

特征加权距离与软子空间学习相结合的文本聚类新方法

王 骏 王士同 邓赵红

(江南大学数字媒体学院 江苏 无锡 214122)

摘 要 文本数据维数高、数据分布稀疏、不同类别的特征相互重叠, 这为聚类分析提出了挑战. 针对文本数据的这一特点, 将特征加权技术与软子空间相结合, 基于模糊聚类的算法框架, 提出了一种适用于高维文本数据的软子空间模糊聚类新方法. 首先, 基于加权范数理论, 提出了新的特征加权距离计算方法. 接着, 将其与软子空间学习的理论框架相结合, 提出了面向模糊聚类的新的目标学习准则. 通过向约束条件中引入熵指数 r , 从而扩展了模糊指数 m 的取值范围, 并给出了物理解释. 基于 Zangwill 收敛定理对算法的全局收敛性给出理论证明. 实验表明, 文中算法可以使软子空间学习和聚类分析同时进行, 其性能比现有的相关算法有了较大的提高.

关键词 模糊聚类; 文本聚类; 软子空间; 特征加权距离; 全局收敛性

中图法分类号 TP311

DOI 号: 10.3724/SP.J.1016.2012.01655

A Novel Text Clustering Algorithm Based on Feature Weighting Distance and Soft Subspace Learning

WANG Jun WANG Shi-Tong DENG Zhao-Hong

(School of Digital Media, Jiangnan University, Wuxi, Jiangsu 214122)

Abstract The text data are characterized by high dimensionality and feature overlapping among different clusters, which is a great challenge for the real-world data mining applications. This paper proposes a novel fuzzy clustering algorithm by integrating the feature weighting metric into the framework of soft subspace learning. Firstly, the feature weighting metric is presented based on the concept of vector norm. Then a novel learning criterion is proposed based on the combination of feature weighting metric and soft subspace clustering. An entropy exponent r is introduced into the constraints so that the span of the fuzzy index m is extended. A physical explanation from the view of the information theory is given. A global convergence theory is also established by applying Zangwill's convergence theorem. At last, experiments are conducted on both synthesis and real text data and the experimental results show that the proposed algorithm can perform tasks of clustering analysis and soft subspace learning simultaneously and obtain better results than some of the existing approaches.

Keywords fuzzy clustering; text clustering; soft subspace; feature weighting distance; global convergence

1 引 言

文本数据具有维数高、数据分布稀疏、不同类别的特征相互重叠等重要特点, 这为实际应用带来了巨大的挑战.

聚类分析^[1]是实现文本挖掘的重要手段. 但是,

在对文本数据进行聚类分析时, 通常需要选择

收稿日期: 2008-09-26; 最终修改稿收到日期: 2009-12-01. 本课题得到国家自然科学基金(60903100, 60975027, 61170122)、江苏省自然科学基金(BK2011417)、江苏“333 高层次人才培养工程”(BRA201142)、中央高校基本科研业务专项资金项目(JUSRP111A38)资助.

王 骏, 男, 1978 年生, 博士, 讲师, 研究方向为数据挖掘、模式识别、机器学习. E-mail: wangjun_sytu@hotmail.com. 王士同, 男, 1964 年生, 教授, 博士生导师, 研究领域为模糊系统、模式识别、生物信息学、数字图像处理. 邓赵红, 男, 1981 年生, 副教授, 研究方向为模糊系统、模式识别、生物信息学.

合适的距离度量来评价样本点之间的相似或相异程度。为了合理计算高维文本数据之间的距离,在聚类过程中引入特征权重来强化重要特征的积极作用,削减冗余特征的不利影响,已经成为一种常用的方法^[2-8]。例如,参考文献[4-5]分别提出了两种基于特征加权的模糊聚类方法。它们的共同点在于,在聚类之前首先使用一个有监督或无监督的学习过程得到能够反映数据集内部结构的特征权重向量。在此基础上,形成特征加权的距离函数。然后基于 FCM 算法框架得到数据集的模糊划分。特征权重的学习过程揭示了数据集里各个类别的结构特征,这有利于后继的聚类过程。但是,在聚类过程中特征权重不再发生变化。因此,聚类过程受先前学习结果的制约。

与上述两种方法不同,Huang 等人^[6-7]提出了自动特征(变量)加权技术。通过向 k -means 或 FCM 中引入特征权重向量来表示整个数据集上各特征的重要程度。通过循环迭代过程,在聚类分析的同时实现了特征权重的无监督学习,这成为处理高维数据的另一种代表性方法。

上述方法的共同点在于,算法在整个特征空间中对特征进行加权,各个类别使用相同的特征权重向量。现实生活中,由于不同主题的文本,其关键词各不相同,每个关键词描述不同主题的能力也有很大差异,因此进行特征权重学习时,整个数据集采用单一的特征权重向量无法与数据集的结构特点相吻合。

近年来,软子空间聚类技术^[9-11]成为学术界的研究热点。其基本思想是,给数据集中的各类别赋予不同的特征权重向量,用来表示聚类过程中各特征对此类别贡献的大小。在聚类过程中,每一维特征对各个类别都有不同的贡献,因此每一类都有不同的特征权重向量,从而在整个特征空间中形成了若干个“软子空间”^[9],聚类过程就是在各个“软子空间”中进行的。Gan 等人^[10]提出的模糊子空间聚类算法 FSC、Jing 等人^[9]提出的熵加权 k 均值聚类算法 EWKM、Chen 等人^[11]提出的特征组 k 均值聚类算法 FG- k -means 就是其典型代表。

令人遗憾的是,现有的软子空间聚类算法大多建立在数据集硬划分的基础上。以 FSC 为例,它具有以下缺点:首先,聚类结果对初始点敏感,如果初始点集选择不合适,会出现若干聚类合并的现象。其原因在于,FSC 本质上采用硬划分方法,这与 k -means 是一样的。因此也不可避免地继承了 k -means 的缺点。其次,数据集的维数过高时,目标函数过早收敛,从而影响了算法的性能。

根据高维本文数据集的结构特点,本文将特征

加权距离、软子空间技术和模糊聚类理论相结合,提出了一种新的熵指数软子空间模糊聚类算法。其特点在于,在完成聚类任务的同时,算法可以自动进行特征权重的无监督学习,这对文本挖掘等应用领域有着重要的意义。从形式上看,加入新的模糊特征权重矩阵 $\mathbf{W}$,使目标函数中出现了两个模糊矩阵,这是目前的研究不曾涉及问题。在约束条件中,模糊隶属度的幂指数从 1 推广到了正实数,模糊指数 m 的取值范围也被推广到 $m > 0$ 。

本文第 2 节回顾 FCM 算法;第 3 节基于加权范数理论,提出特征加权距离的定义;将其与软子空间技术结合,同时向约束条件中引入熵指数 r ,提出本文的新算法,讨论参数的选取,并从信息论的角度给出相应的物理解释;第 4 节应用 Zangwill 全局收敛定理,进一步证明本文算法具有全局收敛性;第 5 节给出实验分析;最后是结论。

2 模糊 C 均值算法 FCM

设 $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ 为欧氏空间 $\mathbf{R}^d$ 中的一个有限数据集, c 为一整数,满足 $2 \leq c < n$ 。用 R^{cn} 表示所有的实 $c \times n$ 的矩阵集合。令 M_{fcn} 表示数据集 X 上划分矩阵的集合, $\mathbf{U} = [u_{ji}] \in R^{cn}$ 是模糊划分矩阵,其每一行元素定义了 X 中的一个模糊聚类, u_{ji} 表示第 i 个样本隶属于聚类 j 的程度,满足如下条件:

$$0 \leq u_{ji} \leq 1, j = 1, 2, \dots, c; i = 1, 2, \dots, n \quad (1a)$$

$$\sum_{j=1}^c u_{ji} = 1, i = 1, 2, \dots, n \quad (1b)$$

$$0 < \sum_{i=1}^n u_{ji} < n, j = 1, 2, \dots, c \quad (1c)$$

$\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c)^T \in R^{cd}$ 表示聚类中心, $\|\cdot\|$ 为 $\mathbf{R}^d$ 上的内积范数, m 为模糊指标, M_{fcn} 为满足式(1)的所有模糊矩阵 $\mathbf{U}$ 的集合。FCM 的目标函数 $J_{\text{FCM}}: M_{fcn} \times R^{cd} \rightarrow \mathbf{R}$ 定义为

$$J_{\text{FCM}} = \sum_{i=1}^n \sum_{j=1}^c (u_{ji})^m \|\mathbf{x}_i - \mathbf{v}_j\|^2, m > 1 \quad (2)$$

根据式(1)、(2),可得 J_{FCM} 取局部极小值的必要条件如下:

$$\mathbf{v}_i = \frac{\sum_{j=1}^n (u_{ji})^m \mathbf{x}_j}{\sum_{j=1}^n (u_{ji})^m}, m > 1 \quad (3a)$$

$$u_{ji} = \frac{\|\mathbf{x}_i - \mathbf{v}_j\|^{-\frac{2}{m-1}}}{\sum_{j=1}^c \|\mathbf{x}_i - \mathbf{v}_j\|^{-\frac{2}{m-1}}}, m > 1 \quad (3b)$$

使用式(3a)、(3b)进行 Picard 迭代,使目标函

数收敛于局部极小点或鞍点,从而得到 X 的一个模糊划分矩阵 U^* .

