SVM 混合协同过滤推荐算法,其思路是首先使用 KNN 方法对用户—项目矩阵中的空缺评分数据进行补齐处理,降低稀疏性,然后将协同过滤问题转换为一个分类问题,使用支持向量机交叉验证对数据进行分类,最后根据分类结果产生推荐列表。

2.2 算法描述

综合两种算法的优势,本文提出基于 KNN-SVM 的混合协同过滤推荐算法,算法包括三个过程:缺失数据填补、数据分类、产生推荐列表。具体算法描述如下:

1)缺失数据填补

a)对于给定的数据集 $T = \{\boldsymbol{x}_i \in \mathrm{R}^n, i = 1, 2, \cdots, l\}$,向量 $\boldsymbol{x}$ 是至少含有一个缺失评分数据的样本点。

(a)将 $\boldsymbol{x}$ 分成两部分:x_c 和 x_{mi},即 $\boldsymbol{x} = (x_c, x_{mi})^{\mathrm{T}}$, x_c 表示 $\boldsymbol{x}$ 中具有观察值的特征属性集,x_{mi} 表示 $\boldsymbol{x}$ 中具有缺失值的特征属性集,其中 $c + mi = n$。

(b)计算 x_c 与 T 中所有样本点的距离,找出 $\boldsymbol{x}$ 的 K 个近邻,对离散属性,使用投票的方法估计 $\boldsymbol{x}$ 的缺失值;对连续属性,用 K 个近邻相应属性值的均值估计 $\boldsymbol{x}$ 的缺失值。K 值根据下面步骤的计算结果进行动态更新,以保证获得最佳填补结果。

b)重复步骤 a),直到每个向量 $\boldsymbol{x}$ 的缺失值都得到填补。

2)数据分类

a)使用支持向量机交叉验证对数据进行分类。本文从基于项目评分和基于用户评分两个角度进行分类。基于项目评分的分类方法将每个项目看做一个单独的分类问题,利用用户数据作为训练集构造分类器。假定目标项目为 I_1,已有 p 个用户对 I_1 进行了评价,那么用户 $U_q (1 \leqslant q \leqslant p)$ 的特征向量为 $\boldsymbol{U}_q = (R_{q,2}, R_{q,3}, \cdots, R_{q,n})^{\mathrm{T}}$,这个特征向量的类标签为 $R_{q,1}$。分类器的任务就是预测所有其他特征向量 $\boldsymbol{U}_f (p+1 \leqslant f \leqslant m)$ 的类标签。基于用户评分的分类方法将每个用户看做一个单独的分类问题,利用产品数据作为训练数据构造分类器。假定目标用户为 U_1,U_1 已经评价了前 l 个项目,那么项目 $I_t (1 \leqslant t \leqslant l)$ 的特征向量为 $\boldsymbol{I}_t = (R_{2,n}, R_{3,n}, \cdots, R_{m,n})^{\mathrm{T}}$,这个特征向量的类标签为 $R_{1,n}$。分类器的任务就是预测所有其他特征向量 $\boldsymbol{I}_s (l+1 \leqslant s \leqslant n)$ 的类标签。为了简化问题,将所有产品分为两类,即喜欢和不喜欢,类标签表示为 +1 和 −1。对于多分类问题,可以通过组合多个 SVM 二值分类器来实现。

b)判断分类结果是否满足预定要求。这里设定为是否超过原始数据的 SVM 校验正确率。如果满足,则输出分类结果;如果不满足,则递增 K 值,进行缺失数据填补操作,反馈循环。

3)产生推荐列表

输出推荐结果。对于目标用户 $\boldsymbol{U}_f (p+1 \leqslant f \leqslant m)$,利用数据分类阶段获得的最优分类器对 $R_{f,s}$ 的取值进行预测。如果为 +1,则表示分类器判定项目 $\boldsymbol{I}_s (l+1 \leqslant s \leqslant n)$ 为用户喜欢的产品。将分类器预测的所有目标用户喜欢的产品形成推荐列表提供给用户。

3 仿真实验

3.1 实验准备

为了验证本文提出的 KNN-SVM 算法在不同数据集上的效果,分别在两个数据集上进行了实验。

第一个数据集使用 Book-Crossing 数据集[12]。该数据集由 Cai-Nicolas Ziegler 通过 Book-Crossing 网站(http://www.bookcrossing.com/)收集,包含了 278 858 个用户对 271 379 本书的 1 149 780 条评分信息,所有的评分值分布在[0,10]区间内,越高的评分值代表越强的用户兴趣。为了便于实验,对评分值进行重新标定,将评分值为 9、10 的标定为 +1,将评分值为 0 ~ 8 的标定为 −1。本文的实验在 Book-Crossing 数据集中前 1 000 个项目的被评分数据上进行。该子数据集总共包含约 3.5 万名用户在这 1 000 个项目上超过 14 万条的评分数据。

