基于用户相关性的动态网络媒体数据无监督特征选择算法

任永功 王玉玲 刘 洋 张 晶
(辽宁师范大学计算机与信息技术学院 辽宁 大连 116029)

摘 要 移动互联网、社交媒体的快速发展,极大推动了各个领域对文本、图像、视频等网络媒体数据处理的需求.该类数据具有高维度、动态更新、内容复杂的特性,增加了特征计算以及分类难度.同时,当前网络媒体数据的特征选择方法主要针对静态数据,并且对数据格式规范性要求较高.针对上述问题,为保证对动态网络媒体数据的实时特征提取,该文提出了一种基于用户相关性的动态网络媒体数据无监督特征选择算法(Unsupervised Feature Selection Algorithm for Dynamic Network Media Based on User Correlation,UFSDUC).首先,对社交网络中的交互用户进行关系分析,作为无监督特征选择的约束条件.然后,利用拉普拉斯算子构建用户相关性的特征选择模型,量化相关用户之间的关系强弱,通过拉格朗日乘子法给出特征模型中最优用户关系的数学方法.最后,基于梯度下降法设定动态网络媒体数据的阈值,用以计算非零特征权值来更新最优特征子集,达到对网络媒体数据进行有效分类的目的.该算法可在保证用户在相关性完整的基础上对动态网络媒体数据进行准确、实时的特征选择.该文采用 3 个标准网络媒体数据集,同时与 5 种目前较为流行的同类型算法进行对比以验证算法的有效性.

关键词 动态网络媒体数据;无监督特征选择;相关性;梯度下降法;关系强弱
中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2018.01517

Unsupervised Feature Selection Algorithm for Dynamic Network Media Data Based on User Correlation

REN Yong-Gong WANG Yu-Ling LIU Yang ZHANG Jing
(*Department of Computer and Information Technology, Liaoning Normal University, Dalian, Liaoning* 116029)

Abstract With the rapid development of the mobile network and social media, more and more Internet multi-media data including texture, image, video and others produce continuously at all times, meanwhile, requirements that learn and apply such data have growth. However, feature calculation and classification efficiency are severely limited, because of the high-dimensional, the complex content and dynamic updating characteristics of Internet multi-media data. Moreover, traditional algorithms mainly solve the feature extraction and classification problem for static multi-media data, and these algorithms require that data format need to conform the specific standard. Aiming to above problems, we proposed an efficient unsupervised feature selection algorithm based on user correlation that is called by UFSDUC (Unsupervised Feature Selection Algorithm for Dynamic Network Media Based on User Correlation) to ensure the feature extraction in real time for the dynamic multi-media data. Firstly, we analyzed user relationships in social networks, and combine the potential social factor to abstract three kinds of relational models

收稿日期:2016-12-12;在线出版日期:2018-01-22.本课题得到国家自然科学基金项目(61373127)、辽宁省高等学校优秀人才支持计划项目(LR2015033)、辽宁省科技计划项目(2013405003)、大连市科技计划项目(2013A16GX116)和辽宁省博士启动基金项目(20170520207)资助.**任永功**,男,1972 年生,博士,教授,主要研究领域为数据库技术、数据挖掘、智能信息计算等.E-mail:ryg@lnnu.edu.cn.**王玉玲**,女,1990 年生,硕士研究生,主要研究方向为数据挖掘.**刘 洋**,男,1986 年生,博士,讲师,主要研究方向为数字图像处理、雷达监测等.**张 晶**,女,1984 年生,博士,讲师,主要研究方向为模式分类、计算机视觉等.

including MFS(Multi-user Follow Same user), SFM(Same user Follow Multi-user), FEO(Follow Each Other). Take such models as the constraint condition for the unsupervised feature selection processing. Secondly, we use Laplace operator with the strength of relationship between users to building the relationship model, and then the lagrangian multiplier method is utilized to obtain the mathematical expression of the optimal relationship in the feature model. Moreover, in the proposed algorithm quantifies the strength of between users, which the more strength of the correlation may be gets the more similar information of the feature of between users. Therefore, our algorithm achieved the optimum solution for the multi-media data of the social network. Finally, we set the threshold of the multi-media data of the social network by utilizing the gradient descent method. This threshold is used to obtain the nonzero feature value, and then update the best subset of features to achieve the efficient performance to classify the multi-media data of the social network. In this paper, contributions of the proposed algorithm can be summarized as follows: (1) different traditional feature select algorithms that each sample need get the classification label, the proposed unsupervised feature selection algorithm can define the feature relationship according to different standards without labeling samples, for instance, the similarity of between samples and the distribution of the local information; (2) the correlative information of users is more stable than the self-users of information, such as the circle of friends once established will stably live in Internet always. Therefore, the proposed method can provides the important constraint condition for the feature extraction of the multi-media data by utilizing the user relevance; (3) the proposed algorithm realizes the feature selection efficiently at real time when the complete user relevance as a precondition. In this paper, we utilize three stander multi-media datasets to verify the proposed algorithm including Sina Weibo dataset, Flicker dataset, Blog Catalog dataset from 'Datatang'. These datasets have many characteristic enhancing the difficult of the feature extraction, such as amount of users, the complex relationship of between users, various categories of users. Moreover, we compare with five popular algorithms to evaluate the performance.

Keywords dynamic network media data; unsupervised feature selection; correlation; gradient descent; tie strength

1 引 言

随着社交媒体服务技术的逐渐成熟和互联网的快速普及,社交网络产品逐渐走进人们的日常生活,如 Facebook、Twitter、微博等等. 多样化的网络媒体软件为人们的社交沟通、娱乐消遣及信息分享提供了极大便利. 在使用这些网络应用的同时会产生数量惊人且种类繁多的用户数据,包括用户个人资料、交互信息、浏览历史、兴趣偏好等等. 在不侵犯个人隐私的情况下充分挖掘和分析这些网络数据背后的信息,利用这些信息来为商业和社会应用提供数据支撑,将会使这些数据的潜在价值得到释放. 然而网络社交媒体数据存在三大难点亟待解决:第一,网络数据动态化,在网络媒体数据应用中,数据规模不断扩大,以数据流的形式动态、快速产生,具有很强的时效性;第二,数据类型复杂化,不仅包括传统的关系数据类型,也包括网页、视频、音频、图片等各种形式的未加工的、半结构化的和非结构化的数据;第三,数据价值密度低,数据量呈指数形式增长的同时,隐藏在海量媒体数据中的有用信息却没有相应比例增长,反而使获取信息的难度加大. 因此,如何准确、实时地获取动态网络媒体数据的特征信息成为一个有意义的问题.

对海量媒体数据进行高效数据挖掘并获取有价值的信息,需要对数据进行合理的组织和分类,去除冗余数据对目标信息的干扰. 网络媒体数据传播过程中,存在格式不统一、表达方式不一致、各种数据资源混杂等问题,进行有效的排序和信息提取非常困难. 特征选择是解决高维网络数据问题的有效方法,能够提高模块学习能力,缓解维灾难. 根据其与分类器的关系可以分为 Filter[1-2] 方法、Wrapper 方

法和 Embedded[3]方法. Filter 方法是根据某一特定准则评估特征优劣，如距离、相容性、从属关系等，选择若干个较优特征构成特征子集. 该方法不依赖具体的学习过程，时间效率较高，但需要设置阈值作为特征选择的标准. Filter 方法获取特征子集的分类性能不仅与特征重要性的计算方法有关，而且与特征搜索策略及特征选择的阈值密切相关. Wrapper 方法依赖于学习过程，将训练样本分成训练子集和测试子集两部分. 利用每组特征子集训练得到的分类器的分类准确率作为该组特征重要性的度量，根据预先定义的学习算法的预测能力来评估被选特征的质量. 该方法的关键是确定搜索策略和评价学习机性能. 为了选择出性能最好的特征子集，需要进行大量计算，不适用于大规模数据挖掘，而且该方法容易产生“过适应”问题，推广性较差. Embedded 方法将特征选择集成在学习机训练过程中，通过目标函数的优化在训练分类器的过程中实现特征选择. 该方法不需要将训练样本分成训练子集和测试子集两部分，减小了评估特征子集的时间开销，可以快速得到最佳特征子集. 但是如何设计合适的函数优化模型是该方法的难点.

社交媒体数据的特征选择是多属性决策问题，对本文研究提出新挑战：(1) 不规则的网络媒体数据通常没有类标签，难以评估特征数据的重要性；(2) 动态产生且无预定义的特征数据不能直接运用无监督特征选择算法进行处理. 虽然目前已经有一些针对社交媒体数据的特征选择算法，但主要是面向静态数据，利用这些算法提取动态网络媒体数据特征会增加计算开销，而且网络数据类型多样，不具有规范的类标签，增大了特征数据提取难度. 因此，利用已有算法对动态网络媒体数据进行特征选择，不能取得较高的时效性能. 基于上述问题，本文提出一种基于用户相关性的动态网络媒体数据无监督特征选择算法(UFSDUC). 不同于传统特征选择算法要求每个实验数据都需要标记类标号，无监督特征选择算法可以在没有类标签的情况下根据不同的标准定义特征关联，如数据相似性或局部信息分布情况[4]等. 社交网络用户的关联信息相比于用户自身产生的数据信息更具有稳定性，比如大多数用户的朋友圈一旦形成都能够稳定存在，因此，可利用网络用户相关性为媒体数据的无监督特征选择提供重要约束条件. 本文首先分析网络用户交互情况并结合潜在社交因素，对用户交互网概括出三种关系类型. 然后，结合关链强度构造关系模型，相关性越强越可能具有相同的特征信息，解决了社交媒体数据最优化问题. 最后，利用梯度下降法计算阈值判定新特征，并根据拟牛顿算法随时间递增更新特征矩阵，计算实时非零权重数值并提取较大值构造最佳特征子集. 基于三个真实网络数据集的实验对比结果，可说明本文算法提高了动态网络媒体数据特征选择的有效性和准确性.

本文第 2 节从监督特征选择，无监督特征选择及静态网络和实时动态网络的角度分析了现有特征选择算法；第 3 节详细阐述了相关性动态网络媒体数据的无监督特征选择算法. 首先，构建了特征选择模型，通过对模型的分析，给出了用户关系和连接项强弱的描述. 然后，根据用户相关性和数据流特性提出模型框架，并在最后部分给出了具体实现步骤；第 4 节使用 3 个真实社交网络媒体数据集与 5 种算法进行对比实验，利用数据分段模拟实时数据更新，验证本文提出算法的有效性；本文的最后部分，总结本文提出算法及创新点，并简述对现有相关算法的改进效果和意义.

2　相关工作

特征选择[5]是将高维空间中的数据通过映射或变换的方式转换到低维空间，然后删除掉冗余和不相关的数据获取特征子集，从而达到降低数据维度的目的. 根据数据有无类标签，特征选择可分为有监督特征选择算法和无监督特征选择算法. 有监督特征选择算法[6]利用类标签里的识别信息，能够从不同数据实体中选出符合类别要求的特征子集. 获得广泛使用的特征选择算法包括 t-test 算法、ReliefF 算法和拓展的 ReliefF 算法[7]. t-test 算法用于比较两组数据是否来自于同一分布，并统计两组数据的方差是否具有显著差异，对数据的整合度要求较高. ReliefF 算法是一种特征权重算法，用于处理目标属性为连续值的回归问题，局限性在于不能有效地去除冗余特征数据. 拓展的 ReliefF 算法能够解决冗余数据，但是算法计算消耗大，不能用于大规模数据挖掘. 真实的社交媒体数据类型多样且容量大，因此无监督特征选择算法越来越受重视，其搜索过程不受限于类标签的缺失，而是根据度量聚类性能[8]产生有效的特征子集. 应用于高维数据中的无监督特征选择算法，不需要考虑额外的约束条件就可以获得

多组特征子集. He 等人[9]提出了一种既可以用于有监督又可以用于无监督的特征选择算法 LapScore. 来自同一类别的数据彼此相近,数据空间的局部结构比全局结构更重要. 根据这一特性,LapScore 算法可以为局部几何数据结构构造最近邻域图模型,用以评估特征数据的重要性. Zhao 等人[10]提出了基于光谱分析的特征选择框架模型(Spectral, SPEC),该框架能够对相似性不同的数据矩阵进行测量,进而提出频谱分析特征选择算法,能够进行有监督特征选择和无监督特征选择的联合研究. Li 等人[11]提出的非负判别特征选择法(Nonnegative Discriminative Feature Selection, NDFS)特征选择算法,在无监督情况下利用光谱聚类获取输入样本的聚类标签,这样能够辨别数据信息. 聚类标签和特征选择矩阵相结合可以获取具有较好辨别力的特征信息. 为了减少冗余和噪声数据,$\ell_{2,1}$范数最小约束增添到目标函数中,保证特征矩阵的稀疏性. Li 等人将聚类分析和稀疏性结构分析相结合提出一个新的无监督特征选择算法引导聚类稀疏结构学习(Clustering-Guided Sparse Structural Learning, CGSSL)[12],非负光谱聚类能够得到精准的输入样本类标签,不仅可以用于特征选择,也可以用于预测被不同特征数据共享的隐藏结构. 在此基础上该作者进一步改进[13],在图像处理和模式识别当中非负光谱分析可以获取更精准的集群输入图像标签,行稀疏模型与 $\ell_{2,p}$范数相结合后的模型更适用于特征选择并且具有较好的鲁棒性.