从数学的角度来看 FCM, 可以把 FCM 看作在式(1)的约束条件下, 求解模糊划分矩阵 U^* , 使式(2)的目标函数取极小值. 这是一个非线性规划问题. 对 FCM 收敛性的研究也是研究的难点之一. 几经反复, Hathaway、Bezdek 等人^[12]终于证明了算法的目标函数最终收敛于其局部最小值点或鞍点.

3 熵指数软子空间模糊聚类算法 EI-FWD-SSFC

3.1 基于加权范数理论的特征加权距离

在数学上, 一个二元实值函数 $dist(x, y)$ 作为拓扑空间中的距离, 应该满足以下 3 个条件:

- (i) 非负性. $dist(x, y) \geq 0, \forall x \neq y, dist(x, x) = 0$;
- (ii) 对称性. $dist(x, y) = dist(y, x)$;
- (iii) 三角不等式. $dist(x, y) \leq dist(x, z) + dist(z, y), \forall z$.

本文的特征加权距离建立在加权范数理论的基础上, 首先介绍加权范数^[13]的定义.

定义 1. 设 $A \in R^{d \times d}$ 是正定矩阵, 对任意 $x \in R^d$ 规定

$$\|x\|_A = \sqrt{x^T A x} \quad (4)$$

则 $\|x\|_A$ 是一种向量范数, 称为加权范数.

定理 1. 设 $x = (x_1, x_2, \dots, x_d)^T, y = (y_1, y_2, \dots, y_d)^T, \omega_1, \omega_2, \dots, \omega_d$ 为正实数, 满足 $\sum_{h=1}^d \omega_h =$

1, 0 则 $wdist(x, y) = \sqrt{\sum_{h=1}^d \omega_h^\alpha (x_h - y_h)^2}$ 是一个距离, 其中 $\alpha > 1$.

证明. 非负性和对称性易证, 下面证明三角不等式成立.

设 $x, y, z \in R^d$, 根据以上加权范数的定义, 有

$$\begin{aligned} wdist(x, y) &= \sqrt{\sum_{h=1}^d \omega_h^\alpha (x_h - y_h)^2} \\ &= \sqrt{(x - y)^T A (x - y)} \\ &= \|x - y\|_A \end{aligned} \quad (5)$$

其中 $A = \text{diag}(\omega_1^\alpha, \omega_2^\alpha, \dots, \omega_d^\alpha)$ 是以 $\omega_1^\alpha, \omega_2^\alpha, \dots, \omega_d^\alpha$ 为对角元的对角矩阵. 显然 A 正定. 根据向量范数的相关性质^[13], 有

$$\begin{aligned} \|x - y\|_A &= \|(x - z) + (z - y)\|_A \\ &\leq \|x - z\|_A + \|z - y\|_A \end{aligned} \quad (6)$$

即 $wdist(x, y) \leq wdist(x, z) + wdist(z, y)$ 成立. 定

理得证.

证毕.

3.2 算法描述

对 FCM 的目标函数(2), 本文作以下改进. 首先, 用本文提出的特征加权距离来改进原 FCM 中的距离计算公式, 并添加适当的正则化项. 由于不同类别所处的软子空间使用不同的特征权重向量, 因此形成了模糊特征权重矩阵 W . 其次, 向约束条件中加入熵指数 r , 使算法更灵活通用.

设 $X = \{x_1, x_2, \dots, x_n\}$ 为欧氏空间 R^d 中的有限数据集. c 为一整数, 满足 $2 \leq c < n$. $V = (v_1, v_2, \dots, v_c)^T \in R^{cd}$ 为 c 个聚类中心. 实数 m, r, α 满足 $m > r > 0, \alpha > 1$. ϵ_w 为任意正实数, ϵ_u 为任意实数. U 为模糊划分矩阵, 满足如下条件:

$$0 \leq u_{ji} \leq 1, j = 1, 2, \dots, c; i = 1, 2, \dots, n \quad (7a)$$

$$\sum_{j=1}^c u_{ji}^r = 1, i = 1, 2, \dots, n; r > 0, r \neq 1 \quad (7b)$$

$$0 < \sum_{i=1}^n u_{ji} < n, j = 1, 2, \dots, c \quad (7c)$$

M_U 为满足以上条件的模糊划分矩阵 U 的集合. $W = [w_{jh}]_{c \times d}$ 为模糊特征权重矩阵, 满足如下条件:

$$0 \leq w_{jh} \leq 1, j = 1, 2, \dots, c; h = 1, 2, \dots, d \quad (8a)$$

$$\sum_{h=1}^d w_{jh} = 1, j = 1, 2, \dots, c \quad (8b)$$

M_W 为满足以上条件的模糊特征权重矩阵的集合. 将特征加权距离与软子空间相结合, 构造一般化的目标函数如下:

$$J_m(W, V, U) = \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m d_{ji} + f\left(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m, \sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha\right) \quad (9a)$$

其中, $d_{ji} = wdist^2(x_i, v_j) = \sum_{h=1}^d w_{jh}^\alpha (x_{ih} - v_{jh})^2$, 定义域 $D_f = \{(W, V, U) | W \in M_W, V \in R^{cd}, U \in M_U\}$.

这里 $f(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m, \sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha)$ 表示基于 $\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m$ 和 $\sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha$ 的正则化项, 通常可以取加法和乘法两种运算. 即

$$f_1\left(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m, \sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha\right) = \epsilon_w \sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha + \epsilon_u \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m$$

$$f_2\left(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m, \sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha\right) = \epsilon \left(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m\right) \left(\sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha\right).$$

本文中, 我们仅研究以 $f_1(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m, \sum_{j=1}^c \sum_{h=1}^d w_{jh}^\alpha)$ 为正则化项的情况. 即

$$J_m(\mathbf{W}, \mathbf{V}, \mathbf{U}) = \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m d_{ji} + \epsilon_w \sum_{j=1}^c \sum_{h=1}^d w_{jh}^a + \epsilon_u \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m \quad (9b)$$

对于以 $f_2(\sum_{j=1}^c \sum_{h=1}^d u_{ji}^m, \sum_{j=1}^c \sum_{h=1}^d w_{jh}^a)$ 为正则化项的情况, 我们将另文发表.

通过 Lagrange 乘子法, 得到在约束条件(7)、(8)下目标函数(9b)取极值的必要条件如下:

$$\mathbf{v}_j = \frac{\sum_{i=1}^n u_{ji}^m \mathbf{x}_i}{\sum_{i=1}^n u_{ji}^m}, \quad j=1, 2, \dots, c \quad (10a)$$

$$w_{jh} = \frac{1}{\sum_{l=1}^d \left[\frac{\sum_{i=1}^n u_{ji}^m (x_{ih} - v_{jl})^2 + \epsilon_w}{\sum_{i=1}^n u_{ji}^m (x_{il} - v_{jl})^2 + \epsilon_w} \right]^{\frac{1}{a-1}}}, \quad j=1, 2, \dots, c; \quad h=1, 2, \dots, d \quad (10b)$$

$$u_{ji} = \left(\frac{d_{ji} + \epsilon_u}{d_{li} + \epsilon_u} \right)^{\frac{1}{m/r-1}}^{-1/r}, \quad j=1, 2, \dots, c; \quad i=1, 2, \dots, n \quad (10c)$$

与 FCM 类似, EI-FWD-SSFC 通过以上 3 个必要条件之间的 Picard 迭代, 使目标函数(9b)收敛, 从而找到数据集 $\mathbf{X}$ 的一个模糊划分.

