文章编号：2095-2163(2019)04-0045-06 中图分类号：TP391.1 文献标志码：A

基于交叉熵与困惑度的 LDA-SVM 主题研究

薛佳奇[1]，杨 凡[2]
（1 西安建筑科技大学 信息与控制工程学院，西安 710055；2 西安建筑科技大学 理学院，西安 710055）

摘 要：目前对于中文影视剧本的分类主要借助人工经验，具有成本高、效率低等特点。当前没有针对中文影视剧本主题自动分类的相关研究，本文将对主题提取进行研究，传统主题生成模型借助于文档和段落、段落和语句、语句和词的相似性，而忽略了文本语句与语句之间的相似性。首先，采用 ISOMAP 方法降低样本集的向量空间维度；其次，提出交叉熵结合困惑度的算法模型，进而确定 LDA 需要提取的最优主题数目；最后，通过剧本-主题的方式，利用 LDA 算法挖掘剧本的隐含主题词，同时利用 SVM 对主题词做出进一步的分类。

关键词：中文影视剧本；ISOMAP 降维；LDA；交叉熵；困惑度；SVM

Research on LDA-SVM subject based on cross entropy and perplexity

XUE Jiaqi[1], YANG Fan[2]
(1 School of Information and Control Engineering, Xi´an University of Architecture and Technology, Xi´an 710055, China;
2 School of Science, Xi´an University of Architecture and Technology, Xi´an 710055, China)

【Abstract】 At present, the classification of Chinese film and television scripts mainly relies on manual experience, which has the characteristics of high cost and low efficiency. There is currently no research on the automatic classification of Chinese film and television scripts. This paper explores the topic extraction. The traditional topic generation model relies on the similarity of documents and paragraphs, paragraphs and sentences, sentences and words, while ignoring the similarity between text statements and statements.Firstly, the ISOMAP method is used to reduce the vector space dimension of the sample set. Secondly, the algorithm model of cross entropy combined with perplexity is proposed to determine the optimal number of topics that LDA needs to extract. Based on the above, through the script-theme method, the script is used to mine implicit subject terms of the script, while using SVM to further classify the subject words.

【Key words】 Chinese film and television script; ISOMAP dimension reduction; LDA; cross entropy; perplexity; SVM

0 引 言

互联网上文本类型数据数量呈现指数式的激增，则使得当今社会各个方面对互联网数据挖掘方法的需求也越来越大[1-2]。与此同时，人们正更加倾向于随时随地浏览信息和观看影视作品，文学剧本的数量也开始急剧上升，也就必然给影视审核人员带来巨大的挑战，即剧审人员需要快速熟知海量剧本的主题。目前，自动化的剧本主题分析鲜有学者进行相关研究，本文即拟对影视剧本的主题词发现展开探讨与论述。

研究可知，剧本与文本同时存在维数过高的问题，因此需要采取降维方法。常见的降维方法有 PCA 降维和 ISOMAP 降维，其中 PCA 降维存在信息丢失问题，故而本文选用了 ISOMAP 降维方法。而研究中，将通过 LDA 来选取主题词，但考虑到 LDA 的参数 K 多会通过困惑度进行计算，本文则有针对性地提出了困惑度与交叉熵结合度的方法。文中对此可做研究分析如下。

1 主题提取相关研究

选择剧本主题特征词时，应选择能代表剧本类别的词作为特征，而在通过向量来表示剧本时，向量空间稀疏和高特征维数问题就是剧本提取特征词的研究热点。针对这一状况，通常需要进行特征降维，降维不仅能够缩减剧本的特征维数，减小模型训练时的迭代次数，也可以消除相似语义的特征，进而提高剧本主题分类的准确率、召回率和效率。相较于英文剧本，中文剧本有着更多的字词组合、更大的编码空间、更稀疏的原始特征空间，更高的矩阵维度等特点，为了获取高效的剧本特征降维方法，不影响剧本主题的分类性能，就需要选取适合于中文影视剧本的降维方法。这里可得研究内容分述如下。