近几年,利用稀疏矩阵正则化降维已广泛应用于特征选择研究,包括多任务特征选择、$\ell_{2,1}$正则化、光谱特征选择,通过稀疏矩阵正则化能够将无监督特征选择嵌入到模型学习过程中. Argyriou 等人[14]提出了 $\ell_{2,1}$范数正则化模型. 在此基础上,Liu 等人[15]将特征选择与 $\ell_{2,1}$正则化相结合解决多维数据任务. Zhao 等人[16]基于 $\ell_{2,1}$范数稀疏回归,提出了光谱特征选择算法,能够有效地选择相关特征以及删除冗余数据. Yang 等人[17]设计了判别分析和 $\ell_{2,1}$范数最小化相结合的无监督特征选择框架,可以从整个特征集合中选择出最具有差别性的特征子集. 上述算法主要适用于同一分布的属性值数据,使用范围较小且效率受限.

利用数据相关性处理特征选择问题不同于传统特征选择方法,它分为一对一和一对多两种关系组合. Tang[18]首次尝试利用相关性解决社交媒体数据,整合网络媒体数据中存在的不同关系类型. 为了进一步改进相关性特征选择的算法,提出了全新的 LinkedFS 模型[19],根据用户在网络上发的帖子作为相关信息辅助模型学习,转发相同的帖子则表明用户相关性强,基于 $\ell_{2,1}$范数正则化处理稀疏数据矩阵,可显著提高特征选择的性能. 但是,LinkedFS 是一种有监督的特征选择算法,只能用于有类标签的网络媒体数据. 基于链接的无监督特征选择法(Linked Unsupervised Feature Selection, LUFS)算法[20]是一种无监督特征选择算法,对社交媒体数据利用模块最大化提取社交维度,进而对关联用户进行聚合处理,提取伪类标签作为无监督特征选择的标准,通过线性判断分析,数学化相关用户的从属关系. 但是该算法忽略关联用户之间的强弱链接,平等对待所有用户关系将增加特征选择的噪声干扰.

前面提到的算法主要是针对某一时刻的静态网络数据. 没有考虑到网络数据的实时动态变化问题,因此很难概括整个网络媒体数据的特征空间. 对于动态网络数据的研究,Alpha-investing 在回归模型里估计新特征的数值来决定该特征是否应该被选择,一旦某个特征被选择,将永远存在[21]. 基于在线流的特征选择算法(Online Streaming Feature Selection, OSFS)[22]提出了一种基于在线特征相关性和冗余分析的最优特征子集方法,如果候选特征与现有特征之间具有强相关关系,则接受该特征数据,但是该算法没有考虑到数据冗余问题. Guo 等人[23]对动态网络提出了一种节点分类方法,考虑到网络结构和节点内容的变化,通过特征选择对节点进行分类,该算法计算消耗较大,对节点进行分类时存在一定的误差. Tang 等人[24]提出了媒体数据流的无监督特征选择算法(Unsupervised Streaming Feature Selection, USFS),利用每个用户的社会背景作为无监督特征选择的约束条件,定时提取媒体数据流中的特征数据. 然而具有相同社会背景的用户不一定具有较强的相关性,该算法忽视了用户的关系强弱,增大了计算开销.

3 UFSDUC 算法

社交网络媒体是由用户、用户信息、用户之间的交互关系共同组成,网络数据的流动主要依赖于用户与用户之间交互活动的参与. 本节首先给出 UFSDUC 的方法框架;然后对用户节点进行分析;

接下来对用户关系进行计算；最后对动态网络媒体数据进行特征选择算法描述.

3.1　模型框架

UFSDUC 算法包括用户相关性建模计算和动态媒体数据特征选择计算两部分，模型框架如图 1 所示. 对用户相关性进行建模计算，首先进行节点分析，从用户的媒体文本数据提取用户特征，根据用户的相互关注行为获取交互向量，探讨用户的背景信息可以发现其社交因素. 然后进行关系分析，用户-特征矩阵和社交因素矩阵可以生成相关用户矩阵，与关系强度计算相结合度量用户之间的关系强弱，并且将三种用户关系类型分别数学化. 用户建模计算部分利用用户相关性作为约束条件对提高无监督特征选择的效率和准确度具有重要意义.

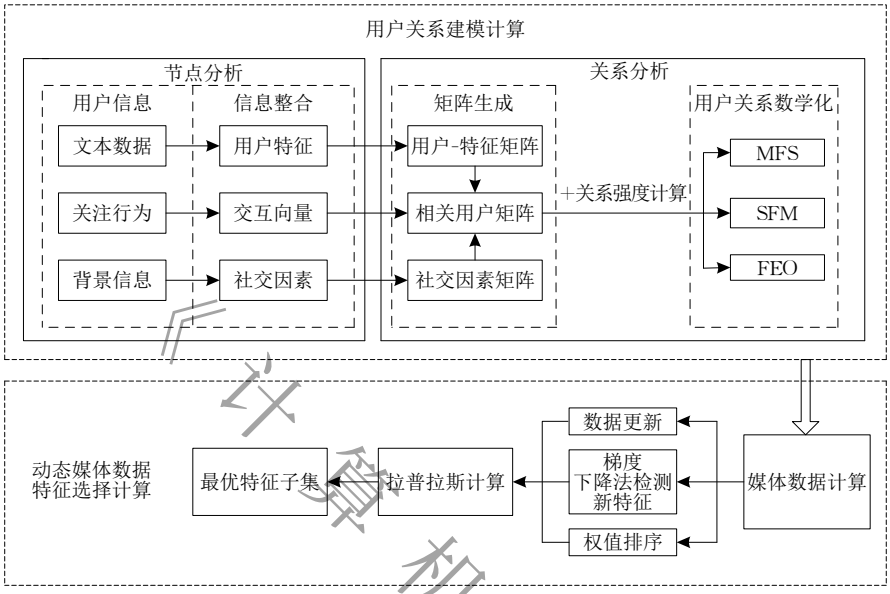


图 1　UFSDUC 算法模型框架

UFSDUC 在动态网络媒体数据特征计算部分，首先，提取动态网络数据的最优特征子集，利用数据分组法模拟媒体数据的动态变化，更新媒体数据；其次，利用梯度下降法检测每个时间段产生的新特征，判断是否接受该特征数据，进而对特征权值计算排序；最后，更新拉普拉斯矩阵，输出最优特征子集.

如图 2(a)所示，假设每个时间间隔只动态产生一个特征数据，到第 t 个时间间隔每个网络用户有 t 个特征，即 $\boldsymbol{f}^t=\{f_1,f_2,\cdots,f_t\}$，则在第 $t+1$ 个时间间隔出现新的特征集合 $\boldsymbol{f}^{t+1}=\{f_1,f_2,\cdots,f_t,f_{t+1}\}$. 社交网络用户集合可标记为 $\boldsymbol{u}=\{u_1,u_2,\cdots,u_n,\cdots\}$，其中 u_n 代表第 n 个网络用户. 在第 t 个时间间隔时，用户集合 $\boldsymbol{u}$ 与特征集合 $\boldsymbol{f}$ 之间的关系表达式为 $\boldsymbol{X}^t=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_t]\in R^{t\times n}$. 如图 2(b)所示，网络媒体用户之间通常存在关注关系(follow relation)，可以利用矩阵 $\boldsymbol{S}\in R^{n\times n}$ 表示用户之间的关注情况，如果 u_j 关注 u_i，则 $\boldsymbol{S}(i,j)=1$；如果 u_j 不关注 u_i，则 $\boldsymbol{S}(i,j)=0$.

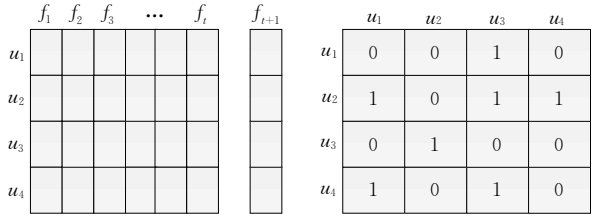


图 2　社会媒体用户特征及关联情况

3.1.1　交互关系

社交网络中的用户有两种典型行为，关注其它用户(following)或被其它用户关注(followed). 图 3(a)为社交媒体用户通过交互活动形成的关系网络，图中已经标出了用户之间的朋友关系或粉丝关系. 网络媒体用户存在三种基本关系类型，同质性[25]和社会影响力[26-27]有助于三种关系的理解：社交网络中志趣相投的用户很可能有关联；反之，有关联的用户很可能志趣相投. 下面对影响特征选择的三种用户关系类型进行详细论述：

多用户关注同一用户(Multi-user Follow Same user，MFS). 如图 3(b)所示，如果两个用户 u_1 和 u_3 关注同一用户 u_4，那么这两个用户 u_1 和 u_3 的微博可能有相似的主题；

同一用户关注多用户(Same user Follow Multi-user，SFM). 如图 3(c)所示，如果两个用户 u_2 和 u_4

被同一个用户 u_1 关注,那么这两个用户 u_2 和 u_4 的微博可能有相似的主题;

用户彼此相互关注(Follow Each Other,FEO).如图 3(d)所示,如果两个用户 u_2 和 u_3 相互关注,说明这两个用户有类似的兴趣爱好,那么这两个用户 u_2 和 u_3 的微博可能有相似的主题.

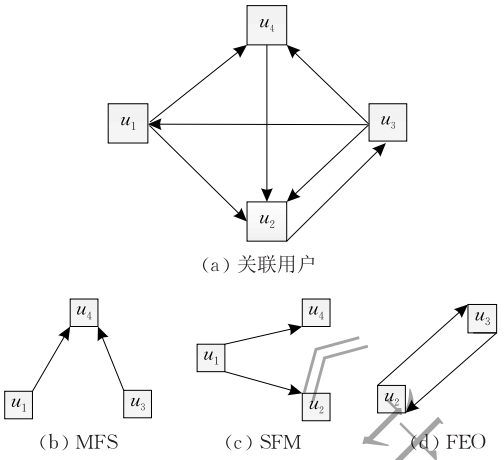


图 3 媒体用户及关系类型

3.1.2 社交因素

社交网络中的用户因为某些因素而相互关联,比如兴趣、教育背景、社会地位等,每种因素都与用户的某些特征或属性有关.实际应用中社交媒体用户之间的关联,主要是因为具有类似的爱好或教育背景等,相似性反应关联用户的特征空间.从相关信息中提取社交因素对特征选择至关重要,本文根据混合成员随机块模型[28]研究用户的潜在社交因素.社交因素以一定概率相互作用形成社会关系,$\boldsymbol{\pi}^i \in R^k$ 表示用户 u_i 存在 k 维潜在社交因素,其中 $\boldsymbol{\pi}_g^i$ 表示用户 u_i 具有社交因素 g 的概率,这说明每个实体可以同时具有不同从属关系的复合社交因素.不同社交因素之间的作用关系存储在 $u_i \rightarrow u_j$ 的矩阵 $\boldsymbol{B}$ 中,每个值的取值范围在 0 到 1 之间.相关信息的产生过程:

(1)每个关联用户 u_i 都设置 k 维向量 $\boldsymbol{\pi}^i \sim \text{Dirichlet}(\theta)$;

(2)每组关联用户 $u_i \rightarrow u_j$,提取关系向量 $\boldsymbol{S}(i,j) \sim \text{Bernoulli}(\boldsymbol{B})$.

根据可扩展推理算法[29]获取 n 个用户实体的有效社交因素 $\boldsymbol{\Pi}=[\boldsymbol{\pi}^1,\boldsymbol{\pi}^2,\cdots,\boldsymbol{\pi}^n]^{\mathrm{T}} \in R^{n\times k}$.

3.2 用户关系计算

本节对三种用户关系进行数学化推导,并结合用户关系强弱的计算方法对相关性用户进行特征选择.

3.2.1 关系数学化

相互关联用户提取潜在社交因素作为约束条件,通过回归模型进行特征选择.根据文献[24],在第 t 个时间间隔,给定用户全部社交因素 $\boldsymbol{\pi}^i$,通过式(1)的最小化问题能够找到最优特征子集:

$$\begin{aligned}\min_{\mathbf{W}^{(t)}}\theta(\mathbf{W}^{(t)}) &= \frac{1}{2}\sum_{i=1}^{k}\left\|\boldsymbol{X}^{(t)\mathrm{T}}(\boldsymbol{w}^{(t)})^i-\boldsymbol{\pi}^i\right\|_2^2+\\ &\quad \alpha\sum_{i=1}^{k}\left\|(\boldsymbol{w}^{(t)})^i\right\|_1\\ &= \frac{1}{2}\left\|\boldsymbol{X}^{(t)\mathrm{T}}\mathbf{W}^{(t)}-\boldsymbol{\Pi}\right\|_{\mathrm{F}}^2+\alpha\left\|\mathbf{W}^{(t)}\right\|_1 \quad (1)\end{aligned}$$

其中 $\|\cdot\|_{\mathrm{F}}$ 表示矩阵的弗罗贝尼乌斯范数.$\boldsymbol{W}^{(t)}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_k] \in \boldsymbol{R}^{t\times k}$ 是映射矩阵,表示在第 t 个时间间隔,为每个用户实体分配的 k 维潜在社交因素向量,$(\boldsymbol{w}^{(t)})^i$ 表示第 i 个社交因素向量的 t 个不同特征权值.α 是损失函数与 ℓ_1 范数之间的权衡参数,ℓ_1 范数的主要作用是使 $(\boldsymbol{w}^{(t)})^i$ 的某些系数恰好为 0,便于选取带有非零系数的特征值.

