多指标推荐的全局邻域模型

吕红亮[1,2]，王劲林[2]，邓峰[2]
(1. 中国科学院研究生院，100049，北京；2. 中国科学院声学研究所
国家网络新媒体工程技术研究中心，100190，北京)

摘要：针对现有的多指标推荐模型预测精度较低、速度较慢的问题，提出一种多指标推荐的全局邻域模型(MGNgbr)．该模型综合全局的打分信息和隐性反馈数据，通过随机梯度下降法学习物品在各个指标上的相似度，选择相似度最高的前k个邻居参与运算并最终预测用户对物品的打分信息．该模型具有预测准确度高、解释性好、计算复杂度低等优点．实验结果表明，该模型的预测准确度和分类准确度均优于现有的平均值融合模型、多维距离模型和多维奇异值分解模型，与多维奇异值分解模型相比，该模型还具有收敛快、运行时间短等优点．

关键词：随机梯度下降法；全局邻域模型；多指标推荐
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2012)11-0098-08

A Global Neighborhood-Based Model with Multi-Criteria Recommendation

LÜ Hongliang[1,2], WANG Jinlin[2], DENG Feng[2]
(1. Graduate University of Chinese Academy of Sciences, Beijing 100049, China; 2. National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: A global neighborhood-based model with multi-criteria recommendation (MGNgbr) is presented to improve the prediction precision and speed of current models. The model synthesizes the global rating information and implicit feedback data, considers the interrelation of criteria, uses the stochastic gradient descent method to learn the similarity of the items on all the criteria, and selects the most k similar neighbors for prediction. MGNgbr has high prediction accuracy with low computational complexity, and good explanatoriness. Experimental results show that MGNgbr produces better predictive accuracy and classification accuracy than the average similarity k-nearest neighbor model (Avg-KNN), the multi-dimension distance model (M-Dist) and multi-dimension singular value decomposition model (MSVD) do. Comparison with the MSVD model shows that the proposed model also has characteristics of fast convergence and short running time.

Keywords: stochastic gradient descent method; global neighborhood-based model; multi-criteria recommendation

个性化技术和推荐系统用来帮助个人找到感兴趣或者相关的产品或服务．现有的推荐系统解决的主要是单一评分下的推荐问题，此时用户只对物品的整体进行评分，如对电影(Netflix, http://www.netflix.com)和书籍(Amazon, http://www.amazon.com)的整体印象打分，对于此类问题的研究已经相对成熟，主要有启发式的推荐方法和基于模型的推荐方法[1]．然而，对于多指标的推荐问题，目前

收稿日期：2012-02-20． 作者简介：吕红亮(1983—)，男，博士生；王劲林(通信作者)，男，研究员． 基金项目：国家"863计划"重点资助项目(2011AA01A102)；国家科技支撑计划资助项目(2011BAH11B04)；中国科学院战略性先导科技专项资助项目(XDA06010302)．
网络出版时间：2012-08-27 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20120827.1110.007.html

并没有成熟的方法. 多指标推荐主要解决已知产品或者服务的多个方面的评价信息、如何给用户推荐合适的产品或服务的问题. 例如：在 Tripadvisor(http://www.tripadvisor.com)的酒店评价系统中，用户从位置、卫生、服务、房间等方面给酒店打分；在大众点评网(http://www.dianping.com)的饭店评价系统中，用户从口味、环境、服务等方面进行评价；在途牛网(http://www.tuniu.com)的旅游线路预订系统中，用户从住宿、交通、导游和行程等方面对线路进行评价. 如果使用现有的单指标推荐算法，则很难利用所有的打分数据. 多指标推荐算法可以利用所有的打分数据，从而得出更准确的推荐. 现有的多指标推荐方法主要有[2]：融合多个指标相似度作为整体相似度的方法；将多维空间距离作为整体相似度的方法；针对各个指标进行预测，然后采用聚集函数计算整体相似度的方法. 但是，这些方法均以指标相互独立为前提，具有较大的局限性. 多维奇异值分解(multi-dimension singular value decomposition，MSVD)技术[3]采用先填充三维矩阵然后分解的方式，考虑了指标之间的关联，但是稀释的打分矩阵经过填充之后增加了大量的数据，造成算法的时间和空间复杂度都比较高，不利于在大规模数据场景下使用.

本文提出一种多指标推荐的全局邻域模型(MGNgbr)，利用三维打分信息来计算物品之间的相似度，并充分利用隐性反馈数据进行参数训练，具有预测准确度高、解释性好等优点，通过选择相似度最高的前 k 个邻居参与运算，保持了较低的计算复杂度. 实验结果表明，MGNgbr 模型具有较高的预测准确度和分类准确度，与 MSVD 技术相比还具有收敛快、运行时间短等优点.

1　相关研究工作

现有的多指标推荐方法很多是从单指标推荐方法扩展而来的. 单指标推荐方法主要有协作过滤的推荐方法和基于内容的推荐方法 2 种[1]，前者是当前使用最广泛的方法. 协作过滤算法又分为启发式算法和基于模型的算法[4]，它们都可以被扩展以支持多指标打分的情况.

