第35卷 第21期 计 算 机 工 程 2009年11月
Vol.35 No.21 Computer Engineering November 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)21—0258—03 文献标识码：A 中图分类号：TP301.6

融合多系统用户信息的协同过滤算法

张付志，张启凤
(燕山大学信息科学与工程学院，秦皇岛 066004)

摘 要：为了提高新用户服务的预测准确率，提出一种融合多系统用户信息的协同过滤算法。该算法通过将多个系统的用户信息融合到低维流形中为用户寻找邻居和推荐项目，并介绍流形学习算法在推荐服务中的应用过程。通过对比实验，结果表明该算法比传统协同过滤算法能更有效、准确地为新用户提供推荐。

关键词：局部不变投影；协同过滤；新用户

Collaborative Filtering Algorithm Fusing Multi-system User Information

ZHANG Fu-zhi, ZHANG Qi-feng
(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

【Abstract】 In order to improve the accuracy of forecast for new-user service, this paper proposes a collaborative filtering algorithm fusing multi-system user information. The algorithm finds neighbor and recommendation item by fusing multi-system user information to low dimensional manifold. It introduces the application process of manifold learning algorithm in recommendation service. Result shows that the proved algorithm can be more effective and accurate than traditional collaborative filtering algorithms for new user by contrastive experiment.

【Key words】 local invariable projection; collaborative filtering; new user

1 概述

随着计算机网络的飞速发展和互联网信息的不断膨胀，人们对个性化服务的要求日益提高。个性化推荐服务一定程度上满足了用户多样性的要求。然而，目前电子商务中的个性化推荐服务是独立发生在单一系统内部，不和其他系统发生联系的。以电影网站为例，假设用户U在电影网站A和B中都进行了注册，当用户U访问过A再访问B时，通常希望在未对电影进行评价的前提下，就能享受B网站提供的个性化推荐服务，即实现对B系统的新用户U推荐，也称为跨系统个性化推荐。

在跨系统个性化服务环境下，用户信息通常分散在多个系统中，因此，如何利用其他系统的用户信息为新用户提供推荐成为研究的重点。文献[1]提出为用户建立统一的用户模型，使不同系统间可相互理解，该理解实质上是多维语义的理解。该方法的难点是建立共同的、可以被广泛接受的词汇，以及每个系统都要遵循的标准。

目前个性化推荐服务都是独立发生在单一系统内部的，不同系统间无法共享用户个性化信息。为了更好地为新用户提供推荐，本文引入流形学习中的非线性降维技术。

2 局部不变投影

流形学习是机器学习中一种新的无监督学习方法，它强调整体结构，通过局部与整体相结合来发现和重建数据的内在规律。假设数据均匀采样于一个高维欧氏空间中的低维流形，流形学习是从高维采样数据中恢复低维流形结构。流形学习旨在发现高维数据集分布的内在规律性，数据集的内在结构有如下特性：由泰勒定理可知，任何可微函数在一点充分小的邻域内满足线性条件，即数据流形是由许多可分割的子流形组成的，数据流形的本征维数沿着流形不断地发生变化，只有局部性才能抓住其根本特性[2]。

局部不变投影是流形学习中的新方法，与其他非线性降维方法相比，局部不变投影在高维数据的维数约简中具有保持数据集几何结构和拓扑结构不变的性质，并继承了线性方法计算方便、快捷的优点[3]。下文介绍局部不变投影算法的具体内容。

局部不变投影是映射数据集 $X=\{x_1,x_2,\cdots,x_n\}$，$x_i\in R^D$ 到数据集 $Y=\{y_1,y_2,\cdots,y_n\}$，$y_i\in R^d$ ($D>d$)。该方法主要包括3个步骤：

(1)局部邻域的选取

为每个 x_i 选择 k 个最近邻点，形成 k 个最近邻域。若 x_j 是 x_i 的最近邻点之一，则记为 $j\sim i$。

(2)求解数据集的内蕴结构

拓扑结构的数学描述为

$$Q\left(\boldsymbol{W}_i^{(1)}\right)=\min_{w_i}\{\left\|x_i-\boldsymbol{X}^{(i)}W_i\right\|^2+\lambda\left\|W_i\right\|^2\}$$

$$\lambda\geq 0,\ \sum_{j=1}^{k}\boldsymbol{W}_{ij}^{(1)}=1 \tag{1}$$

其中，$\boldsymbol{X}^{(i)}$ 为 x_i 的邻域矩阵，$i=1,2,\cdots,N$；$\boldsymbol{W}_i^{(1)}=\left(W_{i1}^{(1)},W_{i2}^{(1)},\cdots,W_{ik}^{(1)}\right)$ 为拓扑结构向量，将 $\boldsymbol{W}_i^{(1)}$ 相应的位置补零后，组成 N 维向量，上述向量的全体组成矩阵 $\boldsymbol{W}^{(1)}$，称为数据集的拓