根据用户之间的交互情况提取了三种关系类型,说明两个用户之间的兴趣关联度可以取决于他们的全部网络文本信息.基于此给出主题兴趣定义.

定义 1(主题兴趣). 根据用户在网络媒体中所发表或转发的文章确定用户兴趣爱好,提取特征数据作为用户关系的链接纽带.用户 u_k 的主题兴趣 $\hat{T}(u_k)$ 定义为式(2):

$$\hat{T}(u_k)=\frac{\sum_{f_i\in \boldsymbol{F}_k}T(f_i)}{|\boldsymbol{F}_k|}=\frac{\sum_{f_i\in \boldsymbol{F}_k}\boldsymbol{W}^{\mathrm{T}}f_i}{|\boldsymbol{F}_k|} \quad (2)$$

其中,$T(f_i)$ 表示一篇网络文章的特征,$|\boldsymbol{F}_k|$ 表示用户 u_k 的所有网络文章的特征集合.利用主题兴趣数学化三种用户关系.

MFS. 该类型用户关系表明多个粉丝用户关注了相同的网络用户,说明这些粉丝之间可能对相同主题感兴趣.为了数学化这种关系类型,首先构造 MFS 的用户关系矩阵为 $\boldsymbol{FI}$,当 u_i 和 u_j 关注相同用户(如 u_k)时,则 $\boldsymbol{FI}(i,j)=1$,否则 $\boldsymbol{FI}(i,j)=0$.$\boldsymbol{FI}$ 可以根据用户关联矩阵 $\boldsymbol{S}$ 计算获得,即 $\boldsymbol{FI}(i,j)=sign(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})$.特征信息的规则化约束为式(3):

$$\begin{aligned}&\frac{1}{2}\sum_{u_k\in \boldsymbol{u}}\sum_{u_i,u_j\in \mathbf{N}_k}\left\|\hat{T}(u_i)-\hat{T}(u_j)\right\|_2^2\\ &=\frac{1}{2}\sum_{i,j}\boldsymbol{FI}(i,j)^{(t)}\left\|(\mathbf{W}^{(t)})^{\mathrm{T}}\boldsymbol{F}^{(t)}\boldsymbol{H}(:,i)^{(t)}-\right.\\ &\quad\left.(\mathbf{W}^{(t)})^{\mathrm{T}}\boldsymbol{F}^{(t)}\boldsymbol{H}(:,j)^{(t)}\right\|_2^2\end{aligned}$$

$$=tr((\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{F}^{(t)}\boldsymbol{H}^{(t)}\boldsymbol{L}_{\boldsymbol{FI}}^{(t)}(\boldsymbol{H}^{(t)})^{\mathrm{T}}(\boldsymbol{F}^{(t)})^{\mathrm{T}}\boldsymbol{W}^{(t)})$$

$$=\left\|(\boldsymbol{W}^{(t)}\boldsymbol{F}^{(t)}\boldsymbol{H}^{(t)})^{\mathrm{T}}(\boldsymbol{L}_{\boldsymbol{FI}}^{(t)})^{\frac{1}{2}}\right\|_2^2 \tag{3}$$

其中，N_k是用户 u_k的粉丝用户集合. 令 $\boldsymbol{H}\in R^{t\times n}$为指示器矩阵，如果用户 u_j含有用户 u_k的特征，那么 $\boldsymbol{H}(i,j)=\dfrac{1}{|\boldsymbol{F}_j|}$. $\boldsymbol{D}_{\boldsymbol{FI}}^{t}$是关系矩阵 $\boldsymbol{FI}$ 的对角线矩阵，所以可以得到 $\boldsymbol{D}_{\boldsymbol{FI}}^{t}(i,i)=\sum_j \boldsymbol{FI}(j,i)^t$，则 $\boldsymbol{L}_{\boldsymbol{FI}}^{t}=\boldsymbol{D}_{\boldsymbol{FI}}^{t}-\boldsymbol{FI}^t$是定义在 $\boldsymbol{FI}$ 上的拉普拉斯矩阵.

给式(1)添加规则约束项，得到 MFS 用户关系类型的最优化特征选择式(4)：

$$\frac{1}{2}\min_{\boldsymbol{W}^{(t)}}\left\|\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)}-\boldsymbol{\Pi}\right\|_{\mathrm{F}}^2+\alpha\left\|\boldsymbol{W}^{(t)}\right\|_1+$$

$$\frac{1}{2}\beta\sum_{u_k\in\boldsymbol{u}}\sum_{u_i,u_j\in N_k}\left\|\hat{T}(u_i)-\hat{T}(u_j)\right\|_2^2 \tag{4}$$

式(4)的第一部分可转化为式(5)：

$$\frac{1}{2}\left\|\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)}-\boldsymbol{\Pi}\right\|_{\mathrm{F}}^2=$$

$$tr(\boldsymbol{W}^{(t)\mathrm{T}}\boldsymbol{X}^{(t)}\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)}-2\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)}+\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi}) \tag{5}$$

结合式(3)和式(5)得到恒等式(6)：

$$tr(\boldsymbol{W}^{(t)\mathrm{T}}\boldsymbol{X}^{(t)}\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)}-2\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)}+\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi})+$$

$$tr((\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{F}^{(t)}\boldsymbol{H}^{(t)}\boldsymbol{L}_{\boldsymbol{FI}}^{(t)}(\boldsymbol{H}^{(t)})^{\mathrm{T}}(\boldsymbol{F}^{(t)})^{\mathrm{T}}\boldsymbol{W}^{(t)})$$

$$=tr(\boldsymbol{W}^{(t)\mathrm{T}}(\boldsymbol{X}^{(t)}\boldsymbol{X}^{(t)\mathrm{T}}+\beta\boldsymbol{F}^{(t)}\boldsymbol{H}^{(t)}\boldsymbol{L}_{\boldsymbol{FI}}^{(t)}\boldsymbol{H}^{(t)\mathrm{T}}\boldsymbol{F}^{(t)\mathrm{T}})\boldsymbol{W}^{(t)}-$$

$$2\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{(t)\mathrm{T}}\boldsymbol{W}^{(t)})$$

$$=tr(\boldsymbol{W}^{(t)\mathrm{T}}\boldsymbol{A}^{t}\boldsymbol{W}^{(t)}-2\boldsymbol{P}^{(t)}\boldsymbol{W}^{(t)}) \tag{6}$$

那么 MFS 用户关系等价于式(7)：

$$\min_{\boldsymbol{W}} tr((\boldsymbol{W}^t)^{\mathrm{T}}\boldsymbol{A}^t\boldsymbol{W}^t-2\boldsymbol{P}^t\boldsymbol{W}^t)+\alpha\left\|\boldsymbol{W}^t\right\|_1 \tag{7}$$

其中，$\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\beta\boldsymbol{F}\boldsymbol{H}\boldsymbol{L}_{\boldsymbol{FI}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}$，$\boldsymbol{P}=\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$.

上述为 MFS 用户关系的数学化证明，SFM 和 FEO 两种用户关系也可以得到与此相似的证明过程.

SFM. 设 $\boldsymbol{FD}$ 为 SFM 的用户关系矩阵，当 u_i和 u_j被相同用户 u_k关注时，则 $\boldsymbol{FD}(i,j)=1$，否则 $\boldsymbol{FD}(i,j)=0$. $\boldsymbol{FD}$ 根据用户关联矩阵 $\boldsymbol{S}$ 计算获得，即 $\boldsymbol{FD}=sign(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}})$，那么 SFM 用户关系数学化等价于式(8)：

$$\min_{\boldsymbol{W}} tr((\boldsymbol{W}^t)^{\mathrm{T}}\boldsymbol{A}^t\boldsymbol{W}^t-2\boldsymbol{P}^t\boldsymbol{W}^t)+\alpha\left\|\boldsymbol{W}^t\right\|_1 \tag{8}$$

其中，$\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\beta\boldsymbol{F}\boldsymbol{H}\boldsymbol{L}_{\boldsymbol{FD}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}$，$\boldsymbol{P}=\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$，$\boldsymbol{L}_{\boldsymbol{FD}}$是定义在 $\boldsymbol{FD}$ 上的拉普拉斯矩阵，推导过程与 MFS 类似，这里不再阐述.

FEO. 设 $\boldsymbol{FE}$ 为 FEO 的用户关系矩阵，当 u_i和 u_j相互关注时，则 $\boldsymbol{FE}(i,j)=1$. $\boldsymbol{FE}$ 根据用户关联矩阵 $\boldsymbol{S}$ 计算获取，即 $\boldsymbol{FE}(i,j)=sign(\boldsymbol{S}\boldsymbol{S})$，那么 FEO 用户关系数学化等价于式(9)：

$$\min_{\boldsymbol{W}} tr((\boldsymbol{W}^t)^{\mathrm{T}}\boldsymbol{A}^t\boldsymbol{W}^t-2\boldsymbol{P}^t\boldsymbol{W}^t)+\alpha\left\|\boldsymbol{W}^t\right\|_1 \tag{9}$$

其中，$\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\beta\boldsymbol{F}\boldsymbol{H}\boldsymbol{L}_{\boldsymbol{FE}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}$，$\boldsymbol{P}=\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$，$\boldsymbol{L}_{\boldsymbol{FE}}$是定义在 $\boldsymbol{FE}$ 上的拉普拉斯矩阵，推导过程与 MFS 类似，这里不再阐述.

仔细观察三个定理可以发现用户相关性特征选择模型相当于解决式(10)的优化问题：

$$\min_{\boldsymbol{W}} tr((\boldsymbol{W}^t)^{\mathrm{T}}\boldsymbol{A}^t\boldsymbol{W}^t-2\boldsymbol{P}^t\boldsymbol{W}^t)+\alpha\left\|\boldsymbol{W}^t\right\|_1 \tag{10}$$

主要不同之处在于三种用户关系类型的关系矩阵不同，不同的用户关系有不同的 $\boldsymbol{A}$ 表达. 当 MFS 类型时，对应用户关系矩阵 $\boldsymbol{FI}$，其中 $\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\beta\boldsymbol{F}\boldsymbol{H}\boldsymbol{L}_{\boldsymbol{FI}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}$，$\boldsymbol{P}=\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$；当 SFM 类型时，对应用户关系矩阵 $\boldsymbol{FD}$，其中 $\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\beta\boldsymbol{F}\boldsymbol{H}\boldsymbol{L}_{\boldsymbol{FD}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}$，$\boldsymbol{P}=\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$；当 FEO 类型时，对应用户关系矩阵 $\boldsymbol{FE}$，其中 $\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\beta\boldsymbol{F}\boldsymbol{H}\boldsymbol{L}_{\boldsymbol{FE}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}$，$\boldsymbol{P}=\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$.

3.2.2　关系强弱预测

强弱预测通过给定一对社交网络上有链接的用户，判断其之间是否具有强关系，区分用户强弱链接有助于社交媒体数据的特征选择. 关系强度预测着重构造现有用户关系的关联强度值，如图 4 所示，社交媒体用户关系的二进制数据转化为数值型数据[30-32]，可以提高特征选择计算效率，本节介绍四种具有代表性的关系强度预测方法.

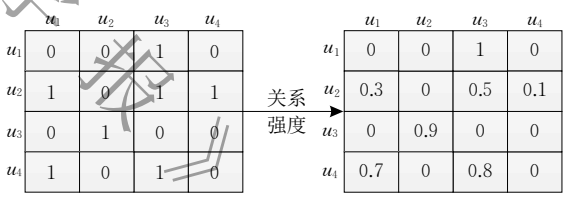


图 4　关系强度预测

结构度量(Structural Measure)用于测量两个用户在社交网络中的距离，用户之间的距离相近则可能存在强有力的关系纽带. 下面是结构度量的代表性方法.

Normalized Common Follower(NCF). 针对关注关系 $u_i\rightarrow u_j$，NCF 正式定义为式(11)：

$$NCF(i,j)=\frac{|\Gamma(i)\bigcap\Gamma(j)|}{\Gamma(i)} \tag{11}$$

其中，$\Gamma(i)=\{x|x\rightarrow u_i\}$是 u_i的粉丝集合.

Jaccard 系数. 只关心用户之间是否具有共同特征，针对关注关系 $u_i\rightarrow u_j$正式定义式(12)：

$$Jaccard(i,j)=\frac{|\Gamma(i)\bigcap\Gamma(j)|}{|\Gamma(i)\bigcup\Gamma(j)|} \tag{12}$$

Katz Score(KS). 针对 $u_i\rightarrow u_j$，将所有从 u_i到

u_j的可能路径按递增顺序排列，通过式(13)的长度衡量确定最短路径：

$$KS(i,j)=\sum_{\ell=1}^{\infty}\beta^{\ell}|path_{i,j}^{\ell}| \tag{13}$$

其中，$path_{i,j}^{\ell}$表示 u_i到 u_j的路径集合，ℓ 表示长度(β值设置为 0.05).

