



一种用于网络用户行为聚类的标签自动生成方法

毕 猛¹, 邵 中¹, 徐 剑²

(1. 沈阳工业大学 软件学院, 沈阳 110023; 2. 东北大学 软件学院, 沈阳 110169)

摘 要: 针对目前多数聚类算法需要事先确定网络用户行为数据规模以及生成的簇标签缺乏明确语义的问题, 提出一种用于网络用户行为聚类分析的簇标签自动生成方法。应用潜在因子模型和矩阵分解方法对原始网络用户行为数据补充缺失值, 根据网络用户行为数据的属性特征进行用户行为聚类并在聚类过程中增加行为特征, 同时利用行为特征信息产生簇标签以提高网络用户行为的聚类准确性。在 Last.fm、Movielens 和 CiteULike 数据集上的实验结果表明, 该方法无需事先确定网络用户行为数据规模, 并且可在保证较高聚类准确率的前提下自动生成语义更明确的簇标签。

关键词: 用户行为; 聚类; 潜在因子模型; 矩阵分解; 标签

开放科学(资源服务)标志码(OSID):



中文引用格式: 毕猛, 邵中, 徐剑. 一种用于网络用户行为聚类的标签自动生成方法[J]. 计算机工程, 2020, 46(10): 81-87.

英文引用格式: BI Meng, SHAO Zhong, XU Jian. An automatic label generation method for clustering of network user behavior[J]. Computer Engineering, 2020, 46(10): 81-87.

An Automatic Label Generation Method for Clustering of Network User Behavior

BI Meng¹, SHAO Zhong¹, XU Jian²

(1. Software College, Shenyang University of Technology, Shenyang 110023, China;

2. Software College, Northeastern University, Shenyang 110169, China)

[Abstract] Existing user behavior clustering methods require the determined size of user behavior data, and the generated cluster labels lack explicit semantics. To solve these problems, this paper proposes an automatic cluster label generation method for clustering analysis of network user behavior. The method applies the Latent Factor Model (LFM) and matrix decomposition method to the raw data of network user behavior for missing value processing. Based on the attribute features of user behavior data, the user behavior cluster is performed and behavior features are added during clustering. At the same time, cluster labels are generated based on behavior feature information to improve the accuracy of user behavior clustering. Experimental results on datasets of Last.fm, Movielens and CiteULike show that the proposed method does not require the determined size of user behavior data, and can automatically generate cluster labels with more explicit semantics while keeping a high clustering accuracy.

[Key words] user behavior; clustering; Latent Factor Model (LFM); matrix decomposition; label

DOI: 10.19678/j.issn.1000-3428.0058973

0 概述

随着云计算、物联网、移动计算等技术的快速发展, 微博、微信及网络直播等自媒体社交公众平台不断涌现, 网络用户正在实现从信息消费者向信息

造者的转变, 然而网络是社会现实的镜像, 网络用户行为不仅会影响虚拟网络, 而且会直接反作用于真实社会, 影响人们的日常生活^[1-3]。对于庞大的网络用户群, 其所表现出的行为各具特点, 若要向这些用户群提供高效、安全、个性化的网络服务, 则需要对

基金项目: 国家自然科学基金(61872069); 中央高校基本科研业务费专项资金(N2017012)。

作者简介: 毕 猛(1982—), 男, 工程师、博士, 主研方向为网络与信息安全、机器学习、数据聚类分析; 邵 中, 副教授; 徐 剑, 副教授、博士生导师。

收稿日期: 2020-07-17

修回日期: 2020-08-17

E-mail: bim@sut.edu.cn

网络用户行为进行分析并建立有效的分析模型。网络用户行为分析综合运用统计学、人工智能、大数据、心理学、社会学以及其他与网络用户行为分析密切相关的理论和方法,对网络用户的构成、特点及其行为活动所表现出的内在规律进行分析,进而对网络用户的行为进行预测、管理和控制,达到为政府、企业及个体用户服务的目的^[4-6]。

聚类算法是数据挖掘的一个重要研究领域,其能够以较高的效率获得全局范围内的数据分布特征,已在网络用户行为分析中得到广泛应用^[7-9],但是目前典型的聚类算法(如 K-means 算法等)更多关注聚类效果,并且在网络用户行为分析过程中存在用户行为数据规模需在聚类前确定,用户行为数据聚类后生成的簇无明确语义的问题。

