鲁棒的模糊方向相似性聚类算法

朱 林,王士同,修 宇
(江南大学 信息工程学院,江苏 无锡 214122)

摘 要:鉴于文本数据具有方向性数据的特征,可利用方向数据的知识完成对文本数据聚类,提出了模糊方向相似性聚类算法 FDSC,继而从竞争学习角度,通过引入隶属度约束函数,并根据拉格朗日优化理论推导出鲁棒的模糊方向相似性聚类算法 RFDSC. 实验结果表明 RFDSC 算法能够快速有效地对文本数据集进行聚类.

关键词:聚类算法;方向相似性;鲁棒性;竞争学习
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-4785(2008)01-0043-08

A robust clustering algorithm with fuzzy directional similarity

ZHU Lin, WANG Shi-tong, XIU Yu
(School of Information Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: One of the important characteristics of text clustering in datasets is that each cluster center in the dataset has a direction that is different from that of all other cluster centers. This directional information should be incorporated in clustering analysis. In this paper, a new robust fuzzy directional similarity clustering algorithm (RFDSC) is proposed by introducing membership constraints. The new objective function was constructed. Finally, the robustness and convergence of the proposed algorithm were analyzed from the viewpoint of competitive learning. Experimental tests of text clustering in datasets using RFDSC demonstrate its effectiveness.

Keywords: clustering algorithm; directional similarity; robustness; competitive learning

聚类分析是无监督模式识别中的一种重要方法,已广泛应用于数据挖掘、图像处理、计算机视觉、生物信息和文本分析. 聚类算法就是将一组分布未知的数据进行分类,其目的是寻找隐藏在数据中的结构,并按照某种相似程度的度量,尽可能地使具有相同性质的数据归于同一类. 针对不同的应用和不同的理论已提出多种各具特色的聚类算法,如划分方法(K-means、Clarans、Frem)、层次方法(Chameleon、Brich)、基于网格方法(WaveCluster、Stng、Clique)、基于密度方法(Dbscan、Optics)等.

近年来大量研究表明,高维数据诸如文本数据及基因表达数据具有方向性数据的特征,可以利用方向数据的知识完成对这类数据的有效聚类. 文献[1-2]分别提出了 2 种不同的针对方向性数据的聚类算法 SPKmeans[1]和 movMF[2],但 2 种算法由于对初始化较敏感,聚类性能有待提高. 文献[3]提出了方向相似性聚类算法,其根据方向分布理论提出了数据的相似性度量,通过集成方向相似性聚类方法和凝聚层次聚类方法解决了聚类中初始化敏感等问题. 但由于该算法将每个样本点均作为不动点,通过迭代求其最优解,在处理高维大量数据如文本数据时,算法速度太慢,不能得到很好的应用.

文中所做工作的意义在于首先提出模糊方向相似性聚类算法(fuzzy directional similarity clustering, FDSC),在此基础之上,从竞争学习角度,通过对目标函数中引入隶属度约束函数,推导出鲁棒的模糊方向相似性聚类算法(robust fuzzy directional similarity clustering, RFDSC),使算法具有更好的收敛性和鲁棒性.

收稿日期:2007-05-14.
基金项目:国家"863"资助项目(2006AA10Z313);国家自然科学基金资助项目(60773206;60704047);国防应用基础研究基金资助项目(A1420461266);教育部科学研究重点基金资助项目(105087).
通讯作者:王士同. E-mail:wxwangst@yahoo.com.cn.

