研究与开发

# 基于自编码器与多模态数据融合的视频推荐方法

顾秋阳[1]，琚春华[2]，吴功兴[2]

（1. 浙江工业大学管理学院，浙江 杭州 310023；2. 浙江工商大学，浙江 杭州 310018）

**摘　要**：现今常用的线性结构视频推荐方法存在推荐结果非个性化、精度低等问题，故开发高精度的个性化视频推荐方法迫在眉睫。提出了一种基于自编码器与多模态数据融合的视频推荐方法，对文本和视觉两种数据模态进行视频推荐。具体来说，所提方法首先使用词袋和 TF-IDF 方法描述文本数据，然后将所得特征与从视觉数据中提取的深层卷积描述符进行融合，使每个视频文档都获得一个多模态描述符，并利用自编码器构造低维稀疏表示。本文使用 3 个真实数据集对所提模型进行了实验，结果表明，与单模态推荐方法相比，所提方法推荐性能明显提升，且所提视频推荐方法的性能优于基准方法。

**关键词**：自编码器；多模态表示；数据融合；视频推荐

**中图分类号**：TP391

**文献标识码**：A

**doi:** 10.11959/j.issn.1000−0801.2021031

# Fusion of auto encoders and multi-modal data based video recommendation method

GU Qiuyang[1], JU Chunhua[2], WU Gongxing[2]

1. Zhejiang University of Technology, School of Management, Hangzhou 310023, China
2. Zhejiang Gongshang University, Hangzhou 310018, China

**Abstract:** Nowadays, the commonly used linear structure video recommendation methods have the problems of non-personalized recommendation results and low accuracy, so it is extremely urgent to develop high-precision personalized video recommendation method. A video recommendation method based on the fusion of autoencoders and multi-modal data was presented. This method fused two data including text and vision for video recommendation. To be specific, the method proposed firstly used bag of words and TF-IDF methods to describe text data, and then fused the obtained features with deep convolutional descriptors extracted from visual data, so that each video document could get a multi-modal descriptors, and constructed low-dimensional sparse representation by autoencoders. Experiments were performed on the proposed model by using three real data sets. The result shows that

**收稿日期**：2020−04−30；**修回日期**：2021−01−30

**通信作者**：顾秋阳，guqiuyang123@163.com

**基金项目**：国家自然科学基金资助项目（No.71571162）；浙江省社会科学规划重点课题项目（No.20NDJC10Z）；国家社会科学基金应急管理体系建设研究专项（No.20VYJ073）；浙江省哲学社会科学重大课题项目（No.20YSXK02ZD）

**Foundation Items:** The National Natural Science Foundation of China (No.71571162), The Social Science Planning Key Project of Zhejiang Province (No.20NDJC10Z), The National Social Science Fund Emergency Management System Construction Research Project (No.20VYJ073), Zhejiang Philosophy and Social Science Major Project (No.20YSXK02ZD)

compared with the single-modal recommendation method, the recommendation results of the proposed method are significantly improved, and the performance is better than the reference method.

**Key words:** autoencoder, multi-modal representation, data fusion, video recommendation

## 1　引言

随着信息技术的不断发展，用户和企业对推荐方法的需求不断提升[1]。推荐方法已在许多领域获得成功应用（包括商品推荐、音乐推荐、电影推荐等），主要通过学习用户历史信息得到其偏好，以生成相关推荐列表，这一过程也被称为 top-*N* 推荐。近年，国内外学者已提出了若干个 top-*N* 推荐算法[2-6]，主要可分为潜空间算法（latent space method，LSM）和基于邻域的算法（又分为基于用户或目标的方法）。现已证明潜空间算法是解决推荐排名问题的最优方法，而邻域算法则能够更好地进行 top-*N* 推荐问题[7]。而基于用户与目标的两种推荐方法中，基于目标的邻域方法是通过预先定义的度量方式计算得到目标间的相似性进行推荐的方法，显然优于基于用户的推荐方法[8]。

但当对目标的描述处于稀疏高维空间时，基于目标的推荐算法存在欧氏距离（由于高维数据的稀疏性，将低维空间中的距离度量函数应用到高维空间时，随着维数的增加，数据对象之间距离的对比性将不复存在，其有效性大大降低）和维数膨胀（当维数越来越多时，数据计算量迅速上升，所需的空间样本数会随维数的增加而呈指数增长，分析和处理多维数据的复杂度和成本呈指数级增长）等方面的困难，也会出现推荐结果非个性化、精度不高等问题[4]。有学者引入了稀疏线性方法（sparse linear method，SLIM）以解决和缓解上述困难和问题。当目标的相关信息（如电影描述等）不断增加时，Ning 等[9]认为应充分利用这些信息，而不是仅关注由用户过去信息组成的偏好。但其关注的始终是目标描述的单一模态，即文本信息模态。而现实网络信息存在多种输入渠道，包括配有情节和架构的视频与带有说明文字和标签的图片等。每种模态的统计特征都大为不同，如何将其有效结合起来进行推荐是现今学术界的一大困难。例如，当给一张图片配上说明文本时，文本往往会对一些从图片中无法直接得到的信息（如地点、人名和事件等）进行描述。故研究如何基于多模态数据融合方法进行目标推荐十分重要。

本文提出了一种基于自编码器与多模态数据融合的视频推荐方法，其中的关键问题是如何对不同模态的信息进行有效组合，以便向用户提供目标信息的推荐。与现今学术界常用的方法不同，本文通过两种数据模态来制定视频推荐方法——视觉（即图像序列等）模态和文本（即标签、标题和描述等）模态，它们构成了视频的基本数据。在本文所提的多模态数据融合推荐方法中，数据由不同的输入模态组成，且各模态的表示方式也各不相同。如图像通常通过像素强度或特征提取器的输出结果进行呈现，它们都为实值且分布密集；相反，文本信息倾向于以离散的稀疏文本向量来呈现，故要找到各模态间的关系非常困难。一个好的多模态表示项目首先需满足在某些模态缺失的情况下，也须能得出最终结果；且其结果必须有利于目标间的相似性计算。

本文所提基于自编码器与多模态数据融合的视频推荐算法可较好地满足以上条件。自编码器是一种自主全连接的单隐层神经网络，其目的是从无标记的数据集中进行学习[11]。传统的自编码器常用于特征学习或维数约简，本文使用自编码器来融合来自多个模态的目标信息。本文的主要贡献如下，首先，设计了一种基于自编码器的多模态推荐架构，有效融合文本与视觉两种模态的

数据，显著提升了视频推荐性能。其次，本文所提视频推荐模型可有效避免单模态数据缺失带来的不能得到推荐结果的情况。再次，本文所提视频推荐方法首先使用词袋和 TF-IDF 方法来描述文本数据，并将所得特征与从视觉数据中提取的深层卷积描述符进行融合，以使每个视频文档都获得一个多模态描述符，并利用自编码器构造低维稀疏表示，有效提升了推荐技术的可行性。最后，基于多个具有明显特征的真实数据集进行了实验，证明了本文所提架构的有效性。

## 2 研究现状

近年，随着电子商务网站的不断流行，学术界在用户信息及产品推荐上已做出了许多突破。可将此类推荐算法分为 3 类：考虑用户个人信息及其偏好的协同过滤算法、基于内容及相似性的算法以及基于用户偏好和内容的算法。

首先对协同过滤视频推荐算法进行介绍。Davidson 等[12]发明了一种针对 Youtube 的视频推荐方法。假设用户观看或点赞了某个视频，该视频就为此用户的相关视频，之后可以使用关联规则基于该用户的相关视频进行推荐。Zhou 等[13]证明了视频推荐是 YouTube 上视频浏览量的最重要来源之一。而 Linden 等[14]在 Amazon 网站上尝试使用协同过滤算法和分组技术，旨在向其客户进行有效的产品推荐。此外，Ahmed 等[15]通过对用户偏好分类来进行视频推荐。

基于内容的过滤算法是现今应用最为广泛的推荐算法[16]。这类方法的一个关键特征是用户建模过程。与其他推荐方法相比，这类算法能够显著提高推荐质量，其根据用户购买的物品对其偏好进行测算。某些内容可能对提高推荐的准确度有较强影响，如标签、消息和多媒体信息等[17]。故基于内容的过滤算法对视频推荐非常重要，这是由于该算法结合了来自 Web2.0 环境中用户的多源信息。现已有几种推荐系统使用了这类算法，其中某些利用文档标题进行推荐[18]，另外一些利用文档中的文本信息进行推荐[19]。