对于以 K-means 为代表的聚类算法,需要事先确定用户行为数据的规模(即簇的个数)才能进行聚类^[10],但在实际应用中,用户行为数据规模很难确定,并且聚类后所产生的簇个数通常为未知,甚至在数据快速增长时簇的个数也可能发生变化。因此,此类聚类算法在面对海量网络用户行为数据时通常聚类效果较差。另外,由于聚类一般不需要有标签的数据,因此生成的簇也没有标签,即没有明确的语义,也无法获知每一个簇所代表的类别^[11]。然而,在实际应用中如果无法逐一遍历每个簇内的数据,则不能确定簇所代表的类别。目前,除了文本聚类算法利用文字出现频率作为标签外,多数聚类算法无法为簇生成有效的标签。而对于海量网络用户行为数据而言,如果生成的簇无标签,则需要用户为每个簇人工赋予标签来达到用户行为分析的目的,但是该任务的工作量非常庞大,而且人工语义标签的生成效率也很低。

针对上述问题,本文结合潜在因子模型(Latent Factor Model, LFM)^[12]和矩阵分解等方法,提出一种用于网络用户行为聚类的簇标签自动生成方法。该方法包括用户行为数据缺失值处理、网络用户行为聚类过程中的簇标签自动生成以及簇标签评价等具体过程,并利用 AP^[13]和 DP-means^[14]等聚类算法及实际用户行为数据进行实验验证。

1 相关工作

目前,已有众多学者利用聚类算法对用户行为进行研究。文献[15]利用 K-means 算法并结合关联规则对银行客户进行聚类以生成簇标签。文献[16]通过分层聚类与卡方检验对文件进行聚类并生成标签。文献[17]将层次聚类算法与文本关键词抽取算法相结合生成微博用户的标签,该标签不包含同义标签,可以从多方面体现出网络用户的兴趣。文献[18]对 K-means 算法进行改进,解决了其依赖初

始聚类中心的问题,在此基础上设计一种面向网络用户行为的喜好标签聚类系统。文献[19]采用文本聚类算法对网络用户的文本内容进行聚类生成标签,并将数据处理为矩阵形式,同时利用匈牙利算法提高生成标签的准确性。文献[20]基于支持向量聚类算法提出一种快速且稳定的簇标签生成方法来提高聚类过程的准确性,解决了因支持向量聚类算法时间复杂度导致生成簇标签准确率降低的问题。

虽然上述方法可以在聚类时生成簇标签,但是其都是针对具体应用场景提出的解决方案,缺乏通用性。为此,本文提出一种用于网络用户行为聚类的簇标签自动生成方法,以解决目前多数聚类算法需要事先确定用户行为数据规模以及生成的簇缺乏明确语义的问题。

2 设计思想

本文依据用户行为数据的属性特征,对用户行为进行聚类,并且在聚类过程中增加行为特征作为参与元素,同时利用这些行为特征信息产生簇标签,提出一种用于网络用户行为聚类的簇标签自动生成方法,具体流程如图 1 所示。

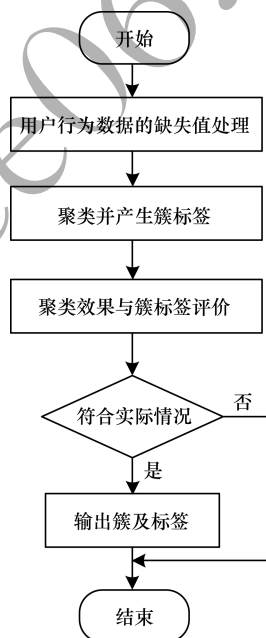


图 1 用于网络用户行为聚类的簇标签自动生成流程

Fig. 1 Procedure of automatic cluster label generation method for clustering of network user behavior

用于网络用户行为聚类的簇标签自动生成方法的具体步骤如下:

步骤 1 用户行为数据的缺失值处理。在当前网络环境下,收集用户行为及其特征。以用户在网络上收听歌曲为例,如果用户通过网络收听说唱类音乐,则说明该用户行为具有时尚、潮流等特征,然而搜集的特征数据存在缺失值情况,例如用户对某