1 方向相似性聚类算法(DSCM)

方向相似性聚类算法[3](directional similarity-based clustering algorithm,DSCM)定义一个新的方向相似度量函数 $S(x_j, w_i)$ 来度量方向数据向量 x_j 和聚类中心向量 w_i 的相似性:$S(x_j, w_i) = e^{kx_j^T w_i}$,$i = 1,2,\cdots,c$ 代表类别下标,$j = 1,2,\cdots,n$ 代表样本下标,k 为尺度参数. $S(x_j, w_i)$ 值越大说明数据向量 x_j 与中心向量 w_i 的相似度越高. DSCM 算法定义聚类目标函数为

$$J_S = \sum_{i=1}^{c}\sum_{j=1}^{n} S(x_j, w_i) = \sum_{i=1}^{c}\sum_{j=1}^{n} e^{kx_j^T w_i}. \quad (1)$$

DSCM 算法通过拉格朗日优化理论构造拉格朗日优化目标函数,推导出中心向量 w_i 的迭代公式,对所有样本经过有限次迭代过程自组织的找到所有不动点,进一步采用凝聚层次聚类算法(agglomerative hierarchical clustering,AHC)分析不动点的层次数,并通过层次聚类图进一步确定最佳聚类数以及类与类之间的关系.

DSCM 算法可以避免初始化敏感的问题,能够对方向性数据进行自组织,并利用 AHC 算法侦测出最优的聚类数. 但对于高维大量数据特别是文本数据,由于样本数很大,对所有样本无法通过 DSCM 算法中的有限次迭代过程,自组织的找到所有不动点. 因此 DSCM 算法的应用受到了很大的局限性.

2 鲁棒的模糊方向相似性聚类算法

2.1 模糊方向相似性聚类算法(FDSC)

在方向相似性聚类算法的基础上,引入模糊隶属度关系,建立如下模糊化的目标函数:

$$J_{FDSC} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m e^{kx_j^T w_i}. \quad (2)$$

目标函数满足条件:$\sum_{i=1}^{c} u_{ij} = 1, 0 < \sum_{j=1}^{n} u_{ij} < n$, $u_{ij} \in [0,1]$,式中:m 为权重指数,$i = 1,2,\cdots,c$ 代表类别下标,$j = 1,2,\cdots,n$ 代表样本下标,k 为尺度参数.

定理1 在 $\sum_{i=1}^{c} u_{ij} = 1, 0 < \sum_{j=1}^{n} u_{ij} < n, u_{ij} \in [0,1], w_i^T w_i = 1$ 条件下,使 FDSC 算法的目标函数 $J_{FDSC} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m (e^{kx_j^T w_i})$ 取极值时:

1) 隶属度 u_{ij} 迭代公式为

$$u_{ij} = (e^{kx_j^T w_i})^{-1/(m-1)} / \sum_{i=1}^{c} (e^{kx_j^T w_i})^{-1/(m-1)}. \quad (3)$$

2) 中心向量 w_i 迭代公式为

$$w_i = \frac{\sum_{j=1}^{n} x_j^T u_{ij}^m e^{kx_j^T w_i}}{\left\| \sum_{j=1}^{n} x_j^T u_{ij}^m e^{kx_j^T w_i} \right\|}. \quad (4)$$

对于式(2)的极值,可以通过利用拉格朗日优化理论推导出模糊方向相似性聚类算法(FDSC)关于隶属度 u_{ij} 的迭代公式(3)和中心向量 w_i 的迭代公式(4)求解. 与传统的 FCM 聚类算法有所不同,按照 FDSC 算法定义,数据向量 x_j 与中心向量 w_i 的相似度越高其隶属度应越大,故式(3)中 $-1/(m-1) > 0$,即权重指数 $m < 1$,同时为了保证算法有较好的收敛性,权重指数 $0 < m < 1$. FDSC 聚类思想可以理解成通过中心向量和隶属度的迭代公式求解最优聚类中心 w_i,使得目标代价函数 J_{FDBC} 最大,即样本间相似度总和最大.