EI-FWD-SSFC 算法描述如下.

算法 1. EI-FWD-SSFC.

输入: D 为数据集; c 为聚类数目

输出: $\mathbf{U}$ 为模糊划分矩阵; $\mathbf{W}$ 为特征权重矩阵;

$\mathbf{V}$ 为聚类中心矩阵

步骤:

1. 随机选择 c 个样本点初始化聚类中心矩阵 $\mathbf{V}^{(0)}$;
2. 通过 $w_{jh} = 1/d (j=1, 2, \dots, c, h=1, 2, \dots, d)$, 初始化 $\mathbf{W}^{(0)}$;
3. 通过式(10c)计算 u_{ji} , 初始化 $\mathbf{U}^{(0)}$;
4. 根据式(9b)计算 $obj_fcn(0) = J_m(\mathbf{W}^{(0)}, \mathbf{V}^{(0)}, \mathbf{U}^{(0)})$;
5. 设置目标函数变化阈值 ϵ_{ps} , 最大迭代次数 max_iter , 迭代次数 $k=1$;
6. while ($k \leq max_iter$)
7. 根据式(10a)更新 $\mathbf{V}^{(k)}$;
8. 根据式(10b)更新 $\mathbf{W}^{(k)}$;
9. 根据式(10c)更新 $\mathbf{U}^{(k)}$;
10. 根据式(9b)计算 $obj_fcn(k) = J_m(\mathbf{W}^{(k)}, \mathbf{Z}^{(k)}, \mathbf{U}^{(k)})$;
11. $error = |obj_fcn(k) - obj_fcn(k-1)|$;
12. if ($error < \epsilon_{ps}$) break;
13. $k = k + 1$;
14. end
15. 输出 $\mathbf{W}, \mathbf{V}$ 和 $\mathbf{U}$.

算法首先通过随机选择 c 个数据点初始化 $\mathbf{V}$, $w_{jh} = 1/d$ 初始化 $\mathbf{W}$, 通过式(10c)来初始化模糊划分矩阵 $\mathbf{U}$. 随后算法通过 Picard 迭代, 依次计算中心点 $\mathbf{V}$ 、模糊特征权重矩阵 $\mathbf{W}$ 以及模糊划分矩阵 $\mathbf{U}$, 直到收敛.

使用式(10c)重新计算模糊隶属度 u_{ji} , 实际上就是根据样本点 $\mathbf{x}_i$ 到各类别中心点 $\mathbf{v}_j$ 的特征加权距离, 重新确定各样本点从属于每个类别的模糊度. 使用式(10b)更新模糊特征权重矩阵 $\mathbf{W}$ 即为对特征权重进行无监督学习的过程, 从而得到更合理的特征权重分布. 在计算 w_{jh} 时考虑了当前迭代中每个样本点从属于每一个聚类的模糊程度, 以及各个样本点到各个聚类中心的距离.

3.3 ϵ_w 和 ϵ_u 参数的选择

对于高维稀疏数据来说, 可能会出现某一维特征上的数据全为 0 的情况, 这时式(10b)中 $\sum_{i=1}^n u_{ji}^m (x_{ih} - v_{jh})^2$ 为 0; 此外, 当数据集某一类中某一维特征上的数据全为 0 时, $\sum_{i=1}^n u_{ji}^m (x_{ih} - v_{jh})^2$ 也可能接近 0. 参数 ϵ_w 的引入避免了式(10b)中分母为 0. 特别的, 当 $\epsilon_w \rightarrow \infty$ 时, 根据式(10b), $w_{jh} \rightarrow 1/d$, 此时算法退化为 FCM.

不妨设当 $h = h_0$ 时, $\sum_{i=1}^n u_{ji}^m (x_{ih_0} - v_{jh_0})^2 = 0, j = 1, 2, \dots, c$. 根据式(10b), 此时有

$$w_{jh_0} = \frac{1}{\sum_{l=1}^d \left[\frac{\sum_{i=1}^n u_{ji}^m (x_{il} - v_{jl})^2}{\epsilon_w} + 1 \right]^{\frac{1}{a-1}}} \quad (11)$$

对于任意 $l=1, 2, \dots, d, 0 \leq \sum_{i=1}^n u_{ji}^m (x_{il} - v_{jl})^2 \leq$

$\sum_{i=1}^n (x_{il} - v_{jl})^2$ 成立, 因此有

$$1/d \leq w_{jh_0} \leq \frac{1}{\sum_{l=1}^d \left[\frac{\sum_{i=1}^n (x_{il} - v_{jl})^2}{\epsilon_w} + 1 \right]^{\frac{1}{a-1}}}, \quad j=1, 2, \dots, c \quad (12)$$

即参数 ϵ_w 对 w_{jh_0} 的上限起着重要的调节作用. 显然, 其上限与数据集的结构特点有关.

聚类过程中, 当聚类中心 $\mathbf{v}_j$ 与某一样本点 $\mathbf{x}_i$ 过于接近而几乎重合时, 它们之间的加权距离 d_{ji} 为 0, ϵ_u 的引入也同样避免了分母为 0, 通常取接近于 0 的任意正实数. 当 ϵ_u 取负值时, 式(9b)可视为采用

了 ϵ 不敏感距离. 当 $\epsilon_u \rightarrow \infty$ 时, 根据式 (10c), $u_{ji} \rightarrow c^{-1/r}$, 此时算法将迅速收敛到整个测试集的中心

$$v_0 = \frac{\sum_{i=1}^n x_i}{n}.$$

3.4 模糊指数 m 和熵指数 r 的大小关系

与 FCM 类似, 实数 m 是模糊划分矩阵 U 的模糊指数. 约束条件中熵指数 r 的引入使 m 的范围由 FCM 中的 $m > 1$ 扩展为 $m > r > 0$. 当 $\frac{m}{r} \rightarrow 1^+$ 时, 算法退化为 FSC; 当 $\frac{m}{r} \rightarrow \infty$ 时, $u_{ji} \rightarrow c^{-1/r}$, 算法将迅速收敛到整个测试集的中心; 当 $m > 0, r > 0$ 且 $m < r$ 时, 由于不能保证目标函数 (9) 是 U 上的凸函数, 算法的收敛性无法保证.

实数 α 是模糊特征权重矩阵 W 的模糊指数. 当 $\alpha \rightarrow 1^+$ 时, 软子空间将演变为硬子空间; 当 $\alpha \rightarrow \infty$ 时, $w_{jh} \rightarrow 1/d$, 即各维特征的权重相等, 此时算法退化为 FCM.

针对 FCM 模糊指标 m 的选取问题, 于剑从理论的角度给出了选取准则^[14-15]. 对于软子空间聚类算法, 由于新增加了计算特征权重矩阵 W 这一步骤, 使得问题更加复杂. 本文根据多个测试集的实验, 当满足 $1.3 \leq \frac{m}{r} \leq 1.8, 2.0 \leq \alpha \leq 3.5$ 时, 本文算法可以取得比较理想的性能. 当然, 更深入的理论方法还有待进一步挖掘.

3.5 对约束条件的解释

在信息论中, 经常使用熵来描述随机变量的不确定性程度. 根据信息论的相关概念, 数据集样本点 x_i 的 Havrda-Charvat 熵定义为^[16]

$$H_{\beta}^{(r)}(x_i) = \frac{1}{1-2^{r-1}} \left(1 - \sum_{j=1}^c u_{ji}^r \right), \quad r > 0 \text{ 且 } r \neq 1 \quad (13)$$

显然, 如果把矩阵 U 看作概率矩阵, 当满足约束条件 (7b) 时, $H_{\beta}^{(r)}(x_i)$ 值为 0, 这意味着数据集 X 中样本点 x_i 从属于各个划分的不确定性最小. 故 EI-FWD-SSFC 算法的实质就是在数据集 X 中关于各个样本点的模糊隶属度 u_{ji} 的 Havrda-Charvat 熵具有最小值的情况下, 求解目标函数 (9b) 的最小值. 此时数据集 X 中各样本点所具有的 Havrda-Charvat 熵总量亦最小.