首歌曲的收听次数通常与喜欢程度成正比,但是未听过该歌曲并不代表不喜欢该歌曲,由于搜集的数据通常只包含用户听过的少部分歌曲,而没有听过的歌曲对原始数据集而言就是缺失值,因此需要对存在缺失值的数据做预处理,从而找到缺失值。

步骤2 在用户行为聚类过程中产生簇标签。将用户行为与行为特征相结合后可对聚类后的簇添加标签,使其具有较明确的语义,例如对于购买最新款iPhone手机的某一类人群可添加时尚、潮流等标签。

步骤3 聚类效果及簇标签评价。评价聚类效果并对生成的簇标签进行检验,同时判断是否符合实际的用户聚类情况。

步骤4 输出符合要求的簇及其对应的标签。

3 用于网络用户行为聚类的簇标签自动生成

3.1 符号描述

在用户行为聚类过程中使用的相关符号及其定义如表1所示。

表1 符号及其定义
Table 1 Symbols and their definitions

符号	定义
m	用户数量
n	用户行为数量
t	簇标签数量
r	簇数量
k	矩阵分解的潜在因子数量
D	用户-行为矩阵
p	用户-潜在因子矩阵
p_i	p 的第 i 个用户
q	行为-潜在因子矩阵
q_i	q 的第 i 个潜在因子
G	用户行为-标签矩阵
H	潜在因子-标签矩阵

3.2 簇标签自动生成过程

3.2.1 用户行为数据的缺失值处理

由于网络中收集的用户行为数据会存在缺失值,因此本文采用潜在因子模型(Latent Factor Model, LFM)对原始用户行为数据进行预处理。LFM是一种矩阵分解算法,将原始高维数据分解为潜在因子矩阵,从而实现稀疏矩阵分解和数据降维。

给定一个大小为 $m \times n$ 的稀疏矩阵 d , m 表示用户数目, n 表示行为数目, k 表示潜在因子个数,通过LFM可以得到一个大小为 $m \times k$ 的矩阵 p 和一个大小为 $k \times n$ 的矩阵 q ,如式(1)所示:

$$d_{mn} = p_{mk} \times q_{kn} \quad (1)$$

LFM计算的核心思想为在迭代过程中,通过优化数据集的矩阵分解误差值来找到合适的 p 与 q ,如式(2)所示:

$$e_{mn} = d_{mn} - q_{nk} \times p_{mk} \quad (2)$$

其中, e_{mn} 表示矩阵分解误差值, k 值若小于原始高维数据的维度,则可实现降维效果。由于通过调整 k 值可以分解出不同的矩阵,因此 k 值会影响矩阵分解的误差结果。

根据收集到的原始数据构建用户-行为矩阵 D ,如式(3)所示:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \approx \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mk} \end{bmatrix} \times \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kn} \end{bmatrix} \quad (3)$$

其中, d_{ij} 表示用户 i 与行为 j 的关系,包括行为执行次数、行为执行程度等,若用户已执行该行为,则不为空,若用户未执行该行为,则为缺失值。由于通过原始数据构建的矩阵 D 为稀疏矩阵,其包含缺失值且维度较高,因此会增加聚类难度及降低聚类效果。

构建用户行为-标签矩阵 G ,如式(4)所示:

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mn} \end{bmatrix} \quad (4)$$

其中,矩阵 G 为包含缺失值的稀疏矩阵,行表示用户,列表示用户行为特征,共有 m 个用户和 n 种行为。

由于 D 为稀疏矩阵,且利用LFM通过潜在因子对矩阵 D 进行矩阵分解可以得到低维矩阵 p 、 q ,因此将 p 、 q^T 相乘可以补全矩阵 D ,如式(5)所示:

$$D \approx p \times q^T \quad (5)$$

对于矩阵 q ,通过式(6)计算得到潜在因子-标签矩阵 H ,并以此作为簇标签自动生成算法的输入。由于 p 为用户-潜在因子矩阵,且潜在因子个数小于 D 的原始维数,因此矩阵分解能够实现降维的效果。

$$H \approx q^T \times G \quad (6)$$

3.2.2 簇标签自动生成算法

簇标签自动生成算法的具体过程如下:

步骤1 将用户-行为矩阵 D 分解为 p 、 q 两个矩阵,再根据 $q \times G$ 构建用户行为潜在因子-标签矩阵 H 。

步骤2 使用聚类算法对矩阵 p 进行聚类,取每个簇的平均值,得到每个簇的簇中心及用户聚类矩阵 C 。

步骤3 将用户聚类矩阵 C 和潜在因子-标签矩阵 H 相乘得到用户标签矩阵 T ,对矩阵 T 内的数值进行排序选出前 t 项作为用户标签。

算法 1 簇标签自动生成算法输入 用户-行为矩阵 D 、用户行为-标签矩阵 G 输出 簇-标签矩阵 Q

```

1.  $(p, q) \leftarrow \text{matrix factorization}(D)$ 
    $H \leftarrow q \times G$ 
2.  $C \leftarrow \text{clustering algorithm}(p)$ 
   for  $i \leftarrow 1$  to  $\text{size}(C)$  do
      $S_i \leftarrow \{p_x \in C_i \mid x = 1, 2, \dots, N\}$ 
      $\text{Center}_i \leftarrow \text{mean}(S_i)$ 
   end
    $Q \leftarrow \text{Center}_i \times H$ 
3.  $\text{Seq} = \text{Sort}(Q)$  // 对矩阵  $Q$  进行排序, 得到每个簇的
   // 标签序列
    $\text{Label}_n = \text{Find}(\text{Seq})$  // 查找前  $n$  个标签

```

在簇标签自动生成过程中利用矩阵分解算法找到潜在因子, 补充矩阵 D 中缺失值的特征, 将 D 分解得到矩阵 p, q^T 。使 p, q^T 相乘得到补全缺失值的矩阵 D' , 但由于该矩阵维度过高, 此时聚类效果不理想, 因此对 p 进行聚类, 将 q^T 与 G 相乘得到潜在因子-标签矩阵 H , 如式(7)所示:

$$q^T \times G = H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1t} \\ h_{21} & h_{22} & \cdots & h_{2t} \\ \vdots & \vdots & & \vdots \\ h_{k1} & h_{k2} & \cdots & h_{kt} \end{bmatrix} \quad (7)$$

在 p 聚类后, 通过两种方式对聚类后的簇自动生成标签:

1) 对簇内的数据取平均值得到每个簇的簇中心, 如式(8)和式(9)所示:

$$S_i = \{x \in \text{center}_i \mid p_x\} \quad (8)$$

$$C_i = \text{means}(S_i) \quad (9)$$

其中, C_i 为每一列中的较大值, 表示该簇具有较强的维度特性, 即对应的潜在因子容易被发现, 因此只要将该簇对应的潜在因子-标签关系矩阵相乘, 即可得到簇-标签矩阵 Q , 计算公式如式(10)所示:

$$Q = C \times H = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1t} \\ q_{21} & q_{22} & \cdots & q_{2t} \\ \vdots & \vdots & & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kt} \end{bmatrix} \quad (10)$$

Q 中的每一列均表示一个簇的标签, 例如赋予第一个簇 n 个标签, 可从 $q_{11}, q_{12}, \dots, q_{1t}$ 中取出最大的 n 个值作为该簇的标签, 或是给定一个阈值, 若超过该阈值, 则视为该簇的标签。

2) 找出所有数据的对应标签, 由于 p 表示用户与潜在因子的关系, 且 H 表示潜在因子与标签的关系, 因此 $p \times H$ 可视为用户与标签的关系, 利用该矩阵统计簇内的标签数量, 并将出现次数最频繁的前 n 个标签作为此簇的标签。

3.2.3 簇标签评价算法

在用户行为聚类过程中每个簇都会生成多个标签, 因此需要给出标签评价方法评估标签与簇之间的适合程度。假设 x 属于 p_{test} 中的一个数据样本, 且

x 被分配到簇 c 中, TP 表示 x 和 c 共同拥有的标签数量, FP 表示 x 没有而 c 拥有的标签数量, FN 表示 x 拥有而 c 却没有的标签数量, TN 表示 x 和 c 都没有的标签数量。

F1 score 是簇标签得分, 其计算公式如式(11)所示:

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

其中, Precision 表示准确率, Recall 表示召回率, 其计算公式如式(12)和式(13)所示:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

将矩阵 p 分为 p_{test} 和 p_{train} 两部分, 选取 80% 作为训练数据集进行聚类并给出标签, 20% 作为测试数据集。簇标签评分的具体过程如下:

步骤 1 根据 p_{train} 聚类结果, 利用相同的聚类算法将 p_{test} 分配到 p_{train} 已有的簇内, 同时根据 $p_{\text{test}} \times H$ 得到测试集中用户所代表的标签。

步骤 2 根据式(12)和式(13)计算准确率与召回率。

步骤 3 根据式(11)计算簇标签得分 F1 score。

算法 2 簇标签评分算法输入 p_{train} 和 p_{test} 的聚类结果 C 及其标签矩阵 T , 潜在因子-标签矩阵 H

输出 簇标签得分 F1 score

```

1.  $S \leftarrow p_{\text{test}} \times H$ 
   assign  $p_{\text{test}}$  into  $C$ 
2.  $TP \leftarrow 0, FP \leftarrow 0, FN \leftarrow 0$ 
   for  $i \leftarrow 1$  to  $N$  do:
     for  $j \in \text{cluster}_i$  do:
        $TP \leftarrow \text{count}(TP + S_j \cap T_i)$ 
        $FP \leftarrow \text{count}(FP + S_j \setminus T_i)$ 
        $FN \leftarrow \text{count}(FN + T_i \setminus S_j)$ 
     end
   end

```

$$\text{Precision} \leftarrow \frac{TP}{TP + FP}$$

$$\text{Recall} \leftarrow \frac{TP}{TP + FN}$$

$$\text{F1 score} \leftarrow \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4 实验结果与分析**4.1 实验数据**

本文实验利用 Last.fm 数据集、Movielens 数据集及 CiteULike 数据集进行测试。

Last.fm 数据集包含用户、歌手以及歌手类型的资料, 包括每个用户收听每个歌手音乐的次数以及用户已标记的歌手类型, 并且不同用户可能对同一个歌手标记不同类型。该数据集中具有 1 892 个用户、

17 632 个歌手及 11 946 种歌手类型,采用用户与歌手关系的数据集构建矩阵 D ,歌手与歌手类型关系的数据集构建矩阵 G 。

CiteULike 数据集包含作者引用论文与每篇论文对应的关键字。将作者引用论文表示为 1,未引用视为缺失值,采用作者与作者引用论文数据集构建矩阵 D ,而论文与论文所对应关键字的关系数据集构建矩阵 G 。该数据集中具有 5 551 个作者、16 980 篇论文以及 46 391 个关键字。

MovieLens 数据集包含用户对不同电影的评分数据,包括自 2000 年起 6 040 个用户对 3 900 部电影产生了 1 000 209 条评分数据。将用户已评分电影表示为 1,未评分视为缺失值。采用用户与电影评分的关系数据集构建矩阵 D ,每部电影所代表的类型构建矩阵 G 。

4.2 结果分析

4.2.1 矩阵分解误差评估

如果矩阵分解误差较大,则会对聚类 and 自动标签生成产生影响。因此,首先对本文矩阵分解方法进行评估。以 Last.fm 数据集为例,采用 LFM 对原始稀疏矩阵进行矩阵分解,潜在因子 k 取值为 4、8、12、16、20、24、28、32、36,潜在因子与矩阵分解误差的关系如图 2 所示。可以看出,随着潜在因子数量的增加,矩阵分解误差呈现下降趋势,初始阶段下降速度较快之后放缓,在潜在因子数量达到 32 后,放缓趋势较为明显。因此,潜在因子数量选择 32 较理想。

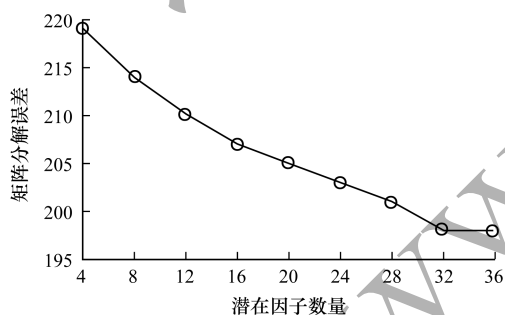


图 2 潜在因子与矩阵分解误差的关系

Fig. 2 The relationship between latent factor and matrix factorization error

4.2.2 聚类效果评估

本文选取潜在因子为 32、矩阵分解误差为 164,分别利用 AP^[13]、DP-means^[14]、K-means^[15]、K-means++^[16]以及层次聚类^[17]算法进行聚类,并使用 Silhouettes 方法对聚类效果进行评估,实验结果如图 3 所示。Silhouettes 用于评价聚类效果的轮廓系数,轮廓系数的取值范围为 $[-1, 1]$,若大于 0 则表示聚类结果可用,若越接近 1 则说明聚类效果越好。在 5 种聚类算法中,