2.2 基于竞争学习理论的隶属度目标函数的构造

竞争学习的学习规则分为 WTA (winner-take-all) 与 WTM (winner-take-more) 2 种,分别称为"硬"竞争学习(hard competitive learning)与"软"竞争学习(soft competitive learning)[4-5]. WTA 规则存在一个重要问题:对于不同初始值的节点,学习过程中可能存在死节点(dead nodes)或利用不充分(under-utilization)的现象. 为此,研究人员提出了 WTM 规则,通过引入模糊隶属度等方法,削弱了学习对节点初始值的依赖,但同时也造成数据的获胜节点不确定,使得部分节点可能偏离其实际的类数据. 文献[6]提出了惩罚对手的竞争学习(RPCL),对于数据集中的每个样本,不仅其获胜节点以一定的学习速率向该数据集"靠近",而且其竞争对手以更小的学习速率(惩罚速率)被"推离"该数据集,从而减小了对学习过程的干扰.

基于惩罚对手的竞争学习(RPCL)的思想,在模糊方向相似性聚类算法 FDSC 的基础之上,文中进一步提出鲁棒性的模糊方向相似性聚类算法 RFDSC. 对于一个单独的样本点 x_j,通过引入隶属度约束函数:$f(u_1, u_2, \cdots, u_c) = \sum_{i=1}^{c} u_i (u_i^{m-1} - 1)$,构造新的目标函数:

$$J_{RFDSC} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m (e^{kx_j^T w_i}) - \sum_{j=1}^{n} a_j \sum_{i=1}^{c} u_{ij} (u_{ij}^{m-1} - 1). \quad (5)$$

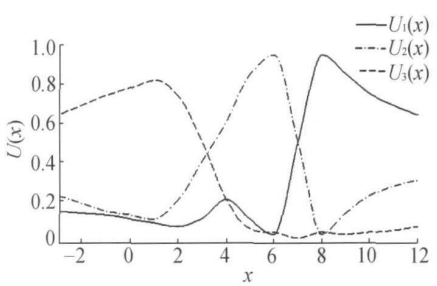
下文将对式(5)中参数 a_j $(1 \le j \le n)$ 的选择做说明. 权重指数 $m \in (0,1)$,目标函数满足 $\sum_{i=1}^{c} u_{ij} = 1, 0 <$

$\sum_{j}^{n} u_{ij} < n, u_{ij} \in [0,1]$.

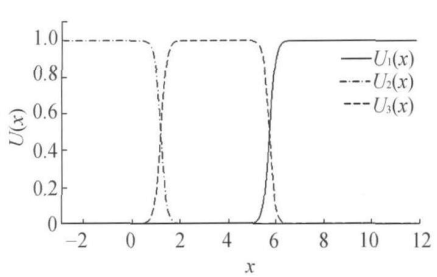
对于单独的样本点 x_j,隶属度约束函数:$f(u_1, u_2, \cdots, u_c) = \sum_{i}^{c} u_i(u_i^{m-1} - 1)$ 有如下结论:

定理 2　对隶属度约束函数 $f(u_1, u_2, \cdots, u_c)$,在 $0 \le u_i \le 1$, $\sum_{i}^{c} u_i = 1$,权重指数 $m \in (0,1)$ 的限制条件下,当 $u_i(i = 1,2,\cdots,c)$ 均为 $1/c$ 时,$f(u_1, u_2, \cdots, u_c)$ 取得最大值 $c^{1-m} - 1$;当 $u_i(i = 1,2,\cdots,c)$ 中一个 u_k 为 1,其他 $u_i(i \ne k)$ 为 0 时,$f(u_1, u_2, \cdots, u_c)$ 取最小值 0,即 $0 \le f(u_1, u_2, \cdots, u_c) \le c^{1-m} - 1$.

通过加入隶属度约束函数,使得当函数 J_{RFDSC} 极大,即 $f(u_1, u_2, \cdots, u_c)$ 极小时,出现某个 u_k 趋向于 1,其他的 $u_i(i \ne k)$ 趋向于 0 的情况. 图 1 显示加入约束函数后,u_i 的明晰含义被显著提升. 改进算法兼顾了硬聚类和模糊聚类的优点,因此对数据集将有更好的鲁棒性.