作者简介：薛佳奇(1993-)，男，硕士研究生，主要研究方向：大数据、人工智能；杨 凡(1995-)，女，硕士研究生，主要研究方向：机器学习、文本分析。

收稿日期：2019-05-18

1.1 PCA 与 ISOMAP 降维

1.1.1 PCA 降维

PCA[3]降维算法是为了去除剧本向量空间中相似的元素,消除维度灾难,从而得到有效的特征空间。PCA 的计算过程详见如下。

设样本数为 n, 特征数为 m, 第 i 维第 k 列的样本值为 $v(i)(k)$, 第 i 维特征的均值是 $\mu(i)$, 第 i 维的标准差为 $\sigma(i)$。

首先,计算每一维特征的均值 $\mu(i)$, 参考计算公式如下:

$$\mu(i)=\frac{\sum_{k=1}^{n} v(i)(k)}{n}, \tag{1}$$

然后,将第 i 维特征的第 k 列的样本值变为 $\overline{v(i)(k)}$, 其计算公式如下:

$$\overline{v(i)(k)}=v(i)(k)-\mu(i), \tag{2}$$

接下来,求取方差归一化,相应数学公式如下:

$$\sigma(i)^2=\frac{1}{n-1}\sum_{k=1}^{n}\overline{v(i)(k)^2}, \tag{3}$$

$$\overline{\overline{v(i)(k)}}=\frac{\overline{v(i)(k)}}{\sigma(i)}, \tag{4}$$

在此基础上,计算协方差矩阵。协方差矩阵的第 h 行第 g 列的维度值的运算将用到如下计算公式:

$$cov(h,g)=E((\overrightarrow{v(h)}-E(\overrightarrow{v(h)}))(\overrightarrow{v(g)}-E(\overrightarrow{v(g)})))=E(\overrightarrow{v(h)}*\overrightarrow{v(g)})-E(\overrightarrow{v(h)})*E(\overrightarrow{v(g)}), \tag{5}$$

过程中,还要求得协方差矩阵的特征值和特征向量。此时,运用 Jacobian method 求解矩阵 $\boldsymbol{\Sigma}$ 的特征值和特征向量,并初始化单位矩阵 $\boldsymbol{P}$。

选择矩阵 $\boldsymbol{\Sigma}$ 中非对角线上绝对值最大的数,记为 $Z(ij)$,求解 θ:$\tan 2\theta=\frac{2Z(ij)}{Z(ii)-Z(jj)}$, 更新矩阵 $\boldsymbol{P}$, 其中, $\boldsymbol{P}(ii)=\cos\theta$,$\boldsymbol{P}(jj)=\cos\theta$,$\boldsymbol{P}(ij)=-\sin\theta$,$\boldsymbol{P}(ji)=\sin\theta$。

将 $\boldsymbol{P}$ 赋值给 $\boldsymbol{P}1$, 即 $\boldsymbol{P}1=\boldsymbol{P}$。重新得到矩阵 $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}=\boldsymbol{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{P}_1$, 重复这一操作,直到矩阵 $\boldsymbol{\Sigma}$ 中非对角线上的绝对值最大的数趋于 0 为止。赋予 $\boldsymbol{B}=\prod_{i=1}^{n}p_i$, 其中 n 为生成的新矩阵 $\boldsymbol{P}$ 的个数。

$\boldsymbol{\Sigma}$ 对角线上的值便为对应的特征值, $\boldsymbol{B}$ 矩阵中列即是特征值对应的特征向量。

将特征值按照从大到小排序,选出前 K 大个特征值。通常情况下,前 K 大特征值之和占总特征值之和的 80%,即用前 K 个特征值来取代矩阵中的 m 个特征。第 j 个 POI 的 $Rank$ 值公式具体如下:

$$Rank(j)=\sum_{i=1}^{K} eigenvalue(i)*\sum_{k=1}^{m}\overline{\overline{v(i)(k)}}*eigenvector(i)(k). \tag{6}$$

1.1.2 ISOMAP 降维

ISOMAP 算法可以进行非线性降维,将高维空间中数据信息映射到低维空间,再通过特征提取方法获得提取后特征,该算法依据多维尺度变换(MDS),将数据点之间原来使用的欧几里得距离替换为测地线距离,保证降维后的数据信息损失最小,同时将高维空间有效映射到低维空间里,在减小计算量的基础上,提高运算速率。