4 EI-FWD-SSFC 的收敛性

参考文献 [10, 12] 中分别证明了 FSC 和 FCM 的收敛性. 但是引入模糊特征权重矩阵 W 后, 目标

函数中出现了两个模糊矩阵. 此外, 向约束条件中引入指数 r 后, 问题发生的变化, 因此有必要重新证明算法的收敛性.

4.1 Zangwill 收敛定理

要判断一个迭代算法是否全局收敛, 只要判断其是否符合 Zangwill 收敛定理的 3 个条件即可. 首先介绍 Zangwill 收敛定理如下.

引理 1 (Zangwill 收敛定理). 假设点 $z^{(0)} \in V$, 由点-集映射 $A: V \rightarrow P(V)$ 定义的迭代算法产生迭代序列 $\{z^{(k)}\}$, 解集 $\Omega \subset V$. 如果:

- (1) $\{z^{(k)}\} \subset \Gamma \subset V$, 其中 Γ 是紧集.
 - (2) 存在连续函数 $J: V \rightarrow \mathbf{R}$, 满足: (a) 若 $z \notin \Omega$, 则对于任意 $y \in A(z)$, $J(y) < J(z)$; (b) 若 $z \in \Omega$, 算法终止, 或对于任意 $y \in A(z)$, $J(y) \leq J(z)$.
 - (3) 若 $z \notin \Omega$, 映射 A 在 z 处是闭的.
- 则算法终止于解集 Ω , 或 $\{z^{(k)}\}$ 的任一收敛子序列的极限是解集 Ω 的一个点.

通常把满足上述引理中条件 (2) 的函数 J 称为是 A 在 V 上的 Zangwill 函数^[17-18].

4.2 EI-FWD-SSFC 全局收敛性证明

在证明之前, 还需引入几个定义和定理.

定义 2. 设 $A: V \rightarrow P(V)$ 是一点-集映射, 如果在点 $x^* \in X$ 处, 有 $\{x^{(k)}\} \subset X, x^{(k)} \rightarrow x^*, y^{(k)} \in A(x^{(k)}), y^{(k)} \rightarrow y^*$, 使得 $y^* \in A(x^*)$, 则称点-集映射 A 在 $x^* \in X$ 处是闭的. 如果它在 X 中的每一点是闭的, 则称点-集映射 A 在 X 上是闭的.

定义 3a. 定义映射 $T_1: M_W \times M_U \rightarrow R^{cd}$:
$$T_1(W, U) = V = (v_1, v_2, \dots, v_c)^T \quad (14a)$$

其中 $v_j = (v_{j1}, v_{j2}, \dots, v_{jd})^T \in R^d, j = 1, 2, \dots, c$, 通过式 (10a) 计算而得.

定义 3b. 定义映射 $T_2: R^{cd} \times M_U \rightarrow M_W$:
$$T_2(V, U) = W = (w_1, w_2, \dots, w_c)^T \quad (14b)$$

其中 $w_j = (w_{j1}, w_{j2}, \dots, w_{jd})^T, j = 1, 2, \dots, c$, 通过式 (10b) 计算而得.

定义 3c. 定义映射 $T_3: M_W \times R^{cd} \rightarrow M_U$:
$$T_3(W, V) = U \quad (14c)$$

其中 $U = [u_{ji}]_{c \times n}$ 通过式 (10c) 计算而得.

定义 4. 定义映射 $T_m: M_W \times R^{cd} \times M_U \rightarrow M_W \times R^{cd} \times M_U$ 如下:
$$T_m(W, V, U) = \{(W^*, V^*, U^*) | V^* = T_1(W, U),$$

$W^* = T_2(V^*, U), U^* = T_3(W^*, V^*)\}.$
定义 5. 如果 $(W^{(k)}, V^{(k)}, U^{(k)}) \in T_m(W^{(k-1)}, V^{(k-1)}, U^{(k-1)})$, $k = 2, 3, \dots$. 这里 $(W^{(1)}, V^{(1)}, U^{(1)})$ 为 $M_W \times R^{cd} \times M_U$ 中的任意值, 则序列 $\{(W^{(k)}, V^{(k)}, U^{(k)})\}$ 称为 EI-FWD-SSFC 迭代序列.

定理 2. 设 $X = \{x_1, x_2, \dots, x_n\}$, 定义 $\Xi(U) = J_m(W^*, V^*, U)$ 为 $\Xi: M_U \rightarrow \mathbf{R}$ 上的函数, 其中 $W^* \in M_W, V^* \in R^{cd}, \alpha > 1, m > r > 0, \epsilon_w > 0, \epsilon_u > 0$ 为定值. 当且仅当 $U = T_3(W^*, V^*)$ 时, U 为 Ξ 在 M_U 上的全局最小值点.

定理 3. 设 $X = \{x_1, x_2, \dots, x_n\}$, 定义 $\Theta(W) = J_m(W, V^*, U^*)$ 为 $\Theta: M_W \rightarrow \mathbf{R}$ 上的函数, 其中 $V^* \in R^{cd}, U^* \in M_U, \alpha > 1, m > r > 0, \epsilon_w > 0, \epsilon_u > 0$ 为定值. 当且仅当 $W = T_2(V^*, U^*)$ 时, W 为 Θ 在 M_W 上的全局最小值点.

定理 4. 设 $X = \{x_1, x_2, \dots, x_n\}$, 定义 $\Psi(V) = J_m(W^*, V, U^*)$ 为 $\Psi: R^{cd} \rightarrow \mathbf{R}$ 上的函数, 其中 $W^* \in M_W, U^* \in M_U, \alpha > 1, m > r > 0, \epsilon_w > 0, \epsilon_u > 0$ 为定值. 当且仅当 $V = T_1(W^*, U^*)$ 时, V 为 Ψ 在 R^{cd} 上的全局最小值点.

定理 5 (EI-FWD-SSFC 收敛定理). 设 $X = \{x_1, x_2, \dots, x_n\}$ 包含至少 c 个独立的点, $c < n$, 定义 $\Omega = \{(W^*, V^*, U^*) \in M_W \times R^{cd} \times M_U \mid$

$$\begin{aligned} J_m(W^*, V^*, U^*) &< J_m(W^*, V^*, U), \forall U^* \neq U; \\ J_m(W^*, V^*, U^*) &< J_m(W, V^*, U^*), \forall W^* \neq W; \\ J_m(W^*, V^*, U^*) &< J_m(W^*, V, U^*), \forall V^* \neq V \} \end{aligned} \quad (15)$$

为如下最优化问题的解集:

$$\begin{aligned} \min J_m(W, V, U), \\ \text{s. t. } U \in M_U, W \in M_W, V \in R^{cd}, \end{aligned}$$

其中, $J_m(W, V, U)$ 如式(9)计算. 令 $(W^{(0)}, V^{(0)}, U^{(0)})$ 为根据映射 T_m 进行迭代的起点, $W^{(0)} \in M_W, V^{(0)} \in R^{cd}, U^{(0)} = T_3(W^{(0)}, V^{(0)})$. 则迭代序列 $\{(W^{(k)}, V^{(k)}, U^{(k)})\}, k=1, 2, \dots$, 终止于解集 Ω 中的一个点 (W^*, V^*, U^*) , 或者存在一子序列收敛于 Ω 中的一个点.

证明.

(1) 设 $\text{conv}(X) = \{x \in \mathbf{R}^d \mid x = \sum_{i=1}^n \alpha_i x_i; \sum_{i=1}^n \alpha_i = 1; \alpha_i \geq 0; x_i \in X \subset \mathbf{R}^d\}$ 为数据集 X 上的凸包. 由式(10a)易得, $v_j^{(k)} \in \text{conv}(X), j=1, 2, \dots, c, k=1, 2, \dots; V = \{v_1, v_2, \dots, v_c\} \in [\text{conv}(X)]^T$. 显然, $(W^{(k)}, V^{(k)}, U^{(k)}) \in M_W \times [\text{conv}(X)]^T \times M_U \subset M_W \times R^{cd} \times M_U, k=1, 2, \dots$. 而 $M_W, \text{conv}(X), M_U$ 均为有界闭集, 故它们均为紧集, $M_W \times [\text{conv}(X)]^T \times M_U$ 亦为紧集.