(a) FDSC 算法的隶属度函数



(b) RFDSC 算法的隶属度函数

图 1　不同的隶属度函数曲线

Fig. 1　Different kinds of membership curves

2.3　RFDSC 算法

对于 RFDSC 算法,首先给出如下定理.

定理 3　在 $\sum_{i}^{c} u_{ij} = 1, 0 < \sum_{j}^{n} u_{ij} < n, u_{ij} < n, u_{ij} \in [0,1], w_i^{\mathrm{T}} w_i = 1$ 以及权重指数 $m \in (0,1)$ 条件下,使 RFDSC 算法的目标函数 $J_{RFDSC} = \sum_{i}^{c}\sum_{j}^{n} u_{ij}^m (e^{k x_j^{\mathrm{T}} w_i}) - \sum_{j}^{n} a_j \sum_{i}^{c} u_{ij}(u_{ij}^{m-1} - 1)$ 取极值时:

1) 隶属度 u_{ij} 迭代公式:

$$u_{ij} = (e^{k x_j^{\mathrm{T}} w_i} - a_j)^{-1/(m-1)} / \sum_{i}^{c} (e^{k x_j^{\mathrm{T}} w_i} - a_j)^{-1/(m-1)}. \tag{6}$$

由于 $e^{k x_j^{\mathrm{T}} w_i} - a_j < 0$,出现负值,隶属度 u_{ij} 也会为负. 故自适应提升 a_j,使得 $e^{k x_j^{\mathrm{T}} w_i} - a_j$ 为正.

定义 $a_j = \min\{e^{k x_j^{\mathrm{T}} w_j} \mid i \in \{1, \cdots, c\}\} - \eta, (\eta > 0)$,代入式(6)从而得到:

$$u_{ij} = (e^{k x_j^{\mathrm{T}} w_i} - \min e^{k x_j^{\mathrm{T}} w^*} + \eta)^{-1/(m-1)} / \sum_{i}^{c} (e^{k x_j^{\mathrm{T}} w_i} - \min e^{k x_j^{\mathrm{T}} w^*} + \eta)^{-1/(m-1)}. \tag{7}$$

2) 中心向量 w_i 迭代公式:

$$w_i = \frac{\sum_{j}^{n} x_j^{\mathrm{T}} u_{ij}^m e^{k x_j^{\mathrm{T}} w_i}}{\left\| \sum_{j}^{n} x_j^{\mathrm{T}} u_{ij}^m e^{k x_j^{\mathrm{T}} w_i} \right\|}. \tag{8}$$

下面对定理 3 中出现的权重指数 m、尺度参数 k 和模糊程度常数 η 的选择做详细说明.

对于权重指数 m,上文已经指出由于与传统的 FCM 聚类算法的不同,按照 FDSC 和 RFDSC 算法定义,数据向量 x_j 与中心向量 w_i 的相似度越高其隶属度应越大,故权重指数 $m < 1$,同时为了保证算法有较好的收敛性,权重指数 $0 < m < 1$. 实验结果表明,当权重指数 m 趋于 1 时,具有较好的鲁棒性和抗噪声能力,故文中权重指数 m 取 0.9.

尺度参数 k 的选择与数据集有关,通过尺度参数 k 的选择可以扩大($k > 1$)或减小($k < 1$)方向数据向量 x_j 和聚类中心向量 w_i 的相似性程度. 实验结果表明,当权重指数 k 取值在 1～5 时,具有较好的收敛性和聚类效果,故文中尺度参数 k 取 3.

