



文章编号:1001-9081(2018)11-3127-05

DOI:10.11772/j.issn.1001-9081.2018041357

基于用户行为特征的多维度文本聚类

黎万英¹, 黄瑞章^{1,2,3*}, 丁志远¹, 陈艳平^{1,2}, 徐立洋¹

(1. 贵州大学 计算机科学与技术学院, 贵阳 550025; 2. 贵州省公共大数据重点实验室(贵州大学), 贵阳 550025;

3. 计算机软件新技术国家重点实验室(南京大学), 南京 210093)

(*通信作者电子邮箱 rzhuang@zju.edu.cn)

摘要:传统多维度文本聚类一般是从文本内容中提取特征,而很少考虑数据中用户与文本的交互信息(如:点赞、转发、评论、关注、引用等行为信息),且传统的多维度文本聚类主要是将多个空间维度线性结合,没能深入考虑每个维度中属性间的关系。为有效利用与文本相关的用户行为信息,提出一种结合用户行为信息的多维度文本聚类模型(MTCUBC)。根据文本间的相似性在不同空间上应该保持一致的原则,该模型将用户行为信息作为文本内容聚类的约束来调节相似度,然后结合度量学习方法来改善文本间的距离,从而提高聚类效果。通过实验表明,与线性结合的多维度聚类相比,MTCUBC模型在高维稀疏数据中表现出明显的优势。

关键词:多维度聚类; 度量学习; 约束; 用户行为特征

中图分类号:TP311.1 **文献标志码:**A

Multi-dimensional text clustering with user behavior characteristics

LI Wanying¹, HUANG Ruizhang^{1,2,3*}, DING Zhiyuan¹, CHEN Yanping^{1,2}, XU Liyang¹

(1. College of Computer Science and Technology, Guizhou University, Guiyang Guizhou 550025, China;

2. Guizhou Provincial Key Laboratory of Public Big Data (Guizhou University), Guiyang Guizhou 550025, China;

3. State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing Jiangsu 210093, China)

Abstract: Traditional multi-dimensional text clustering generally extracts features from text contents, but seldom considers the interaction information between users and text data, such as likes, forwards, reviews, concerns, references, etc. Moreover, the traditional multi-dimension text clustering mainly integrates linearly multiple spatial dimensions and fails to consider the relationship between attributes in each dimension. In order to effectively use text-related user behavior information, a Multi-dimensional Text Clustering with User Behavior Characteristics (MTCUBC) was proposed. According to the principle that the similarity between texts should be consistent in different spaces, the similarity was adjusted by using the user behavior information as the constraints of the text content clustering, and then the distance between the texts was improved by the metric learning method, so that the clustering effect was improved. Extensive experiments conduct and verify that the proposed MTCUBC model is effective, and the results present obvious advantages in high-dimensional sparse data compared to linearly combined multi-dimensional clustering.

Key words: multi-dimensional clustering; metric learning; constraint; user behavior characteristics

0 引言

随着 Twitter、微博等社交媒体的广泛使用,给传统文本内容聚类方法带来挑战。由于社交媒体中存在大量短文本,导致基于文本内容聚类中的特征稀疏问题比较严重。另外,除了文本内容,社交媒体数据还包含很多用户行为信息,如:点赞、转发、评论、关注、引用等(也称“用户行为特征”)。缺少用户行为特征的聚类方法不能对社交媒体数据的分布特征进行建模。除了 Twitter、微博等“新媒体”外,有些传统文本中也包含有用户行为信息,如学术论文中的合作作者和参考文献等。

收稿日期:2018-04-31;修回日期:2018-06-21;录用日期:2018-06-29。基金项目:国家自然科学基金资助项目(61462011);国家自然科学基金重大研究计划项目(91746116);贵州省重大应用基础研究项目(黔科合JZ字[2014]2001);贵州省自然科学基金资助项目(黔科合基础[2018]1035);贵州省科技重大专项计划(黔科合重大专项字[2017]3002)。

作者简介:黎万英(1992—),女,贵州开阳人,硕士研究生,主要研究方向:数据挖掘、文本挖掘、机器学习; 黄瑞章(1979—),女,天津人,副教授,博士,CCF成员,主要研究方向:数据挖掘、文本挖掘、机器学习、信息检索; 丁志远(1993—),男,湖北孝感人,硕士研究生,主要研究方向:数据挖掘、文本挖掘、机器学习; 陈艳平(1980—),男,贵州黔南人,副教授,博士,主要研究方向:人工智能、自然语言处理; 徐立洋(1990—),男,贵州黔南人,硕士研究生,主要研究方向:数据挖掘、文本挖掘、机器学习。