还有一些学者使用混合推荐算法，将协同过滤算法与基于内容的过滤算法相结合。Kim 等[20]证实了这种混合过滤算法能提高推荐质量。在其使用的推荐算法中，使用标签信息来推断用户偏好，尤其当用户（冷启动新用户）的相关信息很少或没有信息时，标签能够为推荐提供一些线索。此外，Yang 等[21]提出了一种混合算法，将评价与电影的文本信息相结合以实现推荐。其中，文本信息由电影标签和流派表示。Zhang 等[28]使用另一种混合算法来进行视频推荐，其利用视频内容（标题和描述）训练神经网络，并利用模糊算子将其结果与用户的个人资料相结合，从而做出推荐。

Bobadilla 等[1]提出利用物品重要性来提高推荐质量。为了衡量目标的重要性，使用一些标准，如通过浏览量和/或评论的数量来判断目标的受欢迎程度等。还通过分析最受欢迎的 Twitter 话题，对 Twitter 上的分享视频进行推荐。Beutel 等[23]利用用户状态（如时间和设备类型等）来提高推荐准确性。电影推荐系统旨在通过使用诸如类型和影片名称等内容信息为其用户推荐新电影。黄立威等[24]通过自深度学习神经网络获得的描述符，与颜色和纹理的描述符相结合实现推荐。参考文献[25-26]使用了神经网络的不同架构来表示影片内容。这些研究的目标是改进目标内容的表现形式，从而提高推荐质量。Cheng 等[27]提出了不同的表示方法，其使用一种基于神经网络来改进用户与目标间关系的表示。

还有一些研究考虑了多种模态维度的数据，如视频帧和文本元数据，并将其结合起来进行推荐，这些方法被称为多模态推荐系统。如 Zhang 等[28]提出了一种深度学习方法，以便结合多种维度的信息（图像、文本和评分等）进行产品推荐。

而 Li 等[29]提出了基于音频和图像特征（颜色）的多模态视频推荐方法。另外，为计算视频间的相似度，使用了余弦距离函数；考虑推荐系统中的信息稀疏，提出了一种基于自编码器的推荐方法；为实现推荐，使用了分层贝叶斯模态的变分自编码器，以使得其能够在潜在概率变化下对每个项目进行表示。

通过上述对现有研究成果的梳理发现，关于视频推荐方法的研究已受到国内外学者的重视并得到丰富的研究成果。上述研究成果虽对本文具有一定的借鉴意义，但也存在一些不足：（1）现有参考文献中的视频推荐方法多使用单模态数据实现视频推荐（如王娜等[30]），较少有将多模态数据融合进行视频推荐的参考文献记录。（2）现有参考文献多只对推荐方法进行了改进，没有考虑其所提方法在高维数据等计算情况中的适用性（如苏赋等[31]），本文利用自编码器构造低维稀疏表示，有效提升了推荐技术的可行性。（3）现有关于基于视觉数据的推荐方法多使用特征提取方法实现推荐（如 Felipe 等[32]），很少有使用词袋和 TF-IDF 方法先进行描述文本数据，并将所得特征与视觉数据中提取的深层卷积描述符进行融合的参考文献记录。

本文认为应设计基于自编码器的多模态推荐架构，有效融合文本与视觉两种模态的数据，显著提升视频推荐性能。使用词袋和 TF-IDF 方法来描述文本数据，并将所得特征与从视觉数据中提取的深层卷积描述符进行融合，使每个视频文档都获得一个多模态描述符，并利用自编码器构造低维稀疏表示，有效提升了推荐技术的可行性。最后，基于多个具有明显特征的真实数据集进行了实验，证明本文所提架构能有效提升推荐的精度与效率。故本文将文本与视觉两种模态数据进行融合，以期在现实生活中提升视频推荐效率，为有关部门进行视频监控与推荐服务提供参考。

## 3　预备知识

### 3.1　top-*N* 推荐方法

本节首先提出了本文涉及的项目推荐问题，并详细描述了本文所提推荐方法中用到的参数符号。本文设 $U$ 为用户集，$V$ 为项目集。$R \subseteq U \times V$ 表示评分矩阵，且如果项目 $v_j$ 与用户 $u_i$ 相关，则 $r_{ij}=1$，否则为 0。用户集和项目集的长度可分别用 $m$ 和 $n$ 表示。

矩阵 $\boldsymbol{R}$ 的第 $i$ 行可用 $r_i^{\mathrm{T}}$ 表示，代表用户 $u_i$ 对所有项目的喜好信息，用 $j^{\mathrm{th}}$ 表示所有用户对项目 $v_j$ 的喜好信息。并用 $l \times |V|$ 得到的矩阵 $\boldsymbol{M}_k$ 表示所有项目的描述信息。其中，$l$ 表示第 $k$ 个模态的特征数量（如在文本模态中使用的单词数量）。矩阵 $\boldsymbol{M}_k$ 的第 $j$ 列表示项目 $v_j$ 的描述性信息特征向量，可用 $\boldsymbol{m}_j^k$ 表示。表 1 为本文使用的主要参数符号及定义。

**表 1　本文使用的参数符号与定义**

| 参数符号 | 参数定义 |
|---|---|
| $\boldsymbol{M}_k$ | 在推荐框架中使用的第 $k$ 种项目模态，即本文中的文本模态和视觉模态 |
| $\boldsymbol{m}_j^k$ | $j$ 项目在 $\boldsymbol{M}_k$ 模态下的表现形式 |
| $U,V$ | 分别表示框架中的用户集和项目集 |
| $u,v$ | 分别表示框架中的任意用户和项目 |
| $m,n$ | 分别表示数据集中的用户数量和项目数量 |
| $\boldsymbol{R}$ | 表示所有用户偏好/项目偏好对的矩阵 |
| $r_{u,v}$ | 用户 $u$ 给予项目 $v$ 的评分 |
| $\lambda,\beta$ | 正则化系数 |
| $\alpha$ | 用于协同和模态信息的平衡参数 |
| $\boldsymbol{B}$ | 项目相似矩阵 |
| $\boldsymbol{S}$ | 聚集系数矩阵 |
| $\boldsymbol{F}$ | 融合 $k$ 个原始项目模态后的项目表示矩阵 |
| $L(*)$ | 自编码器的损耗函数 |
| $\phi(*)$ | 自编码器的激活函数 |
| $W$ | 用于自编码器神经网络结构的权重 |
| $b$ | 用于自编码器神经网络结构的偏权向量 |

目标问题包括根据用户 $i$ 的偏好（根据矩阵 $\boldsymbol{R}$ 推断）和从项目中提取的描述性的边信息确定最符合用户 $i$ 兴趣的前 $N$ 个项目。

**定义 1** （top-$N$ 推荐）给定矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{S}$，并且目标用户 $u_i \in U$，在项目集 $V$ 中为用户 $u_i$ 生成符合其偏好的项目推荐列表。

如 Cremonesi 等[33]所讨论的，本文应考虑评分预测问题，包括预测用户 $u$ 给项目 $i$ 的评分。由此，可基于预测的评分对项目进行排序，以最终生成推荐列表。正如 Cremonesi 等[33]所指出的，top-$N$ 推荐是模拟真实推荐系统的最佳建模方法。故在本文框架中，将为用户生成 top-$N$ 个最相关项目的推荐列表。

### 3.2 稀疏线性方法

本文所提方法是由一类 top-$N$ 推荐问题的算法所驱动的，即 SLIM（sparse linear methods，稀疏线性方法）[5]。在 SLIM 算法中，用户 $u_i$ 对给定项目 $v_j$ 的推荐分数由 $\hat{r}_{ij}$ 表示。此分数可由用户 $u_i$ 评级的项目进行稀疏聚合，具体如式（1）所示。

$$\hat{r}_{ij} = \boldsymbol{r}_i^{\mathrm{T}} \boldsymbol{s}_j \tag{1}$$

如当 $r_{ij} = 0$ 时，而 $\boldsymbol{s}_j$ 为聚集系数大小为 $n$ 的稀疏向量。则该模态可以表示为如式（2）所示。

$$R \sim RS \tag{2}$$

如 $S$ 是 $n \times n$ 的稀疏矩阵的聚集系数，结合式（1）可将其中的第 $j$ 列表示为 $\boldsymbol{s}_j$，而基于式（2）对每一行 $\boldsymbol{r}_i^{\mathrm{T}}$ 表示用户 $u_i$ 对所有项目的推荐得分。目标用户 $u_i$ 的 top-$N$ 推荐列表是根据式（1）中对项目的评分按降序排列计算得到，从而生成 top-$N$ 项目。稀疏线性方法将式（2）中的 $n \times n$ 矩阵 $S$ 作为如式（3）所示正则化优化问题的极小值。

$$\underset{S}{\text{minimize}} \frac{1}{2} \| R - RS \|_{\mathrm{F}}^2 + \frac{\beta}{2} \| S \|_{\mathrm{F}}^2 + \lambda \| S \|_1, \quad \text{subject to } S \geqslant 0 \text{ 且 } \operatorname{diag}(S) = 0 \tag{3}$$