内容度量(Content Measure)是测定用户产生网络数据内容相似度的重要指标[33]. 假设 $\boldsymbol{c}_i$是用户 u_i的支持向量机(SVM)：$\boldsymbol{c}_i=\frac{1}{|\boldsymbol{F}_i|}\sum_{f_j\in\boldsymbol{F}_i}f_j$，其中$|\cdot|$表示集合的大小. 对于关注关系 $u_i\rightarrow u_j$，内容度量定义为式(14)：

$$CS(u_i,u_j)=\frac{\langle\boldsymbol{c}_i,\boldsymbol{c}_j\rangle}{\|\boldsymbol{c}_i\|\|\boldsymbol{c}_j\|} \tag{14}$$

其中$\langle\cdot,\cdot\rangle$表示两个向量的内积，$\|\cdot\|$表示向量的 ℓ_2 范数.

相互影响度量(Interaction Measure, IM)表明社交媒体中只有少数用户具有强关系类型，且彼此间互动非常频繁，可能存在相互转发文章、互写评论等活动[34]. 用户的互动方式集合为$\{I_1,I_2,\cdots,I_m\}$，表示 m 种互动方式，针对关注关系 $u_i\rightarrow u_j$，相互影响度量定义为式(15)：

$$IM(i,j)=\frac{\sum_{k=1}^{m}I_k(i,j)}{\sum_{j=1}^{n}\sum_{k=1}^{m}S(i,j)I_k(i,j)} \tag{15}$$

其中，$I_k\in R^{n\times n}$表示第 k 种互动方式在用户中的交互频率，分子是 u_i和 u_j之间的相互作用频率之和，$S(i,j)$表示有关联的所有用户关系，分母是 u_i的全部相互作用频率之和.

混合度量(Hybrid Measure, HM)是对结构度量、内容度量、相互影响度量的综合考虑. 利用线性关系组合上述三种测量方法定义式(16)：

$$HM(i,j)=\theta_1 SS(i,j)+\theta_2 CS(i,j)+(1-\theta_1-\theta_2)IM(i,j) \tag{16}$$

其中 $0\leq\theta_1\leq1$，$0\leq\theta_2\leq1$，$0\leq\theta_1+\theta_2\leq1$. 当 $\theta_1=1$时，混合度量相当于结构度量；当 $\theta_2=1$ 时，混合度量相当于内容度量；当 $\theta_1=0$，$\theta_2=0$ 时，混合度量相当于相互作用度量.

3.2.3 相关性用户算法实现

结合关系强度对模型进一步研究并给出收敛分析，受文献[3]启发利用拉格朗日函数解决此优化问题：

$$\theta(\mathbf{W}^{(t)})=tr((\boldsymbol{W}^t)^{\mathrm{T}}\boldsymbol{A}^t\boldsymbol{W}^t-2\boldsymbol{P}^t\boldsymbol{W}^t)+\alpha\|\boldsymbol{W}^t\|_1 \tag{17}$$

ℓ_1范数虽然不可微，但是存在次微分，即$\frac{\partial}{\partial W}\|\boldsymbol{W}\|_1=sign(\boldsymbol{W})$，文献[24]利用该思想对 ℓ_1范数进行求导实验. 式(17)求导得到

$$\frac{\partial\theta(\mathbf{W}^{(t)})}{\partial\mathbf{W}^{(t)}}=2\boldsymbol{A}^t\boldsymbol{W}^t-2(\boldsymbol{P}^t)^{\mathrm{T}}+\boldsymbol{\alpha}\boldsymbol{D}_W^t \tag{18}$$

根据 $\lambda\boldsymbol{E}-\boldsymbol{W}$ 可以求得 $\boldsymbol{W}$ 的对角矩阵 $\boldsymbol{D}_W$，$\boldsymbol{E}$ 是单位矩阵. 设导数为 0，得到 $\boldsymbol{W}=(2\boldsymbol{P}-\alpha\boldsymbol{D}_W)(2\boldsymbol{A})^{-1}$，可知 $\boldsymbol{W}$ 取决于 $\boldsymbol{D}_W$. 为全面评价四种关系强弱预测方法的效果，本文对四种方法均进行了实验对比与分析. 用户相关性特征选择模型的优化方法如算法 1 所述.

算法 1. 用户相关性特征选择算法(UCFS).

输入：$\{\boldsymbol{X},\boldsymbol{F},\boldsymbol{S}\}$和期望特征数量 k.

输出：k 个最相关特征.

1. 任意选择一个关系强度预测用来更新邻接矩阵 $\boldsymbol{S}$；
2. 根据选择的假说设置 $\boldsymbol{P}$ 和 $\boldsymbol{A}$；
3. 设置 $n=0$ 并且初始化 $\boldsymbol{D}_{\boldsymbol{W}_n}$ 作为单位矩阵；
4. WHILE (not convergent) DO
   {计算 $\boldsymbol{W}_{n+1}=(2\boldsymbol{P}-\alpha\boldsymbol{D}_{\boldsymbol{W}_n})(2\boldsymbol{A})^{-1}$；
   更新对角线矩阵 $\boldsymbol{D}_{\boldsymbol{W}_{n+1}}$；
   $n=n+1$；}
   END WHILE
5. 根据 $\|\boldsymbol{W}\|_1$分类每个特征，按降序排列选择前 k 个特征值；
6. 输出特征选择结果；
7. END

文献[3]已经证明了任意两个非零常数 x 和 y 都存在不等关系式：

$$\sqrt{x}-\frac{x}{2\sqrt{y}}\leq\sqrt{y}-\frac{y}{2\sqrt{y}} \tag{19}$$

引申到本文可以得到

$$\|\boldsymbol{W}_{n+1}\|_1-\frac{\|\boldsymbol{W}_{n+1}\|_1}{\|\boldsymbol{W}_n\|_1}\leq\|\boldsymbol{W}_n\|_1-\frac{\|\boldsymbol{W}_n\|_1}{\|\boldsymbol{W}_n\|_1} \tag{20}$$

这表明

$$tr(\boldsymbol{W}_{n+1}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W}_{n+1}-2\boldsymbol{P}\boldsymbol{W}_{n+1})+\alpha\sum_i\frac{\|\boldsymbol{W}_{n+1}\|_1}{\|\boldsymbol{W}_n\|_1}\leq tr(\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W}_n-2\boldsymbol{P}\boldsymbol{W}_n)+\alpha\sum_i\frac{\|\boldsymbol{W}_n\|_1}{\|\boldsymbol{W}_n\|_1} \tag{21}$$

进而满足下面的不等式

$$tr(\boldsymbol{W}_{n+1}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W}_{n+1}-2\boldsymbol{P}\boldsymbol{W}_{n+1})+\alpha\sum_i\|\boldsymbol{W}_{n+1}\|_1-\alpha\Big(\sum_i\|\boldsymbol{W}_{n+1}\|_1-\sum_i\frac{\|\boldsymbol{W}_{n+1}\|_1}{2\|\boldsymbol{W}_n\|_1}\Big)\leq tr(\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W}_n-2\boldsymbol{P}\boldsymbol{W}_n)+$$

$$\alpha\sum_i\|\boldsymbol{W}_n\|_1-\alpha\Big(\sum_i\|\boldsymbol{W}_n\|_1-\sum_i\frac{\|\boldsymbol{W}_n\|_1^2}{2\|\boldsymbol{W}_n\|_1}\Big) \tag{22}$$

最后可以证明：

$$tr(\boldsymbol{W}_{n+1}^{\mathrm{T}}(\boldsymbol{A}+\alpha\boldsymbol{D}_{\boldsymbol{W}_n})\boldsymbol{W}_{n+1}-2\boldsymbol{P}\boldsymbol{W}_{n+1})\leq$$
$$tr(\boldsymbol{W}_n^{\mathrm{T}}(\boldsymbol{A}+\alpha\boldsymbol{D}_{\boldsymbol{W}_n})\boldsymbol{W}_n-2\boldsymbol{P}\boldsymbol{W}_n) \tag{23}$$

递归时目标函数单调递减，迭代方法收敛于最优解.

3.3　动态媒体数据特征选择

在已经对媒体用户进行关系分类和数学建模的情况下，进一步考虑网络数据动态变化的时间参数，实现相关性用户的动态网络媒体数据无监督特征选择算法(UFSDUC).

描述动态网络媒体数据特征选择的一般过程，根据时间流动的特性，按照一定的时间间隔设置时间逐渐递增，每个时间段内 UFSDUC 算法决定是否接受新特征，如果该特征纳入最优特征集合，则进一步判断是否需要重新整合特征集，重复该过程直到没有新特征出现.

对用户相关性特征选择优化式(17)进一步改进得到式(24)：

$$\min_{(\boldsymbol{w}^{(t)})^i,i\in k}\theta((\boldsymbol{w}^{(t)})^i)=$$
$$\frac{1}{2}\|\boldsymbol{X}^{(t)\mathrm{T}}(\boldsymbol{w}^{(t)})^i-\boldsymbol{\pi}^i\|_2^2+\alpha\|(\boldsymbol{w}^{(t)})^i\|_1+$$
$$\frac{1}{2}\beta\|(\boldsymbol{W}^{(t)}\boldsymbol{F}^{(t)}\boldsymbol{H}^{(t)})^{\mathrm{T}}(\boldsymbol{L}^{(t)})^{\frac{1}{2}}\|_2^2 \tag{24}$$

其中，$i=1,2,\cdots,k$，在时间间隔 t 时获取特征子集，根据三种关系类型 $\boldsymbol{L}$ 分别对应 $\boldsymbol{L}_{FI}$、$\boldsymbol{L}_{FD}$、$\boldsymbol{L}_{FE}$. 下面介绍在时间间隔 $t+1$ 时，产生的新特征 f_{t+1} 是如何有效执行特征选择. 本节提出的算法功能：(1) 判定新特征；(2) 更新动态媒体数据的特征子集.

3.3.1　判定新特征

根据式(24)，在时间段 $t+1$ 时获取新特征 f_{t+1} 并且添加非零权值 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 到模型中，在 ℓ_1 正则化项上增加 $\alpha\|(\boldsymbol{w}^{(t+1)})_{t+1}^i\|$ 会引发一个处罚. 仅当第一项、第二项、第三项的数据归约超过增加的 $\alpha\|(\boldsymbol{w}^{(t+1)})_{t+1}^i\|$ 处罚时，新特征 f_{t+1} 才会降低整个目标函数值. 本文采用梯度下降法[35]检测新特征，令 $\theta((\boldsymbol{w}^{(t+1)})^i)$ 表示 $t+1$ 时间段的目标函数值：

$$\min_{(\boldsymbol{w}^{(t+1)})^i,i\in k}\theta((\boldsymbol{w}^{(t+1)})^i)=$$
$$\frac{1}{2}\|\boldsymbol{X}^{(t+1)\mathrm{T}}(\boldsymbol{w}^{(t+1)})^i-\boldsymbol{\pi}^i\|_2^2+\alpha\|\boldsymbol{w}^{(t+1)})^i\|_1+$$
$$\frac{1}{2}\beta\|((\boldsymbol{w}^{(t+1)})^i\boldsymbol{F}^{(t+1)}\boldsymbol{H}^{(t+1)})^{\mathrm{T}}(\boldsymbol{L}^{(t+1)})^{\frac{1}{2}}\|_2^2 \tag{25}$$

对公式 $\theta((\boldsymbol{w}^{(t+1)})^i)$ 中的 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 求导：

$$\frac{\partial\theta((\boldsymbol{w}^{(t+1)})^i)}{\partial(\boldsymbol{w}^{(t+1)})_{t+1}^i}=[(\boldsymbol{X}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i-\boldsymbol{\pi}^i)+$$
$$\beta(\boldsymbol{F}^{(t+1)}\boldsymbol{H}^{(t+1)})^{\mathrm{T}}\boldsymbol{L}^{(t+1)}\boldsymbol{H}^{(t+1)}\boldsymbol{F}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i]_{t+1}+$$
$$\alpha sign((\boldsymbol{w}^{(t+1)})_{t+1}^i)$$
$$=[(\boldsymbol{X}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i-\boldsymbol{\pi}^i)+$$
$$\beta(\boldsymbol{F}^{(t+1)}\boldsymbol{H}^{(t+1)})^{\mathrm{T}}\boldsymbol{L}^{(t+1)}\boldsymbol{H}^{(t+1)}\boldsymbol{F}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i]_{t+1}\pm\alpha \tag{26}$$

在式(26)中，ℓ_1 范数项 $\alpha\|(\boldsymbol{w}^{(t+1)})^i\|_1$ 的导数关于 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 不连续，下面讨论其导数符号，即 $sign(\boldsymbol{w}^{(t+1)})_{t+1}^i$. 当有新的特征 f_{t+1} 产生时，首先设置它的特征系数 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 为 0，并将其代入到模块中，如果：

$$[(\boldsymbol{X}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i-\boldsymbol{\pi}^i)+$$
$$\beta(\boldsymbol{F}^{(t+1)}\boldsymbol{H}^{(t+1)})^{\mathrm{T}}\boldsymbol{L}^{(t+1)}\boldsymbol{H}^{(t+1)}\boldsymbol{F}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i]_{t+1}-\alpha>0,$$

则很容易证明 $\dfrac{\partial\theta((\boldsymbol{w}^{(t+1)})^i)}{\partial(\boldsymbol{w}^{(t+1)})_{t+1}^i}>0$. 为了降低目标函数 $\theta((\boldsymbol{w}^{(t+1)})^i)$ 的值，需要减小 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 使其消极，并且 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 为负数.