K-means++、DP-means 以及 AP 算法无需事先确定用户行为数据规模,轮廓系数均大于需要实现确定用户行为数据规模的 K-means 与层次聚类算法,聚类效果更好。由于处理的数据集不同,各算法聚类效果存在差异,例如在 Last.fm 数据集中,AP 算法的轮廓系数为 0.354,但在 CiteULike 数据集中轮廓系数为 0.427。虽然整体轮廓系数均未超过 0.6,但本文方法的主要目的是完成聚类过程的标签自动生成,因此该结果是可以接受的。

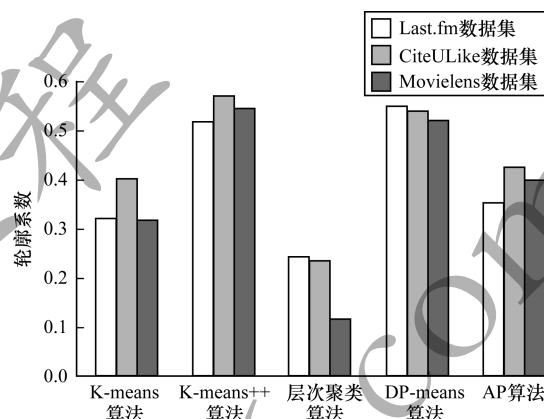


图 3 5 种算法的聚类效果对比

Fig. 3 Comparison of clustering effects of five algorithms

4.2.3 簇标签评分评估

本文将同一个歌手被标记的所有类型进行求和,被标记次数较多的类型作为该歌手偏向类型的标签。采用歌手与歌手类型关系的 user_taggedartists.dat 数据集构建矩阵 G 。同时将用户收听歌手音乐的次数通过数据标准化转化为 1 分~5 分的喜欢程度,若未听过则不评分并将其作为缺失值,采用用户与歌手关系的 user_artists.dat 数据集构建矩阵 D 。在聚类完成后进行矩阵运算得到簇-标签矩阵 Q 并对 Q 进行排序,得到每个簇的前 n 个标签,而标签评分通过 F1 score 进行表示,其值越接近 1 表示聚类效果越好。实验在 CiteULike 数据集上使用 AP 算法进行测试,由于采用的标签数量不同,因此计算得到的 F1 score 也不同,如图 4 所示。

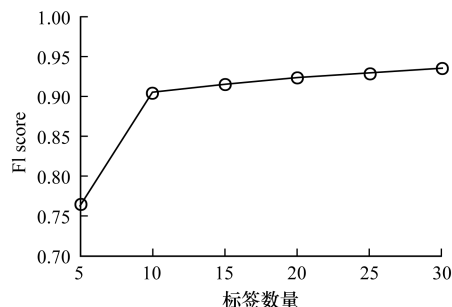


图 4 标签数量对 F1 score 的影响

Fig. 4 The impact of the number of tags on F1 score

在图 4 中,当标签数量为 5 时,通常会取得较常见的标签,但簇标签评分较低,簇的语义较模糊,随着标签数量的增多,簇标签评分越来越高,会取得区别能力较高的标签,使簇的语义更明确。当标签数量为 10~30 时,F1 score 增加效果不明显,因此本文取标签数量为 10。

通过综合轮廓系数(聚类评分)和 F1 score(标签评分)两项聚类性能指标(如表 2 所示),证明本文方法可在保证一定聚类效果的前提下生成语义更明确的簇标签,且标签评分均大于 0.4。经对比分析可以看出,不同数据集下各聚类算法的聚类评分及标签评分不同,无须事先确定用户行为规模的 K-means++、DP-means 以及 AP 算法的两项指标均大于 K-means 和层次聚类算法。