模糊程度常数 η 的选择同样与数据集有关,影响模糊划分的模糊程度,值越大模糊程度越高;值越小模糊程度越低,越趋进于硬聚类. 当趋于无穷时,样本点对每一类的隶属度均为 $1/c$. 通过对模糊程度常数的控制,使得样本点对距离最近的类中心施加最大的吸引力,这个力越大,类中心收敛速度也越快;样本点也可以对相似性较小的类中心有微弱的吸引力,避免聚类过程出现死节点,保证了算法具有很好的鲁棒性[7]. 由于 RFDSC 算法采用了竞争学习规则,通过加入隶属度约束函数显著提升隶属度 u_{ij} 明晰含义,兼顾了硬聚类和模糊聚类的优点,因此在高维大量数据情况下,RFDSC 算法比 FDSC 算法的鲁棒性和收敛速度都有了很大提高. 由于 RFDSC 算法对赢者的吸引力的增强作用随 η 减小

而增大,对赢者对手的抑制作用随η增大而减小,所以需要选取一个合适的η值.实验结果表明,η的取值取所有方向数据向量 x 到聚类中心向量 w_i 的最大相似性的0.01～0.2左右较为合适,故文中模糊程度常数η取0.15.

基于定理3,得到具有更好鲁棒性的模糊方向相似性聚类算法 RFDSC.下面给出 RFDSC 算法的完整描述:

1)初始化聚类数目 $c(2 \leq c \leq N)$,阈值 $\varepsilon(l=1$ 为迭代次数)、权重指数 m、尺度参数 k、模糊程度常数 η和最大迭代次数 T,并随机产生并归一化中心向量 w_i^1;

2)根据式(7)计算样本的隶属度值 u_{ij}^{l+1};

3)根据式(8)计算归一化中心向量 w_i^{l+1};

4)如果 $\| u_{ij}^{l+1} - u_{ij}^{(l)} \| < \varepsilon$或者 $l > T$ 则停止,算法结束,否则 $l = l+1$,跳转2);

5)当算法收敛,得到各类的中心向量 w_i 和样本集模糊隶属度 u_{ij}.

3 实验结果及分析

在本节中首先介绍聚类性能的评价标准,然后分别对文本数据集进行说明,最后通过使用 SPK-means[1]、soft-movMF[2]、FDSC 和 RFDSC 4种相似性聚类算法对文本数据集进行测试,以检测各种聚类算法的性能.由于 DSCM 算法在处理高维大量数据时,算法速度太慢,对文本数据集的有所有样本无法通过有限次迭代过程,自组织的找到所有不动点,因此没有对 DSCM 算法进行比较.

3.1 算法性能评价准则

对于聚类结果有效性的度量,可以分为面向分类的和面向相似性的2种度量方式.第1种度量方式,如互信息(normalized mutual information, NMI)、纯度和 F 度量,这些度量评估聚类结果包含单个类的对象的程度.第2种度量方式,如平均准确率 AA(averaged accuracy or rand index)、Jaccard 度量,这些方法度量在多大程度上,同一个类的2个对象在同一簇中,或相反[8].而衡量传统信息检索系统的性能参数:召回率(Recall)和精度(Precision)也是衡量分类算法性能的常用指标.然而聚类的过程中并不存在自动分类类别与手工分类类别确定的一一对应关系,因而无法像分类一样直接以精度和召回率作为评价标准[9].为此在文中采用互信息 NMI[10]和平均准确率 AA[11-12]分别作为算法的面向分类和面向相似性的2种评价标准.假设 X 代表已知的文本类标随机变量,Y 代表聚类结果的类标随机变量,互信息 NMI 公式定义为 $\mathrm{NMI}(X,Y) = \frac{I(X;Y)}{\sqrt{H(X) \cdot H(Y)}}$,$I(X;Y)$ 为变量 X,Y 的互信息量,$H(X)$,$H(Y)$ 为变量 X 和 Y 的熵.平均准确率 AA 的公式定义为:$\mathrm{AA}(X,Y) = \frac{a+b}{n \times (n-1)/2}$,其中 a 表示任意2个样本在 X、Y 中同属于一类的个数,b 表示任意2个样本都不属于同一类的个数,n 表示数据集的样本个数.互信息 NMI 和平均准确率 AA 的范围均为$[0,1]$,值越高表示聚类结果越准确,值越小2种划分差距越大,在 X、Y 完全一致的情况下,NMI 和 AA 值为1.