其中，$\| S \|_1$ 表示 $S$ 的乘积范数，$\| \cdot \|_{\mathrm{F}}$ 表示矩阵范数，$\beta$ 和 $\lambda$ 表示正则化参数，这些约束避免平凡解（$\boldsymbol{S}$ 表示单位矩阵），也确保了不使用 $r_{ij}$ 计算 $\hat{r}_{ij}$。SLIM 仅使用评分矩阵 $\boldsymbol{R}$ 来计算聚集系数矩阵 $\boldsymbol{S}$ 的通用架构，而 SSLIM 系列算法加入了描述性信息以改进 SLIM 推荐的质量。故除矩阵 $R$ 外，SSLIM 在学习 $\boldsymbol{S}$ 时也将描述性项目矩阵 $\boldsymbol{M}$ 纳入考虑范围。而其目标主要是使矩阵 $\boldsymbol{S}$ 满足以下两种关系：（1）$\hat{R} \sim RS$，（2）$\hat{B} \sim BS$。具体如式（4）所示。

$$\underset{S}{\text{minimize}} \| R - RS \|_{\mathrm{F}}^2 + \frac{a}{2} \| B - BS \|_{\mathrm{F}}^2 + \frac{\beta}{2} S \|_{\mathrm{F}}^2 + \lambda S \|_1, \text{subject to } S \geqslant 0 \text{ 且 } \operatorname{diag}(S) = 0 \tag{4}$$

其中，参数 $\alpha$ 平衡了协同过滤（矩阵 $\boldsymbol{R}$）和边信息（矩阵 $\boldsymbol{B}$）的使用，$\beta$ 和 $\lambda$ 表示正则化参数。由于矩阵 $\boldsymbol{S}$ 的各列是独立的，故两种算法的优化问题都可以解耦为一组优化问题，具体如式（5）所示。

$$\underset{S}{\text{minimize}} \| r_i - Rs_i \|_2^2 + \frac{\alpha}{2} \| b_i - Bs_i \|_2^2 + \frac{\beta}{2} \| s_i \|_2 + \lambda \| s_i \|_1, \text{subject to } S \geqslant 0 \text{且} \operatorname{diag}(S) = 0 \tag{5}$$

值得注意的是，SLIM 只能用于处理单模态数据，不能直接用于多模态信息处理。故在本文中，对 SLIM 进行了改善，引入了一种新的推荐算法，使其能够同时对多种模态数据进行处理。

### 3.3 自编码器

自编码器作为无监督全连接的隐层神经网络，其目的是对未标记的数据集进行学习[11]。本文希望在训练中将输入数据复制到输出时，得到的隐藏层能够呈现有用的属性。本文将隐藏层中的表示同来自不同模态的信息进行融合，以得到一个新的项目表示，用于构建代表用户偏好的推荐模型。自编码通常是由输入层、隐藏层和重构层组成的前馈神经网络进行搭建，隐藏层将目标值设置为与输入层相同。通用自编码器框架如图 1 所示。

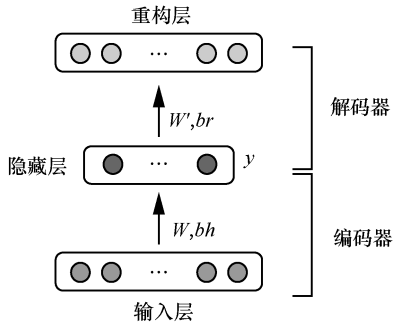



图 1　通用自编码器框架

故自编码器网络由两部分组成，即编码器函数和解码器函数。编码器函数为一个定性映射函数 $h=h(v)$，其可将输入 $v\in R^{d_v}$ 转换为隐藏表示 $y\in R^{d_h}$，具体如式（6）所示。

$$y=h(v)=\phi_h(Wx+b_h) \tag{6}$$

而解码器则是将得到的表示 $y$ 映射回原始输入重构层。故存在另一个映射函数 $g$ 定义为如式（7）所示。

$$r=g(y)=\phi_g(W'h+b_r) \tag{7}$$

其中，$\phi_h$ 和 $\phi_g$ 分别表示编码器和解码器的激活功能。而网络参数 $W\in R^{d_h}\times R^{d_v}$ 和 $W'\in R^{d_v}\times R^{d_h}$ 分别表示编码器和解码器加权矩阵。而编码器和解码器的偏差向量分别为 $b_h\in R^{d_h}$ 和 $b_r\in R^{d_v}$。

如前所述，如自编码器只能简单地将 $g(h(v))=v$ 映射到所有实例中。本文对自编码器进行了设计，使其不能完美地进行复制。接下来本文将详细介绍用来融合多模态项目表示的 3 种自编码器：欠完备自编码器、稀疏自编码器和去噪自编码器。

3.3.1　欠完备自编码器

欠完备自编码器试图通过使隐藏层的维度小于输入层的维度来找到有用的数据表示。故其目的是迫使自编码器学习一个欠完备的表示，在此过程中捕获训练数据最显著的特征。这一学习过程可以简单描述为最小化损失函数如式（8）所示。

$$L(\boldsymbol{v},g(h(\boldsymbol{v}))) \tag{8}$$

其中，$L(*)$ 为一个损失函数，惩罚 $g(h(\boldsymbol{v}))$ 与 $V$ 的差异。使用的是典型的损失函数均方误差。而欠完备自编码器可能存在的限制为：如果编码器和解码器函数是非线性的，可能无法从隐藏层获得有效学习。与限制编码器和解码器函数的复杂性不同，本文选择使用损失函数来对自编码器进行正则化，以获得所需的其他特性。在这种情况下，其特性包括表示的稀疏性以及对噪声或丢失输入的鲁棒性。

3.3.2　稀疏自编码器

训练稀疏自编码器来执行带有稀疏惩罚的复制任务，最终得到学习有用特征的模态。通过这种方式，项目表示通过正则化后变得稀疏，由此自编码器能够找到数据集的统计特征，不再是一个简单的标识函数。除重构误差外，还可以通过在隐藏层 $y$ 上添加稀疏性惩罚 $\Omega(y)$ 来实现稀疏性。本文将新构建的损失函数定义为：

$$L(\boldsymbol{v},g(h(\boldsymbol{v})))+\Omega(y) \tag{9}$$

其中，$g(y)$ 表示解码器输出，而 $y=h(v)$ 表示编码器输出。惩罚 $\Omega(y)$ 作为一个正则化项目添加到前馈网络中，而前馈网络的原始任务是将输入复制到输出。通常稀疏自编码器用于学习其他任务的特征，如分类等。在本文中使用它们通过融合各种模态信息来学习项目的融合表示。

3.3.3　去噪自编码器

在去噪自编码器中，本文提出了一种不同于增加惩罚系数 $\Omega$ 的方法，即以自编码器接收数据点的损坏版本作为输入，并通过训练预测原始版本，即以未损坏的数据点作为其输出。故去噪自编码器可以最小化为损耗函数如式（10）所示。

$$L(\boldsymbol{v},g(h(\tilde{\boldsymbol{v}}))) \tag{10}$$

其中，$\tilde{\boldsymbol{v}}$ 为被某种形式的噪声破坏的 $\boldsymbol{v}$ 的副本。故这类自动编码器用于还原损坏数据，而不是简单地复制它们的输入。

### 3.4 卷积神经网络

卷积神经网络（convolutional neural network，CNN）为一种特殊的神经网络，其目的是在网格状拓扑结构中处理数据[34]。卷积神经网络已成功地应用于图像[35]、图形[36]和时间序列[37]数据处理中。卷积网络是在其至少一层中使用卷积的神经网络。故其训练方法包括通过卷积滤波器层反向传播、恢复和池化等其他操作。卷积神经网络的关键数学运算称为卷积运算，其为一种专门学习数据局部平稳属性的线性运算。

## 4 模型构建

本节描述了本文所提推荐方法的基本架构，以基于项目的多种模态和用户偏好（即评分）向用户进行项目推荐。本模型将项目的不同多模态高维模态移动到一个低维潜空间中，接收不同模态的输入项目（在本文中项目为视频）。本文假定项目会随少数解释变量而共同变化，故不能直接被观察到，本文将这些解释变量称作隐性因素[38]。图 2 表示将多维数据映射到公共潜空间的思想。其中，图 2 左边为嵌入文本和图像的形式从其来源到一个潜在空间，而图 2 右边为项目和用户映射到公共空间的示例。

本文提出用降维方法根据特定的准则从高维数据中发现和提取这些潜在因素。通过将这些项目的表示映射到一个低维潜在空间中，并自动降低了模态的稀疏性。然后利用这种新的低维项目表示来计算项目对间的相似度，最终提高推荐质量。