表 2 3 个数据集的算法聚类性能比较

Table 2 Comparison of algorithm clustering performance in three datasets

算法	Last. fm		CiteULike		Movielens	
	聚类评分	标签评分	聚类评分	标签评分	聚类评分	标签评分
K-means	0.323	0.424	0.401	0.575	0.318	0.559
K-means++	0.519	0.749	0.572	0.728	0.545	0.923
层次聚类	0.244	0.450	0.237	0.488	0.116	0.524
DP-means	0.553	0.741	0.542	0.781	0.521	0.691
AP 算法	0.354	0.761	0.427	0.904	0.402	0.934

为证明本文方法的聚类效果符合用户实际情况,利用 Last. fm 数据集通过 AP 算法对本文方法的聚类结果及簇标签进行分析,结果如表 3 所示。通过聚类后共形成 4 个簇,每簇取前 10 个标签作为簇标签,可以看出簇 1 的用户喜好偏向独立摇滚乐,簇 2 的用户喜好偏向 80 年代经典摇滚乐,簇 3 的用户喜好偏向流行电子音乐,簇 4 的用户喜好偏向前卫摇滚乐。通过人工检查这些数据,发现聚类结果与实际情况相符合,进一步证明了本文方法的有效性。

表 3 Last. fm 数据集中簇所对应的标签

Table 3 The labels corresponding to the clusters in Last. fm dataset

簇 1	簇 2	簇 3	簇 4
indie rock	classic rock	electronic	progressive rock
British	British	britpop	60s
ambient	80s	classic rock	British
progressive rock	indie	pop	thrash metal
experimental	electronic	British	80s
alternative rock	alternative	alternative rock	metal
indie	dance	indie rock	heavy metal
electronic	rock	alternative	hard rock
rock	female vocalists	Indie	rock
alternative	pop	rock	classic rock

4.2.4 缺失值补充效果评估

在进行聚类前,为提高生成标签的准确性,需要对用户行为数据进行补充缺失值处理。同时,原始数据维度较大,若对原始数据直接进行聚类,则聚类效率较低。以 Last. fm 数据集为例,原始数据经去重复用户处理后,矩阵 D 大小为 $1\ 877 \times 376$, LFM 矩阵分解后的矩阵 p 大小为 $1\ 877 \times 32$,从数据维度上可以看出对经补充缺失值处理后的矩阵 p 聚类效率优于对原始矩阵直接进行聚类。实验在 Movielens 数据集上利用 K-means、K-means++、层次聚类、DP-means 及 AP 算法分别在有无补充缺失值的条件下进行聚类,实验结果如图 5 所示。

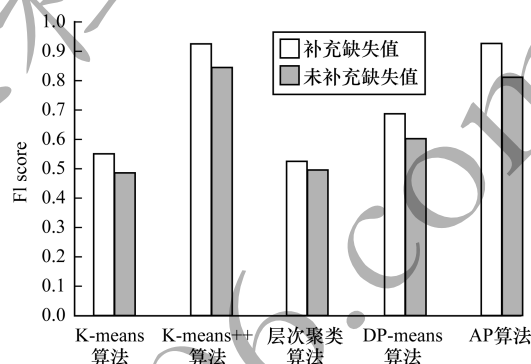


图 5 缺失值对 F1 score 的影响

Fig. 5 The impact of missing values on F1 score

从图 5 可以看出,经补充缺失值后 F1 score 提高了 4%~15%,从而提升生成标签的准确性。通过一系列实验结果分析可知,本文方法可在保证聚类效果的前提下,在聚类过程中自动生成语义更明确的簇标签。

5 结束语

现有聚类算法在对网络用户行为进行分析时需要预先确定用户行为数据规模,并且生成的簇也缺少语义特征。为此,本文对网络用户行为数据特征进行聚类分析,并应用 LFM、矩阵分解等方法,提出一种用于网络用户行为聚类的簇标签自动生成方法。实验结果表明,本文方法无需事先确定用户行为数据规模,在聚类过程中可以同时产生簇标签,且所生成的簇标签符合用户行为的实际语义。但由于本文方法对用户行为数据有一定的格式要求,因此后续将对此进行改进,扩展其应用范围并提升通用性。

参考文献

- [1] YANG Jie, QIAO Yuanyuan, ZHANG Xinyu, et al. Characterizing user behavior in mobile Internet [J]. IEEE Transactions on Emerging Topics in Computing, 2015, 3(1): 95-106.