3.2 实验数据集说明

实验采用 20-Newsgroups[13]的数据集及部分来自 CLUTO[14]文本聚类工具箱的10种数据集.数据集含有的样本数从204个到19 949不等,数据维数最小的为5 832维,最大的为43 586维,实际聚类数最小的为3个,最大的为20个,从以上特征看出这些数据集很好的反映了不同文本数据集所具有的特征.其中 NG20 数据平均地选自来自20个不同新闻组,经过的 Bow toolkit[15]对 20-Newsgroups 文本进行了预处理后含有19 949个向量文本数据.NG17-19 是 NG20 数据的一个子集,以往的聚类算法对该数据集的聚类结果表明,因为类与类之间有重叠导致对该数据集的聚类难度较高.其他数据均来自 CLUTO 工具箱[14],这些数据集均已经过预处理为向量文本数据.数据集的详细说明见表1.

表1 文本数据集的简要说明

Table 1 Summary of text datasets

Data	Source	n_d	n_w	K	balance
NG20	20-Newsgroups	19 949	43 586	20	0.991
NG17-19	3 overlapping/ subgroups from NG20	2 998	15 810	3	0.998
ohscal	OHSUMED-233445	11 162	11 465	10	0.437
klb	WebACE	2 340	21 839	6	0.043
hitech	San Jose Mercury(TREC)	2 301	10 080	6	0.192
la12	LA Times(TREC)	6 279	31 472	6	0.282
tr11	TREC	414	6 429	9	0.046
tr23	TREC	204	5 832	6	0.066
tr41	TREC	878	7 454	10	0.037
tr45	TREC	690	8 261	10	0.088

注:n_d 代表文本数(样本数),n_w 代表词项(维数),k 代表文本实际类别数,balance 是数据的平衡系数即包含最少文本数的类与包含最多文本数的类中的文本数之比,这个值反映了数据集内类与类之间的平衡性.

在使用方向性聚类算法分析各数据集前,以上数据都经过 L_2 归一化处理.

3.3　实验结果及分析

为了测试算法性能,对以上数据集分别采用 SPKmeans、soft-movMF、FDSC 和 RFDSC 4 种相似性聚类算法的结果进行比较. 为保证实验的公平性,所有算法均采用随机初始化的策略选取初始聚类中心,并对每种算法进行 20 次聚类实验后取平均值作为最终实验结果.

首先,对 FDSC 算法和 RFDSC 算法采用不同大小的尺度参数 k、模糊程度常数 η 和权重指数 m 来测试其对聚类结果的影响.

表 2、表 3 分别给出了 FDSC 算法采用不同大小的尺度参数 k 针对 NG17-19 和 tr45 数据集进行聚类实验后得到的互信息 NMI 的均值和标准差及平均准确率 AA 的均值. 实验中权重指数 m 取 0.9.

表 2　尺度参数 k 变化时 FDSC 算法针对 NG17-19 数据集的实验结果

Table 2　The results of FDSC on NG17-19 datasets with the change of k

k	1	2	3	4
互信息	0.326 ±0.05	0.410 ±0.05	0.431 ±0.04	0.394 ±0.05
平均准确率	0.596	0.677	0.682	0.642

表 3　尺度参数 k 变化时 FDSC 算法针对 tr45 数据集的实验结果

Table 3　The results of FDSC on tr45 datasets with the change of k

k	1	2	3	4
互信息	0.564 ±0.03	0.630 ±0.03	0.649 ±0.03	0.594 ±0.04
平均准确率	0.862	0.876	0.893	0.845

表 4、表 5 分别给出了 RFDSC 算法采用不同大小的模糊程度常数 η 针对 NG17-19 和 tr45 数据集进行聚类实验后得到的互信息 NMI 的均值和标准差及平均准确率 AA 的均值. 实验中权重指数 m 取 0.9,尺度参数 k 取 3.