为了在文本内容的基础上有效利用用户行为特征进行聚类,本文提出了结合用户行为特征的多维度文本聚类(Multi-dimensional Text Clustering with User Behavior Characteristics, MTCUBC)模型,该模型主要针对传统多维度聚类中存在的两个问题:1)传统多维度聚类主要使用文本内容和超链接等“静态特征”,缺乏对用户行为特征的有效利用;2)传统多维度聚类只是简单地将多个维度空间进行线性叠加,没有考虑不同维度空间的差异性。本文提出的 MTCUBC 模型根据文本相似性在不同空间上应该保持一致的原则,将用户行为特征作为约束(constraints)来辅助聚类。同时,采用度量学习(metric learning)方法来精确地调节每个属性值,从而提高聚



类的效果。

本文通过两个数据集对模型进行验证。实验结果表明,该方法与单维度文本聚类方法相比有明显的改进,与多维度聚类方法比较也有明显的提升。

1 相关工作

在相关研究中,多维度聚类备受关注,例如:在网络中,广泛使用图像、音频、超链接、文本等不同类型的特征来进行聚类。在早期的研究中,数据的标注信息经常被用作多维度聚类的约束。例如,文献[1]提出了一种将有标记和未标记样本进行合并的方法;文献[2]提出了一种成对约束(pairwise)的聚类框架,使用标记数据作为约束来指导聚类过程;文献[3~4]均研究了带标记和未标记样本聚类结果的影响。这类方法的实现需要部分数据带有标注信息,难以在无标注数据集中使用。

在约束聚类中,文献[1,5]通过修改聚类目标函数实现多维度空间聚类,其中,文献[1]在 K-Means 算法的基础上,提出如何修改利用流行的聚类算法应用到实际问题中,文献[5]提出半监督聚类算法,将部分未标记的数据进行聚类,然后用已知类来预测未来样本点的类别。文献[6]提出了成对约束聚类方法。这类模型均使用约束来修改聚类的目标函数,但由于特征的高维稀疏性而影响聚类效果,没有考虑能改善文本间距离测量的距离度量方法。

文献[7]提出学习距离矩阵的方法;文献[8]给出成对约束的聚类框架,并提出一种主动选择成对约束的方法来改进聚类效果;文献[9]将社会行为信息的相似性嵌入到视觉空间中;为了提高多维度聚类的结果,文献[10]使用相似的文章来学习距离矩阵;文献[11]提出使用相似的文章作为约束,为每个类学习出一个距离矩阵的聚类方法;文献[12]提出一种可以同时进行多维度聚类和特征选择的方法,该方法主要是针对高维数据稀疏问题;文献[13]提出基于隐马可夫条件随机场的多维度聚类框架;文献[14]提出一种多维度 K-Means 聚类算法,该算法为每个维度指定一个权重,使其与聚类结果相关联,其中维度用给定的内核矩阵表示,并且内核的加权组合与簇并行学习;文献[15]提出一种多类型的网络文本聚类(Multi-type Features based Web Document Clustering, MFRC),为每一个特征空间设置一个权重值。

在成对约束聚类中,文献[10]结合梯度下降和迭代过程来学习马氏矩阵(Mahalanobis metric);文献[16]提出冗余成分分析算法,它只用必连约束来学习马氏距离,但是这些矩阵学习方法都只训练出一个距离矩阵。

与前面的模型相比,本文提出的 MTCUBC 模型允许度量学习方法利用成对约束和无标签数据来学习出多个距离矩阵,使实验结果有很大提升。

2 模型的提出

2.1 社会特征的学习

本文将社会维度中的用户行为特征嵌入到词维度聚类中,因此,如何将复杂、无结构的用户行为特征加入到有结构的词维度空间是本文首先要解决的问题。用论文中作者和参考文献间的关系来举例说明, $x_i = \{a_1, a_2, \dots, a_n\}$ 表示社会维度中第 i 篇论文中出现的作者和论文引用参考文献中的

作者的表示,其中作者出现为 1,不出现为 0,并且当前论文中的作者及参考文献中的作者只取前三位。社会维度中,作者间的相似性对于词维度的聚类有帮助,为了证明这一点,本文收集了两个数据,以 Aminer 论文集^[17]为例进行说明,该数据集收集了 3000 篇论文,统计有 8812 个词和 22373 个作者。对于每篇论文中的作者,若他们引用的文章都有较高的相似性、较低的差异性,则这些文章中的作者有相同的爱好领域。本文抽象出社会维度中的作者特征和词维度中的词特征向量来计算它们之间的相似度,假如要给作者推荐参考文献,除了从文章的内容以外,社会相似度也是一个可靠、可以考虑的指标。

$\chi = \{x_i\}_{i=1}^K$ 是一组数据点集合, x_{id} 是向量 x_i 的第 d 个元素, $\{\mu_h\}_{h=1}^K$ 表示 K 个聚类中心点, l_i 表示数据点 x_i 所属的类标号,其中 $l_i \in \{1, 2, \dots, K\}$ 。传统 K-means 算法将数据集 χ 划分为 K 个簇 $\{\chi_h\}_{h=1}^K$ 使目标函数 $\sum \|x_i - \mu_{l_i}\|^2$ 的值最小,其中,点 x_i 和对应的聚类中心 μ_{l_i} 的欧氏距离为 $\|x_i - \mu_{l_i}\|^2 = (x_i - \mu_{l_i})^\top (x_i - \mu_{l_i})$ 。