基金项目：河北省自然科学基金资助项目(F2008000877)
作者简介：张付志(1964－)，男，教授、博士生导师，主研方向：Web数据库技术，网络资源共享与管理，智能网络信息处理，网络与信息安全；张启凤，硕士研究生
收稿日期：2009-05-18 **E-mail：**zhangqifeng83@163.com

扑结构矩阵；由于 λ 的选择对结果的影响不大，因此取 $\lambda = 2C_{\max}k/N$，$C_{\max}$ 为 $\boldsymbol{X}$ 的协方差矩阵 $Var(\boldsymbol{X})$ 的最大特征值[3]。

几何结构的数学描述为

$$W_{ij}^{(2)} = \begin{cases} \exp[-\|x_i - x_j\|^2 / t] & i \sim j \text{或} j \sim i \\ 0 & \text{其他情况} \end{cases} \tag{2}$$

其中，$t > 0$ 为给定的常数；$\left(W_{i1}^{(2)}, W_{i2}^{(2)}, \cdots, W_{iN}^{(2)}\right)$ 是几何结构向量，此类向量的全体组成矩阵 $\boldsymbol{W}^{(2)}$，称为数据集的几何结构矩阵。

(3)保持数据内蕴结构不变的降维

令 $\boldsymbol{M}^{(1)} = \left(I - \boldsymbol{W}^{(1)}\right)^{\mathrm{T}}\left(I - \boldsymbol{W}^{(1)}\right)$，$\boldsymbol{M}^{(2)} = \boldsymbol{D} - \boldsymbol{W}^{(2)}$

$\boldsymbol{M} = \boldsymbol{X}\left(l_1\boldsymbol{M}^{(1)} + l_2\boldsymbol{M}^{(2)}\right)\boldsymbol{X}^{\mathrm{T}}$

其中，$\boldsymbol{D}$ 为对角矩阵 $\boldsymbol{D}_{ii} = \sum_j W_{ij}^2$；$l_1, l_2 \geq 0$ 为事先给定的常数，通常取 $l_1 = l_2 = 1$；$\boldsymbol{M}$ 的最小 d 个特征值对应的特征向量为 $\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_d$。因此，最优嵌入映射为 $y_i = \boldsymbol{A}^{\mathrm{T}}x_i$，$y_i \in R^d$，$\boldsymbol{A} \in R^{D \times d}$。

3 2个子流形的融合

将 2 个数据集 $s_x \equiv \{\boldsymbol{x}_i \in R^n, i = 1,2,\cdots,l_x\}$，$s_y \equiv \{\boldsymbol{y}_i \in R^m, i = 1,2,\cdots,l_y\}$ 分别看作 2 个子流形，通过流形融合得到低维流形。假设数据集的前 c 个点有对应关系，即 2 个流形的前 c 个点成对排列。在实际应用中，$\boldsymbol{x}_i$，$\boldsymbol{y}_i$ 分别代表用户 u_i 在系统 A 和系统 B 中的用户向量，假设 $x_i \to y_i$，$i = 1,2,\cdots,c$，该对应关系是已知的。

为了找到 s_x 和 s_y 中数据点的低维嵌入点，在 s_x 和 s_y 上分别构建图 G_x 和 G_y，$\boldsymbol{f}$ 和 $\boldsymbol{g}$ 是图 G_x 和 G_y 上的实值函数。定义组合函数 $\boldsymbol{h} = \left(\boldsymbol{f}^{\mathrm{T}}, \boldsymbol{g}^{\mathrm{T}}\right)^{\mathrm{T}}$，为了将子流形上有对应关系的点映射到低维流形的同一点上，最小化 $\tilde{C}(\boldsymbol{h}) = \dfrac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}}{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}}$，$\mathrm{s.t}\sum_i \boldsymbol{h}_i = 0$ 来计算嵌入函数 f 和 g。其中，$\boldsymbol{L}$ 为联合图 $G \equiv G_x \cup G_y$ 的拉普拉斯算子。该最优化问题可以通过求解矩阵 $\boldsymbol{L}$ 的特征向量来解决[4]。