同理，如果：

$$[(\boldsymbol{X}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i-\boldsymbol{\pi}^i)+$$
$$\beta(\boldsymbol{F}^{(t+1)}\boldsymbol{H}^{(t+1)})^{\mathrm{T}}\boldsymbol{L}^{(t+1)}\boldsymbol{H}^{(t+1)}\boldsymbol{F}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i]_{t+1}+\alpha<0,$$

则 $\dfrac{\partial\theta((\boldsymbol{w}^{(t+1)})^i)}{\partial(\boldsymbol{w}^{(t+1)})_{t+1}^i}<0$，即 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 为正数.

如果前两个条件都不满足，则说明不可能通过改变 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 来降低目标函数 $\theta((\boldsymbol{w}^{(t+1)})^i)$ 的值. 对于新特征 f_{t+1} 需要检查：

$$[(\boldsymbol{X}^{(t+1)})^{\mathrm{T}}(\boldsymbol{X}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i-\boldsymbol{\pi}^i)+$$
$$\beta(\boldsymbol{F}^{(t+1)}\boldsymbol{H}^{(t+1)})^{\mathrm{T}}\boldsymbol{L}^{(t+1)}\boldsymbol{H}^{(t+1)}\boldsymbol{F}^{(t+1)}(\boldsymbol{w}^{(t+1)})^i]_{t+1}>\alpha \tag{27}$$

若满足式(27)，则说明新特征数据 f_{t+1} 能够降低目标函数 $\theta((\boldsymbol{w}^{(t+1)})^i)$ 的值，那么将新特征列入到特征子集中.

3.3.2　更新特征子集

当不断有新特征产生时，应该考虑是否有必要更新特征子集，因为新数据能够更好地代表用户的即时特性.

当新特征增添到模型中，利用现有特征权重优化式(25)，最优化过程会使一些特征权重为 0. 如果特征权重为 0 说明该特征不能够降低目标函数值并且该特征可以被删除. 接下来探讨如何处理式(25)的优化问题，目标函数是关于 $(\boldsymbol{w}^{(t+1)})_{t+1}^i$ 的凸函数，可以求导得到式(26)获得最优解. 本文利用文献[36]的求解方法(Broyden Fletcher Goldfarb Shanno,

BFGS)算法,仅需要在每次迭代时计算目标函数的梯度.

方程式(25)的最小化问题可以泛化 $\min f(x)$, $x\in\mathbb{R}^n$,每次迭代时,更新最优解 x: $x_{m+1}=x_m-\delta_m\boldsymbol{H}_m g_m$,其中 $\boldsymbol{H}_m=\boldsymbol{B}_m^{-1}$, $\boldsymbol{B}_m$ 是 Hessian 矩阵的近似值, $g_m=\nabla f(x_m)$ 是倾斜度. 向量 $\boldsymbol{s}_m$ 和 $\boldsymbol{c}_m$ 为: $\boldsymbol{s}_m=x_{m+1}-x_m$, $\boldsymbol{c}_m=g_{m+1}-g_m$. 下面的 Hession 函数近似满足于正切方程: $\boldsymbol{B}_{m+1}\boldsymbol{s}_m=\boldsymbol{c}_m$. 通过正切函数可以得到曲率条件: $\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{B}_{m+1}\boldsymbol{s}_m=\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m>0$. 如果曲率条件得到满足,在正切方程中 $\boldsymbol{B}_{m+1}$ 至少有一个解决方法,通过下面的方式能够更新:

$$\boldsymbol{B}_{m+1}=\boldsymbol{B}_m+\frac{\boldsymbol{c}_m\boldsymbol{c}_m^{\mathrm{T}}}{\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{s}_m}-\frac{\boldsymbol{B}_m\boldsymbol{s}_m\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{B}_m}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{B}_m\boldsymbol{s}_m} \tag{28}$$

$\boldsymbol{H}_{m+1}$ 可以通过 Sherman-Morrison 公式获得更新:

$$\boldsymbol{H}_{m+1}=\boldsymbol{H}_m-\frac{\boldsymbol{s}_m\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{H}_m+\boldsymbol{H}_m\boldsymbol{c}_m\boldsymbol{s}_m^{\mathrm{T}}}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m}+\left(1+\frac{\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{H}_m\boldsymbol{c}_m}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m}\right)\frac{\boldsymbol{s}_m\boldsymbol{s}_m^{\mathrm{T}}}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m} \tag{29}$$

在第 $t+1$ 个时间段内通过解决全部 k 个子问题,获取稀疏系数矩阵 $\boldsymbol{W}=[(\boldsymbol{w}^{(t+1)})^1,\cdots,(\boldsymbol{w}^{(t+1)})^k]$. 分别解决每个子问题,使每个 $(\boldsymbol{w}^{(t+1)})^i$ 非零权重的数量不一定相同. 对于每个特征 f_j,如果任意 k 个特征权重系数 $(\boldsymbol{w}^{(t+1)})_j^i$ 是非零的,该特征就被包含于最终模块中,否则就排除该特征. 如果 f_j 被选择,则在时间段 $t+1$ 的特征得分是:

$$FScore(j)^{(t+1)}=\max((\boldsymbol{w}^{(t+1)})_j^1,\cdots,(\boldsymbol{w}^{(t+1)})_j^k) \tag{30}$$

根据特征值降序排列对被选特征进行分类,特征值越大该特征重要性越大.

算法 2. 相关性流媒体数据的无监督特征选择算法(UFSDUC).

输入:时间步长 $t+1$ 时的 f_{t+1},时间步长 t 时的特征权重矩阵 $\boldsymbol{W}^t$,关联信息 $\boldsymbol{S}$,起始点 x_0,Hessian 近似矩阵 $\boldsymbol{H}_0$,潜在社交因素数目 k.

输出:在时间步长 $t+1$ 选择的特征子集 $\boldsymbol{f}^{(t+1)'}$.

1. 从 $\boldsymbol{S}$ 中获得潜在社交因素 $\boldsymbol{\Pi}$;
2. FOR(每个社交潜在因素 $\boldsymbol{\pi}^i$)
3. { 根据式(26)对 f_{t+1} 计算梯度 g;
4. IF(abs(g)>α)
5. { 增加特征 f_{t+1} 到最优特征集合;
6. $m\leftarrow 0$;
7. 倾斜度 $g_m=\nabla f(x_m)$;
8. WHILE ($\|g_m\|>\varepsilon$)
9. { 获得向量 $\boldsymbol{p}_m=-\boldsymbol{H}_m g_m$;
10. 计算 $x_{m+1}=x_m+\delta_m p_m$,其中 δ_m 的选择通过线索满足曲率条件;
11. $\boldsymbol{g}_{m+1}=\nabla f(x_{m+1})$;
12. $\boldsymbol{s}_m=x_{m+1}-x_m$;
13. $\boldsymbol{c}_m=\boldsymbol{g}_{m+1}-\boldsymbol{g}_m$;
14. $\boldsymbol{H}_{m+1}=\boldsymbol{H}_m-\frac{\boldsymbol{s}_m\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{H}_m+\boldsymbol{H}_m\boldsymbol{c}_m\boldsymbol{s}_m^{\mathrm{T}}}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m}+\left(1+\frac{\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{H}_m\boldsymbol{c}_m}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m}\right)\frac{\boldsymbol{s}_m\boldsymbol{s}_m^{\mathrm{T}}}{\boldsymbol{s}_m^{\mathrm{T}}\boldsymbol{c}_m}$;
15. $m\leftarrow m+1$; } }
16. IF(特征 f_{t+1} 被选择)
17. { 更新拉普拉斯矩阵 $\boldsymbol{L}^{(t+1)}$;
18. 根据式(30)获取特征数值;
19. 根据数值排列特征并且更新特征集合 $f^{(t+1)'}$; }}
20. RETURN $\boldsymbol{f}^{(t+1)'}$

算法 2 对动态网络媒体数据的每个新特征 f_{t+1} 执行有效的无监督特征选择. 第 1 行应用关联信息 $\boldsymbol{S}$ 获得社交因素矩阵 $\boldsymbol{\Pi}$. 算法第 2~13 行检测新特征和现有特征,尤其是对每个子问题验证梯度条件,这步骤决定了是否接受该新特征(第 8 行). 如果条件满足(第 9 行),新特征输入到模型中(第 10 行),并利用现有特征权重对模型再次优化(第 11~16 行). 最后,更新拉普拉斯矩阵(第 17 行),计算特征权值,更新最优特征子集.

时间复杂度分析. 假设所有流动特征的数目是 t,最后获取的特征数目是 s,则更新拉普拉斯矩阵的时间消耗是 $O(n^2st)$. 在每个时间段内检查式(28)的梯度情况,时间复杂度是 $O(n^2kst)$,利用特征权重优化模型的时间复杂度是 $O(n^2s^2t)$. 所以整个 UFSDUC 算法的综合时间复杂度为 $O(n^2st)+O(n^2kst)+O(n^2s^2t)$. 因为 $k\ll t$,并且 $s\ll t$,所以整个算法的渐进时间复杂度可以看作是 $O(n^2s^2t)$.

4 实验结果与分析

通过在真实的网络社交媒体数据集上进行对比实验,进一步验证 UFSDUC 算法的有效性和准确性. 本节由实验数据、实验设置、准确性分析、关系强度影响和参数有效性分析 5 部分组成. 所有实验在主频为 2.8GHz CPU 和 4GB 内存的 PC 机上完成,操作系统为 Windows 7,用 Matlab 语言在 MATLAB 2011b 环境下实现 UFSDUC 算法.

4.1 实验数据

实验采用 3 个标准社交媒体数据集,分别是由数据堂提供的 SinaWeibo 数据集、Flickr 数据集和 BlogCatalog 数据集. SinaWeibo 是提供微型博客服务类的社交网站. 用户可通过 Web 网页、手机客户

端等方式发布消息或上传图片，还提供评论和转发等功能供用户交流. Flickr 是雅虎旗下管理和共享图片的网站，用户根据图片类别表下载自己感兴趣的图片，是提供网络社群服务的平台，特点是基于社交网络人际关系的拓展与内容的组织. BlogCatalog 提供博客用户和博文管理的社区网站，用户在预定义目录下注册账号，是用户进行分享和交流的一个网络平台.

这三个数据集的显著特点是用户数量多，用户与用户之间的交互记录复杂，产生的用户数据种类多样. 为了更好的完成实验，利用数据集成和数据归约对数据进行预处理，得到本实验所需要的数据集合，集合的详细信息如表 1 所示.

表 1　数据集详细信息

		原始特征数	用户数	类别数	链接关系数	平均等级数
数据集	SinaWeibo	8693	3493	6	27593	34.65
	Flickr	9866	2163	9	25877	31.92
	BlogCatalog	7683	4046	5	24735	35.37

4.2　实验设置

本文应用准确度 ACC(Accuracy)和归一化互信息 NMI(Normalized Mutual Information)作为评定无监督特征选择算法的性能指标. 准确度是最常见的性能指标，表示在一定条件下多次测定的平均值与真值相符合的程度. $ACC=(TP+TN)/(P+N)$，即被正确分类的特征数据除以所有的特征数据，通常来说准确度越高，分类器性能越好. 归一化互信息常用来度量两个聚类结果的相近程度，给定两个簇 A 和 B，NMI 定义为 $NMI(A,B)=\dfrac{MI(A,B)}{\max(H(A),H(B))}$，其中 $NMI(A,B)$的值在 0～1 之间，值越大说明聚类效果越好.

下面列举实验中对比的所有算法：

(1) LapScore，利用拉普拉斯分数值有效衡量各个特征的权重，优先选择权重较大值；

(2) SPEC，利用光谱分析测量特征相关性，并制作光谱图表实现特征选择算法；

(3) NDFS，通过联合正频谱分析和 $\ell_{2,1}$ 范数正则化进行特征选择；

(4) LinkedFS，根据用户的网络帖子的相关性处理稀疏数据矩阵；

(5) USFS，处理动态数据的无监督特征选择算法.

实验中，不仅采用算法的运行时间和建模时间评估效率，还对比了不同关系强度预测方法对 UFSDUC 算法的影响，以及详细研究了各个参数对特征选择性能的影响，进而选出最佳参数值.

4.3　准确性分析

本节比较 UFSDUC 与 5 种对比算法在特征选择准确性上的表现. 文献[18]提出根据聚类性能可以进行特征选择质量评估，特征子集与目标概念越相关，利用该子集训练得到分类器的准确性越好.

4.3.1　用户相关性特征选择

用户之间的关联信息比用户自身产生的数据信息更稳定，本文假定用户关系不在短时间内随时间变化. 首先对用户相关性特征选择算法(UCFS)进行评估. 将每个数据集都分成大小不同的测试子集 $\{D_5, D_{25}, D_{50}, D_{100}\}$，相当于整个数据集的 5%、25%、50%、100%. 在每个测试子集中依次设置 100、200、300 个特征数据，分别计算每种算法的聚类性能，实验对比结果如表 2～表 4 所示.