1.1　基于启发式算法的扩展思路

1.1.1　*单指标推荐的启发式算法*　单指标推荐系统希望找到物品 i 对于用户 u 的效用函数 $f(u,i)$.

假设有 N 个用户、M 个物品，用 r_{ui} 表示用户 u 对物品 i 的实际评分，$\hat{r}_{ui}$ 表示推荐系统预测的 u 对 i 的打分. 通过用户对和物品 i 相似的物品 j 的打分来预测对物品 i 的打分，用 $s(i,j)$ 表示物品 i 和 j 的相似度，U_i 表示所有对物品 i 打过分的用户集合，I_u 表示用户 u 打过分的物品的集合，$\tilde{r}_i$ 表示物品 i 的平均打分，则基于物品的最近邻算法(Item-Based KNN)[5]主要可以表示为

$$\hat{r}_{ui}=\tilde{r}_i+k\sum_{j\in I_j}s(i,j)(r_{uj}-\tilde{r}_j) \tag{1}$$

$s(i,j)$ 的计算方法主要有 Pearson 相关系数法和余弦法. 常用的 Pearson 相关系数法的计算方法如下

$$s(i,j)=\frac{\sum_{u\in U_{ij}}(r_{ui}-\tilde{r}_i)(r_{uj}-\tilde{r}_j)}{\left(\sum_{u\in U_{ij}}(r_{ui}-\tilde{r}_i)^2\right)^{1/2}\left(\sum_{u\in U_{ij}}(r_{uj}-\tilde{r}_j)^2\right)^{1/2}} \tag{2}$$

式中：U_{ij} 表示同时对物品 i 和 j 打分的用户集合.

1.1.2　*支持多指标评分的启发式算法*　对于多指标评分的推荐系统，用户对某一物品的评分是一个向量

$$\boldsymbol{R}:\quad \boldsymbol{R}_{\text{users}}\times\boldsymbol{R}_{\text{items}}\rightarrow\boldsymbol{R}_0\times\boldsymbol{R}_1\times\cdots\times\boldsymbol{R}_k$$

文献[6]扩展了上述算法，使其能够支持多指标打分的情况，具体方法如下.

(1)分别针对每个指标计算相似度，然后再将这些相似度进行叠加得到整体的相似度. 叠加的方法有取平均值法(如式(3))、取最低值法(如式(4))等，还可以通过聚集函数的方式获得整体相似度.

$$s_{\text{avg}}(i,j)=\frac{1}{k+1}\sum_{c=0}^{k}s_c(i,j) \tag{3}$$

$$s_{\min}(i,j)=\min_{c=0,1,\cdots,k}s_c(i,j) \tag{4}$$

(2)利用多维空间中的距离来计算相似度，如曼哈顿距离、欧拉距离和切比雪夫距离. 物品 i 和物品 j 对于用户 u 的评分距离可以用下面的公式计算

$$d(r_{ui},r_{uj})=\left(\sum_{c=0}^{k}|r_{ui}-r_{uj}|\right)^{1/m} \tag{5}$$

式中：$m=1$ 时 $d(r_{ui},r_{uj})$ 是曼哈顿距离；$m=2$ 时 $d(r_{ui},r_{uj})$ 是欧拉距离. 物品 i 和 j 的距离可以用下面的公式计算

$$d(i,j)=\frac{1}{|U_{ij}|}\sum_{u\in U_{ij}}d(r_{ui},r_{uj}) \tag{6}$$

式中：$|U_{ij}|$ 是归一化参数. 2 个 Item 的距离越远，相似度应该越小，相似度和距离是反比关系. 定义物品 i 和物品 j 的整体相似度

$$s(i,j)=\frac{1}{d(i,j)+1} \tag{7}$$

多指标推荐的启发式算法存在的共同问题是，都认为各个指标之间的打分是相互独立的，而这与实际情况经常不符，从而影响了预测的准确度.

1.2 基于多维矩阵分解的多指标推荐算法

矩阵分解技术(matrix decomposition，MD)是一种常用的基于模型的推荐方法，奇异值分解(SVD)是其中最重要的一种分解方式[7]. 矩阵分解技术已广泛运用于推荐系统，因为它能够显著提高预测的准确度. 矩阵分解的基本思想是将打分矩阵 $\boldsymbol{R}$ 分解为 $\boldsymbol{U}$ 和 $\boldsymbol{I}$ 两个矩阵，$\boldsymbol{R}=\boldsymbol{U}\boldsymbol{I}$，$\boldsymbol{U}$ 代表用户特征矩阵，$\boldsymbol{I}$ 代表物品特征矩阵.