ISOMAP 算法引进了邻域图,距离很近的点可以用欧氏距离来代替,较远的点可通过最短路径算出距离,在此基础上进行降维保距。邻域图中相邻且靠近的点之间存在连接,而与之相反的便不存在连接,因此计算 2 个点之间的距离问题就是测地线距离计算问题,也即演变成了邻域图中 2 点之间的最短路径计算问题,最短路径的计算常采用经典 Floyd 算法或 Dijkstra 算法。

假设高维度空间 R^n 中 $X=(x_1,x_2,\ldots,x_m)$, 低维度子空间 R^d 中 $Y=(y_1,y_2,\ldots,y_m)$,d_m 为测地线距离, d 为欧几里得距离,对 X 中的数据进行预处理,找到一个满足公式(7)的线性嵌入函数 f:

$$\min_f\sum_{i,j}(d_m(X_i,X_j)-d(f(X_i),f(X_j))^2, \tag{7}$$

应用多维缩放算法对目标函数进行优化,原始 Data 的内积形式通过测地线距离转化为新的矩阵。令 $\boldsymbol{D}_G$ 为最短路径矩阵, $\boldsymbol{D}_Y$ 为降维子空间中的欧几里得距离矩阵, $\boldsymbol{\tau}(\boldsymbol{D}_G)$,$\boldsymbol{\tau}(\boldsymbol{D}_Y)$ 为内积矩阵,目标函数如下:

$$\min_f\|\boldsymbol{\tau}(\boldsymbol{D}_G)-\boldsymbol{\tau}(\boldsymbol{D}_Y)\|_{L_2}, \tag{8}$$

线性映射函数为:$f(x)=\boldsymbol{a}^{\mathrm{T}}x$, 其中 $y_i=f(x_i)$ 和 $\boldsymbol{Y}=(y_1,y_2,\ldots,y_m)=\boldsymbol{a}^{\mathrm{T}}x$, 得到 $\boldsymbol{\tau}(\boldsymbol{D}_Y)=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{X}$, 因此,求解 $f(x)$ 最优解变换为如下函数形式:

$$\boldsymbol{a}^*=\min_a\|\boldsymbol{\tau}(\boldsymbol{D}_G)-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{X}\|^2, \tag{9}$$

其中:

$$\|\boldsymbol{\tau}(\boldsymbol{D}_G)-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|^2=tr(\boldsymbol{\tau}(\boldsymbol{D}_G)-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X})(\boldsymbol{\tau}(\boldsymbol{D}_G)-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}})=tr(\boldsymbol{\tau}(\boldsymbol{D}_G)\boldsymbol{\tau}(\boldsymbol{D}_G)^{\mathrm{T}}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}\boldsymbol{\tau}(\boldsymbol{D}_G)^{\mathrm{T}}-\boldsymbol{\tau}(\boldsymbol{D}_G)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}), \tag{10}$$

其中, $\boldsymbol{a}$ 代表投影的方向,并不影响实际大小,因此,添加如下限制条件:

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}=1,$$

得到：

$$tr(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}) = tr(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}) = 1,$$

因此，目标函数(8)可进一步改写为：

$$\arg\max_{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}=1} \boldsymbol{a}^{T}\boldsymbol{X}\boldsymbol{\tau}(\boldsymbol{D}_G)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}, \tag{11}$$

其中，$\boldsymbol{a}_i(i = 1,2,\ldots,l)$ 是影响目标函数求得最小值的向量，目标向量求解等同于求解式(12)的特征解，即：

$$\boldsymbol{X}[\boldsymbol{\tau}(\boldsymbol{D}_G)]\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} = \lambda\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}. \tag{12}$$

令 $\boldsymbol{A} = [a_1, a_2, \ldots, a_i]$，$:x \rightarrow y = \boldsymbol{A}^{\mathrm{T}}x$，$l$ 维的向量 $\boldsymbol{Y}$ 表示更高维度的数据 x，$\boldsymbol{A}$ 为变换矩阵。本文将在第 2 实验部分对比上述 2 种降维算法。