2.2 基于约束的聚类算法

由于有标签的样本比较少,文献[16]考虑到使用约束方法比提供有标签的数据更现实,因此使用相似文章对来学习距离,虽然类标签可能未知,但用户仍然可以指定对应样本是否属于同一个簇,所以约束的方法也比类标签更通用。

针对传统聚类方法不能直接利用约束的问题,基于约束的聚类算法是使用标注数据作为约束来辅助聚类,但本文中没有标注好的样本,是通过在社会维度(social)聚类后,挑选具有较高相似度的文章对作为约束来辅助词(word)维度聚类,此时挑选的文章对类别是确定的。利用约束将目标函数和约束条件结合起来可以解决这一问题。

设 M 表示是一组必须关联的文章,其中 $(x_i, x_j) \in M$ 表示 x_i 和 x_j 应该被聚到同一个类中。让 $W = \{\omega_{ij}\}$ 作为 M 中违反约束的惩罚因子,所以目标是最小化下面的目标函数:

$$J_{ckm} = \sum_{x_i \in \chi} (\|x_i - \mu_{l_i}\|^2 + \sum_{(x_i, x_j) \in M} \omega_{ij} I[l_i \neq l_j]) \quad (1)$$

其中: I 为指示函数, $I[\text{true}] = 1$ 和 $I[\text{false}] = 0$ 。 l_i 和 l_j 在社会维度中有较高的相似度,在词向量维度聚类中应该被聚到同一个类中,否则将会受到相应惩罚。

2.3 基于度量学习聚类

为调节向量中每个属性的贡献度,本模型不是统一地给定某个权重参数,而是通过一个度量学习矩阵来给每个属性一个权重,这样更能满足每个属性间的差异性。

成对约束方法对聚类结果有一定的促进作用,它能用于调整潜在的文本距离矩阵,从而有效代表用户在特定领域中的相似性。结合式(1)可以用对称正定矩阵 A 来参数化欧氏距离: $\|x_i - x_j\|_A = \sqrt{(x_i - x_j)^\top (x_i - x_j)}$,同样的参数化方法以前也用于文献[10]。如果权重矩阵 A 是严格意义上的对角矩阵,那么它会通过不同的权重对每个维度中相应的特征进行缩放;否则,就会与原始特征线性结合而产生新的特征。

在调整矩阵聚类相关工作中,文献[10, 17]调整矩阵权重同时满足最小化必连(must-linked)样本间的距离和最大化必不连(cannot-kinked)样本间的距离,而且存在基本的限制:对于所有的类都只能用同一个矩阵,允许每个簇 h 有单独的



一个权重矩阵 A_h , 可以证明, 在这种广义 K-means 模型下, 完整的数据最大对数似然函数等价于最小化目标函数:

$$J_{\text{mkn}} = \sum_{x_i \in \chi} (\|x_i - \mu_{l_i}\|_{A_{l_i}}^2 - \lg(\det(A_{l_i}))) \quad (2)$$

其中: 第二项是第 l_i 个高斯与协方差矩阵 $A_{l_i}^{-1}$ 的正态常数。

2.4 MTCUBC 模型

结合式(1)和(2), 即结合约束与度量学习方法。在较少约束违反的情况下, 最小化在学习矩阵下的聚类分散度, 可以得到目标函数:

$$\begin{aligned} J_{\text{comb}} = & \sum_{x_i \in \chi} (\|x_i - \mu_{l_i}\|_{A_{l_i}}^2 - \lg(\det(A_{l_i}))) + \\ & \sum_{(x_i, x_j) \in M} \omega_{ij} I[l_i \neq l_j] \end{aligned} \quad (3)$$

如果统一约束开销 ω_{ij} , 所有约束违背都平等对待; 然而, 在必连(must-link)集合中, 对那些违背约束且离得远的点的惩罚应该大于那些违背约束且离得相对较近的点。直观地说: 如果两个必连点根据当前的距离度量方法相距很远, 则度量是非常不充分的, 并且需要对它进行严格的修改。由于两个簇中参与了同一个必连的违背行为, 相应的惩罚应该影响到两个簇的度量, 这可以通过对式(3)的第二部分乘以一个惩罚值来实现, 该惩罚值表示为:

$$f_M(x_i, x_j) = \frac{1}{2} \|x_i - x_j\|_{A_{l_i}}^2 + \frac{1}{2} \|x_i - x_j\|_{A_{l_j}}^2 \quad (4)$$