李秋丹等人[8]使用一种新型的 MSVD 方法来解决多指标推荐问题，并将这种方法应用于一个饭店推荐系统，取得了不错的效果. 这种方法的具体步骤如下：

(1)利用打分数据构建打分矩阵 $\mathscr{A}$；

(2)对 $\mathscr{A}$ 进行展开，得到 $\boldsymbol{A}_1$、$\boldsymbol{A}_2$、$\boldsymbol{A}_3$；

(3)针对展开矩阵 $\boldsymbol{A}_1$、$\boldsymbol{A}_2$、$\boldsymbol{A}_3$ 进行矩阵分解

$$\boldsymbol{A}_1=\boldsymbol{U}^{(1)}\cdot\boldsymbol{S}_1\cdot\boldsymbol{V}_1^{\mathrm{T}} \tag{8}$$

$$\boldsymbol{A}_2=\boldsymbol{U}^{(2)}\cdot\boldsymbol{S}_2\cdot\boldsymbol{V}_2^{\mathrm{T}} \tag{9}$$

$$\boldsymbol{A}_3=\boldsymbol{U}^{(3)}\cdot\boldsymbol{S}_3\cdot\boldsymbol{V}_3^{\mathrm{T}} \tag{10}$$

(4)构造最核心的多维矩阵(tensor)

$$\mathscr{S}=\mathscr{A}\times_1\boldsymbol{U}_{c1}^{(1)^{\mathrm{T}}}\times_2\boldsymbol{U}_{c2}^{(2)^{\mathrm{T}}}\times_3\boldsymbol{U}_{c3}^{(3)^{\mathrm{T}}} \tag{11}$$

(5)根据 $\mathscr{S}$ 构造出最核心的打分矩阵，并据此预测未知打分值.

但是，这种方法存在的问题是：需要进行矩阵填充，从而造成数据量过大，算法复杂度过高，同时人工填充的数据还会降低预测精度.

1.3 单指标推荐的基础预测模型

基础预测模型是 Koren 等人[9]提出的一种基础模型，是单指标推荐问题中很多复杂方法的基础. 这种模型的基本思想是，有些用户的打分比其他用户的高，有些物品更容易得到高分. 用户 u 对物品 i 打分的预测值 b_{ui} 可以表示为

$$b_{ui}=\mu+b_u+b_i \tag{12}$$

式中：μ 是全局打分的平均值；b_u 是用户偏置项，用来描述用户的打分偏好；b_i 是物品偏置项，用来描述物品的平均得分相对于所有物品是偏高还是偏低. 可以通过解决以下最小二乘问题来得到 b_u 和 b_i 的值

$$\min_{b_u,b_i}\sum_{(u,i)\in\kappa}(r_{ui}-\mu-b_u-b_i)^2+\lambda_1\Big(\sum_u b_u^2+\sum_i b_i^2\Big) \tag{13}$$

式中第一项 $\sum_{(u,i)\in\kappa}(r_{ui}-\mu-b_u-b_i)^2$ 表示要找到$\{b_u\mid u\in U\}$和$\{b_i\mid i\in I\}$使预测打分误差较小，后面的正则项 $\lambda_1\Big(\sum_u b_u^2+\sum_i b_i^2\Big)$ 是通过惩罚参数的方式防止过拟合.

本文的算法以此模型为基础，逐步优化扩展，从而得到一个预测效果好、计算复杂度低的模型.

2 多指标推荐的全局邻域模型

2.1 多指标推荐的基础预测模型

式(12)描述的基础预测模型不能直接用于多指标，因为在多指标推荐系统中，除了有用户偏置项 b_u 和物品偏置项 b_i 之外，还存在指标偏置. 指标偏置描述的是某一指标更容易获得高分还是低分，用 b_c 来表示. 考虑指标偏置的因素后，用户 u 在物品 i 的指标 c 上的预测打分 b_{uic} 可以用下式表示

$$b_{uic}=\mu+b_u+b_i+b_c \tag{14}$$

可以通过解决以下的最小二乘问题来估计 b_u、b_i 和 b_c 的值

$$\min_{b^*}\sum_{(u,i,c)\in\kappa}(r_{uic}-\mu-b_u-b_i-b_c)^2+\lambda_1\Big(\sum_u b_u^2+\sum_i b_i^2+\sum_c b_c^2\Big) \tag{15}$$

2.2 多指标推荐的邻域模型

公式(14)可称为多指标基础预测器，它仅仅描述了用户自身的偏置、物品本身的偏置和指标的偏置，而并没有考虑用户对特定物品的偏好. 单指标推荐算法中，基于物品的最近邻算法[5,10](Item-based KNN)是一种比较常用也比较有效的算法，它的基本思想是用户会喜欢与当前喜欢的物品相似度较高的物品. 在衡量物品之间的相似度时，一般使用 Pearson 相关系数 ρ_{ij}，具体的计算方法见式(2).