1.2　交叉熵与困惑度

1.2.1　交叉熵

在统计学中，利用困惑度评价模型的性能优劣，能够给测试数据得出更高概率值的算法显然更好[4]，即困惑值越小，模型对实验的文本数据有更好的预测能力，因此困惑值与剧本潜在主题数量呈反比。在 LDA 主题模型中，困惑度计算公式可表示如下：

$$Perplexity(D) = \exp\left\{\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}, \tag{13}$$

其中，D 表示语料库中的测试集，共 M 篇文档；N_d 表示每篇文档 d 中的单词数；w_d 表示文档 d 中的词数；$p(w_d)$ 表示文档中词 w_d 产生的概率。

1.2.2　交叉熵结合困惑度方法

在计算主题相似度时，目前常用的方法有：Kullback-Leibler 散度(KL 散度)[5]、Jensen-Shanon 散度(JS 散度)[6]、交叉熵(Cross Entropy，CE)。其中，KL 散度不满足对称性和三角不等式，JS 散度也不能很好地衡量每个真实主题和预测的主题之间的相似性，因此本文选取交叉熵作为衡量剧本各个主题间相似度的标准。在交叉熵的基础上，将随机变量方差的概念引入到潜在主题空间中，即可衡量主题空间的整体差异性[7]。主题方差 $Var(T)$ 是各个主题分别与其均值之间的距离平方和的平均数。主题方差的计算方法详述如下。

先计算求出主题-词概率分布 ϕ 均值 $\bar{\phi}$；再利用未曾应用于剧本主题的交叉熵来得到各个主题间的方差，数学公式可写作如下形式：

$$Var(T) = \frac{\sum_{i=1}^{K}[H_{CE}(T_i, \bar{\phi})]^2}{K}, \tag{14}$$

其中，T 表示 LDA 抽取的主题；K 表示主题数目；CE 表示交叉熵。

$Var(T)$ 可以计算得到隐藏主题之间的稳固性，$Var(T)$ 越大，稳固性越好，主题易于分类。困惑度可以用来作为模型预测能力评价指标，过分追求指标值会导致主题数偏大，因此可将二者相结合。由此提出如下的 $Perplexity - Var$ 指标的公式：

$$Perplexity - Var(D_{test}) = \frac{Perplexity(D_{test})}{Var(Test)}. \tag{15}$$

其中，D_{test} 为测试数据集；$Perplexity(D_{test})$ 是困惑度；$Var(T)$ 是隐藏主题间的方差。

$Perplexity - Var$ 指标含义是：从以上关系式分析得出，$Perplexity - Var$ 值最小时，则寻求的 LDA 主题模型为最优。

1.3　LDA 主题模型

LDA 模型可以提取出研究篇章中的隐含主题，通过主题、词频生成文档，因此属于生成模型。针对剧本，使用 LDA 模型可以生成主题，提取剧本的隐含语义并对剧本进行形式化的表示。假设剧本集 D 包含 M 篇剧本，每篇剧本的长度是 N_i，在 LDA 模型中，LDA 概率图模型如图 1 所示。完整的文档生成步骤参见如下。

(1) 抽取剧本 - 主题分布 $\theta_i \sim Dirichlet(\alpha)$，$i = 1,2,\ldots,M$。

(2) 抽取主题词分布 $\varphi_i \sim Dirichlet(\beta)$，$k = 1,2,\ldots,K$。

(3) 对于剧本中的每个词 $i \in \{1,2,\ldots,M\}$，$j \in \{1,2,\ldots,N_i\}$。

(4) 抽取主题 $Z_{ij} \sim Multinomial(\theta_i)$，抽取词 $\omega_{ij} \sim Multinomial(\varphi_{z_{ij}})$。

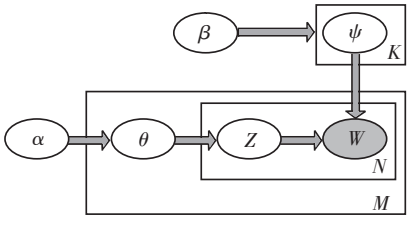

图 1　LDA 概率图模型

Fig. 1　LDA probability map model

图 1 中，M 表示剧本数量，N 表示单篇剧本中词的数量，K 表示主题数量，W 表示剧本集中的所有词，Z 表示所有主题；参数 θ 表示文档-主题分布，由 Dirichlet 先验知识 α 控制产生；ψ 表示主题-词分布，由 Dirichlet 先验知识 β 控制产生；矩形表示连续重复过程，外层矩形表示从 Dirichlet 分布中为剧本