图 3 表示本文所提推荐方法的框架。首先，本文假设使用的数据集包含以下内容：用户偏好历史记录，即用户对视频进行的评分记录；项目集，即 $k$ 种不同的模态表示形式。然后将该数据集分成两个数据集，即训练集和测试集。当处于训练模式时，本文所提推荐方法构建了一个推荐模型，该模型表示用户的偏好模式；而在测试模式中，使用先前训练过的推荐模型向用户推荐不在训练集中的项目。

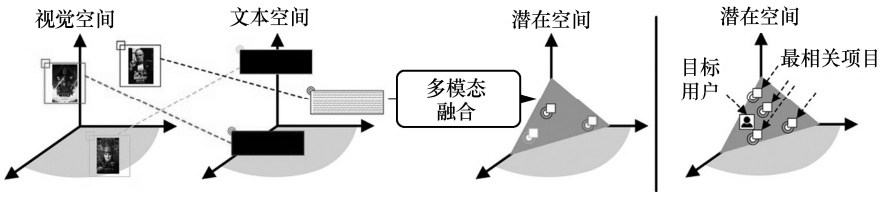



图 2　多模态数据融合在推荐系统中的应用实例

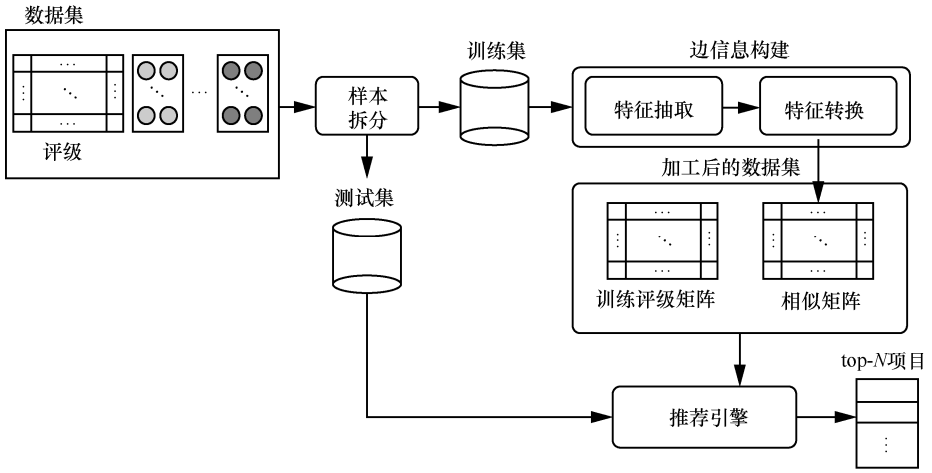



图 3　多模式数据融合的推荐框架

其框架中包含两个模块：边信息构建和推荐引擎。边信息构建通过处理项目的视觉和文本特征并计算其间的相似性来创建边信息矩阵。推荐引擎为用户生成了一个 top-$N$ 推荐列表。

### 4.1　边信息构建模块

边信息构建模块包含与项目表示相关的两个重要任务。首先，本文详细说明如何根据原始特征来表示项目。故本文展示了如何组合不同模态的信息，各个模态都以原始特征表示，从而为数据集中的所有项目提供一个新的表示形式。最后，该模块输出的是边信息矩阵，正如在第 3.2 节提到的，稀疏线性推荐模态引入了项目间的相似性矩阵（具体如第 4.2 节所述），提高了推荐的质量。

#### 4.1.1　特征提取

本文考虑了两种模态来对项目进行表示，即文本表示和视觉表示。本文的重点是视频推荐，故将视频的标题和情节看作文本信息，并将视频帧看作视觉信息。

首先对文本描述建模进行介绍。关于文本表示，在提取视频文本数据（即文本和情节）后，使用词袋模型（bag of words representation，BoW）[39]表示该文本。为实现这一点，定义 $\boldsymbol{m}_j^1=[w_{j1},w_{j2},w_{j3},\cdots,w_{jn}]$ 为视频 $j$ 的文本表示方法，其中，$w_{jt}$ 表示单词 $t$ 在视频 $j$ 中的权重。为计算权重 $w_{jt}$，首先应计算视频 $j$ 中每个单词 $t$ 中的搜索词频率 $\mathrm{TF}_{jt}$ 和逆文档频率 $\mathrm{IDF}_t$，计算方法如式（11）和式（12）所示。

$$\mathrm{TF}_{jt}=1+\mathrm{lb}\left(f_{jt}\right) \tag{11}$$

$$\mathrm{IDF}_t=\mathrm{lb}\left(\frac{n}{n_t}\right) \tag{12}$$

其中，$n$ 表示训练集中的视频数量，$n_t$ 表示出现单词 $t$ 的视频数量，$f_{jt}$ 是单词 $t$ 在视频 $j$ 中的出现频率。最后，根据式（13）计算权重 $w_{jt}$。

$$w_{jt}=\mathrm{TF}_{jt}\times\mathrm{IDF}_t \tag{13}$$

其次，对视觉帧建模进行介绍。关于视觉特征，本文首先提取给定视频 $q$ 的帧集。由于每个视频需要用相同数量的特征进行表示，故考虑集合中最小的视频，以均匀的采样间隔从 $q$ 中选择 $p$ 帧。以使有一个视频的代表性样本，同时减少此特征提取过程的计算成本。

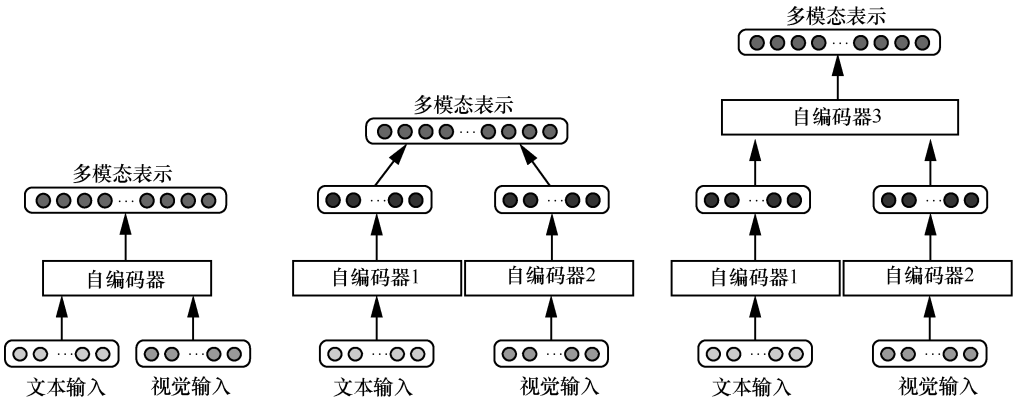
现在每个视频都具有 $p$ 帧，即 $\{f_1,f_2,\cdots,f_p\}$，然后本文使用深度神经网络来学习每个视频中帧的表示。正如第 3.4 节讨论的，本文选择使用卷积神经网络是因为其代表了在几种情况下图像表示和处理的最前沿技术。与参考文献[40]的做法相同，本文使用 Tensorflow 深度学习框架进行帧特征计算，该框架通过 ImageNet 数据集中的 1 200 000 幅图像和 1 000 个类别进行预训练[41]。然后，将向量串联起来，以得到表示视频 $v_j$ 的视觉特征向量 $\boldsymbol{m}_j^2=[\boldsymbol{v}_j^1\boldsymbol{v}_j^2,\cdots,\boldsymbol{v}_j^p]$。

#### 4.1.2　特征转换

在这个模块中，通过融合多个模态来执行特征转换，为项目构建一个新的表示形式。在执行该融合后，每个视频被表示为矩阵 $\boldsymbol{F}$ 中的行 $f_i$。基于自编码器与多模态数据融合的推荐框架如图 4 所示，本文提出了 3 种基于自编码器的体系结构来进行这种新项目表示的学习。自编码器由于其简单性和高效性，能够在各种条件下进行特征学习，本文在多模态数据融合的推荐推荐方法中利用了这种特性。

关于本文提出的 3 个融合多模态数据的体系结构，其原理是利用模态间和模态内的语义相关性。在 3 个体系结构中，都计算了一个潜在的项目表示，这些项目在一个公共空间中组合为不同的模态。

共享单层自编码器（shared single-layered autoencoder，S-SLAE）架构指直接地使用自编码器进行新数据表示学习的架构。本文将文本和视觉模态连接起来，作为特定自编码器的输入。并将该自编码器的隐藏层作为输入项的新表示，以