将 ρ_{ij} 扩展到多指标环境 ρ_{ijc}，则物品 i 和物品 j 在指标 c 上的相似度可以用下式表示

$$\rho_{ijc}=\frac{\sum_{u\in U_{ijc}}(r_{uic}-\tilde{r}_{ic})(r_{ujc}-\tilde{r}_{jc})}{\Big[\sum_{u\in U_{ijc}}(r_{uic}-\tilde{r}_{ic})^2\Big]^{1/2}\Big[\sum_{u\in U_{ijc}}(r_{ujc}-\tilde{r}_{jc})^2\Big]^{1/2}} \tag{16}$$

由于有很多打分是未知的，在计算 ρ_{ijc} 时仅基于同时对物品 i 和物品 j 在指标 c 上打分的用户集合 U_{ijc}. U_{ijc} 中的用户数目越多，则 ρ_{ijc} 的计算越可靠. 如果考虑到共同评价的人数对相似度的影响，可以用下面的公式对 ρ_{ijc} 进行优化，得到物品 i 和物品 j 在指标 c 上的最终相似度

$$s_{ijc} \stackrel{\text{def}}{=} \frac{n_{ijc}}{n_{ijc}+\lambda_2}\rho_{ijc} \tag{17}$$

式中：n_{ijc} 是物品 i 和物品 j 在指标 c 上共同打分的用户数量；λ_2 是一个常量.

综合基础预测器(式(14))和式(17)得到的相似度，即可使用邻域模型来预测用户 u 对物品 i 在指标 c 上的打分 $\hat{r}_{uic}$，即

$$\hat{r}_{uic} = b_{uic} + \frac{\sum\limits_{j\in S^k(i;u,c)} s_{ijc}(r_{ujc}-b_{ujc})}{\sum\limits_{j\in S^k(i;u,c)} s_{ijc}} \tag{18}$$

式中：变量 k 表示邻居的数量；$S^k(i;u,c)$ 表示用户 u 在指标 c 上打过分的物品的集合中，与物品 i 最相似的 k 个物品的集合.

2.3　多指标推荐的全局邻域模型

从式(16)可以看出，为了使物品 i 和 j 在指标 c 上的相似度 s_{ijc} 的计算只与同时对物品 i 和 j 在指标 c 上打分的用户的集合 U_{ijc} 相关，并获得全局的优化效果，希望能够抛弃与少量用户相关的权重，而选择一个全局的权重 w_{ijc}，这个全局权重将通过启发式学习的方式从用户打分数据中获得. 引入 w_{ijc} 后，用户 u 对 i 在指标 c 上的预测打分 $\hat{r}_{uic}$ 可以表示为

$$\hat{r}_{uic} = b_{uic} + \sum_{j\in R(u,c)} (r_{ujc}-b_{ujc})w_{ijc} \tag{19}$$

式中：$R(u,c)$ 表示用户 u 在指标 c 上打过分的物品的集合.

2.4　考虑隐性反馈的多指标推荐的全局邻域模型

公式(19)仅仅考虑了用户打过分的物品，而实际上面对如此众多的物品，用户打分是有一定选择的，一个物品是否被用户打分也反映了用户的一部分偏好，这是用户的一种隐性的反馈. 考虑到用户的隐性反馈，式(19)可以优化为

$$\hat{r}_{uic} = b_{uic} + \sum_{j\in R(u,c)} (r_{ujc}-b_{ujc})w_{ijc} + \sum_{j\in N(u,c)} d_{ijc} \tag{20}$$

式中：$N(u,c)$ 表示用户 u 在指标 c 上提供了反馈的物品的集合. 公式(20)有一个特征，即如果用户 u 在指标 c 上提供了更多的显性反馈($|R(u,c)|$较大)，或者提供了更多的隐性反馈($|N(u,c)|$较大)，则用户的预测打分 $\hat{r}_{uic}$ 会偏离基础预测器更远. 一般来说，这种情况对推荐系统是一个好现象，对于提供更多反馈的用户，我们愿意承担更大的风险来推荐不是那么流行的物品，而对于提供反馈较少的用户，我们倾向于提供更加保守的预测. 但是，从经验来看，公式(20)夸大了反馈较多的用户和反馈较少的用户之间的差别[9].

一种更折中的方案是采用下面的模型

$$\hat{r}_{uic} = b_{uic} + |R(u)|^{-1/2} \sum_{j\in R(u,c)} (r_{ujc}-b_{ujc})w_{ijc} + |N(u)|^{-1/2} \sum_{j\in N(u,c)} d_{ijc} \tag{21}$$

这个模型的计算复杂度较高，如果去掉与 i 相似度较低的物品，只考虑用户评过分的物品集合 $R(u,c)$ 中与 i 最相似的 k 个物品，则可以降低其计算复杂度. 如果用 $S^k(i,c)$ 表示与物品 i 在指标 c 上最相似的 k 个物品的集合，相似度用式(17)定义的 s_{ijc} 来衡量，定义

$$R^k(i;u,c) = R(u,c) \cap S^k(i,c)$$
$$N^k(i;u,c) = N(u,c) \cap S^k(i,c)$$

则模型(21)只考虑 k 个最相似邻居时可以表示为

$$\hat{r}_{uic} = \mu + b_u + b_i + b_c + |R^k(i;u,c)|^{-1/2} \sum_{j\in R^k(i;u,c)} (r_{ujc}-b_{ujc})w_{ijc} + |N^k(i;u,c)|^{-1/2} \sum_{j\in N^k(i;u,c)} d_{ijc} \tag{22}$$

当 $k=\infty$ 的时候，模型(22)和模型(21)完全一致，但是当 k 取其他值的时候，模型(22)可以减少参与运算的数据数量.