集 D 中的每篇剧本反复抽取主题分布,内层矩形表示从主题分布中反复抽样产生剧本 d 的词。

2 实验结果及分析

2.1 实验数据与处理

本文的数据来源于互联网资源,共计 317 篇外国剧本。该数据集是 PDF 格式,利用程序将 PDF 格式剧本文件转化为实验所需要的 txt 剧本格式,通过人工标注将 317 篇剧本分为 20 种类别,分别是爱情、传记、动作、犯罪、歌舞、记录、家庭、惊悚、剧情、科幻等。

首先,分词;然后,通过停用词表过滤掉剧本中的一些无关词,将剧本文字形式转化为 TD-IDF 的向量形式,使用 TF-IDF 算法;最后,将 TF-IDF 向量矩阵进行降维,降维后的 TF-IDF 作为 LDA 的输入参数。

2.2 基于 ISOMAP 的 TF-IDF 降维实验

SVM 模型中的输入是数据,因此本文可任选向量空间模型,权重采用 TF-IDF 权重值,但由于剧本转化为 TF-IDF 时维数达到了 50 万,超出了普通计算机的运算能力,故而仍需继续降维。而降维时,在保证信息损失最少的同时,同时还要保证可靠的计算效率。通过实验对比来观测 PCA 降维与 ISOMAP 降维的处理时间的对比,将高维数据降到 2 维,再聚类为 10 类,最终可得各种降维算法处理时间的结果对比如图 2 所示。

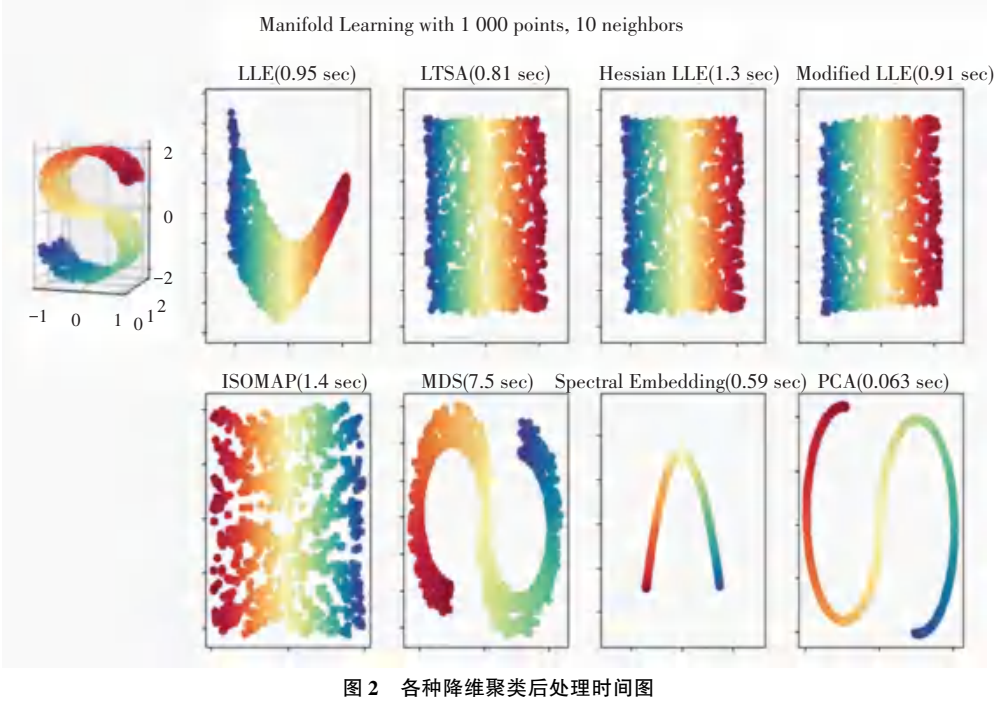

图 2 各种降维聚类后处理时间图

Fig. 2 Various time graph after dimension reduction and clustering processing

由图 2 可以看出,ISOMAP 算法的处理时间要好于 PCA 算法,但是聚类效果明显优于 PCA,如此就降低了信息的丢失率。故而,对于剧本特征降维,本文选择了 ISOMAP 算法。