- [2] REN Xingyi, SONG Meina, SONG Junde. Point-of-Interest recommendation based on the user check-in behavior [J]. Chinese Journal of Computers, 2017, 40(1):28-51. (in Chinese)
任星怡,宋美娜,宋俊德.基于用户签到行为的兴趣点推荐[J].计算机学报,2017,40(1):28-51.
- [3] SUN Lifeng, WANG Xiaoyan, WANG Zhi, et al. Social-aware video recommendation for online social groups [J]. IEEE Transactions on Multimedia, 2017, 19(3):609-618.
- [4] KANG Guosheng, LIU Jianxun, TANG Mingdong, et al. An effective Web service ranking method via exploring user behavior [J]. IEEE Transactions on Network and Service Management, 2015, 12(4):554-564.
- [5] BERNASCHINA C, BRAMBILLA M, MAURI A, et al. A big data analysis framework for model-based Web user behavior analytics [C]//Proceedings of International Conference on Web Engineering. Berlin, Germany: Springer, 2017:98-114.
- [6] WAN Ming, SHANG Wenli, ZENG Peng. Double behavior characteristics for one-class classification anomaly detection in networked control systems [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(12):3011-3023.
- [7] SU Qiubin, JIA Zhihao, LU Lu. Research on user behavior clustering algorithm based on mobile application [J]. Journal of Intelligent & Fuzzy Systems, 2018, 35(2):1291-1300.
- [8] JOSHI D J, SUPEKAR N, CHAUHAN R, et al. Modeling and detecting change in user behavior through his social media posting using cluster analysis [C]//Proceedings of the 4th ACM IKDD Conference on Data Sciences. New York, USA: ACM Press, 2017:1-10.
- [9] ZHANG Liping, DENG Song, LI Shiyue. Analysis of power consumer behavior based on the complementation of K-means and DBSCAN [C]//Proceedings of 2017 IEEE Conference on Energy Internet and Energy System Integration. Washington D. C., USA: IEEE Press, 2017:26-28.
- [10] ZHANG Jing, DUAN Fu. Improved K-means algorithm with meliorated initial centers [J]. Computer Engineering and Design, 2013, 34(5):1691-1694. (in Chinese)
张靖,段富.优化初始聚类中心的改进 K-means 算法[J].计算机工程与设计,2013,34(5):1691-1694.
- [11] ZHOU Zhihua. Machine learning [J]. Beijing: Tsinghua University Press, 2016. (in Chinese)
周志华.机器学习[M].北京:清华大学出版社,2016.
- [12] KOREN Y. Collaborative filtering with temporal dynamics [J]. Communications of the ACM, 2010, 53(4):89-97.
- [13] GUAN R C, SHI X H, MARCHESE M, et al. Text clustering with seeds affinity propagation [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(4):627-637.
- [14] BRIAN K. Revisiting k-means: new algorithms via Bayesian nonparametrics [EB/OL]. [2020-06-17]. <https://arxiv.org/abs/1111.0352>.
- [15] FARAJIAN M A, MOHAMMADI S. Mining the banking customer behavior using clustering and association rules methods [J]. International Journal of Industrial Engineering and Production Research, 2010, 21(4):239-245.
- [16] TREERATPITUK P, CALLAN J. Automatically labeling hierarchical clusters [C]//Proceedings of 2006 International Conference on Digital Government Research. New York, USA: ACM Press, 2006:167-176.
- [17] LÜ Haiyan, ZHANG Jie, WANG Lina. Automatic generation of micro-blog user tags based on clustering analysis [J]. Electronic Design Engineering, 2015, 23(7):67-69, 73. (in Chinese)
吕海燕,张杰,王丽娜.基于聚类分析的微博用户标签自动生成[J].电子设计工程,2015,23(7):67-69,73.
- [18] ZHONG Qing. Research on mobile Internet user behavior classification based on preference tag [J]. Mobile Communications, 2016(9):93-96. (in Chinese)
钟庆.基于喜好标签的移动互联网用户行为分类研究[J].移动通信,2016(9):93-96.
- [19] MOHAN A J. An approach to improving cluster labeling and evaluation [D]. Twin Cities, USA: University of Minnesota, 2014.
- [20] LI Huina. A fast and stable cluster labeling method for support vector clustering [J]. Journal of Computers, 2013, 8(12):3251-3256.

编辑 陆燕菲