模型(22)即为最终预测模型，可以用来预测用户对物品的打分，并根据这个打分为用户提供合适的推荐. 这个模型还可以提供快速的在线推荐，因为大量的参数估计的运算都是离线完成的，在线预测则需要使用式(22)进行计算. 由于采用了邻域模型作为基础，本算法具有较好的解释性.

设计这个模型的主要目的是利用全局的打分信息对用户的偏好进行刻画，这是 Adomavicius 等人[6]的模型所缺乏的，并且该模型还考虑到了隐性反馈信息，信息利用量大，因此可获得更高的预测准确度. 从以上的描述可以知道，式(22)描述的是一种多指标推荐系统下的全局邻域模型(multi-criteria global neighborhood-based model)，我们称之为 MGNgbr 模型.

如果用 e_{uic} 来描述预测值 $\hat{r}_{uic}$ 和实际值 r_{uic} 之间的误差，即令 $e_{uic}=r_{uic}-\hat{r}_{uic}$，则模型(22)中的参数 b_u、b_i、b_c、w_{ijc} 和 d_{ijc} 可以通过最小化如下损失函数来获得

$$C(\kappa) = \sum_{(u,i,c)\in\kappa} \Big(e_{uic}^2 + \lambda\Big(b_u^2 + b_i^2 + b_c^2 + \sum_{j\in R^k(i;u,c)} w_{ijc}^2 + \sum_{j\in N^k(i;u,c)} d_{ijc}^2\Big)\Big) \tag{23}$$

式中:κ 表示整个训练集的已知打分数据;$\lambda\left(b_u^2+b_i^2+b_c^2+\sum_{j\in K^k(i;u,c)}w_{ijc}^2+\sum_{j\in N^k(i;u,c)}d_{ijc}^2\right)$ 用来防止训练出现过拟合. 公式(23)可以通过随机梯度下降法进行优化:首先对 b_u、b_i、b_c、w_{ijc} 和 d_{ijc} 求偏导

$$\frac{\partial C}{\partial b_u}=-2e_{uic}+2\lambda b_u \tag{24}$$

$$\frac{\partial C}{\partial b_i}=-2e_{uic}+2\lambda b_i \tag{25}$$

$$\frac{\partial C}{\partial b_c}=-2e_{uic}+2\lambda b_c \tag{26}$$

$$\forall j\in R^k(i;u,c):$$

$$\frac{\partial C}{\partial w_{ijc}}=-2\mid R^k(i;u,c)\mid^{-1/2}\cdot(r_{ujc}-b_{ujc})e_{uic}+2\lambda w_{ijc} \tag{27}$$

$$\frac{\partial C}{\partial d_{ijc}}=-2\mid N^k(i;u,c)\mid^{-1/2}e_{uic}+2\lambda d_{ijc} \tag{28}$$

然后,利用随机梯度下降法对 b_u、b_i、b_c、w_{ijc} 和 d_{ijc} 进行迭代更新

$$b_u\leftarrow b_u+\eta(e_{uic}-\lambda b_u) \tag{29}$$

$$b_i\leftarrow b_i+\eta(e_{uic}-\lambda b_i) \tag{30}$$

$$b_c\leftarrow b_c+\eta(e_{uic}-\lambda b_c) \tag{31}$$

$$\forall j\in R^k(i;u,c):$$

$$w_{ijc}\leftarrow w_{ijc}+\eta(\mid R^k(i;u,c)\mid^{-1/2}\cdot(r_{ujc}-b_{ujc})e_{uic}-\lambda w_{ijc}) \tag{32}$$

$$d_{ijc}\leftarrow d_{ijc}+\eta(\mid N^k(i;u,c)\mid^{-1/2}e_{uic}-\lambda d_{ijc}) \tag{33}$$

以上各式中:η 表示学习速率,即一次学习过程中各个参数的更新速率.