(2) J_m 为 T_m 在 X 上的 Zangwill 函数. 由于 $V^{(k+1)} = T_1(W^{(k)}, U^{(k)})$, 根据定理 4, 有 $J_m(W^{(k)}, V^{(k+1)}, U^{(k)}) \leq J_m(W^{(k)}, V^{(k)}, U^{(k)})$. 同理, 根据定理 3, 可得如下不等式: $J_m(W^{(k+1)}, V^{(k+1)}, U^{(k)}) \leq J_m(W^{(k)}, V^{(k+1)}, U^{(k)})$; 根据定理 2, 可得如下不等式: $J_m(W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) \leq J_m(W^{(k+1)}, V^{(k+1)}, U^{(k)})$, 即

$$J_m(W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) \leq J_m(W^{(k)}, V^{(k)}, U^{(k)}) \quad (16)$$

下面证 $(W^{(k)}, V^{(k)}, U^{(k)}) \notin \Omega$ 时, 取严格不等号. 用反证法. 假设存在 $(W^{(k)}, V^{(k)}, U^{(k)}) \notin \Omega$, 使 $J_m(W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) = J_m(W^{(k)}, V^{(k)}, U^{(k)})$ 成立, 则必有

$$\begin{aligned} J_m(W^{(k)}, V^{(k)}, U^{(k)}) &= J_m(W^{(k)}, V^{(k+1)}, U^{(k)}) \\ &= J_m(W^{(k+1)}, V^{(k+1)}, U^{(k)}) \\ &= J_m(W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) \end{aligned} \quad (17)$$

根据定理 4, 有 $V^{(k+1)} = T_1(W^{(k)}, U^{(k)})$ 为 $\Psi(V) = J_m(W^{(k)}, V, U^{(k)})$ 的全局最小值点. 又根据式(17)中 $J_m(W^{(k)}, V^{(k)}, U^{(k)}) = J_m(W^{(k)}, V^{(k+1)}, U^{(k)})$ 成立, 根据全局最小值点的唯一性, 有 $V^{(k)} = V^{(k+1)}$. 同理可得, $W^{(k)} = W^{(k+1)}, U^{(k)} = U^{(k+1)}$. 所以有

$$(W^{(k)}, V^{(k)}, U^{(k)}) = (W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) \quad (18)$$

由于 $(W^{(k)}, V^{(k)}, U^{(k)}) \notin \Omega$, 由解集 Ω 的定义, 存在以下 3 种情况:

① 存在 $V^* \neq V^{(k)}$, 使

$$J_m(W^{(k)}, V^{(k)}, U^{(k)}) \geq J_m(W^{(k)}, V^*, U^{(k)}) \quad (19a)$$

成立;

② 存在 $W^* \neq W^{(k)}$, 使

$$J_m(W^{(k)}, V^{(k)}, U^{(k)}) \geq J_m(W^*, V^{(k)}, U^{(k)}) \quad (19b)$$

成立;

③ 存在 $U^* \neq U^{(k)}$, 使

$$J_m(W^{(k)}, V^{(k)}, U^{(k)}) \geq J_m(W^{(k)}, V^{(k)}, U^*) \quad (19c)$$

成立.

下面以①为例进行讨论. 当存在 $V^* \neq V^{(k)}$, 使式(19a)成立时, 根据 $V^{(k)} = V^{(k+1)}$, 有 $V^* \neq V^{(k+1)}$. 根据定理 4, $V^{(k+1)} = T_1(W^{(k)}, U^{(k)})$, 有

$$\begin{aligned} J_m(W^{(k)}, V^*, U^{(k)}) &> J_m(W^{(k)}, V^{(k+1)}, U^{(k)}) \\ &= J_m(W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) \end{aligned} \quad (20)$$

综合式(19a)、(20), 有

$$J_m(W^{(k)}, V^{(k)}, U^{(k)}) > J_m(W^{(k+1)}, V^{(k+1)}, U^{(k+1)}) \quad (21)$$

这与式(17)矛盾.

同理, 根据②、③均可推出矛盾. 故假设不成立, 即当 $(W^{(k)}, V^{(k)}, U^{(k)}) \notin \Omega$ 时, 式(16)取严格不等号. 故 J_m 是 T_m 在 X 上的 Zangwill 函数.

(3) 根据定义 3, 显然 $T_1: M_W \times M_U \rightarrow R^{cd}$ 和 $T_2: R^{cd} \times M_U \rightarrow M_W$ 是连续映射. 下面根据定义 2 证明 $T_3: M_W \times R^{cd} \rightarrow M_U$ 是闭映射. 不妨设

$$(W^{(k)}, V^{(k)}) \rightarrow (W^*, V^*), k \rightarrow \infty,$$

$$U^{(k)} = T_3(W^{(k)}, V^{(k)}), k=1, 2, \dots,$$

$$U^{(k)} \rightarrow U^*, k \rightarrow \infty.$$

因为

$$\begin{aligned} u_{ji}^* &= \lim_{k \rightarrow \infty} u_{ji}^{(k)} = \lim_{k \rightarrow \infty} \frac{1}{\left(\sum_{h=1}^c \left(\frac{d_{ji}^{(k)} + \epsilon_u}{d_{jh}^{(k)} + \epsilon_u} \right)^{\frac{1}{m/r-1}} \right)^{1/r}} \\ &= \frac{1}{\left(\sum_{h=1}^c \left[\frac{\lim_{k \rightarrow \infty} (d_{ji}^{(k)} + \epsilon_u)}{\lim_{k \rightarrow \infty} (d_{jh}^{(k)} + \epsilon_u)} \right]^{\frac{1}{m/r-1}} \right)^{1/r}} \end{aligned} \quad (22)$$

又

$$\begin{aligned} \lim_{k \rightarrow \infty} (d_{ji}^{(k)} + \epsilon_u) &= \lim_{k \rightarrow \infty} d_{ji}^{(k)} + \epsilon_u \\ &= \lim_{k \rightarrow \infty} \sum_{h=1}^d (\omega_{jh}^{(k)})^a (x_{ih} - v_{jh}^{(k)})^2 + \epsilon_u \\ &= \sum_{h=1}^d (\lim_{k \rightarrow \infty} \omega_{jh}^{(k)})^a (x_{ih} - \lim_{k \rightarrow \infty} v_{jh}^{(k)})^2 + \epsilon_u \\ &= \sum_{h=1}^d (\omega_{jh}^*)^a (x_{ih} - v_{jh}^*)^2 + \epsilon_u > 0 \end{aligned} \quad (23)$$

即

$$U^* = T_3(W^*, V^*) \quad (24)$$

成立. 所以, $T_3: M_W \times R^{cd} \rightarrow M_U$ 为闭映射. 易得 $T_m: M_W \times R^{cd} \times M_U \rightarrow M_W \times R^{cd} \times M_U$ 亦为闭映射^[10]. 当然在 $\{D_f - \Omega\}$ 上是闭映射, 其中 D_f 如式(9(b))定义. 综合上述, 定理 5 得证. 证毕.

5 实 验

5.1 模拟数据实验

本实验用来验证算法 EI-FWD-SSFC 在进行有

效聚类的时候, 可以成功地发现各类别包含的重要特征. 所使用的模拟数据集采用参考文献[19]的方法生成, 由 1000 个包含 50 个特征的本点构成, 包含两类, 具有以下特点:

- (1) 它包含了不止一个类别.
- (2) 在各类别的相关特征所构成的特征子空间中, 数据点的坐标值较为集中; 在各类别的无关特征所构成的特征子空间中, 数据点的坐标值在某一区间均匀分布, 并且接近于 0.
- (3) 不同类别所具有的特征子集相互重合. 模拟数据的这种空间分布特点与文本数据的结构是类似的. 相关信息见表 1.

表 1 模拟数据集信息		
类别	相关特征号	Points
1	1, 3, 11, 16, 17, 19, 21, 22, 24, 25, 27, 29, 31, 33, 35, 37, 38, 39, 40, 44, 49, 50	726
2	1, 2, 3, 5, 8, 10, 11, 16, 18, 19, 20, 21, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 44	274

本实验中, 参数设置如下: $m = 1.5$, $r = 1.1$, $\alpha = 3$, $\epsilon_u = 10^{-14}$, $\epsilon_w = 0.1$.