通过实验得到 4 组数据,将得到的稀疏矩阵维数降为 1 000 维、3 000 维、5 000 维、10 000 维。对这 4 组数据使用带有高斯核函数的 SVM 训练模型,并以训练语料测试分类准确率,研究得到的结果见表 1。

表 1 PCA 与 ISOMAP 降维后 SVM 准确率

Tab. 1 SVM accuracy after PCA and ISOMAP dimension reduction %

	1 000 维	3 000 维	5 000 维	10 000 维
PCA	92.4	93.2	91.6	86.4
ISOMAP	93.5	94.7	92.3	89.1

PCA 与 ISOMAP 降维对比结果曲线如图 3 所示。根据表 1 与图 3 的结果,当 PCA 与 ISOMAP 降到 3 000 维的时候,分类的准确率最高,同时可以证明,在剧本分类中,使用 ISOMAP 在特征降维方面要优于 PCA 降维,因此本实验中选取降维后的维数为 3 000维。在图 3 中,PCA 降维至 5 000 维之后,基本呈一条直线,考虑到 PCA 降维时可能造成大量信息损失,会使得分类准确率大致呈现线性下降趋势。

2.3 基于交叉熵与困惑度的最优主题数实验

研究中,根据困惑度、以及困惑度与交叉熵相结合的算法,并结合各种分类器进行对比实验,通过仿真来验证该算法的优越性。在进行对比实验时,将降维算法加以统一,LDA 主题个数寻优实验选择

PCA 降维，同样，选择 TF-IDF 特征向量加权算法；SVM 的核函数，选择高斯核函数。定义困惑度计算得到的主题数为 *Perp_K*，定义困惑度和交叉熵相结合的主题数量为 *PerpSimla_K*，通过本文提出的交叉熵与困惑度计算公式分别得到最优主题个数，*Perp_K* = 200，*PerpSimla_K* = 230。不同主题数的分类器的准确率见表 2。

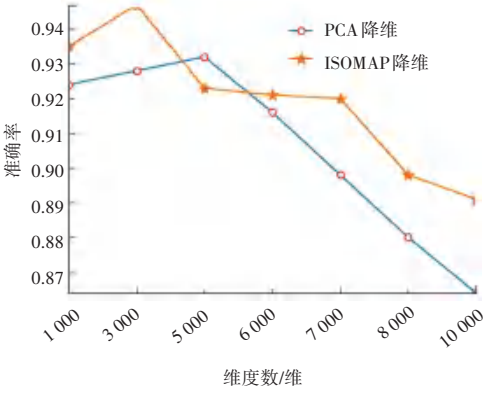

图 3　PCA 与 ISOMAP 降维对比图

Fig. 3　Comparison of PCA and ISOMAP dimension reduction

表 2　不同主题数的分类器的准确率

Tab. 2　Accuracy of classifiers with different subject numbers %

主题数	贝叶斯	KNN	SVM
Perp_K = 200	85.7	88.3	90.1
PerpSimla_K = 230	87.5	89.5	93.2

由表 2 得到的结果数据显示，利用交叉熵与困惑度结合的方法，使得各个分类器的分类准确率明显高于单独使用困惑度方法，困惑度计算可以为主题数量的确定提供有效参考，但并未能够保障构造得到最优分类器。因此需要进一步的仿真研究验证最优主题数是否准确且有效，需要将 LDA 的主题个数 *K* 值范围设置在经验数值 50~450 之间。交叉熵和困惑度结合下的不同主题数的对比结果值如图 4 所示。