第二个数据集取自 MovieLens 数据集[13]。该数据集由明尼苏达大学 GroupLens 研究小组通过 MovieLens 网站(http://movielens.umn.edu)收集,包含了 943 位用户对 1 682 部电影的 100 000 条 1 ~ 5 分的评分数据,每位用户至少对 20 部电影进行了评分。与第一个数据集一样,对评分值进行重新标定,将评分值为 4、5 的标定为 +1,将评分值为 1 ~ 3 的标定为 −1。本文从 MovieLens 数据集上随机抽取 100、200、300 位用户的评分数据组成三个数据集,分别记为 TDS100、TDS200、TDS300。

实验在 WEKA[14] 平台上进行,支持向量机算法使用广泛应用的 LIBSVM[15]。实验按照 80% ~ 20% 的比例拆分数据集,构造训练—测试数据。实验中支持向量机采用高斯核函数,这是应用最广泛的核函数之一。实验相关的模型参数通过交叉验证方式获得,对比算法也采用最优的参数以确保数据比较的合理性。

3.2 实验结果与分析

为评估填充预处理对支持向量机算法的影响,采用平均正确率来验证分类效果。每组数据集被随机拆分十次构造训练集和测试集,在不同的数据集上分别执行标准支持向量机算法

和本文提出的算法,将多次实验结果进行平均并比较,结果以百分数的形式在表 1 中给出。

表 1　分类正确率比较　/%

实验数据集	标准支持向量机		本文方法	
	基于项目评分的分类方法	基于用户评分的分类方法	基于项目评分的分类方法	基于用户评分的分类方法
Book-Crossing	70.218	70.692	80.256	82.372
TDS100	60.358	60.517	66.875	66.751
TDS200	62.349	62.425	73.135	73.532
TDS300	61.867	61.928	73.865	73.782

通过分析实验结果可以很明显地看出,本文方法在多组数据中均获得较好的分类精度,平均正确率提升在 6% ~11% 之间。在 Book-Crossing 数据集上的结果明显地体现了本文方法的优势,经过 KNN 的稀疏数据预填充处理,推荐精度分别提升了 10% 和 11%;在 MovieLens 数据集上,分类精度也得到了提升,KNN-SVM 在 TDS100 上平均正确率最小,而在 TDS300 上的平均正确率最大,说明随着样本容量的增加,算法预测的准确性随之小幅提高,使得对评分的预测更加准确,推荐质量也随之提高。

4　结束语

本文旨在分析如何提高 SVM 在高稀疏数据集上进行分类的精度,在实验部分重点与 SVM 的结果进行了比较。文献[9]在协同过滤框架下将 KNN 与 SVM 进行了对比,实验结果表明,SVM 的分类性能优于传统的协同过滤算法,因而本文并未进行与 KNN 算法的比较。本文方法增加了数据预处理部分,降低了算法的时间效率,这对于推荐算法的应用来说是一个不利因素,因此本文下一步的工作将集中在填补缺失数据的同时删减恶意数据,从而在保证精度的前提下提高算法效率和健壮性。

参考文献:

[1] UNGAR L H, FOSTER D P. Clustering methods for collaborative filtering[C]//Proc of Workshop on Recommendation Systems. California: AAAI Press, 1998: 11-15.

[2] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proc of 14th Conference on Uncertainty in Artificial Intelligence. 1998: 43-52.

[3] ROBLES V, LARRANAGA P, MENASALVAS E, *et al*. Improvement of naive bayes collaborative filtering using interval estimation[C]// Proc of IEEE/WIC International Conference on Web Intelligence. Washington DC: IEEE Computer Society, 2003: 168-174.

[4] BASU C, HIRSH H, COHEN W. Recommendation as classification: using social and content-based information in recommendation[C]// Proc of the 15th National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 1998: 714-720.

[5] VAPNIK V N. Statistical learning theory[M]. New York: Wiley-Interscience, 1998: 35-53.

[6] VAPNIK V N. The nature of statistical learning theory[M]. 2nd ed. Germany: Springer, 2000: 24-37.

[7] XIA Zhong-hang, DONG Yu-lin, XING Guang-ming. Support vector machines for collaborative filtering[C]//Proc of the 44th Annual Southeast Regional Conference. New York: ACM Press, 2006: 169-174.

[8] ZHANG Tong, IYENGAR V S. Recommender systems using linear classifiers[J]. Journal of Machine Learning Research, 2002, 2: 313-334.

[9] GRČAR M, FORTUNA B, MLADENIČD, *et al*. KNN versus SVM in the collaborative filtering framework[C]//Proc of Data Science and Classification. 2006: 251-260.

[10] SARWAR B, KARYPIS G, KONSTAN J, *et al*. Item-based collaborative filtering recommendation algorithms[C]//Proc of the 10th International Conference on World Wide Web. New York: ACM Press, 2001: 285-295.

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

[12] ZIEGLER C, DBIS F. Book-crossing dataset[EB/OL]. (2006-10-05). http://www.informatik.unl-freiburg.de/~cziegler/BX/

[13] Grouplens Research. Movielens data sets[EB/OL]. (2011-08-24). http://www.grouplens.org/node/73.

[14] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques[M]. 2nd ed. Massachusetts: Morgan Kaufmann, 2005: 56-66.

[15] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [EB/OL]. (2001-10-26)[2007-05-12]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.