(a) 共享单层自编码架构　(b) 独立单层自编码器架构　(c) 双层自编码器架构

图 4　基于自编码器与多模态数据融合的推荐框架

保持模态间的语义关系。最后得到了一个自编码器，在一个单一层次的体系结构中，可以在模态间共享。

独立单层自编码器（independent single-layer autoencoder，I-SLAE）架构指使用自编码器来探索数据中的模态内语义关系。具体地说，每一个模态都是用不同的编码器来处理的，即对每一个模态都有一个不同的特征学习过程。将从每个自编码器学习到的新表示进行连接，以构建新的项目表示。此处的每个模态都有独立的自编码器，且仍然在体系结构中保留一个单层。

双层自编码器（two-layered autoencoder，TLAE）架构指通过同时利用模态间和模态内的语义关系来结合前述两种体系的结构。为了实现这一目的，本文提出了一个双层体系结构，以不同的方式进行信息融合。在底层，和独立单层自编码体系一样，使用不同的自编码器来学习模态间的语义关系；在顶层，先将学习到的表示连接起来，并使用不同的自编码器来探索模态间的语义关系。

值得注意的是，在独立单层自编码器和双层自编码器中，能够并行化模态内学习的步骤，这在处理大规模数据集时非常重要。最后，本文所提自编码器中的所有训练都为无监督的，而在项目无标签的情况下，这是本文的一个重要优势。

### 4.2　推荐模块

本文结合前文中讨论的概念对推荐模块进行详细介绍。首先，定义 SLIM 的聚集系数矩阵 $\boldsymbol{S}$ 表示项目特征空间的函数。故可直接在矩阵 $\boldsymbol{B}_{ij}$（也称为边信息矩阵）中展现项目间的相似性。这种方法的优点是能够捕获项目间基于特征的关系。值得关注的是，本文所提方法不像 SLIM 那样只考虑一种模态，而是能够考虑任意数量的项目模态，这是本文的重要优势之一，这在数据稀疏和冷启动的情况下尤为明显。

首先，本文利用矩阵 $\boldsymbol{F}$ 给出的项目表示形式，构造项目间的相似性矩阵 $\boldsymbol{B}$。如前所述，矩阵 $\boldsymbol{F}$ 通过使用自编码器融合模态，拥有项目的新表示形式。值得关注的是，本文使用新的项目表示形式可以避免原始项描述的稀疏性（如用于文本模态的词袋等）。最初提出稀疏线性模型时，使用一次范数来控制模型复杂度和避免过拟合。本文还使用此范数，通过在非稀疏维度中表示项目来强调稀疏性问题。

本文所提推荐方法与 SLIM 的一个关键区别在于如何学习矩阵 $\boldsymbol{S}$。算法 1 详细描述了本文学习矩阵 $\boldsymbol{S}$ 的算法，然后在测试模式下使用该矩阵计算推荐值。根据式（1）计算测试推荐。故在测试模式下，不会显示用户已看到的项目，即在训

练模式中用于学习用户偏好的项目。本文所提算法 1 接收与 SLIM 相同的矩阵作为输入，即评级矩阵 $\boldsymbol{R}$ 和描述性项目矩阵 $\boldsymbol{F}$。$f_i$ 为矩阵 $\boldsymbol{F}$ 的行，指的是物品 $i$ 的融合表示。可将新表示形式看作通过将项目特征映射到新的空间中来探索模态间的语义关系。

**算法 1**　本文所提多模态数据融合推荐方法

**Require**　Matrices R and F, and item similarity function g

（1）**for** $i=1,\cdots,n$ **do** ▷ Computing items similarities

（2）**for** $j=1,\cdots,n$ **do**

（3）$B_{ij} \leftarrow g(f_i, f_j)$

（4）end for

（5）end for

（6）**for** $j=1,\cdots,n$ **do** ▷ Computing the aggregation

coefficient matrix

（7）$S \leftarrow \underset{S}{\operatorname{argmin}} \| |r_i - Rs_i|_2^2 + \frac{\alpha}{2} b_i - Bs_{i2}^2 + \frac{\beta}{2} s_{i2} + \lambda s_{i1}$

（8）subject to $S \geqslant 0$ and $\operatorname{diag}(S) = 0$

（9）end for

本文所提推荐方法是灵活的，这是由于可设置输入为：（1）通过矩阵 $\boldsymbol{F}$ 融合的任何模态；（2）项目间的相似度函数 $g$（此处使用余弦函数进行相似度计算）。该算法的工作分两个步骤：第一步，根据给定的相似度函数 $g$ 识别项目对间的相似性（第 1~5 行）。第二步，计算聚集系数矩阵 $\boldsymbol{S}$，并在此遵循 SLIM 的优化方案，运用坐标下降法来求得矩阵 $\boldsymbol{S}$（第 6~9 行）。

## 5　实验结果与分析

### 5.1　数据说明

本文实验中使用了来自电影推荐领域常用的 3 个真实数据集。其中包括两个版本的 MovieLens 数据集（关于电影评分的数据集）：MovieLens-1M 和 MovieLens-10M。除此之外，本文团队利用 Python 工具分别利用 API 爬取豆瓣的真实用户评分数据集作为实验的基础数据，爬取时间为 2019 年 8 月 19 日至 2020 年 1 月 24 日。表 2 为经过数据预处理后的数据集统计结果。MovieLens-1M 数据集包括来自 3 582 个视频的 6 383 个用户的 1 023 839 条评分数据。MovieLens-10M 数据集包含来自 8 923 个视频的 75 124 个用户的 11 293 834 条评级数据。豆瓣数据集包含 2 694 个用户对 2 445 个视频的 601 543 条评级数据。在上述数据集中，为创建用户—项目矩阵，当用户对该视频的评分至少为 4 星及 4 星以上时，则认为该视频与该用户相关，并从数据库中提取相关电影的预告片进行分析，其中豆瓣数据集中的视频长度较短。

### 5.2　评估方法

本文在 top-$N$ 推荐任务中对 $N \in [1,10]$ 的所有推荐策略进行了评估。本文使用五重交叉验证程序来评估所提架构的性能。使用 Rapidminer 数据挖掘工具随机选取各用户评级数据的 20%作为测试集，并将剩下的 80%用户数据作为训练集。为了验证本文所提推荐方法的性能，使用威尔科克森符号秩检验（Wilcoxon signed-rank test）来评估

**表 2　实验使用数据集统计**

| 数据集名称 | 视频数量 | 用户数量 | 评级数量 | 帧数/视频 | 字数/视频 |
|---|---|---|---|---|---|
| MovieLens-1M | 3 582 | 6 383 | 1 023 839 | 2 943 | 24 |
| MovieLens-10M | 8 923 | 75 124 | 11 293 834 | 2 856 | 21 |
| 豆瓣 | 2 445 | 2 694 | 601 543 | 1 933 | 17 |

推荐结果的统计学意义，这是由于其为一种非参数统计假设检验程序，不需要预先知道样本的分布情况。

本文参照参考文献[42]的做法，选取 top-$N$ 的归一化折现累积增益（NDCG@$N$）作为判断标准，其常用于衡量推荐与理想排名间的接近程度，具体计算方法如式（14）所示。在此得出的为 100 次迭代后的平均度量结果。

$$\mathrm{NDCG@}N=\frac{1}{S}\sum_{i=1}^{N}\left(\frac{r_i}{\mathrm{lb}(i+1)}\right) \tag{14}$$

其中，$r_i$ 表示真值，当用户推荐的项目位于排名 $i$ 时 $r_i=1$，否则 $r_i=0$。$S$ 表示归一化因子，等于折现累积增益，即式（14）中的理想排序之和。

本文对第 3.3 节中介绍的 3 种自编码器（欠完备自编码器、稀疏自编码器和去噪自编码器）进行了实例化。还测试了原始特征的级联，本文称之为非自编码特征融合（non-autoencoder feature fusion，NOAE）。最后，为从视频帧中进行原始视觉特征的提取，本文使用了现今最先进的卷积网络，即 AOP 框架下的 interception-V3 算法[40]。为实现这一点，本文使用倒数第二个全连接层的输出表示视频的单个帧，其维度为 2 048。

### 5.3 基准方法

为比较本文所提基准方法与经典推荐方法和现今前沿推荐方法相比的优劣，选取下列推荐方法作为基准进行比较。

（1）贝叶斯个性化排名矩阵分解推荐方法（the Bayesian personalized ranking matrix factorization recommendation approach，BPRMF）是目前最先进的基于排名的 top-$N$ 推荐方法之一[5]。

（2）加权正则化矩阵分解（weighted regularized matrix factorization，WRMF）将用户评级作为二进制值，并考虑观察到的评级和未观察到的评级的不同机密值，从而对该模型进行拟合[2]。

（3）稀疏线性方法（sparse linear method，SLIM）使用一个稀疏线性模型来进行项目推荐，通过其他项目的集合来计算新项目的评分[9]。

（4）协同主题回归（collaborative topic regression，CTR）模型是一种同时进行主题建模和协同过滤的生成式推荐模型。该方法侧重于在学习语义主题的过程中从项目描述中挖掘文本信息[6]。