设图 G_x 和 G_y 的拉普拉斯算子分别为 $\boldsymbol{L}_x = \begin{bmatrix} \boldsymbol{L}_{cc}^x & \boldsymbol{L}_{cs}^x \\ \boldsymbol{L}_{sc}^x & \boldsymbol{L}_{ss}^x \end{bmatrix}$，$\boldsymbol{L}_y = \begin{bmatrix} \boldsymbol{L}_{cc}^y & \boldsymbol{L}_{cs}^y \\ \boldsymbol{L}_{sc}^y & \boldsymbol{L}_{ss}^y \end{bmatrix}$，联合图的拉普拉斯算子 $\boldsymbol{L}_{xy}$ 由式(3)计算。式中前 c 个点成对排列，其余 s 个点没有对应关系。

$$\boldsymbol{L}_{xy} = \begin{bmatrix} \boldsymbol{L}_x^{cc} + \boldsymbol{L}_y^{cc} & \boldsymbol{L}_x^{cs} & \boldsymbol{L}_y^{cs} \\ \boldsymbol{L}_x^{sc} & \boldsymbol{L}_x^{ss} & 0 \\ \boldsymbol{L}_y^{sc} & 0 & \boldsymbol{L}_y^{ss} \end{bmatrix} \tag{3}$$

2 个数据集联合图的拉普拉斯算子的特征向量矩阵称为低维流形，低维流形的维数即特征向量的个数。融合后的低维流形的前 c 个点为 2 个流形中有对应关系的点的低维嵌入点，低维流形中其余的点称为孤立点是 2 个流形中没有对应关系点的低维嵌入点。

4 融合多系统用户信息的协同过滤算法

融合多系统用户信息的协同过滤算法的主要思想是将 2 个系统的用户向量映射到低维流形上，共同注册用户映射到同一点，这样就能够在低维流形上更加有效地寻找邻居，为用户提供推荐。算法主要分为 5 个步骤：

(1)邻居选择：采用协同过滤算法中的皮尔森相关系数。式(4)为每个用户寻找到 k 个最近邻居：

$$sim(i,j) = \frac{\sum_{c \in I_{ij}}\left(R_{i,c} - \bar{R}_i\right)\left(R_{j,c} - \bar{R}_j\right)}{\sqrt{\sum_{c \in I_{ij}}\left(R_{i,c} - \bar{R}_i\right)^2}\sqrt{\sum_{c \in I_{ij}}\left(R_{j,c} - \bar{R}_j\right)^2}} \tag{4}$$

(2)计算重构权值矩阵：用局部不变投影为每个邻居计算重构权值，形成权值矩阵。

(3)计算低维流形：计算联合图拉普拉斯算子 $\boldsymbol{L}_{xy}$ 的特征向量得到低维流形。

(4)计算孤立点的原像：为孤立点在低维流形上寻找最近邻居，并计算重构权值矩阵，同步骤(1)和步骤(2)。保持权值不变，预测出孤立点的原像。

(5)产生推荐：由式(5)为目标用户预测评分并产生推荐：

$$P_{a,j} = \bar{r}_a + \sum_{i=1}^N sim(a,i) \times \left(r_{ij} - \bar{r}_i\right) / \sum_{i=1}^N sim(a,i) \tag{5}$$

融合多系统用户信息的协同过滤算法的流程如图 1 所示。

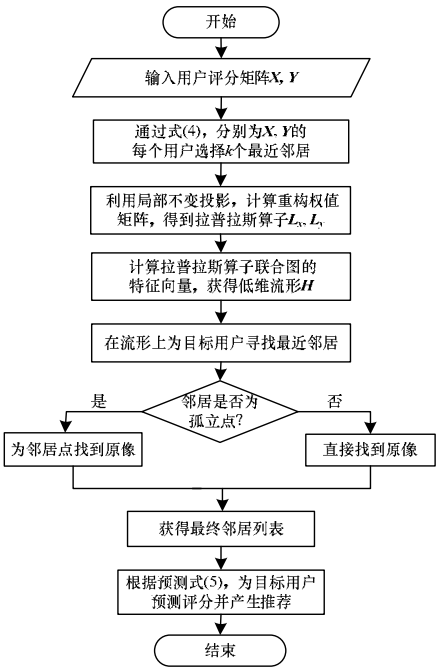

图 1　算法流程

上述算法的核心是计算低维流形和孤立点的原像，下文给出具体的算法描述。

算法 1 计算低维流形 $\boldsymbol{H}$

输入 用户评分矩阵 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，用户数量分别为 n_x，n_y，每一列对应一个用户，前 c 列有对应关系，k 是邻居数目，d 是流形的维数。