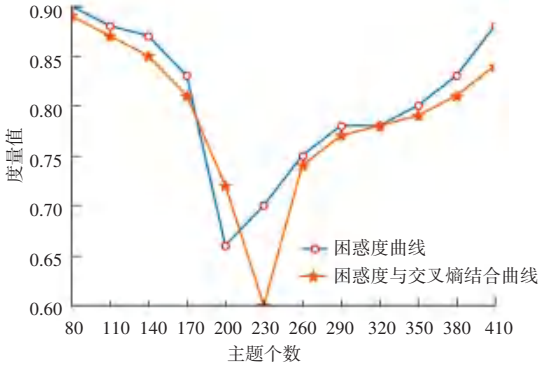

图 4　交叉熵和困惑度结合下的不同主题数的值

Fig. 4　Values of different subject numbers under the combination of cross entropy and confusion

由图 4 与表 2 可以得知，基于困惑度与交叉熵结合的方法，得到的最优主题数明显优于单纯基于困惑度计算剧本最优主题数。在接下来的部分实验中将会采用此方法，进行 LDA 主题提取。

2.4　LDA 隐含主题特征词提取

一个主题下有大量相近的词，一个词也会依附于不同的主题，这些词语和该主题有很强的相关性，也正是这些词语共同定义了这一主题。对于一篇剧本来说，通常是由若干个主题生成。综上分析可知，LDA 主题模型，能够发现隐含的主题。对降维过后的数据，进行 LDA 主题提取，以确保更低的维数，进而提取更准确的特征，后续即以 LDA 提取的特征作为 SVM 的输入。

由于剧本数量多，因此采用了 stem 图（火彩梗图）。此处，显示了前 3 篇剧本的可能的主题词的概率大小。运行结果如图 5 所示。

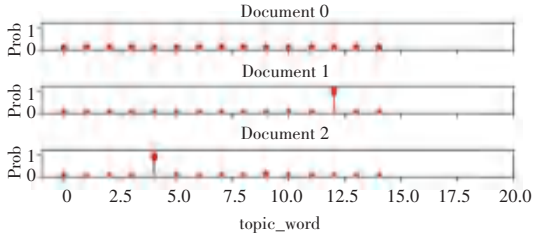

图 5　前 3 篇剧本的主题词概率图

Fig. 5　The probability map of the subject words in the first three scripts

最终，运行得到的 3 篇文档的主题词具体如下：

* Topic 0

- 近景 胳臂 中景 远景 特写 前景 后景 柜台 盯住 指指 二人 侦探 望望 摇拍 起居室 侍者 男子 手枪 艾达 后拉

* Topic 1

- 日本 东京 刑警 冈田 店员 道路 警察署 少年 西佳敬 爱德华 久间 科长 辛苦 罗尼 大海 日元 同上 玛丽 混蛋 太郎

* Topic 2

- worldcinema 蒙蒂 起居室 学生 库尔特 史蒂夫 泰亚 戴维 劳丽 布卢 莱尔 斯莱特 玛莉 scripts 特里 约翰 斯蒂芬 2011 雅各布 公寓

2.5　实验结果分析

由前文的实验部分确定了 LDA 的 *K* 值，紧接着将提取的特征向量，输入到各类分类器中，用来验证融合核函数对于剧本主题分类的优越性。

本节将从 KNN、贝叶斯以及向量机分类器进行对比实验。在 python 环境里，SVM 的模型参数可以

选择自定义的核函数。各类分类器对比实验结果见表 3。

表 3 各类分类器对比实验结果

Tab. 3 Comparison of experimental results of various types of classifiers %

	KNN	贝叶斯	SVM(多项式核函数)	SVM(线性核)	SVM(高斯核)
准确率	89.5	87.5	86.9	94.3	93.2
召回率	84.3	85.4	83.5	90.2	88.7

由表 3 可以看出,线性核的准确率逼近融合核,验证了从低维映射到高维线性可分的理论,而且由于使用了 ISOMAP 降维方法以及 LDA,使得特征空间基本处于线性可分的状态。同时表 3 给出的实验结果还验证了,相比其它核函数和分类器而言,SVM 核函数对剧本及其它文本分类能够获得更好的研究效果。