（5）协同变分自编码器推荐方法（collaborative variational autoencoder recommendation approach，CVAE）是一种将变分自编码器集成到概率矩阵分解中的层次贝叶斯模型，用其学习概率潜在变量来表示项目内容信息[41]。

（6）联合表示学习（joint representation learning，JRL）是一种最新的推荐方法，将多个证据源结合起来，以产生与产品推荐有关的前 $N$ 个项目。

本文为每种模态的数据来源（图像、评论文本和评分等）创建了深层表示。为对文本进行表示，使用了 Le 等[42]提出的方法，并使用 Jia 等[43]提出的方法对图形进行表示，即由在包含 1 200 000 张 ImageNet 的图像数据集中进行预先训练的 Caffe 深度学习网络进行表示。本文对用户 $u$ 对所有物品的评分及所有用户对项目 $i$ 的评分进行了深度学习表示。为了结合这些数据，增加了一个新的层，然后结合成对学习排序方法生成一个 top-$N$ 推荐列表。

### 5.4 实验结果

#### 5.4.1 总体绩效

为比较本文所提推荐方法与经典、先进基准推荐方法，本文参考文献[44-45]的做法，使用 NDCG@$N$ 值对测试数据集中的不同推荐方法进行分析。将本文所提推荐方法表示为 AE-MDF（autoencoder-multimodal data fusion），且在实验中均使用其最佳实例化结果（$N\in[1,10]$）。图 5 表示测试数据集中不同方法的 NDCG@$N$ 值。

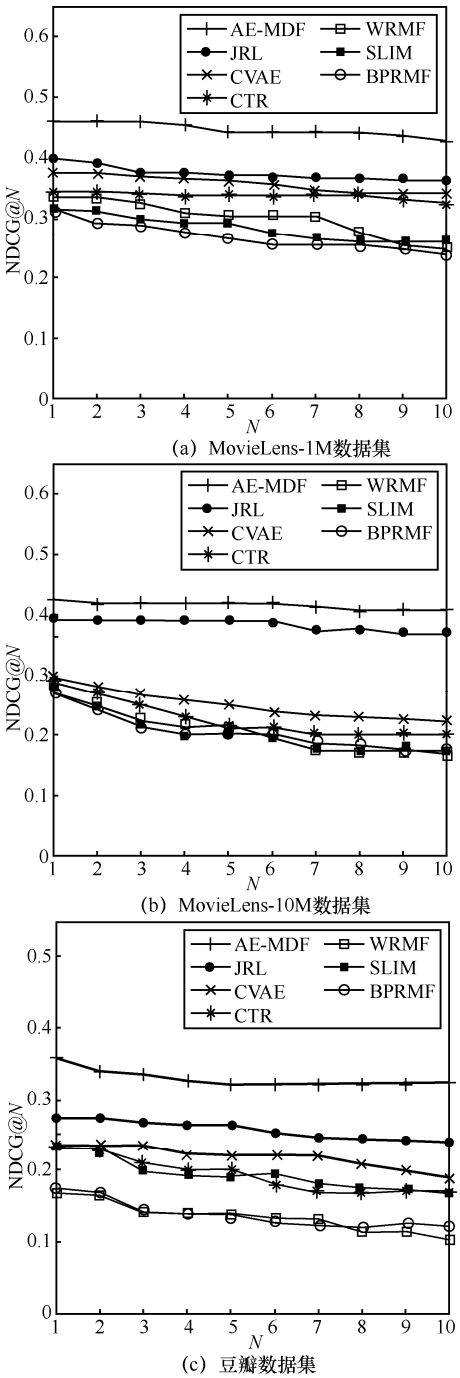



图 5　测试数据集中不同推荐方法的 NDCG@$N$ 值

由图 5 可知，本文所提 AE-MDF 推荐方法较其他推荐方法具有更好的性能。JRL 方法也具有较好的表现，而 BPRMF 方法性能较差，这是由于其只考虑了用户评分，而忽略了项目描述内容。本文所提 AE-MDF 推荐方法与基准方法 JRL 相比，性能在不同的数据集中提升了 7%~32% 不等。尤其是在最大的数据集 MovieLens-10M 数据集中，JRL 方法与本文所提 AE-MDF 方法性能最为接近。而图 6 的不同数据集中的视频推荐算法精度也证明了本文所提视频推荐方法的优越性。

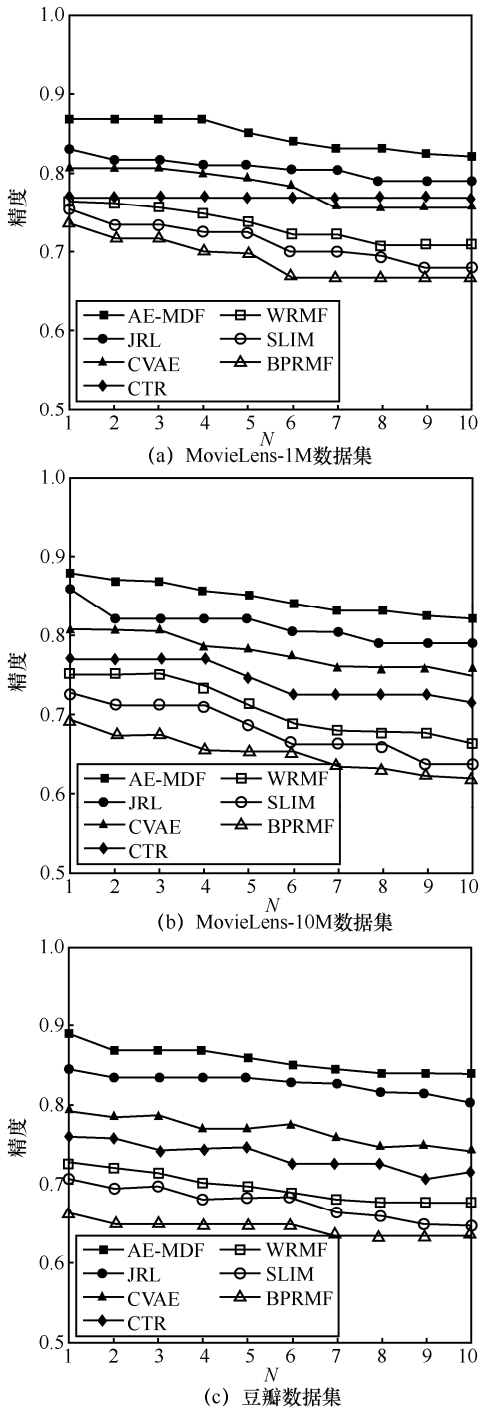



图 6　测试数据集中不同推荐方法的精度值

图 7 表示使用不同模态表示（文本、视觉、文本+视觉）时，本文所提 AE-MDF 方法在不同数据集中的 NDCG@$N$ 值($N \in [1,10]$)。图 7 中，当使用多模态数据时，本文所提 AE-MDF 方法比使用单一模态（文本或视觉）具有更好的性能（较基于视觉进行推荐性能提升了 14.4%~60.6%）。

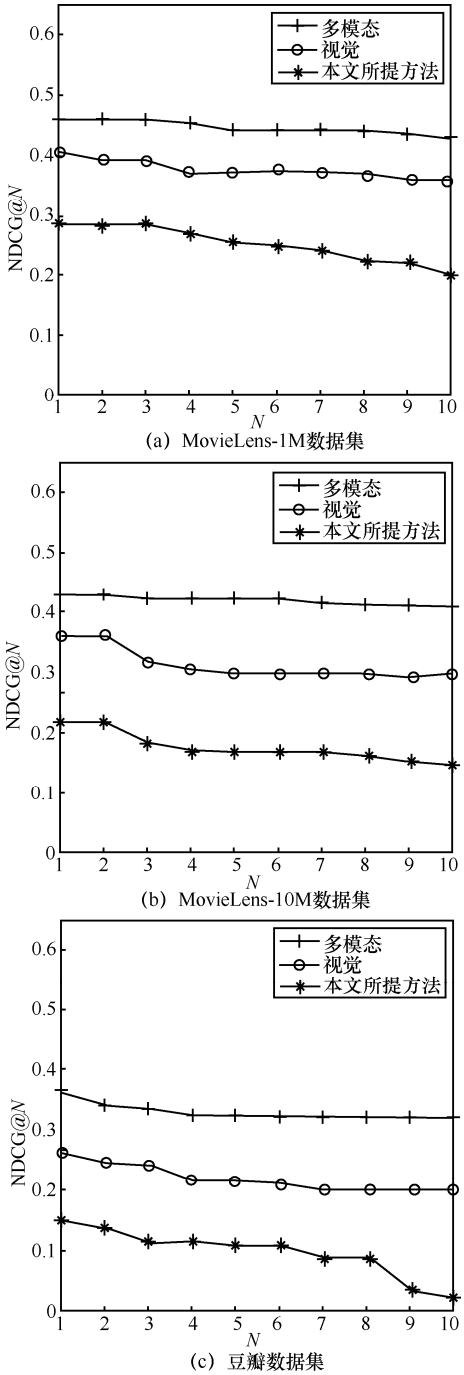



图 7　使用不同模态数据实例化的情况下的 NDCG@$N$ 值

由图 7 还可看出，在所有数据集中基于视觉推荐的表现要比基于文本的推荐效果更好。另外，基于文本推荐可能有助于改进推荐结果，这是由于多模态协同推荐比单独使用两种模式效果更好。这一点在豆瓣数据集中表现得尤为明显。图 8 报告了不同数据集中使用不同模态数据实例化的情况下的 Precision 值，由结果可知本文所提多模态融合的视频推荐方法相对具有更优的精度。