(1)以项目的平均值来填充 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的缺失值，得到 $\boldsymbol{x}_{norm}$ 和 $\boldsymbol{y}_{norm}$。

(2)由式(4)为每个 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 选择 k 个最近邻居。

(3)利用局部不变投影中的权值计算式(1)分别求解 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的拓扑结构矩阵 $\boldsymbol{W}_x^1$ 和 $\boldsymbol{W}_y^1$，再用式(2)计算几何结构矩阵 $\boldsymbol{W}_x^2$ 和 $\boldsymbol{W}_y^2$，令 $\boldsymbol{W}_x = a\boldsymbol{W}_x^1 + (1-a)\boldsymbol{W}_x^2$，$\boldsymbol{W}_y = a\boldsymbol{W}_y^1 + (1-a)\boldsymbol{W}_y^2$。其中，$a$ 的取值范围为 0~1，步长取 0.1，通过实验寻找到 a 的最优值；t 可以取任何正数，为计算简便，将 t 值固定，设 t=1。

(4)根据拉普拉斯特征映射[5]中的计算公式

$$\boldsymbol{L}_{ij}=\begin{cases}\sum_{j\sim i}\boldsymbol{W}_{ij} & \text{if} \quad i=j\\ -\boldsymbol{W}_{ij} & \text{if} \quad i\sim j\\ 0 & \text{otherwise}\end{cases}$$ 分别为 $\boldsymbol{W}_x$ 和 $\boldsymbol{W}_y$ 计算图拉普拉斯算子 $\boldsymbol{L}_x$，$\boldsymbol{L}_y$。

(5)由式(3)计算联合图的拉普拉斯算子，s 代表没有对应关系的孤立点。

(6)计算 $\boldsymbol{L}_{xy}$ 的特征向量，得到低维流形 $\boldsymbol{H}$。$\boldsymbol{H}$ 的维数为 $(n_x+n_y-c)\times d$。唯一的参数为流形维数 d。

输出 低维流形 $\boldsymbol{H}$。

算法 2 计算孤立点的原像

输入 流形 $\boldsymbol{H}$，x_{norm}，y_{norm}，n_x，n_y，c，邻居数目 $k1$。n_x-c, n_y-c 分别代表 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的孤立点。$\boldsymbol{H}(i)$ 表示流形上的第 i 个 d 维点。

(1)将前 c 个点和后面的 n_y-c 个点组合起来构成子流形 $\boldsymbol{H}_y$。

(2)For $i=(c+1)to(n_x)$ do $\hat{x}_i=H(i)$，在子流形 $\boldsymbol{H}_y$ 上找到 $\hat{x}_i$ 的 $k1$ 个最近邻居 $\hat{y}_r$。采用算法 1 中的权值计算方法计算重构权值矩阵 $\boldsymbol{W}^*=(w_r)_{r=1}^k$。计算 $\boldsymbol{F}(\hat{x}_i)=\sum_r w_r y_r$，$\hat{x}_s(i-c)=\boldsymbol{F}(\hat{x}_i)$，$y_r$ 表示 $\hat{y}_r$ 在 y_{norm} 中的原像。

End for

(3)把 $\boldsymbol{x}$ 替换为 $\boldsymbol{y}$，对 $\boldsymbol{y}$ 重复上述过程，计算 $\hat{y}_s$。

输出 原像 $\hat{x}_s$，$\hat{y}_s$。

5 实验

5.1 实验数据集

实验数据集为 MovieLens 网站上的 943 个用户对 1 682 部影片的 10^6 条评分记录，用户评分数据集的稀疏等级为 1-100 000/(943×1 682)=0.937 0。数据集被分成 2 个子集，即将用户评分矩阵分成 2 个子矩阵 840×943 和 842×943。随机选择 5%作为项目重叠率。用户重叠数即有对应关系的用户数从 0 变到 800，最后 143 个用户作为测试集(目标用户)。假设目标用户已评分项目为 10，取流形维数 d=6，最近邻居数 k=36，流形上的邻居数 $k1$=55。