3 结束语

本文首先将剧本集向量化,得到向量空间;传统的文本向量空间,通常是利用词频作为分析的依据。而剧本向量空间,采用 TF-IDF 算法得到词语加权向量空间。对比了 PCA 与 ISOMAP 降维效果,通过实验发现 PCA 与 ISOMAP 相比有着更快的执行速率,而 ISOMAP 有着更好的降维效果,因此在更大程度上有效提升了剧本主题的分类准确率。提出交叉熵结合困惑度的方法,通过实验表明,提出的交叉熵结合困惑度的方法,可以显著改善剧本主题词的个数不准确问题,进而提高剧本主题分类准确率。本文不足之处在于,没有对 SVM 核函数做进一步的实验研究,未来工作将是利用核函数融合进行深入的探讨与分析。

参考文献

[1] WU Xindong, ZHU Xingquan, WU Gongqing, et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(1):97-107.

[2] LAZER D,KENNEDY R, KING G, et al. The parable of Google Flu:Traps in big data analysis[J]. Science, 2014, 343(6176): 1203-1205.

[3] 刘海旭. 基于 PCA 和 LDA 的文本分类系统设计与实现[D]. 北京:北京邮电大学,2013.

[4] 裘友荣. 相对熵在图像去噪中的应用[J]. 遥感信息, 2018, 33(3): 124-129.

[5] 孔锐, 施泽生, 郭立, 等. 利用组合核函数提高核主分量分析的性能[J]. 中国图象图形学报, 2004, 9(1): 40-45.

[6] 牟华英. 脑电信号特征提取的算法研究[D]. 广州:华南理工大学, 2010.

[7] 李强. 基于主题模型的中文情感分类方法研究[D]. 杭州:杭州电子科技大学,2016.

[8] 田象明. 基于视频流的车牌识别系统设计[D]. 西安:西安电子科技大学, 2017.

(上接第 44 页)

象,NMNN 能够发现位于稀疏区域和数据集内边缘的对象,NSNN 检测出的数据对象相对比较集中。

5 结束语

本文从概念上对目前存在的近邻技术进行了对比分析,对各种近邻技术的特点进行了剖析。结合自然最近邻概念,提出了自然逆最近邻、自然互最近邻和自然共享最近邻 3 种算法,并给出了 3 种算法的定义和算法描述。对提出的算法在离群检测应用中进行了实验对比,实验结果表明自然逆最近邻和自然互最近邻能够有效地检测出局部和全局离群点,自然共享最近邻与数据集中数据的相对分布密度有关,检测出的对象相对比较集中。

参考文献

[1] 毋雪雁, 王水花, 张煜东. K 最近邻算法理论与应用综述[J]. 计算机工程与应用, 2017, 53(21): 1-7.

[2] 黄文明, 莫阳. 基于文本加权 KNN 算法的中文垃圾短信过滤[J]. 计算机工程, 2017, 43(3): 193-199.

[3] 张旭, 蒋建国, 洪日昌, 等. 基于朴素贝叶斯 K 近邻的快速图像分类算法[J]. 北京航空航天大学学报, 2015, 41(2): 302-310.

[4] 管建, 亚娟, 王立功. K 近邻分类指导的区域迭代图割算法研究[J]. 计算机应用与软件,2018,35(11):237- 244,265.

[5] 卢伟胜, 郭躬德, 严宣辉, 等. SMwKnn:基于类别子空间距离加权的互 k 近邻算法[J]. 计算机科学, 2014, 41(2):166-169.

[6] 苏晓珂, 郑远攀, 万仁霞. 基于共享最近邻的离群检测算法[J]. 计算机应用研究, 2012, 29(7):2426-2428,2453.

[7] KORN F, MUTHUKRISHNAN S. Influence sets based on reverse nearest neighbor queries [J]. ACM SIGMOD Record, 2000, 29(2): 201-212.

[8] GAO Yunjun, LIU Qing, MIAO Xiaoye, et al. Reverse k -nearest neighbor search in the presence of obstacles [J]. Information Sciences, 2016, 330:274-292.

[9] 邹咸林. 自然最近邻居在高维数据结构学习中的应用[D]. 重庆: 重庆大学, 2011.

[10] 朱庆生, 唐汇, 冯骥. 一种基于自然最近邻的离群检测算法[J]. 计算机科学, 2014, 41(3): 276-278, 305.