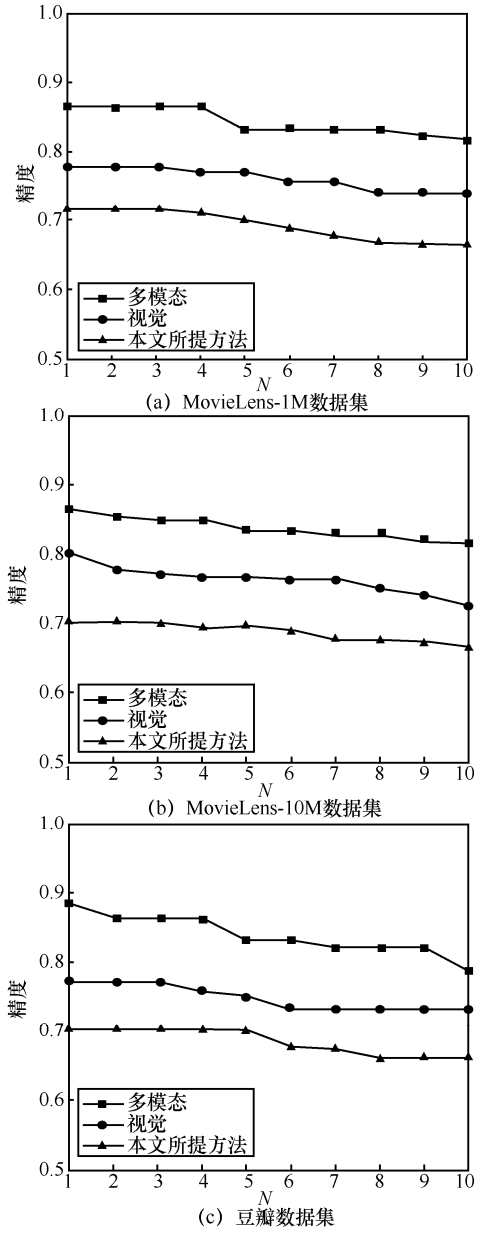



图 8　使用不同模态数据实例化的情况下的精度值

表 3　本文所提推荐方法在使用不同自编码器时的 NDCG@10 值

| 融合策略 | MovieLens-1M 数据集 | | | MovieLens-10M 数据集 | | | 豆瓣数据集 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 文本 | 视觉 | 多模态 | 文本 | 视觉 | 多模态 | 文本 | 视觉 | 多模态 |
| NOAE | 0.153 | 0.224 | 0.228 | 0.124 | 0.238 | 0.237 | 0.001 | 0.231 | 0.239 |
| UAE | 0.182 | 0.352 | 0.369 | 0.139 | **0.381** | **0.326** | 0.001 | 0.192 | 0.211 |
| DAE | 0.104 | 0.279 | 0.342 | 0.127 | 0.273 | 0.273 | 0.001 | 0.179 | 0.228 |
| SAE | **0.215** | **0.376** | **0.413** | **0.152** | 0.285 | 0.319 | **0.024** | **0.257** | **0.362** |

注：加粗项为每列最佳项。

### 5.4.2　自编码器类型

为呈现不同类型的自编码器对本文所提 AE-MDF 推荐方法的影响，在最简单的共享单层自编码架构下进行实验。表 3 列出了本文所提模型使用不同自编码器时（欠完备自编码器（UAE）、去噪自编码器（DAE）和稀疏自编码器（SAE）），在不同数据集中的 NDCG@10 值。本文还在不使用自编码器（NOAE）的情况下提供了不同模态数据的 NDCG@10 值。

从表 3 结果可知，对单独模态而言，使用视觉信息时的性能要优于使用文本描述时的性能。另外，在使用多模态数据时，文本信息与视觉信息相结合效果要优于使用任何单独模态的性能。在 MovieLens-1M 和豆瓣数据集中，稀疏自编码器在所有模态上都取得了比其他方法更好的表现；而对于 MovieLens-10M 数据集，虽然稀疏自编码器表现较好，但欠完备自编码器在视觉模态和多模态表示方面表现优于任何其他自编码器。

### 5.4.3　数据融合结构

本文对使用不同组合的自编码器对多模态数据金融融合的方式进行实验。如第 4.1.2 节所述，本文提出了 3 种基于自编码器的不同结构对多模态数据进行融合。为实现这一点，本文使用了稀疏自动编码器，这是由于其在第 5.4.2 节中的性能效果最好。图 9 表示本文所提 AE-MDF 推荐方法在不同特征转换结构下的性能，同时给出了仅使用多模态数据（即当不使用自动编码器进行不同模态的数据融合）的 NDCG@$N$ 值。

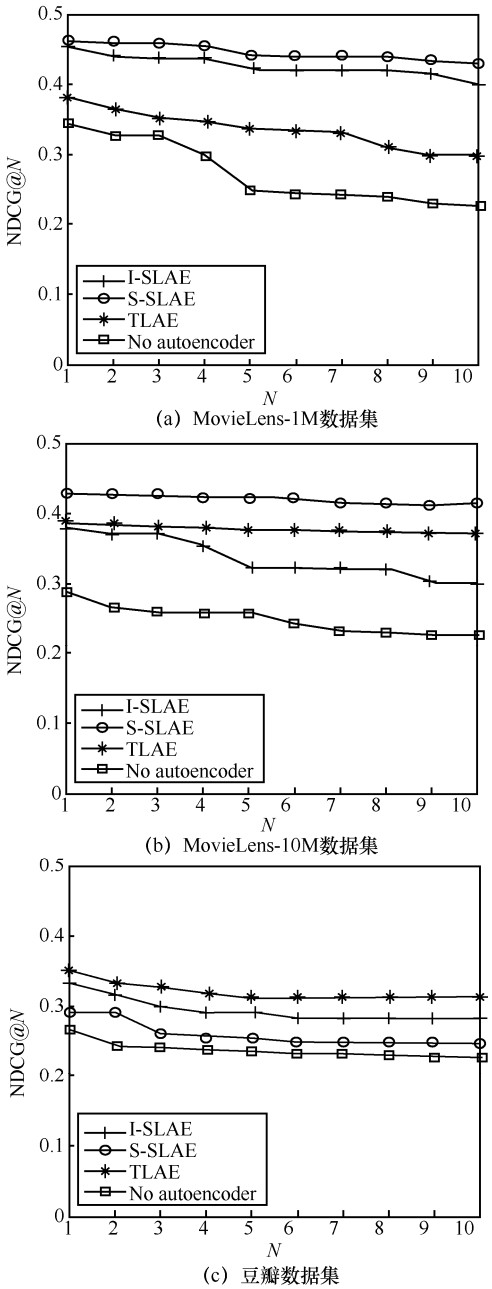



图 9　不同特征转换结构的 NDCG@$N$ 值

当对 MoiveLens-1M 和 MoiveLens-10M 数据集进行处理时，最佳策略是使用独立单层自编码器来挖掘数据中的模态语义关系，这是由于在这两张数据集中有足够多的视频信息可供提取模态信息。当对豆瓣数据集进行处理时，使用双层自编码器架构可获得最优结果。在此情况下，可利用模态内和模态间的语义关系克服该数据集内视频太短的问题。

## 6 结束语

本文提出了一种基于自编码器与多模态数据融合的推荐方法。提出 3 种架构来进行项目的多模态表示（本文使用的模态包括文本和视觉，其中本文使用 Inception 卷积神经网络算法进行视觉模态特征提取），并展示了如何使用多模态数据融合项目表示来提高推荐的质量。值得注意的是，本文提出的视频推荐方法非常灵活，除文本与视觉还可以使用其他类型的模态数据进行融合推荐。最后，在 3 个不同特征的真实数据集进行实验，对本文所提架构进行验证。实验结果表明，本文所提 AE-MDF 推荐方法优于其他基准算法。

尽管本文已提出了上述具有重要意义的发现，但还是具有一些局限性，其中一些可能会为未来的进一步研究指明方向：首先，可对图像的其他原始特征表示进行研究，并分析其对本文所提方法性能的影响。其次，可尝试增加其他模态（如音频等）的数据到模型中，以进一步提升其对视频的推荐性能。最后，可针对其他推荐领域（如社交网络、产品和音乐等）进行研究和算法优化。

## 参考文献：


[1] BOBADILLA J, ORTEGA F, HERNANDO A, et al. Recommender systems survey[J]. Knowledge-based systems, 2013(46): 109-132.