图 1(a)和图 1(b)分别给出了聚类结果各类中每一维特征的权重. 对照表 1, 我们发现, 在聚类结果中, 权重大的特征恰为表 1 所列的生成数据时各类别的相关特征, 可见本文算法可以成功地发现数据集各类别包含的相关特征.

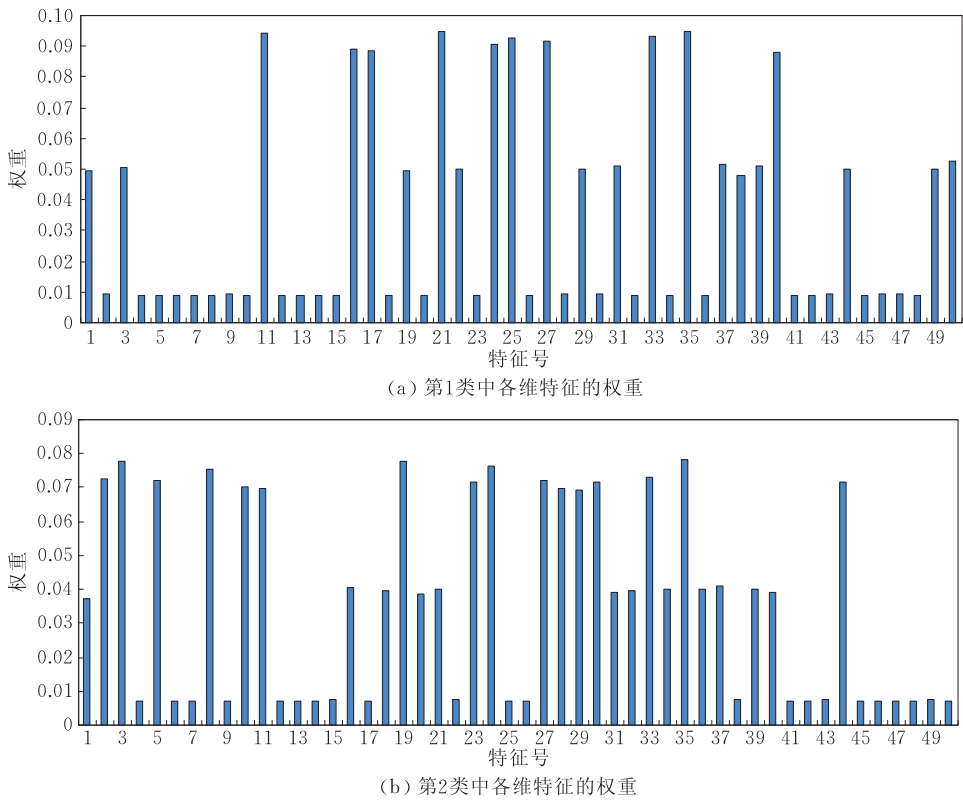


图 1

5.2 真实数据实验

5.2.1 测试样本集

文本聚类算法的性能受多个因素影响,比如测试集包含类别的数量、各个类别之间语义上的差别、每个类别的大小等等.为研究这些因素如何影响本文算法,我们基于标准文本数据集 classic、k1b、NG20^① 建立测试样本集.通过计算 $tf \cdot idf$,去除

idf 值过高和过低的特征.

本文建立如表 2 所示的 9 个测试集,这 9 个测试集分为 3 个系列.每一系列中,依次向前一测试集增加 1~2 个类别,从而构成新测试集.每个测试集所包含的类别数量由字母代号后的数字给出.此外,我们兼顾平衡数据集与非平衡数据集对算法性能的影响,K4 和 K6 为非平衡数据集,其它的都是平衡数据集.

表 2 测试集信息

数据源	测试集	包含类别	样本点总数	特征数量	每类样本点数量
classic	C2	1,2	200	2458	类中包含 100 个样本点
	C3	1,2,3	300	2539	
	C4	1,2,3,4	400	3134	
k1b	K2	1,3	200	3341	第 5 类包含 60 个样本点, 其它每类包含 100 个样本点
	K4	1,2,3,5	360	4847	
	K6	1,2,3,4,5,6	560	6327	
NG20	NG2	4,17	200	6163	每类中包含 100 个样本点
	NG4	4,8,13,17	400	8476	
	NG6	4,8,10,13,16,17	600	11263	

5.2.2 评估指标与测试平台

本文从 $RandIndex$ ^[20] 指标和算法的运行时间这两个方面来评估算法的性能.

$RandIndex$ 指标用来评价聚类结果与外部准则(external criterion)相吻合的程度.对于所给点集 $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$,假设 $U^{(e)} = \{u_1^{(e)}, u_2^{(e)}, \dots, u_R^{(e)}\}$ 和 $U = \{u_1, u_2, \dots, u_C\}$ 是数据集 X 上两种不同的划分,对于 $1 \leq i \neq i' \leq R, 1 \leq j \neq j' \leq C$,有 $\bigcup_{i=1}^R u_i^{(e)} = S = \bigcup_{j=1}^C u_j$, $u_i^{(e)} \cap u_{i'}^{(e)} = \emptyset = u_j \cap u_{j'}$.若 $U^{(e)}$ 是外部准则, U 是聚类结果,令 a 为在 $U^{(e)}$ 和 U 中均出现在同一类中的点对的数量, b 为在 $U^{(e)}$ 中出现在同一类中、而在 U 中出现在不同类中的点对的数量, c 为在 $U^{(e)}$ 中出现在不同类中、而在 U 中出现在同一类中的点对的数量, d 为在 $U^{(e)}$ 和 U 中均出现在不同类中的点对数量,则 $RandIndex$ 通过如下式子进行计算:

$$RandIndex = \frac{a+d}{a+b+c+d} \tag{25}$$

显然, $RandIndex$ 值介于 $[0,1]$ 之间.当两种分类情况完全一致时, $RandIndex$ 的值为 1.

本文所用实验平台如表 3 所示.

表 3 实验平台

结构	配置
CPU	Intel T7100
内存	1 GB
操作系统	Vista Home Edition
软件平台	Matlab 7.0

5.2.3 本文算法和 FSC 性能比较

本实验考查经过充分迭代后,本文算法和 FSC

性能的差异.为了避免迭代终止阈值的影响,我们设置算法迭代次数为 50 次,使算法充分迭代后收敛.本文算法参数设置为 $m = 1.5, r = 1.1, \alpha = 3, \epsilon_u = 10^{-14}, \epsilon_w = 0.1$; FSC 参数设置为 $\alpha = 3.5, \epsilon = 0.1$.在每个测试集上分别运行算法 10 次,得到聚类结果 $RandIndex$ 指标的平均值以及每次迭代所需要的平均时间,如表 4 所示.

表 4 本文算法和 FSC 聚类结果比较

测试集	$RandIndex$		每次迭代的平均时间/s	
	本文算法	FSC	本文算法	FSC
C2	0.8359	0.5175	0.18	0.18
C3	0.6708	0.3860	0.41	0.41
C4	0.6828	0.2972	0.89	0.90
K2	0.8485	0.4991	0.26	0.28
K4	0.7954	0.4020	1.25	1.26
K6	0.6422	0.2831	3.72	3.76
NG2	0.7371	0.4995	0.46	0.45
NG4	0.6661	0.2824	2.47	2.38
NG6	0.7445	0.2729	7.04	7.06

从表 4 中我们不难发现,对于具有上千个特征的高维文本数据,本文算法在不增加时间开销的前提下,可以取得比 FSC 更好的效果,这说明使用模糊划分矩阵 U 来代替 FSC 中的二值划分矩阵有效地提高了算法的性能.

下面我们以目标函数的取值变化为研究对象,分析 FSC 性能下降的原因.以测试集 NG2 为例,图 2 给出运行本文算法和 FSC 时目标函数的变化过程,其它测试集也可得到类似结果.