表 2　SinaWeibo 数据集中不同算法的聚类结果

数据集	特征数	算法					
		LapScore	SPEC	NDFS	LinkedFS	USFS	UCFS
D_5	100	22.56	24.95	25.35	25.12	24.86	**28.78**
	200	23.35	25.23	27.53	26.89	26.45	**29.96**
	300	25.43	26.93	28.43	28.95	28.92	**33.48**
D_{25}	100	23.98	23.48	24.81	25.82	25.36	**29.17**
	200	24.79	25.87	26.64	27.46	28.14	**31.34**
	300	25.55	26.47	27.06	28.44	29.06	**32.49**
D_{50}	100	25.19	25.01	26.11	26.21	27.15	**31.68**
	200	26.68	26.14	27.24	28.68	28.67	**34.16**
	300	27.37	28.96	29.85	30.66	31.55	**34.74**
D_{100}	100	27.15	28.63	29.78	30.52	30.85	**33.47**
	200	28.34	29.31	30.14	32.11	32.46	**34.19**
	300	29.06	30.69	31.05	34.29	35.05	**35.72**

表 3 Flickr 数据集中不同算法的聚类结果

数据集	特征数	算法					
		LapScore	SPEC	NDFS	LinkedFS	USFS	UCFS
D_5	100	24.01	24.67	25.02	26.58	26.16	**27.88**
	200	25.98	26.65	28.43	28.16	28.05	**29.21**
	300	30.71	29.54	30.67	29.34	30.42	**32.72**
D_{25}	100	25.37	26.41	25.63	27.18	27.05	**28.76**
	200	27.66	27.24	28.65	29.28	30.81	**30.54**
	300	28.04	29.02	29.07	29.86	**32.42**	31.67
D_{50}	100	28.96	29.93	29.78	28.91	29.08	**31.87**
	200	30.05	30.71	30.55	31.05	30.75	**32.27**
	300	31.88	32.35	32.49	32.74	32.85	**33.79**
D_{100}	100	28.82	29.08	30.94	31.16	32.14	**32.16**
	200	30.61	30.27	33.07	33.48	33.52	**34.17**
	300	31.02	32.29	33.82	34.07	34.27	**34.84**

表 4 BlogCatalog 数据集中不同算法的聚类结果

数据集	特征数	算法					
		LapScore	SPEC	NDFS	LinkedFS	USFS	UCFS
D_5	100	21.75	23.68	25.82	26.06	27.02	**28.34**
	200	24.05	25.98	26.24	27.69	29.44	**30.05**
	300	27.75	27.35	28.51	29.45	30.85	**32.76**
D_{25}	100	23.35	24.68	26.43	28.94	28.19	**30.74**
	200	25.79	26.68	28.41	29.18	31.33	**32.87**
	300	28.34	27.38	29.33	30.74	32.12	**33.07**
D_{50}	100	26.55	24.95	27.55	28.15	28.75	**31.27**
	200	27.14	28.78	29.35	30.61	31.22	**33.87**
	300	29.15	30.38	30.67	32.49	32.84	**34.06**
D_{100}	100	26.83	26.16	28.09	29.07	30.94	**33.86**
	200	29.75	30.31	31.88	31.62	32.06	**35.07**
	300	30.24	31.69	32.41	33.43	34.21	**35.84**

从实验结果的整体趋势来看，随着被选特征数据的增加，六种算法的聚类性能都得到提高. 五种对比算法的实验结果相似，其中 LinkedFS 和 USFS 算法的执行效果相对较好，因为这两种对比算法考虑到用户关系，有利于用户相关性的特征选择. 与五种对比算法相比，本文提出的 UCFS 算法在三个数据集上具有持续优越性，即使最低实验准确度也达到了 27.88%，但是仔细观察表格可以发现 UCFS 算法随着数据集的增大，聚类准确度的增加量逐渐减小. 比如在 SinaWeibo 数据集，4 个测试子集中的最大准确度增长量分别是 4.70%、3.32%、3.06%、2.25%，也就是说数据集的存储量对聚类性能有一定的影响.

4.3.2 动态数据特征选择

本节考虑网络媒体数据的动态特性，对基于用户相关性的动态网络媒体数据无监督特征选择算法(UFSDUC)进行性能评定. 实验中为了模拟动态数据的时序特征，一个合理的设想就是将数据分成数据块，按照固定时间间隔均匀采样，模拟动态网络媒体数据的顺序变化，这种方式可以代表数据流随时间变化的特性. 将所有网络媒体数据集按照时序变化划分 9 个子数据集，依次对每个子集进行处理，相当于媒体数据的动态变化. 在每个组合中分别将 LapScore、SPEC、NDFS、LinkedFS 与 USFS、UFSDUC 算法均进行特征选择的对比研究，根据 6 种特征选择算法获取相同特征数据量的实验结果进行比较. 实验获得特征选择结果后，利用 k-means 算法在所选特征数据的基础上进行聚类，重复 k-means 算法 20 次计算平均值. 聚类结果利用 ACC 和 NMI 值进行评价，ACC 和 NMI 值越大代表特征选择性能越好，实验结果如表 5～表 7 所示.

在 3 个数据集上 UFSDUC 相比于其它 5 种特征选择算法获得了较好的性能提升，从实验结果中可见，在数据集 SinaWeibo 和 BlogCatalog 上，本文算法的 ACC 值和 NMI 值始终优于其它 5 种同类算法. 由于 USFS 也是面向媒体数据流的研究，所以在 Flickr 数据集的 70%数据分组中获得了优于 UFSDUC 算法的结果，但是 USFS 算法具有不稳定性. 传统无监督特征选择算法是针对独立同分布数据，在媒体数据流中效果不明显. UFSDUC 算法利用关联信息进行无监督动态媒体数据的特征选择，当特征信息较少时关联信息可以更好地为特征选择补充数据信息.

表 5　SinaWeibo 数据集中不同特征选择算法的性能比较

	ACC 值								
	20%(250)	30%(260)	40%(270)	50%(280)	60%(290)	70%(300)	80%(310)	90%(320)	100%(330)
LapScore	25.18	26.39	27.18	29.73	30.73	31.26	30.94	29.62	29.05
SPEC	27.63	27.82	28.57	29.84	31.51	31.67	30.59	30.41	29.48
NDFS	26.68	26.26	29.85	30.27	30.61	32.51	32.63	30.88	30.59
LinkedFS	27.44	27.35	28.78	31.58	32.47	31.84	32.08	31.49	31.55
USFS	31.38	32.07	33.64	32.75	32.81	31.47	30.06	31.85	30.16
UFSDUC	**35.02**	**34.86**	**36.74**	**36.85**	**36.85**	**36.96**	**37.06**	**37.14**	**37.14**
	NMI 值								
	20%(250)	30%(260)	40%(270)	50%(280)	60%(290)	70%(300)	80%(310)	90%(320)	100%(330)
LapScore	0.0682	0.0834	0.1024	0.1057	0.1234	0.1384	0.1294	0.1152	0.1086
SPEC	0.0715	0.0994	0.1085	0.1078	0.1348	0.1407	0.1217	0.1314	0.1242
NDFS	0.0686	0.0954	0.1109	0.1185	0.1274	0.1439	0.1485	0.1341	0.1264
LinkedFS	0.0975	0.1057	0.1094	0.1216	0.1307	0.1224	0.1382	0.1275	0.1251
USFS	0.1258	0.1276	0.1341	0.1435	0.1484	0.1361	0.1307	0.1382	0.1325
UFSDUC	**0.1534**	**0.1364**	**0.1546**	**0.1576**	**0.1586**	**0.1591**	**0.1594**	**0.1624**	**0.1642**

表 6　Flickr 数据集中不同特征选择算法的性能比较

	ACC 值								
	20%(650)	30%(660)	40%(670)	50%(670)	60%(670)	70%(670)	80%(670)	90%(670)	100%(670)
LapScore	22.44	22.39	21.18	20.73	19.73	17.26	15.94	15.62	14.05
SPEC	23.63	23.82	23.57	21.84	21.51	20.67	19.59	17.41	16.48
NDFS	22.68	23.26	21.85	21.27	20.61	18.51	18.63	17.88	17.59
LinkedFS	25.34	26.12	24.55	22.49	21.65	20.08	20.76	19.61	18.45
USFS	26.48	26.76	25.16	24.08	25.74	**28.04**	25.62	25.89	25.73
UFSDUC	**29.02**	**28.86**	**28.84**	**27.85**	**27.85**	27.85	**27.85**	27.85	**27.85**
	NMI 值								
	20%(650)	30%(660)	40%(670)	50%(670)	60%(670)	70%(670)	80%(670)	90%(670)	100%(670)
LapScore	0.0816	0.0752	0.0686	0.0523	0.0511	0.0476	0.0412	0.0383	0.0323
SPEC	0.0945	0.0974	0.0728	0.0673	0.0614	0.0573	0.0488	0.0413	0.0402
NDFS	0.0867	0.0945	0.0685	0.0634	0.0586	0.0423	0.0476	0.0332	0.0301
LinkedFS	0.1163	0.1244	0.1362	0.1251	0.1105	0.1046	0.1184	0.0864	0.0781
USFS	0.1232	0.1271	0.1378	0.1276	0.1325	**0.1485**	0.1274	0.1297	0.1341
UFSDUC	**0.1367**	**0.1456**	**0.1394**	**0.1374**	**0.1374**	0.1374	**0.1374**	**0.1374**	**0.1374**

表 7　BlogCatalog 数据集中不同特征选择算法的性能比较

	ACC 值								
	20%(500)	30%(510)	40%(520)	50%(530)	60%(540)	70%(540)	80%(540)	90%(540)	100%(540)
LapScore	24.67	25.17	25.04	24.46	22.85	21.06	20.04	19.16	19.22
SPEC	23.35	24.81	25.37	23.85	22.18	21.47	19.82	18.08	17.44
NDFS	25.84	26.06	26.34	24.87	23.31	21.35	20.44	19.32	20.46
LinkedFS	28.35	27.12	26.03	24.16	23.45	22.61	22.03	20.74	19.32
USFS	29.11	29.04	28.62	26.73	26.21	25.85	26.07	26.34	26.25
UFSDUC	**30.42**	**28.15**	**29.84**	**28.05**	**27.74**	**27.48**	**27.48**	**27.48**	**27.48**
	NMI 值								
	20%(500)	30%(510)	40%(520)	50%(530)	60%(540)	70%(540)	80%(540)	90%(540)	100%(540)
LapScore	0.0734	0.0834	0.0747	0.0615	0.0585	0.0441	0.0417	0.0405	0.0539
SPEC	0.0805	0.0804	0.0834	0.0591	0.0524	0.0424	0.0408	0.0377	0.0338
NDFS	0.0957	0.1005	0.1012	0.0794	0.0701	0.0718	0.0547	0.0388	0.0445
LinkedFS	0.1236	0.1207	0.1125	0.1046	0.0975	0.0764	0.0731	0.0622	0.0541
USFS	0.1434	0.1364	0.1272	0.1127	0.1004	0.0817	0.0915	0.0947	0.0831
UFSDUC	**0.1637**	**0.1391**	**0.1408**	**0.1355**	**0.1239**	**0.1139**	**0.1139**	**0.1139**	**0.1139**

对于 LapScore 算法、SPEC 算法、LinkedFS 和 NDFS 算法来说，特征数据超过一定值会导致聚类性能降低，而 USFS 算法和 UFSDUC 算法都是处理动态数据的无监督特征选择算法，聚类性能随数据特征增加能够保持相对稳定. UFSDUC 算法基于用户的关系类型进行特征选择，效率要高于 USFS 算法，并且处理大量动态媒体数据特征时，能存储部分有效相关特征，性能稳定.

根据观察可知，在 3 个媒体数据集上，UFSDUC 算法在实验开始部分就接受新特征，到实验后期获

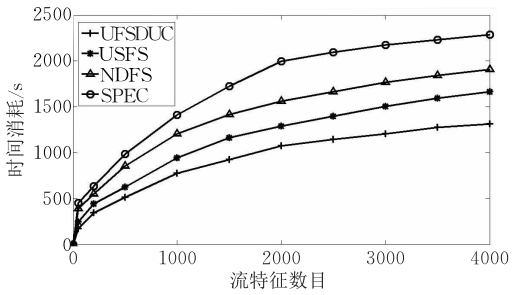
取新特征更新最优特征子集就会变得越来越困难，因为现有数据已能够提供足够代表特征子集的信息.当最优特征数据达到最大数量时，新特征将不再被接受.例如，在三个数据集中，当特征数据分别达到320、670和540时模型不再接受新特征，甚至因为数据量较大、时间消耗多而出现聚类性能退化的现象.从实验结果可以看出，LapScore算法、SPEC算法、NDFS算法、LinkedFS和USFS算法，随着特征数量的增加而导致聚类性能降低，而UFSDUC算法在媒体数据流的特征数据达到一定值时聚类性能保持稳定.以Flickr数据集为例，数据比例从50%变化到100%时，聚类结果相对平稳.