[2] HU Y, KOREN Y, VOLINSKY C. Collaborative filtering for implicit feedback datasets[C]//Proceedings of 2008 Eighth IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2008: 263-272.

[3] LI Z, PENG J Y, GENG G H, et al. Video recommendation based on multi-modal information and multiple kernel[J]. Multimedia Tools and Applications, 2015, 74(13): 4599-4616.

[4] NING X, KARYPIS G. Slim: sparse linear methods for top-n recommender systems[C]//Proceedings of 2011 IEEE 11th International Conference on Data Mining. Piscataway: IEEE Press, 2011: 497-506.

[5] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. Bpr: Bayesian personalized ranking from implicit feedback. UAI'09[J]. Arlington, Virginia, United States, 2009: 452-461.

[6] WANG C, BLEI D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 448-456.

[7] KABBUR S, NING X, KARYPIS G. Fism: factored item similarity models for top-*N* recommender systems[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 659-667.

[8] DESHPANDE M, KARYPIS G. Item-based top-n recommendation algorithms[J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177.

[9] NING X, KARYPIS G. Sparse linear methods with side information for top-*N* recommendations[C]//Proceedings of the Sixth ACM Conference on Recommender Systems. New York: ACM Press, 2012: 155-162.

[10] 任永功, 杨柳, 刘洋. 基于热扩散影响力传播的社交网络个性化推荐算法[J]. 模式识别与人工智能, 2019, 32(8): 746-757.
REN Y G, YANG L, LIU Y. Heat diffusion influence propagation based personalized recommendation algorithm for social network[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(8): 746-757.

[11] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[12] DAVIDSON J, LIEBALD B, LIU J, et al. The YouTube video recommendation system[C]//Proceedings of the Fourth ACM Conference on Recommender Systems. New York: ACM Press, 2010: 293-296.

[13] ZHOU R, KHEMMARAT S, GAO L. The impact of YouTube recommendation system on video views[C]//Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2010: 404-410.

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

[15] AHMED M, IMTIAZ M T, KHAN R. Movie recommendation

system using clustering and pattern recognition network[C]//Proceedings of 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). Piscataway: IEEE Press, 2018: 143-147.

[16] BEEL J, GIPP B, LANGER S, et al. paper recommender systems: a literature survey[J]. International Journal on Digital Libraries, 2016, 17(4): 305-338.

[17] BOBADILLA J, HERNANDO A, ORTEGA F, et al. Collaborative filtering based on significances[J]. Information Sciences, 2012, 185(1): 1-17.

[18] LAO N, COHEN W W. Relational retrieval using a combination of path-constrained random walks[J]. Machine Learning, 2010, 81(1): 53-67.

[19] NASCIMENTO C, LAENDER A H F, DA SILVA A S, et al. A source independent framework for research paper recommendation[C]//Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. New York: ACM Press, 2011: 297-306.

[20] KIM H N, JI A T, HA I, et al. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation[J]. Electronic Commerce Research and Applications, 2010, 9(1): 73-83.

[21] YANG C, CHEN X, LIU L, et al. A hybrid movie recommendation method based on social similarity and item attributes[C]// Proceedings of International Conference on Sensing and Imaging. Heidelberg: Springer, 2018: 275-285.

[22] CHRISTAKOU C, VRETTOS S, STAFYLOPATIS A. A hybrid movie recommender system based on neural networks[J]. International Journal on Artificial Intelligence Tools, 2007, 16(5): 771-792.

[23] BEUTEL A, COVINGTON P, JAIN S, et al. Latent cross: making use of context in recurrent recommender systems[C]//Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2018: 46-54.

[24] 黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述[J]. 计算机学报, 2018, 41(7): 1619-1647.
HUANG L W, JIANG B T, LV S Y, et al. A review of recommendation Systems based on deep learning[J]. Journal of Computer Science, 2018, 41(7): 1619-1647.

[25] COVINGTON P, ADAMS J, SARGIN E. Deep neural networks for youtube recommendations[C]//Proceedings of the 10th ACM Conference on Recommender Systems. New York: ACM Press, 2016: 191-198.

[26] FAN Y, WANG Y, YU H, et al. Movie recommendation based on visual features of trailers[C]//Proceedings of International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. Heidelberg: Springer, 2017: 242-253.

[27] CHENG H T, KOC L, HARMSEN J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. [S.l.:s.n.], 2016: 7-10.

[28] ZHANG Y, AI Q, CHEN X, et al. Joint representation learning for top-*N* recommendation with heterogeneous information sources[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM Press, 2017: 1449-1458.

[29] LI Z, PENG J Y, GENG G H, et al. Video recommendation based on multi-modal information and multiple kernel[J]. Multimedia Tools and Applications, 2015, 74(13): 4599-4616.

[30] 王娜, 何晓明, 刘志强, 等. 一种基于用户播放行为序列的个性化视频推荐策略[J]. 计算机学报, 2020, 43(1): 123-135.
WANG N, HE X M, LIU Z Q, et al. Personalized video recommendation strategy based on user's playback behavior sequence[J]. Journal of Computer Science, 2020, 43(1): 123-135.

[31] 苏赋, 吕沁, 罗仁泽. 基于深度学习的图像分类研究综述[J]. 电信科学, 2019, 35(11): 58-74.
SU F, LV Q, LUO R Z. Review of image classification based on deep learning[J]. Telecommunications Science, 2019, 35(11): 58-74.

[32] FELIPE L A, CONCEIC A L C, Pádua A L A C, et al Multimodal data fusion framework based on autoencoders for top-*N* recommender systems[J]. Applied Intelligence, 2019, 49(9). 3267-3282.

[33] CREMONESI P, KOREN Y, TURRIN R. Performance of recommender algorithms on top-*N* recommendation tasks[C]// Proceedings of the fourth ACM conference on Recommender systems. New York: ACM Press, 2010: 39-46.

[34] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.

[35] NASCIMENTO G, LARANIEIRA C, BRAZ V, et al. A robust indoor scene recognition method based on sparse representation[C]//Proceedings of Iberoamerican Congress on Pattern Recognition. Heidelberg: Springer, 2017: 408-415.

[36] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems.[S.l.:s.n.], 2016: 3844-3852.

[37] YANG J, NGUYEN M N, SAN P P, et al. Deep convolutional neural networks on multichannel time series for human activity recognition[C]//Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence. [S.l.:s.n.],2015: 3995-4001.

[38] CUNNINGHAM J P, BYRON M Y. Dimensionality reduction for large-scale neural recordings[J]. Nature Neuroscience, 2014, 17(11): 1500-1509.

[39] BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval[M]. New York: ACM Press, 1999.

[40] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the

inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2818-2826.

[41] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[42] LE Q, MIKOLOY T. Distributed representations of sentences and documents[C]//Proceedings of International Conference on Machine Learning. [S.l.:s.n.], 2014: 1188-1196.

[43] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 675-678.

[44] 胡春华, 童小芹, 梁伟. 基于信任和不信任关系的实值受限玻尔兹曼机推荐算法[J]. 系统工程理论与实践, 2019, 39(7): 1817-1830.
HU C H, TONG X X, LIANG W. The real-value restricted Boltzmann machine recommendation algorithm based on trust-distrust relationship[J]. Systems Engineering-Theory & Practice, 2019, 39(7): 1817-1830.

[45] 肖云鹏, 孙华超, 戴天骥, 等. 一种基于云模型的社交网络推荐系统评分预测方法[J]. 电子学报, 2018, 46(7): 1762-1767.
XIAO Y P, SUN H C, DAI T J, et al. A rating prediction method based on cloud model in social recommendation system[J]. Acta Electronica Sinica, 2018, 46(7): 1762-1767.

[作者简介]

**顾秋阳**（1995− ），男，浙江工商大学博士生，主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。

**琚春华**（1962− ），男，博士，浙江工商大学教授、博士生导师，主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。

**吴功兴**（1974− ），男，博士，浙江工商大学副教授，主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。