MGNgbr 算法的具体流程描述见图 1.

for count=1,…,maxStep ＃迭代过程
for u=1,…,m do
for c=1,…,r do
for $i\in R(u,c)$ do
用公式(22)计算 $\hat{r}_{uic}$;
$e_{uic}\leftarrow r_{uic}-\hat{r}_{uic}$
$b_u\leftarrow b_u+\eta(e_{uic}-\lambda b_u)$;
$b_i\leftarrow b_i+\eta(e_{uic}-\lambda b_i)$;
$b_c\leftarrow b_c+\eta(e_{uic}-\lambda b_c)$;
for $j\in R^k(i;u,c)$ do
$w_{ijc}\leftarrow w_{ijc}+\eta(\mid R^k(i;u,c)\mid^{-1/2}(r_{ujc}-b_{ujc})e_{uic}-\lambda w_{ijc})$;
$d_{ijc}\leftarrow d_{ijc}+\eta(\mid N^k(i;u,c)\mid^{-1/2}e_{uic}-\lambda d_{ijc})$
return $\{b_u,b_i,b_c,w_{ijc},d_{ijc}\mid u\in U,\ i\in I,\ c\in C\}$

图 1 MGNgbr 算法描述

根据图 1 的算法描述可知,MGNgbr 模型的计算复杂度为 $O(k|\kappa|)$,仅与训练数据中的样本数量 $|\kappa|$ 和邻居数量 k 有关;存储复杂度为 $O(|I|^2|C|)$,与物品的数量 $|I|$ 和指标的数量 $|C|$ 相关.

3 实验评估

3.1 测试数据集

为了评估本文提出的 MGNgbr 模型,采用离线测试的方法,选取 Tripadvisor 的酒店打分数据和雅虎电影(http://movies.yahoo.com/)的评分数据作为测试数据集.

Tripadvisor 是一个基于 Web2.0 的旅游向导网站,帮助用户收集旅游服务信息,用户可以在上面查看其他用户对酒店的评价信息,也可以发布自己对酒店的评价. 这里打分数据是从 Tripadvisor 官方网站抓取的,每个打分数据包括整体、价值、房间、位置、卫生、前台、服务、商务服务等 8 个指标的分数,分值从 1(最不喜欢)到 5(最喜欢). 为了保证数据的有效性,仅仅选择打分超过 15 个酒店的用户的打分数据,最后得到的数据集属性如表 1 所示.

表 1 实验数据集属性

数据集	用户数	物品数	指标数	打分值	稀疏度/%
Tripadvisor	8 215	1 850	8	260 850	1.71
雅虎电影	4 652	4 244	5	115 173	0.58

雅虎电影是一个提供电影综合信息的网站,包括电影基本信息、预告片、票房数据等,还提供用户评论功能,用户可以查看和发布对电影的评价信息. 雅虎电影中的评价信息包括对剧情(story)、表演(acting)、导演(direction)、画面(visual)和整体(overall)的评分信息,分值从 A+(最喜欢)到 F(最不喜欢). 为了便于运算,将 A+到 F 映射成为 13 到 1,同时为了使数据不过于稀疏,挑选打分超过 10 部电影的用户的数据,最后得到的雅虎电影的数据集属性如表 1 所示.

3.2 实验数据分析

推荐系统的评估指标主要有预测准确度、分类准确度和算法效率[11]. 预测准确度主要用平均绝对误差 E_{ma} 或均方根误差 E_{rms} 来表示,分类准确度主要用准确率和召回率来衡量,而算法的效率主要通过算法运行时间,以及通过对算法时间复杂度和空间复杂度的分析得到.

3.2.1 预测准确度 平均绝对误差和均方根误差

是常用的评价算法预测准确度的方法，定义为

$$E_{\mathrm{ma}}=\frac{1}{|\kappa_t|}\sum_{(u,i,c)\in\kappa_t}|r_{uic}-\hat{r}_{uic}| \tag{34}$$

$$E_{\mathrm{rms}}=\left(\frac{1}{|\kappa_t|}\sum_{(u,i,c)\in\kappa_t}|r_{uic}-\hat{r}_{uic}|^2\right)^{1/2} \tag{35}$$

式中：κ_t 为测试集；$|\kappa_t|$ 为测试集内的打分数目. 推荐算法的准确度是测试集中所有预测打分的准确度的平均值. 平均绝对误差具有计算方法简单、易于理解的优点，而均方根误差在求和之前对系统预测打分与用户打分的误差进行平方，因此打分误差越大，其对均方根误差的影响会比对平均绝对误差的影响更大，故本实验采用均方根误差来衡量各种算法的预测准确度.

3.2.2　*分类准确度*　分类准确度衡量推荐系统分类是否准确的能力，常用的指标是准确率和召回率. 令 $R(u)$ 表示根据用户在训练集上的行为作出的推荐列表，$T(u)$ 表示测试集中用户的喜好列表，则推荐结果的准确率定义为

$$P=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}R(u)} \tag{36}$$

推荐结果的召回率定义为

$$R=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}T(u)} \tag{37}$$

对于 Tripadvisor 的数据集，本文定义用户评分为 4～5 分表示喜欢，1～3 分表示不喜欢；对于雅虎电影的数据，定义用户评分为 10～13 分表示喜欢，1～9 分表示不喜欢.

3.3　实验过程

针对上面提到的 3 个指标分别进行测试. 在实验过程中，首先把 tripadvisor 和雅虎电影的数据集都平均分成 10 份，随机挑选其中若干份作为训练集，对模型进行训练；然后将剩下的数据作为测试集，用来评价每个模型的表现. 每组实验重复进行 5 次后取均值，以消除随机误差. 实验所用的机器是一台 Dell R710 服务器，操作系统为 RHEL AS4，CPU 的型号为 Intel Xeon E5520(8 核)，内存 16 GB.