4.3.3 时间评估

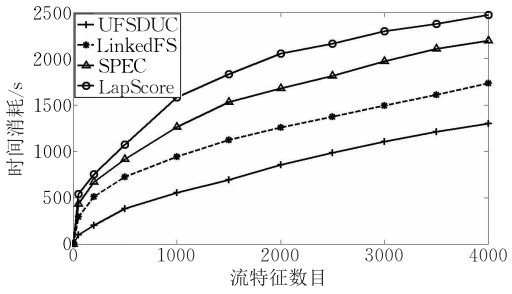
比较六种不同算法的运行时间来验证本文提出UFSDUC算法的有效性.因为LapScore、SPEC、NDFS和LinkedFS算法不是针对动态数据而设计，所以在每个时间步长内都需要运行一下特征选择过程.USFS算法和UFSDUC算法虽然都是面向动态数据而设计，但是本文的算法根据不同的关系类型进行有针对性的特征选择，时间效率较高.由于对比算法过多，所以本节在三个不同的数据集中进行时间评估时，会任选三种对比算法与本文提出的算法进行比较.我们设置的累积运行时间阈值大约是2500 s，五种对比算法的运行时间均超过1400 s.可以观察到所本文提出的UFSDUC算法明显优于其它算法，每个特征的平均运行时间分别是0.46 s～0.55 s之间.当其它算法的累积运行时间到达阈值时，记录UFSDUC的累积运行时间，结果显示在图5.

此外，计算UFSDUC算法在3个数据集上执行特征选择时获取特征子集的建模时间和聚类时间，如图6所示.UFSDUC的建模时间和聚类时间明显低于其它特征选择算法，平均时间消耗分别约为14.27 s和18.37 s.优于其它算法的主要原因是根据用户相关性构造特征选择模型能节约时间，提高准确度，达到最优聚类效果.

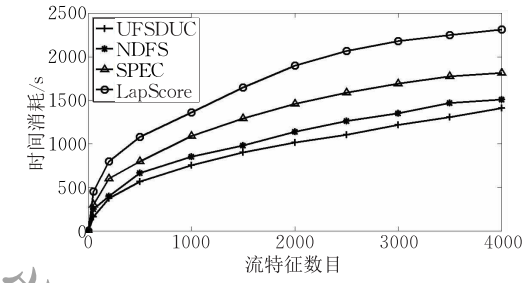
同时，本文将提出的UFSDUC算法与对比算法进行复杂度的比较.在五种对比算法中选取的三种有代表性的算法，分别是无监督算法SPEC，研究相关用户算法LinkedFS和研究数据流算法USFS.SPEC算法的渐进时间复杂度是$O(n^3r)$，LinkedFS算法的渐进时间复杂度是$O(n^2k+nN)$，USFS算法的渐进时间复杂度是$O(n^2m+n)$，本文的UFSDUC算法的渐进时间复杂度是$O(n^2s^2t)$.通过实验结果与分析可知，本文算法的时间复杂度优于对比算法，说明本文算法在实验运行中具有一定的优越性.



(a) SinaWeibo数据集上的运行时间

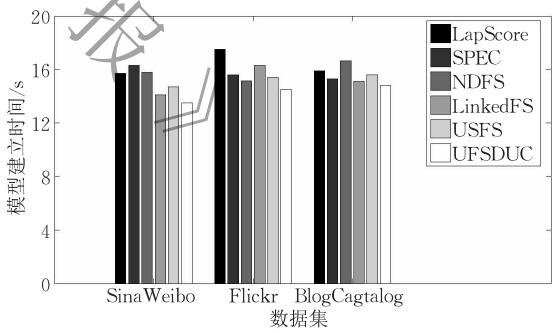


(b) Flickr数据集上的运行时间

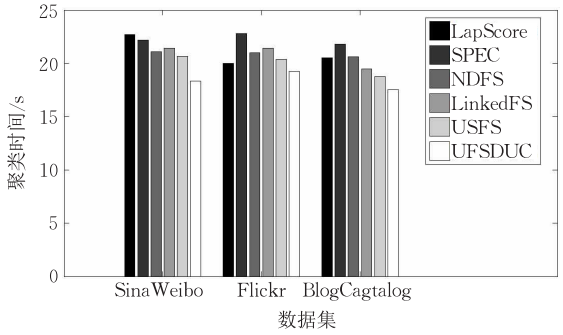


(c) BlogCatalog数据集上的运行时间

图5 数据流特征选择的时间消耗对比



(a)建立模型时间



(b)聚类时间

图6 建模时间和聚类时间

4.4 关系强度影响

本节研究用户关系强度对 UFSDUC 算法性能的影响，实验中特征数目设定为 300，为了节约空间只介绍 T_5 和 T_{100} 的实验结果，如表 8～表 10 所示.

表 8　SinaWeibo 数据集中关系强度对 UFSDUC 的影响

度量方法		UFSDUC	
		T_5	T_{100}
无度量方法		38.18	47.05
结构度量	NCF	42.71	54.82
	Jaccard	42.04	54.95
	KS	42.88	53.76
内容度量		43.87	54.29
相互影响度量		45.92	55.77
混合度量	$\theta_1=0.4,\theta_2=0.6$	46.97	55.85
	$\theta_1=0.7,\theta_2=0$	48.06	57.16
	$\theta_1=0,\theta_2=0.5$	48.27	57.46
	$\theta_1=0.3,\theta_2=0.6$	47.96	57.55

表 9　Flickr 数据集中关系强度对 UFSDUC 的影响

度量方法		UFSDUC	
		T_5	T_{100}
无度量方法		42.64	51.35
结构度量	NCF	44.65	53.54
	Jaccard	44.81	53.01
	KS	44.22	53.28
内容度量		45.53	54.85
相互影响度量		46.98	55.64
混合度量	$\theta_1=0.3,\theta_2=0.7$	47.37	56.62
	$\theta_1=0.1,\theta_2=0$	47.98	56.26
	$\theta_1=0,\theta_2=0.4$	47.65	56.89
	$\theta_1=0.1,\theta_2=0.4$	48.17	58.54

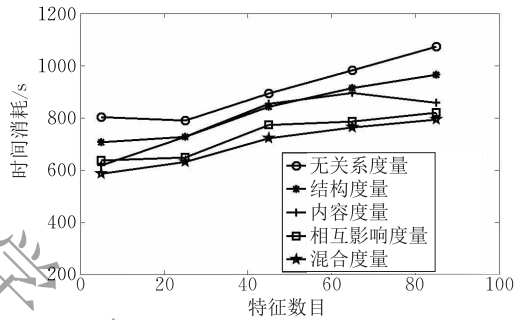
表 10　BlogCatalog 数据集中关系强度对 UFSDUC 的影响

度量方法		UFSDUC	
		T_5	T_{100}
无度量方法		40.43	49.91
结构度量	NCF	43.71	52.06
	Jaccard	43.56	52.78
	KS	43.96	52.24
内容度量		44.74	53.61
相互影响度量		45.74	54.27
混合度量	$\theta_1=0.2,\theta_2=0.8$	46.54	55.34

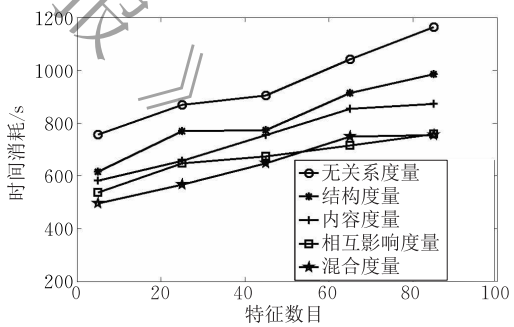
在 SinaWeibo、Flichr 和 BlogCatalog 数据集上分别系统评价结构度量、内容度量、相互影响度量对实验性能的影响. 其中对混合度量的实验设置有所不同，BlogCatalog 数据集的 θ_1 步长为 0.1，从 0 逐渐递增到 1，且 $\theta_2=1-\theta_1$，SinaWeibo 和 Flickr 数据集中 θ_1 和 θ_2 都设为步长为 0.1，从 0 逐渐递增到 1.

从整体来看，有关系强度的 UFSDUC 算法性能始终优于无关系强度的 UFSDUC 算法，在 SinaWeibo 数据集中，混合度量(Hybrid Measure)可以得到 6.11%的相对提高. 数据集 T_{100} 比 T_5 实验效果好，选择特征数据越多越有利于实验结果的准确度. 结构度量(Structural Measure)的三个代表性方法中 KS 的执行效果最好，Jaccard 和 NCF 是定义用户的本地网络(本地信息)，而 KS 是计算整个社交网络(全局信息)，这说明在关系强度预测方面对于结构度量全局信息比本地信息更重要. 相互作用度量获得良好的计算性能说明用户之间的相互作用越多就越可能具有相似兴趣. 在关系强度预测方面内容度量非常重要，也暗示了这种度量的方法实现了较高的精确度. 混合度量在所有度量方法中执行效果最好. 在 Flickr 和 BlogCatalog 数据集里，适当的结合度量方法能够有效提高实验性能. 当结构度量、内容度量和相互作用度量三种方法结合时最好性能得到实现，暗示了这三种方法彼此信息互补. 这些观察表明，多种关系强度预测与用户关系类别相结合能够提高 UFSDUC 算法的特征选择性能.

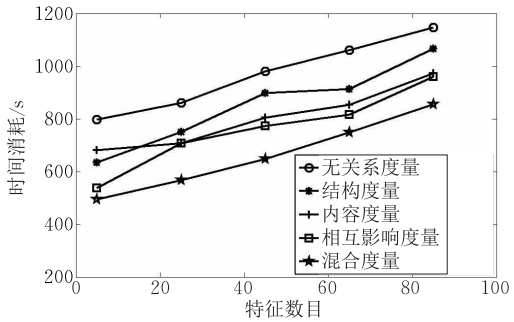
三个数据集中各关系强度对运行时间的影响结果如图 7 所示，从图可知没有考虑关系强度的算法



(a) Sina Weibo数据集中各方法时间对比



(b) Flickr数据集中各方法时间对比



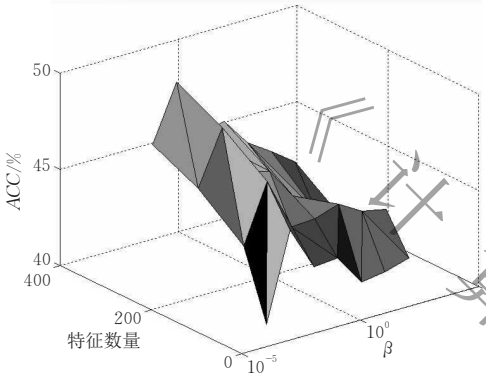
(c) BlogCatalog数据集中各方法时间对比

图 7　关系强度的时间消耗比较

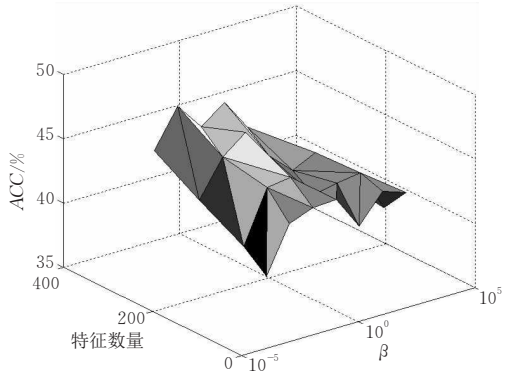
运行时间比具有关系强度的算法运行时间长，在几种度量方法中混合度量的运行时间最短，实验效果最好，另外三种关系度量方法的运行时间相差不大.

4.5 参数有效性分析

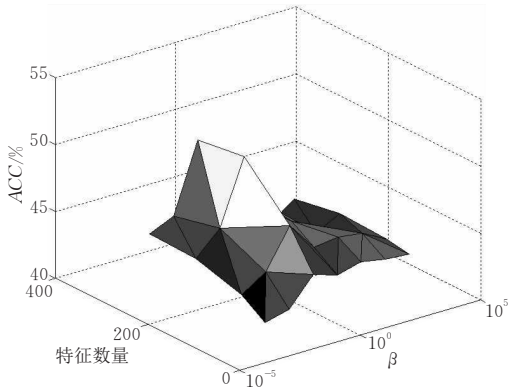
UCFS 算法里一个重要参数是 β，它决定用户关系对特征选择的影响，能够显示用户关系类型的重要程度. 另一个重要参数是被选特征的数量，它将影响模型的运行时间，进而影响模型性能. 在本小节分别研究 MFS、SFM、FEO 三种用户关系类型是如何随着 β 和被选特征数据量的变化而影响 UCFS 的性能，β 分别取值{0.001,0.01,0.1,1,10,100,1000}，特征数据量分别是{50,100,200,300}.