加入惩罚值后的目标函数如下:

$$\begin{aligned} J_{\text{mtcubc}} = & \sum_{x_i \in \chi} (\|x_i - \mu_{l_i}\|_{A_{l_i}}^2 - \lg(\det(A_{l_i}))) + \\ & \sum_{(x_i, x_j) \in M} \omega_{ij} f_M(x_i, x_j) I[l_i \neq l_j] \end{aligned} \quad (5)$$

惩罚仅针对那些违背约束的样本。 M 是社会维度中满足一定相似度的样本。忽略不相连(cannot-link)的样本, 因为在社会维度中不同作者的两篇文章, 在词向量维度中其内容可能相似。从式(5)可以看出, 公式由两部分组成, 其中惩罚因子 ω_{ij} 为每个约束提供一个权重, 同时该约束也体现公式中两部分之间的相对重要性。

3 算法求解

给定一组数据 χ , 一组必连样本 M 和相应的惩罚因子 W 集合, 类的个数 $K, \{\chi_h\}_{h=1}^K$ 表示将数据集 χ 划分为 K 互不相交的簇, A 为权重矩阵。该算法结合了约束和度量学习方法, 在聚类初始化过程中使用约束把每个数据点分配到各个簇中去。在每次迭代过程中用重新估计的矩阵 A 去调整样本间的距离。该算法的伪代码如下所示。

输入 数据集 $\chi = \{x_i\}_{i=1}^N$, 类的个数 K , 违背约束的惩罚因子 ω_{ij} , 约束中必连样本 $M = \{(x_i, x_j)\}$ 。

输出 通过最小化目标函数 J_{mtcubc} 将数据集 χ 聚类到 K 个不相交的簇: $\{\chi_h\}_{h=1}^K$ 。

步骤 1 初始化 K 个类的中心点 $\{\mu_h^{(0)}\}_{h=1}^K, t = 0$; 先从必连集合中挑选一个点作为第一个中心点, 然后从非必连集合中找到离第一个中心点最远的点作为第二个中心点, 第三个中心点为离前两个点的距离之和最远的点。

步骤 2 重复下面的过程直到收敛:

① 簇的中心点为 $\{\mu_h^{(t)}\}_{h=1}^K$, 给数据 $\{x_i\}_{i=1}^N$ 中的点分配类标签 $\{l_i^{(t+1)}\}_{i=1}^N$ 使目标函数到达最小值。

② 根据上一步结果的类标签 $\{l_i^{(t+1)}\}_{i=1}^N$, 重新计算类的中心点 $\{\mu_h^{(t+1)}\}_{h=1}^K$ 。重新估计度量矩阵 A_h 。

③ $t = t + 1$ 。

EM 算法和复杂度分析如下。

算法主要分两个步骤: 一是从社会维度中挑选相似对, 二是计算词维度中特征间的距离。在第一步中, 要计算两向量间的距离, 时间复杂的在 $O(m^2I)$ 级别上, 其中 I 为迭代次数, m 表示论文的数量。第二步中, 时间复杂度 $O(Ikm)$, 其中 I 为迭代的次数, k 为类的个数, m 为论文数量。

K-Means 算法对初始化和类的个数比较敏感, 好的初始中心点对 K-Means 聚类算法很重要。首先介绍初始化方法, 本实验采用的方法是先在必连集合中随机挑选一个点作为初始化中心点, 在必连以外的集合中挑选第二个中心点且离第一个中心点的距离最远, 同理, 第三个中心点是所有样本点中距离前两个中心点距离最远的点。这样做可以加快数据的聚类收敛速度, 由于 K-Means 对初始点比较敏感, 该方法也能一定程度上促进聚类。

期望最大(Expectation Maximization, EM)算法是在概率模型中寻找参数最大似然估计的算法。求解式(5)中的矩阵 A , 同时要找到聚类的最优效果, 不断迭代优化聚类结果, 该过程就是一个 EM 算法过程。下面具体介绍 EM 的实现过程。

E 步 在文献[14,18]中, 通过使用能够代表当前类的样本点用于更新数据点的分布。在简单的 K-Means 中, 聚类过程中是没有交互的, 本文的方法是在 E 步和 M 步中不断交替: E 步是将每个样本点分配到每个类中, M 步将重新估计中心点和度量学习距离矩阵, 在本文模型作用下使所有样本点到各自的类中心点距离之和最小。

值得注意的是, 这个分配步骤是依赖于顺序的, 因为 M 的子集和每个类有关, 可能会改变样本点的分配。做了随机分配的实验: 每个点会分配到离它最近的类中去, 同时也会涉及到最少数量的约束对。实验表明, 分配的顺序不会导致聚类质量的显著差异, 所以在评估中使用随机分配的策略。

在 E 步骤中, 样本点的分配遵循的原则是保持目标函数 J_{mtcubc} 最小, 因此, 当所有的点都被重新分配时, 目标函数 J_{mtcubc} 相比上一次将会减少或是保持不变。简而言之, 结合成对约束和度量学习来指导聚类过程使聚类达到更好的效果。