① <http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/datasets.tar.gz>, 2009

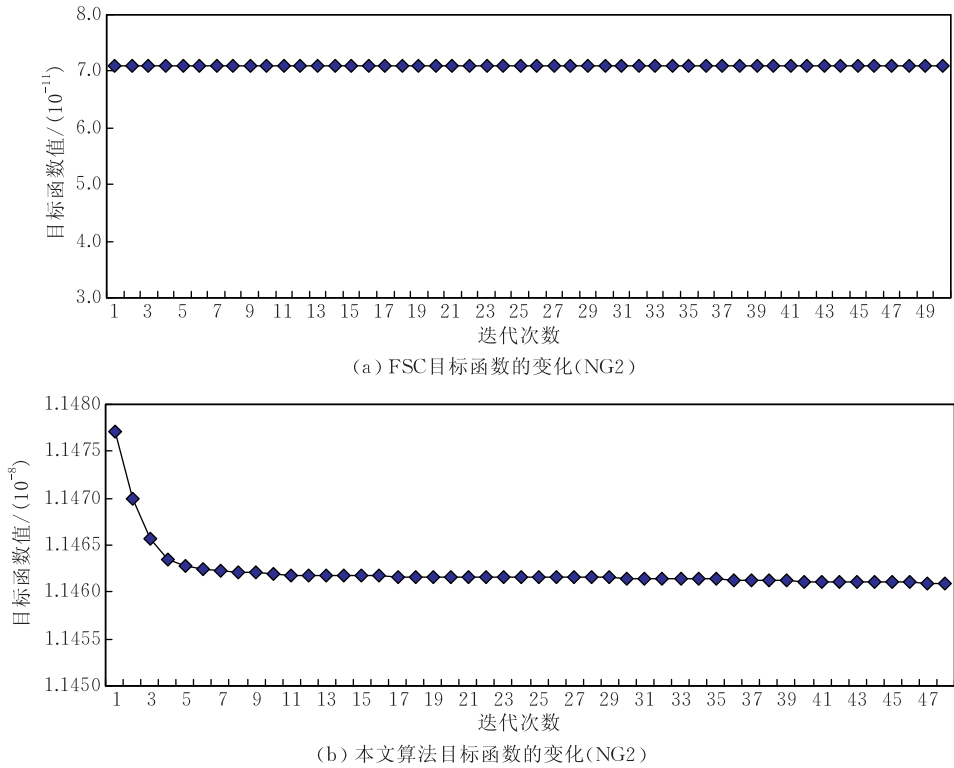


图 2

如图 2(b)中,本文算法的目标函数随着迭代的推进而递减,经过 10 次左右的迭代后基本达到平衡状态,这一过程也是算法寻找数据集上最优划分的过程.而在迭代过程中,FSC 目标函数的取值始终不发生变化(如图 2(a)),寻找更优划分的过程也因此而停止.

下面我们分析原因.以数据集 NG2 为例,对其进行预处理后,数据集是一个具有 6163 维的高维矩阵.特征数量过多,即 d 值过大,使用 $w_{jh}=1/d$ ($j=1,2,\cdots,c,h=1,2,\cdots,d$)对特征权重矩阵 $\mathbf{W}$ 进行初始化^[10], w_{jh} 值就会非常小.FSC 更新二值划分矩阵时使用的是最近邻原则,可以将 u_{ji} 的更新公式看成是一个对 w_{jh} 变化不敏感的滤波器.当 w_{jh} 值很小时,其绝对变化量也非常小,这不足以使 d_{ji} 发生很大的变化从而使 u_{ji} 值也随之变化.当划分矩阵 $\mathbf{U}$ 停止更新时,目标函数逐渐减小的过程也因此终止,算法也过早地陷于局部极值而不能自拔.而本文算法采用式(10c)更新模糊划分矩阵 $\mathbf{U}$,其定义域为连续空间, $\mathbf{W},\mathbf{V}$ 任何细微变化都会使 $\mathbf{U}$ 进行更新,从而使聚类结果更趋于合理,所以迭代过程可以向前推进.

5.2.4 特征数量的变化对不同特征加权算法的影响

本实验研究算法在充分迭代的前提下,特征数量的变化对不同特征加权算法的影响.使用表 2 中

最为复杂的测试集 NG6,指定迭代次数为 50 次,依次改变特征数量,从 200 递增到 3000.分别运行本文算法($m=1.5, r=1.1, \alpha=3, \epsilon_u=10^{-14}, \epsilon_w=0.1$)、WKMeans($\beta=2$)^[6]、EWKM($\gamma=10\,000$)^[9]、FSC($\alpha=3.5, \epsilon=0.1$)^[10]各 10 次.采用实验平台如表 3 所示,得到平均运行时间和 $RandIndex$ 指标分别如图 3(a)和图 3(b)所示.

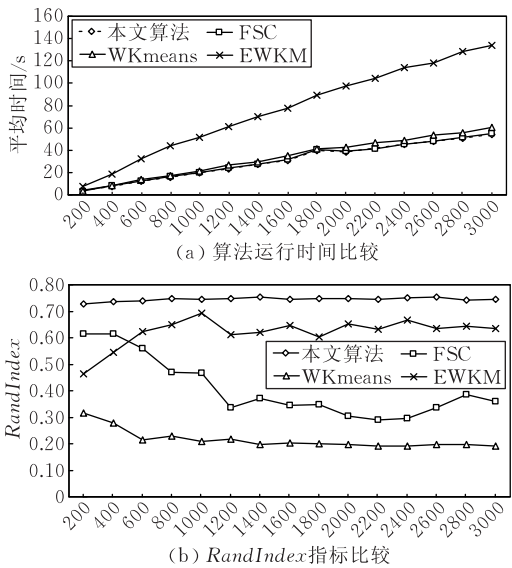


图 3

从图 3(a)中我们可以看出,在相同迭代次数的前提下,EWKM 需要耗费更多的时间,虽然 4 个算

法每次迭代的时间复杂度均为 $O(nkd)$, 其中 n, k, d 分别为样本点、类别、特征的数量。但是, 熵运算需要更大的运算量, 因此 EWKM 需要更多的时间。

图 3(b) 给出了在特征数量逐渐增加的情况下, 分别运行这 4 个算法所得结果的 *RandIndex* 值。比较本文算法和 WKmeans, 充分体现了特征加权与软子空间相结合的优势: 本文算法在不同的软子空间中进行聚类, 不同的类别使用不同的权重向量进行特征加权; 而 WKmeans 算法虽然进行了特征加权, 但是没有考虑到不同类别中特征权重的差异, 所以聚类效果不理想。

将本文算法与 FSC 相比, 显然, 用模糊划分矩阵代替二值划分矩阵, 有利于算法性能的提高。随着特征数量的增加, FSC 聚类结果的 *RandIndex* 指标有逐步减小的趋势。此外, FSC 和 EWKM 均为硬聚类算法 k -means 与软子空间相结合的产物, 其效果不如本算法。

另一方面, 对于特征数量的增加, 本文算法的聚类结果较其它算法更加平稳, 其 *RandIndex* 指标维持在 0.75 左右, 可见本文算法对特征数量的变化表现出了较强的鲁棒性。

6 结 论

将特征加权距离与软子空间相结合, 本文提出了一种适用于高维文本数据的模糊聚类新算法。其特点是: 算法在进行聚类的时候, 可以成功地进行特征权重的无监督学习, 这对于文本数据挖掘等应用有着重要的意义。从形式上讲, 本文算法的目标函数中出现了两个模糊矩阵; 此外目标函数的约束条件中引入了熵指数, 从而扩大了模糊指数 m 和 α 的选取范围。本文从信息论的角度, 给出了其物理解释; 并应用 Zangwill 定理就算法的全局收敛性给出了理论证明。通过在模拟数据集和真实数据集上的实验表明, 本文算法的性能较现有的特征加权算法有了很大的提高。