5.2 评价标准

实验以黄金标准和受欢迎推荐为 2 个基准，系统能够达到的最好预测称为黄金标准，即一个系统已知 2 个系统的所有数据，用皮尔森相关系数或奇异值分解来计算预测。系统能够达到的最低要求称为受欢迎推荐，即系统对目标用户一无所知，将平均评价分最高的那些资源推荐给目标用户。传统的协同过滤算法对初次使用系统的新用户只能提供非个性化的推荐，即受欢迎推荐。

实验评价标准包括预测准确性评价标准平均绝对误差(Mean Absolute Error, MAE)和推荐结果准确性评价标准 Top-N 推荐的等级分。

(1)*MAE*：*MAE* 用于测量预测值与实际评价值之间的偏差，假设目标用户已对 N 个项目进行评分，对于 N 个评价值-预测值对 $<p_i,q_i>$，$MAE=\dfrac{\sum_{i=1}^{N}|p_i-q_i|}{N}$。*MAE* 值越小，预测越准确。

(2)Top-N 推荐的等级分：$R_{score}=100\times(\sum R/\sum R_{\max})$，$R_{score}$ 取值范围为 0~100。推荐等级分越大，推荐结果越准确。

5.3 实验结果及分析

考虑局部不变投影算法中权重参数 a 对 *MAE* 和推荐等级分的影响，在使用融合算法预测时，找出最优 a 值的范围。随机抽取 200 个、400 个、800 个用户作为跨系统用户，实验结果如图 2、图 3 所示。由此可知，a 的取值范围在 0.6~0.8 是最优的。跨系统用户越多，该趋势越明显。

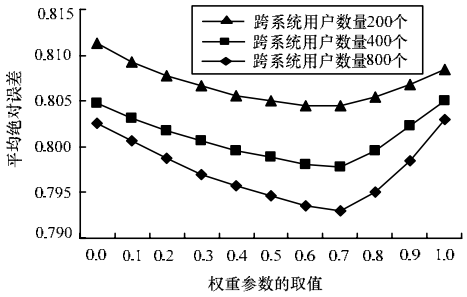

图 2　权重参数对绝对预计误差的影响

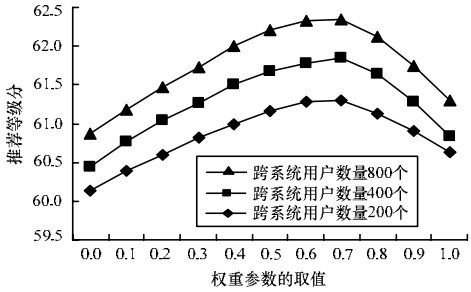

图 3　权重参数对推荐等级分的影响

图 4、图 5 将受欢迎推荐、黄金标准、融合算法三者进行比较。为了说明融合算法能解决传统基于用户的协同过滤算法的新用户的预测问题，做了 2 个实验比较这 2 种方法，其中，a 取值为 0.7。

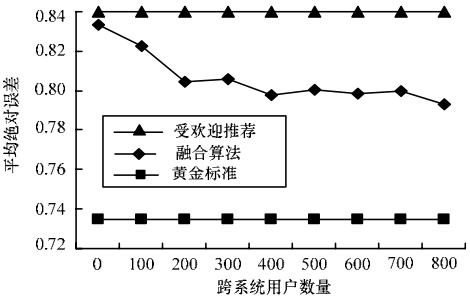

图 4　跨系统用户数量对绝对估计误差的影响

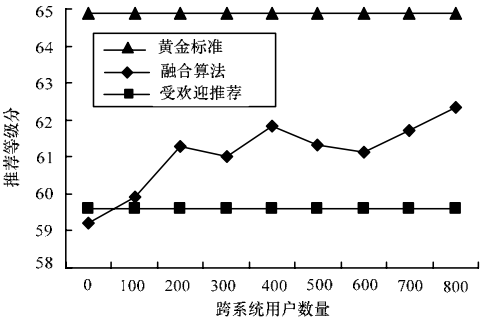

图 5　跨系统用户数量对推荐等级分的影响

由此可知，在预测准确性和推荐准确性上融合算法明显优于受欢迎推荐即传统协同过滤算法。跨系统用户越多，融合算法的性能越好，越接近黄金标准，它很大程度上提高了

（下转第 263 页）