M 步 用每个类中的所有点 x_h 重新估计当前的类中心 μ_h , 因此, 每个簇中的分布对于目标函数 J_{mtcubc} 来说都是最小的。因为约束违背值依赖于类的分配, 而成对约束不参与中心点的重新估计步骤, 所以这些在 M 步都不会发生, 因此, 只有 J_{mtcubc} 的第一个距离分量最小化, 重新估计样本中心点的步骤实际上与 K-Means 算法类似。

M 步 中针对度量学习, 矩阵 $\{A_h\}_{h=1}^K$ 被重新估计以减小目标函数 J_{mtcubc} 。每次更新局部矩阵 A_h 是通过求偏导数 $\frac{\partial J_{\text{mtcubc}}}{\partial A_h}$ 并设置为 0 而得到:

$$\begin{aligned} A_h = & |\chi_h| \left(\sum_{x_i \in \chi_h} (x_i - \mu_h)(x_i - \mu_h)^T + \right. \\ & \left. \sum_{(x_i, x_j) \in M_h} \frac{1}{2} \omega_{ij} (x_i - x_j)(x_i - x_j)^T I[l_i \neq l_j] \right)^{-1} \end{aligned} \quad (6)$$



其中 M_h 是必连约束的子集,包含当前分配给第 h 个簇的点。

由于实际的高维或大数据集中估计整个矩阵 A 可能会使计算花费相当大,在这种情况下,可以使用对角矩阵 $\alpha = \text{diag}(A)$,它等价于特征加权,而使用全矩阵会导致相应的特征生成。对于该对角矩阵 A ,第 d 个对角元素为 $\alpha_{dd}^{(h)}$,簇 h 中其对应的第 d 个特征的权重为:

$$\begin{aligned} \alpha_{dd}^{(h)} &= |\chi_h| \left(\sum_{x_i \in \chi_h} (\mathbf{x}_{id} - \boldsymbol{\mu}_{hd})^2 + \right. \\ &\quad \left. \sum_{(x_i, x_j) \in M_h} \frac{1}{2} \omega_{ij} f_M(\mathbf{x}_{id}, \mathbf{x}_{jd}) I[l_i \neq l_j] \right)^{-1} \end{aligned} \quad (7)$$

因为每个 A_h 是式(7)中的协方差矩阵之和的逆,其和不能为奇异值,如果其中任何一个元素为奇异值时,可以通过添加单位矩阵乘以矩阵 A^{-1} 的迹(trace)的一部分来加以限制: $A_h^{-1} = A_h^{-1} + \varepsilon \text{tr}(A_h^{-1}) I$ 。从直观上看,距离学习修改了聚类变形算法使得相似的点离得更近。

4 实验及分析

本文在两个数据集上验证 MTCUBC 模型的有效性,实验结合了用户行为特征和文本词向量空间的特征,使用社会特征辅助词向量空间的聚类。

4.1 数据集

用向量空间模型表示文本特征时,很难对稀疏高维数据集进行聚类,因为聚类算法容易遇到局部最优而停止迭代,从而导致聚类质量差。在以往的研究中,文献[19]在文本集上用 SP-K-means (SParse K-means) 算法,其文本集大小比单词空间的维数小,可以看出,在大多数初始化过程中,集群之间的文档迁移很少,这导致算法收敛后的聚类质量较差。这种现象在许多实际应用中出现,例如,当将搜索结果聚类到网络搜索引擎中时,通常聚类中的网页数量是数以百万计的,然而,特征空间的维度,对应于所有网页中的单词的数量是成千上万的,而且因为它只包含少量的所有可能的单词,所有每个网页都是稀疏的,在这种情况下,度量学习结合成对约束方法就可以凸显它的优势,并且可以显著提高聚类的质量。为了证明 MTCUBC 模型中度量学习文本聚类的有效性,本文使用了具有稀疏性、高维特征的 Aminer 论文数据集^[17] 和 NIPS Papers 数据集^[20]。

Aminer 论文数据集和 NIPS Papers 论文数据集,收集了大量的关于计算机科学的学术论文信息,两个数据集都包含论文信息、论文引用、作者信息和合作者信息等,每篇文章都包含 paperID、authorID、摘要和参考文献等属性。在数据预处理中,本文只考虑每篇论文的前三个作者,同样,参考文献的作者也是只取前三个。在两个数据集中,将词向量维度和社会维度分别用 view1、view2 表示。实验数据集的统计信息如表 1 所示。

表 1 Aminer 和 NIPS Paper 数据集信息

Tab. 1 Detailed information of Aminer dataset and NIPS Paper dataset

数据集	维度	文章数(D)	元素数(N)	类数(K)
Aminer	词维度	3 000	8 812	3
	社会维度	3 000	22 373	3
NIPS	词维度	5 000	16 846	3
	社会维度	5 000	38 375	3