3.3.1　*预测准确度实验*　在实验过程中，分别使用文献[6]提到的平均值融合模型(Avg-KNN)、多维距离模型(M-Dist)和文献[3]提到的 MSVD 方法，以及本文提出的考虑隐性反馈的全局邻域模型(MGNgbr)进行评测，并比较它们在均方根误差指标上的表现. 通过实验，测得 Avg-KNN 模型中邻居数取 120 时效果最好，故这里取 $k=120$；M-Dist 模型采用较常用的欧拉距离进行计算. 对于 MGNgbr 模型，邻居数取 100 时效果最好，故这里取 $k=100$，并且式(23)中的正则化因子 λ 取 0.01，式(29)～(33)中的学习速率因子 η 取 0.05.

评测方法如下.

(1)将目标数据根据用户随机挑出 α 作为测试集，$(1-\alpha)$作为训练集. α 从 0.1 变化到 0.9，评价算法在不同稀疏度数据下的表现.

(2)利用训练集的数据，分别运用 Avg-KNN、M-Dist、MSVD 和 MGNgbr 模型进行训练，然后预测测试集中各个数据的打分，并计算所有打分的均方根误差.

(3)针对 Tripadvisor 的数据集和雅虎电影的数据集分别得到的结果，用 Matlab 进行拟合，结果如图 2、图 3 所示.

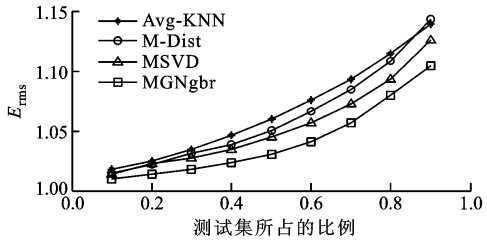

图 2　Tripadvisor 数据集的均方根误差比较

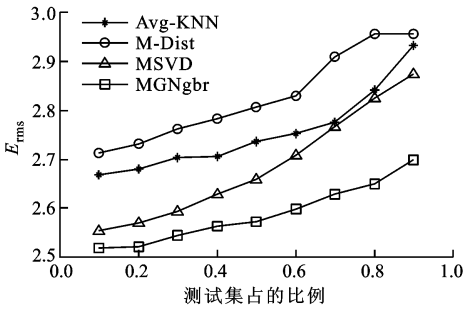

图 3　雅虎电影数据集的均方根误差比较

可以看出，无论是使用 Tripadvisor 的数据集还是雅虎电影的数据集，MGNgbr 模型都具有较大的优势，特别是在数据稀疏度比较高的情况下，MGNgbr 模型的优势更大，而实际的数据往往都是稀疏度比较高的数据.

3.3.2　*准确率-召回率曲线比较*　在数据稀疏度固定的情况下，考察不同方法的准确率和召回率曲线，观察它们的表现. 具体方法是，先对物品进行预测打分，然后对打分进行排序，在进行 top-N[12] 推荐时，返回预测打分结果最高的前 N 个物品. 可以通过提

高 N 的值来提高召回率.

取 $\alpha=0.2$,分别考察 Avg-KNN、M-Dist、MSVD 和 MGNgbr 模型在 Tripadvisor 数据集和雅虎电影数据集上的表现,最终得到的准确率-召回率曲线如图 4 和图 5 所示. 由这 2 个图中可以看出:MGNgbr 模型具有较好的表现,即在相同的查准率下具有较高的查全率,而在相同的查全率下又有着较高的查准率,因而在进行 top-N 推荐的时候能达到较好的效果;MSVD 模型优于 Avg-KNN 和 M-Dist 模型,这是因为 MGNgbr 和 MSVD 模型均从全局利用打分信息,而不是孤立地利用各个指标的打分信息. MGNgbr 的效果更好的原因是它还利用了隐性反馈信息,信息挖掘得更充分.