在实验过程中, 我们发现合理选择 m, r, α 的值对算法性能有较大影响, 因此, 如何合理选取软子空间聚类算法中的参数将是我们进一步要解决的问题。

参 考 文 献

- [1] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981
- [2] Green P E, Carmone F J, Kim J. A preliminary study of optimal variable weighting in k -means clustering. Journal of Classification, 1990, 7(2): 271-285
- [3] Makarenkov V, Legendre P. Optimal variable weighting for ultrametric and additive trees and k -means partitioning: Methods and software. Journal of Classification, 2001, 18(2): 245-271
- [4] Wang Li-Juan, Guan Shou-Yi, Wang Xiao-Long, Wang Xi-Zhao. Fuzzy C Mean algorithm based on feature weights. Chinese Journal of Computers, 2006, 29(10): 1797-1803(in Chinese)
(王丽娟, 关守义, 王晓龙, 王熙照. 基于属性权重的 Fuzzy C Mean 算法. 计算机学报, 2006, 29(10): 1797-1803)
- [5] Li Jie, Gao Xin-Bo, Jiao Li-Cheng. A new feature weighted fuzzy clustering algorithm. Acta Electron Sinica, 2006, 34(1): 89-92(in Chinese)
(李洁, 高新波, 焦李成. 基于特征加权的模糊聚类新算法. 电子学报, 2006, 34(1): 89-92)
- [6] Huang J Z, Ng M K, Rong H, Li Z. Automated variable weighting in k -means type clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 1-12
- [7] Wang Qiang, Ye Yunming, Huang J Z. Fuzzy k -means with variable weighting in high dimensional data analysis//Proceedings of the 9th International Conference on Web-Age Information Management. Zhangjiajie, China, 2008: 365-372
- [8] Modha D S, Spangler W S. Feature weighting in k -means clustering. Machine Learning, 2003, 52(3): 217-237
- [9] Jing Liping, Ng M K, Huang J Z. An entropy weighting k -means algorithm for subspace clustering of high-dimensional sparse data. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1026-1041
- [10] Gan G, Wu J. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm. Pattern Recognition, 2008, 41(6): 1939-1947
- [11] Chen X, Ye Y, Xu X, Huang J Z. A feature group weighting method for subspace clustering of high-dimensional data. Pattern Recognition, 2012, 45(1): 434-436
- [12] Hathaway R, Bezdek J, Tucker W. An improved convergence theorem for the fuzzy c-means clustering algorithms//Bezdek J. Analysis of Fuzzy Information, vol. III. Boca Raton: CRC Press, 1987: 123-131
- [13] Xu Zhong, Zhang Kai-Yuan et al. Introduction to Matrix Theory. Beijing: Science Press, 2004(in Chinese)
(徐仲, 张凯院等. 矩阵论简明教程. 北京: 科学出版社, 2004)
- [14] Yu Jian. On the fuzziness index of the FCM algorithms. Chinese Journal of Computers, 2003, 26(8): 968-973(in Chinese)
(于剑. 论模糊 C 均值算法的模糊指标. 计算机学报, 2003, 26(8): 968-973)
- [15] Yu J, Cheng Q, Huang H. Analysis of the weighting exponent in the FCM. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2004, 34(1): 634-638
- [16] Havrda J H, Charvat F. Quantification methods of classification processes: Concepts of structural α -entropy. Kybernetika, 1967, 3: 30-35
- [17] Zang W. Nonlinear Programming: A Unified Approach. Englewood Cliffs, NJ: Prentice-Hall, 1969
- [18] Luenberger D G. Linear and Nonlinear Programming. 2nd

Edition. Reading, Mass; Addison-Wesley, 1984

- [19] Aggarwal C C, Procopiuc C M, Wolf J L, Yu P S, Park J S. Fast algorithms for projected clustering//Proceedings of the 1999 ACM SIGMOD International Conference on Manage-

ment of Data. ACM Press, 1999; 61-72

- [20] Rand W M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 1971, 66: 846-850



WANG Jun, born in 1978, Ph. D., lecturer. His research interests include data mining, pattern recognition.

WANG Shi-Tong, born in 1964, professor, Ph. D. supervisor. His research interests include fuzzy system, pattern recognition, bioinformatics, digital image processing.

DENG Zhao-Hong, born in 1981, associate professor. His research interests include fuzzy system, pattern recognition, bioinformatics.

附 录. 定理 2、定理 3 证明.

在证明这两条定理之前,首先引入一条引理.

引理. 设 M_{fcn} 为满足式(1)的所有矩阵的集合, $\mathbf{A} = [\lambda_{ji}]_{c \times n} \in M_{fcn}$, c_{ji} 为正实数,其中 $j=1,2,\dots,c$, $i=1,2,\dots,n$.

$m>1$,设函数 $\varphi: M_{fcn} \rightarrow \mathbf{R}$ 如下定义: $\varphi(\mathbf{A}) = \sum_{i=1}^n \sum_{j=1}^c (\lambda_{ji})^m c_{ji}$, 则 $\mathbf{A}^*$ 为 φ 的严格全局极小值点当且仅当 $\mathbf{A}^*$ 如下计算:

$$\lambda_{ji}^* = \frac{1}{\sum_{k=1}^c \left[\frac{c_{ji}}{c_{ki}} \right]^{\frac{1}{m-1}}}, j=1,2,\dots,c, i=1,2,\dots,n \quad (\text{附 } 1)$$

定理 2 证明如下:

证明. 当 $\mathbf{W}^*, \alpha$ 为定值时,如式(9)定义的 $J_m(\mathbf{W}^*, \mathbf{V}^*, \mathbf{U})$ 的第 1 项为常数. 因此, $\mathcal{E}(\mathbf{U})$ 取极值,当且仅当如下函数取极值:

$$\mathcal{E}_1(\mathbf{U}) = \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m d_{ji}^* + \epsilon_u \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m = \sum_{j=1}^c \sum_{i=1}^n u_{ji}^m (\epsilon_u + d_{ji}^*) \quad (\text{附 } 2)$$

$$\text{其中 } d_{ji}^* = \text{wdist}^2(\mathbf{x}_i, \mathbf{v}_j) = \sum_{h=1}^d (w_{jh}^*)^\alpha (x_{ih} - v_{jh}^*)^2.$$

令 $t_{ji} = u_{ji}^r$, 则 $\mathbf{T} = [t_{ji}]_{c \times n} \in M_{fcn}$, M_{fcn} 为满足式(1)的所有模糊矩阵的集合. 求 $\mathcal{E}_1(\mathbf{U})$ 在 M_U 上极值问题,转化为在 M_{fcn} 上求函数

$$F(\mathbf{T}) = \sum_{j=1}^c \sum_{i=1}^n (t_{ji})^{m/r} (d_{ji}^* + \epsilon_u) \quad (\text{附 } 3)$$

的极值问题. 因为 $m>r>0$, 所以 $m/r>1$. 令 $c_{ji} = d_{ji}^* + \epsilon_u > 0$, 根据引理 2, 易证

$$t_{ji} = \frac{1}{\sum_{l=1}^c \left(\frac{d_{ji}^* + \epsilon_u}{d_{li}^* + \epsilon_u} \right)^{\frac{1}{m/r-1}}}, j=1,2,\dots,c \quad (\text{附 } 4)$$

为式(附 3)的全局极小值点. 由 $u_{ji}^* = (t_{ji}^*)^{1/r}$ 可得式(10c)为式(附 2)的全局最小值点. 定理得证. 证毕.

用类似方法即可证定理 3.

Background

Text data is a typical example of high dimensional data in data mining applications. It is characterized by high dimensionality and feature overlapping among different clusters. However, the conventional clustering algorithms cannot cluster text data effectively because they consider each feature equally and cannot discover the inner cluster structures successfully.

Usually, clustering of high dimensional data requires special treatment according to the inner structure of the dataset. Up to now, many approaches have been proposed to tackle this problem. One typical approach is introducing the mechanism of feature weighting into the clustering process. However, this strategy assigns a weight to each feature for the entire dataset and this does not agree with the inner structure of text data. In text data, each cluster is identified by a subset of features and different clusters are identified by different subsets of features. Soft subspace clustering approaches aim at finding clusters from subspaces of data instead of the entire data space, which agree with the characteristics of text data well. However, most of the soft sub-

space clustering algorithms are established on the hard partition of the dataset and suffers from the problem of early convergence when applied on high dimensional text data.

To solve this problem, we propose a novel fuzzy clustering algorithm by integrating the feature weighting metric into the framework of soft subspace learning. Firstly, a feature weighting metric is presented based on the concept of vector norm. Then a novel learning criterion is proposed. Meanwhile, an entropy exponent r is introduced into the constraints so that the span of the fuzzy index m is extended. A global convergence theory is also established by applying Zangwill's convergence theorem.

This research work is supported by the National Natural Science Foundation of China under Grants 60903100, 60975027, 61170122, and by the Natural Science Foundation of Jiangsu Province under Grant BK2011417, Jiangsu 333 Expert Engineering Grant (BRA201142), the Fundamental Research Funds for the Central Universities (Grant JUSRP111A38).