4.2 聚类评估

本文主要采用 NMI (Normalized Mutual Information) 作为

聚类算法的度量标准。NMI 的定义如下:

$$NMI = \frac{I(C;L)}{(H(C) + H(L))/2} \quad (8)$$

算法聚类后的簇集合表示为 $C = \{c_1, c_2, \dots, c_k\}$, 标准的聚类标签表示为: $L = \{l_1, l_2, \dots, l_j\}$ 。其中 $I(X;Y) = H(X) - H(X|Y)$ 表示随机变量间的互信息, $H(X)$ 为 X 的熵, $H(X|Y)$ 为在给定 Y 时 X 的条件熵。 NMI 取值范围为 $0 \sim 1$, 值越大说明聚类效果越好。

4.3 实验结果与分析

实验表明,在两个数据集中,社会维度特征的加入对 MTCUBC 算法的实验结果有一定的影响,从而证明结合社会维度约束的度量学习方法的 NMI 值比单独的社会维度或词维度的聚类结果好;同时 MTCUBC 模型的结果也比基于特征选择的加权多视角聚类 (Weighted Multi-view Clustering with Feature Selection, WMCF) 模型^[12]、多视角聚类 (Multi-view Clustering)^[21] 中的多视角 EM 算法 (Multi-View EM, MVEM) 的结果好,其中 WMCF 算法出自文献 [12],它提出一种可以同时进行多维度聚类和特征选择的方法,该方法主要是针对高维数据稀疏提出的解决办法。

由表 2 可以看出: MTCUBC 模型与相对应的单维度 (social-single 和 word-single) 聚类结果相比有明显的提升,提升效果达到 10 个百分点到 14 个百分点;与其他两个多维度算法 (WMCF 和 WMCF) 聚类结果相比提升 7 个百分点。

表 2 MTCUBC 模型在两个数据集上与单维度和多维度方法的对比实验

Tab. 2 Comparison with several single and multi-view algorithms on two datasets

方法	Aminer	NIPS Paper	方法	Aminer	NIPS Paper
social-single	0.710	0.680	WMCF	0.753	0.725
word-single	0.679	0.651	MTCUBC	0.821	0.793
MVEM	0.772	0.752			

图 1 表示 MTCUBC 模型在两个数据集中,几个多维度聚类模型的聚类效果与加入约束对数量间的关系。可以看出,在没有约束的情况下: WMCF 模型变成传统的 K-means 算法,MVEM 模型变成单维度的文本聚类,MTCUBC 模型的结果优于 WMCF 模型,因为从式(5)可以看出,MTCUBC 模型由两部分组成,第二部分的约束不起作用,度量学习矩阵的结果取决于第一部分,其结果优于 WMCF 模型。此时,单维度的 MVEM 模型优于 MTCUBC 模型,MVEM 聚类算法是一个基于概率的算法,准确度高于 MTCUBC 模型。

在加入约束后,每个方法的聚类效果都有所提升,在加入 2000 个左右约束对时,聚类的提升效果不明显,分析原因是加入的这些约束对中,很大比例的约束对相似度比较高,能被分到同一个类中,或是不能被分到同一个簇中但受到的惩罚比较小,没能改变其最初的聚类结果。

在约束对超过 2000 时,聚类效果提升比较明显,在约束对数量达到 10000 对左右时算法处于收敛状态。在没有约束时,MVEM 算法效果比本文 MTCUBC 模型好,但随着约束对数量的增加,本文模型 MTCUBC 的聚类效果超越了 WMCF 算法。整个过程中 MTCUBC 算法都比 WMCF 算法的效果好,WMCF 算法是使用一个合适的参数,把多个维度线性叠加,而本文模型使用度量学习方法将多个维度结合,度量矩阵



能影响到每个向量中的元素,而非简单的维度之间的叠加关系。

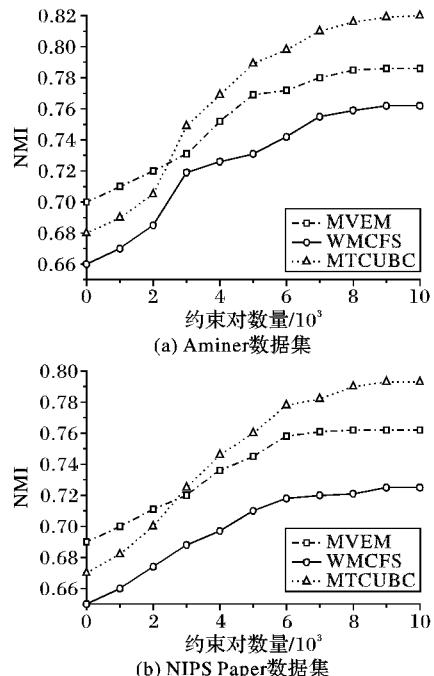


图 1 MTCUBC 模型与几个多维度算法在两个数据集上 NMI 对比

Fig. 1 NMI comparison of MTCUBC and several multi-view algorithm on two datasets

5 结语

为提高聚类效果,除利用文本自身内容外,还充分利用和文本内容相关的用户行为信息,带约束的聚类方法结合度量学习方法来改善传统多维度聚类中不同维度线性结合问题,使得用户行为信息在聚类过程中充分发挥作用;同时,每个簇中都会学习出一个度量矩阵,改善多个类共用一个度量矩阵的情况。本文中对惩罚的开销值 ω_i 有待细化,深究每一维数据的权重值。