(a) β对MFS用户关系的影响



(b) β对SFM用户关系的影响



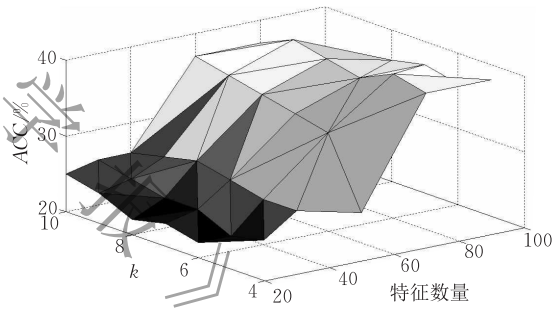
(c) β对FEO用户关系的影响

图 8 β 对不同用户关系类型的影响

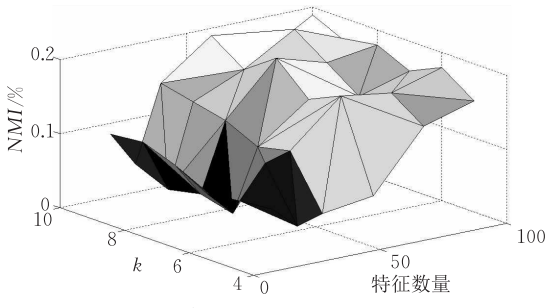
为了节约空间，仅针对 SinaWeibo 数据集进行参数研究. 观察图 8 可以发现，随着 β 的增大三种关系类型的 *ACC* 值都是先增加后降低，其中 MFS 和 SFM 两种用户关系类型实验结果相似，几乎都在 $\beta=0.01$ 时达到峰值，FEO 在 $\beta=0.1$ 时到达峰值. 所以 FEO 的稳定性能要强于 MFS 和 SFM. 与参数 β 相比，特征数量对特征选择性能的影响更加敏感，目前还没有较明确的算法来准确描述如何确定特征数量以达到性能最优.

接下来讨论 UFSDUC 算法中的三个重要参数：社交潜在因素数目 k，参数 α 和 β. 为了研究这三个参数的有效性，每次仅改变一个参数，另外两个参数固定不变，随着特征数据量的变化观察参数对特征选择的影响，同样在 SinaWeibo 数据集中进行参数研究.

本文首先改变社交因子 k，使其从 5 到 10 依次变化，另外两个参数 $\alpha=10$，$\beta=0.1$，根据 *ACC* 和 *NMI* 判断聚类性能结果如图 9 所示. 当社交隐藏因子数目接近于集群数目时聚类效果最好，在 SinaWeibo 数据集中 $k=9$ 时实验结果最好.



(a) *ACC*

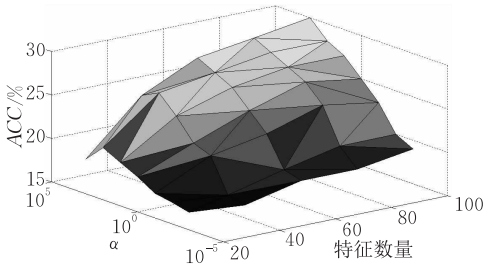


(b) *NMI*

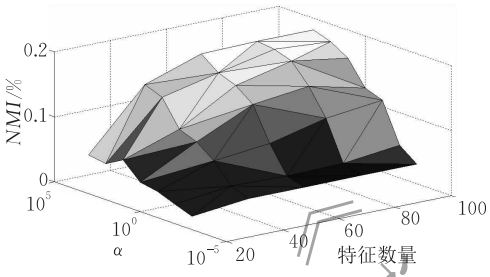
图 9 k 对特征选择的影响

评定参数 α 控制模型稀疏性的效果，改变 α 分别为{0.001,0.01,0.1,1,10,100,1000}，固定 $k=6$，$\beta=0.1$，参数 α 和特征数目之间的性能差异显示如图 10 所示. 随着 α 的增加，聚类性能快速增加，当达到 10～1000 之间保持相对稳定. α 表明新特征不

容易通过梯度测试，因此已被选择的特征彼此之间更相关且更有意义.



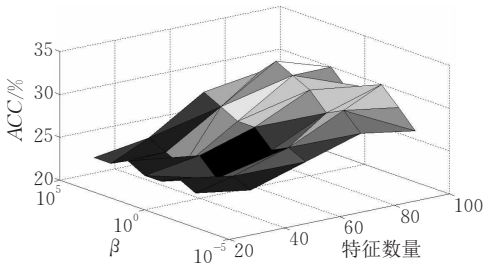
(a) *ACC*



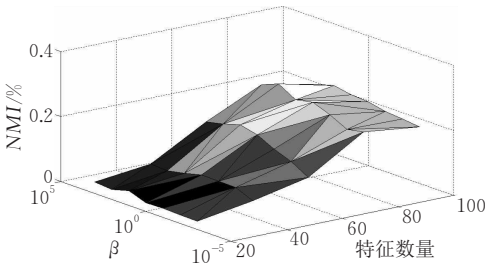
(b) *NMI*

图 10　α对特征选择的影响

研究参数 β 对模型稳健性的影响，与 α 类似 β 分别为{0.001,0.01,0.1,1,10,100,1000}，固定 $k=6$，$\alpha=10$，结果显示在图 11，可以看到特征数目对聚类性能的影响要大于 β 对聚类性能的影响.



(a) *ACC*



(b) *NMI*

图 11　β对特征选择的影响

5　结　论

特征选择是数据挖掘领域的重要技术，而社交媒体数据的高维性和动态性会干扰特征选择的计算效率，目前现有的无监督特征选择算法基本是面向静态数据，不适用于动态网络媒体数据. 本文提出的基于用户相关性的流媒体数据无监督特征选择算法有效解决了上述问题. 利用媒体用户之间的关联将用户关系分成三种类别，并对三种用户类别进行数学建模. 然后基于弗罗贝尼乌斯范数和强弱关系预测相结合获取最优化相关性特征矩阵. 最后引入时间变量进行流媒体数据的无监督特征选择，实时计算非零权重数值，提取较大值构造最佳特征子集. 在真实数据集中与传统无监督特征选择算法相比，本文提出的算法能在较短时间内对数据进行聚类，精确度和稳定性都可以达到较高水平，对媒体用户和流媒体数据的研究有重要价值. 下一步的主要研究将结合本文提出的算法，通过特征选择和机器学习研究媒体用户的情感倾向，解决网络用户管理和高维数据管理的难题.

参　考　文　献

[1] Gu Quan-Quan, Li Zhen-Hui, Han Jia-Wei. Generalized fisher score for feature selection//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Catalina Island, USA, 2010: 266-273

[2] Peng Han-Chuan, Long Fu-Hui, Ding Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[3] Nie Fei, Huang Heng, Cai X. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization//Proceedings of the 26th International Conference on Data Engineering. Chicago, USA, 2010: 1813-1821

[4] Deng Cai, Zheng Chi-Yuan, He Xiao-Fei. Unsupervised feature selection for multi-cluster data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Columbia, USA, 2010: 333-343

[5] Lui Huan, Motoda H. Computational Methods of Feature Selection. London, UK: CRC PRESS, 2008

[6] Nie Fei-Ping, Xiang Shi-Ming. Trace ratio criterion for feature selection//Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Chicago, USA, 2008: 671-676

[7] Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of relieff and rrelieff. Machine Learning, 2003, 53(1): 23-69

[8] Dy J G, Brodley C E. Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(3): 373-378

[9] He Xiao-Fei, Cai Deng, Niyogi Partha. Laplacian score for feature selection//Proceedings of the Advances in Neural Information Processing Systems. Columbia, Canada, 2006:

507-514

[10] Zhao Zheng, Liu Huan. Spectral feature selection for supervised and unsupervised learning//Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA, 2007: 1151-1157

[11] Li Ze-Chao, Yang Yi. Unsupervised feature selection using nonnegative spectral analysis//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 1026-1032

[12] Li Ze-Chao, Liu Jing, Yang Yi, Zhou Xiao-Fang. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2138-2150

[13] Li Ze-Chao, Tang J H. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Transactions on Image Processing, 2015, 24(12): 5343-5355

[14] Argyriou A, Evgeniou T, Massimiliano Pontil. Multi-task feature learning//Proceedings of the Neural Information Processing System. Cambridge, UK, 2007: 41-48

[15] Liu Jun, Ji Shui-Wang, Ye Jie-Ping. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization//Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal, Canada, 2009: 339-348

[16] Zhao Zheng, Wang Lei. Efficient spectral feature selection with minimum redundancy//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Georgia, USA, 2010: 1-6

[17] Yang Yi, Shen Heng-Tao. $L_{2,1}$-norm regularized discriminative feature selection for unsupervised learning//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1589-1594

[18] Tang Ji-Liang, Liu Huan. Feature selection with linked data in social media. SDM, 2012, 16(2): 118-128

[19] Tang Ji-Liang, Liu H. Feature selection for social media data. ACM Transactions on Knowledge Discovery from Data, 2014, 8(4): 19-46

[20] Tang Ji-Liang, Liu Huan. Unsupervised feature selection for linked social media data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery. Beijing, China, 2012: 904-912

[21] Zhou Jing, Foster Dean. Streaming feature selection using alpha-investing//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2005: 384-393

[22] Wu Xin-Dong, Yu Kui, Wang Hao. Online streaming feature selection//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 1159-1166

[23] Guo T, Zhu X Q. Snoc: Streaming network node classification//Proceedings of the IEEE International Conference on Data Mining. Shenzhen, China, 2014: 150-159

[24] Tang Ji-Liang, Liu Huan. Unsupervised streaming feature selection in social media//Proceedings of the ACM International Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1041-1050

[25] McPherson M, Lovin L S, Cook J M. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001, 27(1): 415-444

[26] Marsden P V, Friedkin N E. Network studies of social influence. Sociological Methods and Research, 1993, 22(1): 127-151

[27] Morris S A. Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. Journal of the Association for Information Science and Technology, 2005, 56(12): 1250-1273

[28] Airoldi E M, Blei D M, Fienberg S E. Mixed membership stochastic blockmodels. Machine Learning Research, 2008, 12(6): 33-40

[29] Gopalan P, Gerrish S M. Scalable inference of overlapping communities. Advances in Neural Information Processing Systems, 2012, 3(21): 2249-2257

[30] Tang Ji-Liang, Wang Xu-Fei, Liu Huan. Integrating social media data for community detection//Proceedings of the International Conference on Modeling and Mining Ubiquitous Social Media. Boston, USA, 2011: 1-20

[31] Macskassy S A, Provost F. Classification in networked data: a toolkit and a univariate case study. Machine Learning Research, 2007, 8(3): 935-983

[32] Gao Hui-Ji, Tang Ji-Liang, Liu Huan. Exploring social-historical ties on location-based social networks. Association for the Advancement of Artificial Intelligence, 2012, 5(12): 104-115

[33] Tang Ji-Liang, Gao Hu-Ji, Liu Huan. mTrust: Discerning multi-faceted trust in a connected world//Proceedings of the ACM International Conference on Web Search and Data Mining. Washington, USA, 2012:93-102

[34] Xiang, Neville J, Rogati M. Modeling relationship strength in online social networks//Proceedings of the International Conference on World Wide Web. North Carolina, USA, 2010: 981-990

[35] Perkins S, Lacker K, Theiler J. Grafting: Fast, incremental feature selection by gradient descent in function space. Machine Learning Research, 2003, 3(3): 1333-1356

[36] Boyd S, Vandenberghe L. Convex optimization. IEEE Transactions on Automatic Control, 2006, 51(11): 1859

REN Yong-Gong, born in 1972, Ph. D., professor. His research interests include database technology, data mining and intelligent information calculation.

WANG Yu-Ling, born in 1990, M. S. candidate. Her research interest is data mining.

LIU Yang, born in 1986, Ph. D., lecturer. His research interests include digital image and radar retrieval.

ZHANG Jing, born in 1984, Ph. D., lecturer. Her research interests include pattern classification and computer vision.

Background

In recent years, social network software becomes increasingly popular among young people. In social network, users can make friends, sharing their writings or mutual comment on Weibo and so on. The explosive growth of social media users brings about massive amounts of high-dimensional data. Feature extraction is effective in preparing high dimensional data for data analytics. As a powerful tool to solve Big Data, feature extraction has been widely studied in recent years.

Most traditional algorithms are based on supervised feature selection, which is not always available. What we research in the real social media data is often lack of class label. So this paper mainly studies the unsupervised feature selection. In addition to high-dimensional and complexity of data, dynamic is another important problem to be solved. Comparing to static data, researching streaming media data is more significant and widespread. Although there are many studies about streaming data, those algorithms are mainly aim to supervised feature selection. However, streaming media data is constantly update with time and the processing streaming data efficiency's limited by the application of traditional feature selection algorithm.

Aiming above problem, we proposed a novel location recommendation algorithm that unsupervised feature selection algorithm for streaming media based on user correlation. As the connection, user correlation can solve the problem of unsupervised feature selection lack of class labels. In our paper, user relationship can be divided into three types that can be considered the constraint condition. In social media environment, there exist immense amount of users and media objects, and the media data are rapidly changing every day. To tackle this problem above, we set the time interval and each time segment only dynamic generates a characteristic data. Utilize gradient descent to calculate the threshold value of streaming media data that can measure the new feature. Real-time calculate non-zero feature weights and extract greater value construct optimal feature subset.

This work is supported by the National Natural Science Foundation of China (No. 61373127), the Program for Liaoning Excellent Talents in University (No. LR2015033), the Science and Technology Plan Project of Liaoning Province (No. 2013405003), and the Science and Technology Plan Project of Dalian (No. 2013A16GX116).