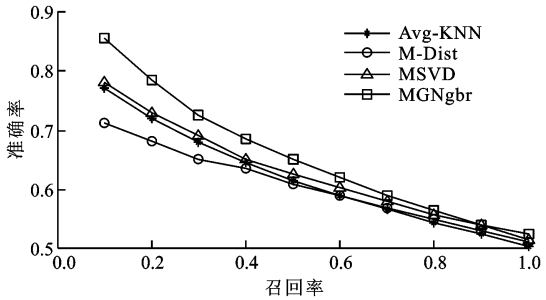

图 4 Tripadvisor 数据集的准确率-召回率曲线

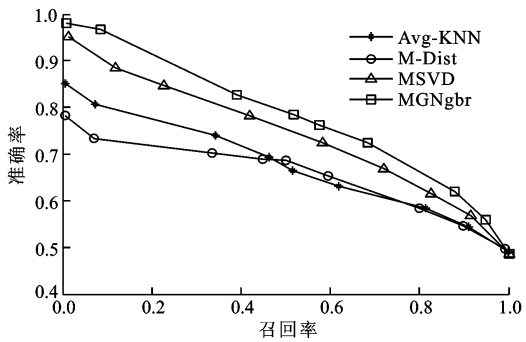

图 5 雅虎电影数据集的准确率-召回率曲线

3.3.3 *算法的时间复杂度和空间复杂度比较* 根据 2.4 节的分析,可知 MGNgbr 模型的时间复杂度为 $O(K|\kappa|)$,与 Avg-KNN、M-Dist 模型的时间复杂度基本相同,而 MSVD 模型需事先对矩阵进行填充,然后对矩阵进行分解,由于打分矩阵往往都是很稀疏的,故其时间复杂度远远高于另 3 种模型. MGNgbr 模型在实际运行过程中经 15 次迭代就可以收敛,运行时间为 30~45 s;Avg-KNN 模型的实际运行时间为 33~60 s;M-Dist 模型的实际运行运行时间为 60~100 s;MSVD 模型的实际运行时间则都在 20 min 以上. MGNgbr 算法的时间复杂度较低还有一个重要原因,就是仅仅选择相似度最高的前 k 个邻居参与运算.

MGNgbr、Avg-KNN 和 M-Dist 模型所消耗的最大空间就是相似度矩阵,其空间复杂度为 $O(|I|^2|C|)$,MSVD 模型占用空间最大的是经填充后的三维矩阵,空间复杂度为 $O(|U||I||C|)$. 各模型占用空间大小的关系取决于用户和物品的相对数量:如果物品数量多于用户数量,则 MGNgbr、Avg-KNN 和 M-Dist 模型消耗的空间大;如果用户数量多于物品数量,则 MSVD 模型消耗的空间大. 在本文的实验中,都是用户的数量多于物品的数量.

4 结束语

对于多指标打分环境下的推荐问题,现有的 Avg-KNN 模型和 M-Dist 模型的预测准确度较差,而 MSVD 模型的计算复杂度较高. 本文提出了一种多指标推荐的全局邻域模型(MGNgbr),用来解决多指标打分预测问题. 此模型综合全局的打分信息,并考虑隐性反馈,通过随机梯度下降法学习物品在各个指标上的相似度,从而预测用户对物品的打分. 实验结果表明,本文提出的 MGNgbr 模型与现有的 Avg-KNN 模型、M-Dist 模型和 MSVD 模型相比,具有较高的预测准确度和分类准确度,且时间复杂度与 Avg-KNN 和 M-Dist 模型基本持平,但是低于 MSVD 的时间复杂度.

本文采用的基于邻域的方法必须要存储物品各个指标之间的相似度,在物品数量很多的情况下,需要的存储空间很大,因此如何降低存储空间将是未来需要深入研究的问题. 多指标评价系统需要用户提供更多的反馈信息,用户进行反馈的成本较高,从而造成数据更加稀疏,因此,如何解决稀疏度较高场景下的多指标推荐问题,也值得进一步研究.

参考文献:

[1] 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.

[2] ADOMAVICIUS G, MANOUSELIS N, KWON Y. Multi-criteria recommender systems [M]// Recommender Systems Handbook: A Complete Guide for Research Scientists & Practitioners. New York, USA: Springer, 2010: 767-801.

[3] LI Qiudan, WANG Chunheng, GENG Guanggang.

Improving personalized services in mobile commerce by a novel multicriteria rating approach [C]// Proceedings of the 17th International Conference on World Wide Web. New York, USA: ACM, 2008:1235-1236.

[4] SU Xiaoyuan, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J/OL]. Advances in Artificial Intelligence, 2009, 2009: 421425 [2011-12-12]. http://dl.acm.org/citation.cfm? id=1722966.

[5] LINDEN G, SMITH B, YORK J. Amazon.com recommendations: item-to-item collaborative filtering [J]. IEEE Trans Internet Computing, 2003, 7(1): 76-80.

[6] ADOMAVICIUS G, KWON Y O. New recommendation techniques for multicriteria rating systems [J]. IEEE Trans Intelligent Systems, 2007, 22(3): 48-55.

[7] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42(8): 30-37.

[8] LI Qiudan, WANG Chunheng, GENG Guanggang, et al. A novel collaborative filtering-based framework for personalized services in m-commerce [C]// Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM, 2007:1251-1252.

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

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

[11] 刘建国，周涛，郭强，等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学, 2009, 6(3): 1-10.
LIU Jian-guo, ZHOU Tao, GUO Qiang, et al. Overview of the evaluated algorithms for the personal recommendation systems [J]. Complex Systems and Complexity Science, 2009, 6(3): 1-10.

[12] DESHPANDE M, KARYPIS G. Item-based top-*n* recommendation algorithms [J]. ACM Transactions on Information Systems, 2004, 22(1): 177-187.

（编辑　葛赵青　武红江）