在未来的工作中,为得到更准确的聚类结果和充分利用社会信息,拟研究如何利用更多空间聚类结果来互相辅助提升聚类效果。例如,可以将社会维度、词维度、主题维度等特征综合利用,并使用双向辅助作用来提高聚类结果。

参考文献 (References)

- [1] WAGSTAFF K, CARDIE C, ROGERS S. Constrained K -means clustering with background knowledge[C]// Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2001: 577 – 584.
- [2] BASU S, BANERJEE A, MOONEY R J. Semi-supervised clustering by seeding[C]// Proceedings of the 19th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2002: 27 – 34.
- [3] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]// Proceedings of the 11th Annual Conference on Computational Learning Theory. New York: ACM, 1998: 92 – 100.
- [4] JOACHIMS T. Transductive inference for text classification using support vector machines[C]// Proceedings of the 16th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 1999: 200 – 209.
- [5] DEMIRIZ A, BENNETT K P, EMBRECHTS M J. Semi-supervised clustering using genetic algorithms[EB/OL]. [2018-03-20]. https://www.researchgate.net/profile/M_Embrechts/publication/2395752_Semi-Supervised_Clustering_Using_Genetic_Algorithms/links/0c9605203c771a5687000000/Semi-Supervised-Clustering-Using-Genetic-Algorithms.pdf.
- [6] BANSAL N, BLUM A, CHAWLA S. Correlation clustering[C]// Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science. Piscataway, NJ: IEEE, 2002: 238 – 247.
- [7] SCHULTZ M, JOACHIMS T. Learning a distance metric from relative comparisons[EB/OL]. [2018-03-20]. <http://papers.nips.cc/paper/2366-learning-a-distance-metric-from-relative-comparisons.pdf>.
- [8] BASU S, BANERJEE A, MOONEY R J. Active semi-supervision for pairwise constrained clustering[EB/OL]. [2018-03-20]. <http://www.cs.utexas.edu/users/ai-lab/pubs/semi-sdm-04.pdf>.
- [9] LIU S, CUI P, ZHU W, et al. Social embedding image distance learning[C]// Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM, 2014: 617 – 626.
- [10] XING E P, NG A Y, JORDAN M I, et al. Distance metric learning, with application to clustering with side-information[C]// Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 521 – 528.
- [11] BILENKO M, BASU S, MOONEY R J. Integrating constraints and metric learning in semi-supervised clustering[C]// Proceedings of the 21st International Conference on Machine Learning. New York: ACM, 2004: 11.
- [12] XU Y M, WANG C D, LAI J H. Weighted multi-view clustering with feature selection[J]. Pattern Recognition, 2016, 53: 25 – 35.
- [13] BASU S, BILENKO M, MOONEY R J. A probabilistic framework for semi-supervised clustering[C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 59 – 68.
- [14] TZORTZIS G, LIKAS A. Kernel-based weighted multi-view clustering[C]// Proceedings of the 12th International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2012: 675 – 684.
- [15] HUANG S, XUE G R, ZHANG B Y, et al. Multi-type features based Web document clustering[C]// Proceedings of the 5th International Conference on Web Information Systems Engineering, LNCS 3306. Berlin: Springer, 2004: 253 – 265.
- [16] BAR-HILLEL A, HERTZ T, SHENTAL N, et al. Learning distance functions using equivalence relations[C]// Proceedings of the 20th International Conference on Machine Learning. Washington, DC: IEEE Computer Society, 2003: 11 – 18.
- [17] TANG J, ZHANG J, YAO L, et al. Arnetminer: Extraction and mining of academic social networks[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990 – 998.
- [18] DOMENICONI C. Locally adaptive techniques for pattern classification[D]. Riverside: University of California, 2002.

(下转第 3138 页)



- based clustering method[C]// Proceedings of the 18th International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2006: 912–915.
- [3] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. Berlin: Springer Science & Business Media, 2013: 80–86.
- [4] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: Analysis and an algorithm[C]// Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge, MA: MIT Press, 2002: 849–856.
- [5] 丁祥武, 郭涛, 王梅, 等. 一种大规模分类数据聚类算法及其并行实现[J]. 计算机研究与发展, 2016, 53(5): 1063–1071. (DING X W, GUO T, WANG M, et al. A clustering algorithm for large-scale categorical data and its parallel implementation[J]. Journal of Computer Research and Development, 2016, 53(5): 1063–1071.)
- [6] 姜火文, 曾国荪, 马海英. 面向表数据发布隐私保护的贪心聚类匿名方法[J]. 软件学报, 2017, 28(2): 341–351. (JIANG H W, ZENG G S, MA H Y. Greedy clustering-anonymity method for privacy preservation of table data-publishing[J]. Journal of Software, 2017, 28(2): 341–351.)
- [7] SHIRKHORSHIDI A S, AGHABOZORGI S, WAH T Y, et al. Big data clustering: a review[C]// Proceedings of the 2014 International Conference on Computational Science and Its Applications. Berlin: Springer, 2014: 707–720.
- [8] ALTHOFF T, ULGES A, DENGEL A. Balanced clustering for content-based image browsing[J]. Series of the Gesellschaft für Informatik, 2011(1): 27–30.
- [9] DU Z, LIU Y, QIAN D. An energy-efficient balanced clustering algorithm for wireless sensor networks[C]// Proceedings of the 2009 Wireless Communications, Networking and Mobile Computing. Piscataway, NJ: IEEE, 2009: 1–4.
- [10] ALOISE D, DESHPANDE A, HANSEN P, et al. NP-hardness of Euclidean sum-of-squares clustering[J]. Machine Learning, 2009, 75(2): 245–248.
- [11] MALINEN M I, FRÄNTI P. Balanced k -means for clustering[C]// Proceedings of the 2014 Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Berlin: Springer, 2014: 32–41.
- [12] LIU H, HAN J, NIE F, ET AL. Balanced clustering with least square regression[C]// Proceedings of the 31th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2017: 2231–2237.
- [13] BRADLEY P S, BENNETT K P, DEMIRIZ A. Constrained k -means clustering[EB/OL].[2018-03-01]. <http://machinelearning102.pbworks.com/f/ConstrainedKMeanstr-2000-65.pdf>.
- [14] KUHN H W. The Hungarian method for the assignment problem [J]. Naval Research Logistics, 2005, 52(1): 7–21.
- [15] BANERJEE A, GHOSH J. On scaling up balanced clustering algorithms[C]// Proceedings of the 2002 SIAM International Conference on Data Mining. Columbus, Ohio: SIAM, 2002: 333–349.
- [16] HAJEK B. Cooling schedules for optimal annealing[J]. Mathematics of Operations Research, 1988, 13(2): 311–329.
- [17] CAI D, HE X, HAN J. Document clustering using locality preserving indexing[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(12): 1624–1637.
- [18] STREHL A, CHOSH J. Knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 33(3): 583–617.

This work is partially supported by the National Natural Science Foundation of China (61562014, 61763007, U1711263), the Natural Science Foundation of Guangxi (2015GXNSFAA139303), the Project of Guangxi Key Laboratory of Trusted Software, the Director Fund Project of Guangxi Key Laboratory of Automatic Testing and Instrument (YQ17111).

TANG Haibo, borned in 1993, M. S. candidate. His research interests include distributed data management.

LIN Yuming, borned in 1978, Ph. D., associate professor. His research interests include opinion mining, massive data management.

LI You, borned in 1980, M. S., associate professor. Her research include text mining.

CAI Guoyong, borned in 1971, Ph. D., professor. His research interests include social media data mining, machine learning, trusted software.

(上接第3131页)

- [19] DHILLON I S, GUAN Y. Information theoretic clustering of sparse co-occurrence data[C]// Proceedings of the 3rd IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2003: 517.
- [20] NIPS[EB/OL].[2018-01-04]. <https://www.kaggle.com/benhamner/nips-papers/data>.
- [21] BICKEL S, SCHEFFER T. Multi-view clustering[C]// Proceedings of the 4th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2004: 19–26.

cial Science and Technology Projects of Guizhou Province ([2017]3002), the Science and Technology Project of Guizhou Province ([2018]1035).

LI Wanying, born in 1992, M. S. candidate. Her research interests include data mining, text mining, machine learning.

HUANG Ruizhang, born in 1979, Ph. D., associate professor. Her research interests include data mining, text mining, machine learning, information retrieval.

DING Zhiyuan, born in 1993, M. S. candidate. His research interests include data mining, text mining, machine learning.

CHEN Yanping, born in 1980, Ph. D., associate professor. His research interests include artificial intelligence, natural language processing.

XU Liyang, born in 1990, M. S. candidate. His research interests include data mining, text mining, machine learning.

This work is partially supported by the National Natural Science Foundation of China (61462011), the Major Research Program of the National Natural Science Foundation of China (91746116), the Major Applied Basic Research Program of Guizhou Province (JZ20142001), the Major Spec-