表 4　模糊程度常数 η 变化时 RFDSC 算法针对 NG17-19 数据集的实验结果

Table 4　The results of RFDSC on NG17-19 datasets with the change of η

η	0.05	0.1	0.15	0.2
互信息	0.378 ±0.08	0.397 ±0.08	0.418 ±0.05	0.399 ±0.09
平均准确率	0.645	0.647	0.674	0.673

表 5　模糊程度常数 η 变化时 RFDSC 算法针对 tr45 数据集的实验结果

Table 5　The results of RFDSC on tr45 datasets with the change of η

η	0.05	0.1	0.15	0.2
互信息	0.657 ±0.05	0.661 ±0.05	0.673 ±0.05	0.640 ±0.04
平均准确率	0.894	0.895	0.899	0.884

表 6、表 7 分别给出了 RFDSC 算法采用不同大小的权重指数 m 针对 NG17-19 和 tr45 数据集进行聚类实验后得到的互信息 NMI 的均值和标准差及平均准确率 AA 的均值. 实验中模糊程度常数 η 取 0.15,尺度参数 k 取 3.

表 6　权重指数 m 变化时 RFDSC 算法针对 NG17-19 数据集的实验结果

Table 6　The results of RFDSC on NG17-19 datasets with the change of m

m	0.85	0.9	0.95
互信息	0.397 ±0.06	0.418 ±0.05	0.375 ±0.10
平均准确率	0.675	0.674	0.640

表 7　权重指数 m 变化时 RFDSC 算法针对 tr45 数据集的实验结果

Table 7　The results of RFDSC on tr45 datasets with the change of m

m	0.85	0.9	0.95
互信息	0.649 ±0.04	0.674 ±0.03	0.649 ±0.04
平均准确率	0.885	0.902	0.887

从上述实验结果可以看出,当尺度参数 k 取 3、模糊程度常数 η 取 0.15、权重指数 m 取 0.9 时,FDSC 算法和 RFDSC 算法取得比较好的聚类结果.

下面给出 SPKmeans、soft-movMF、FDSC 和 RFDSC 4 种相似性聚类算法对上述 10 种文本数据集的聚类结果.

表 8、表 9 给出了 4 种算法进行聚类实验后得到的互信息 NMI 的均值和标准差.

表8 针对 NG20、NG17-19、ohscal、k1b、hitech 数据集的互信息值

Table 8 NMI results on NG20, NG17-19, ohscal, k1b, hitech datasets

	NG20	NG17-19	ohscal	k1b	hitech
SPKmeans	0.548±0.03	0.303±0.09	0.437±0.02	0.579±0.05	0.282±0.02
soft-movMF	0.570±0.02	0.390±0.10	0.442±0.02	0.607±0.04	0.292±0.02
FDSC	0.443±0.02	0.431±0.08	0.433±0.01	0.602±0.03	0.288±0.01
RFDSC	0.564±0.02	0.418±0.05	0.447±0.01	0.624±0.03	0.298±0.01

表9 针对 la12、tr11、tr23、tr41、tr45 数据集的互信息值

Table 9 NMI results on la12, tr11, tr23, tr41, tr45 datasets

	la12	tr11	tr23	tr41	tr45
SPKmeans	0.523±0.03	0.535±0.05	0.269±0.05	0.585±0.05	0.657±0.07
soft-movMF	0.535±0.04	0.602±0.04	0.366±0.04	0.627±0.04	0.660±0.04
FDSC	0.471±0.02	0.641±0.02	0.386±0.02	0.645±0.03	0.636±0.03
RFDSC	0.522±0.03	0.646±0.02	0.349±0.02	0.654±0.03	0.674±0.03

表10、表11给出了4种算法进行聚类实验后得到的平均准确率 AA 的均值.

表10 针对 NG20、NG17-19、ohscal、k1b、hitech 数据集的平均准确率

Table 10 AA results on NG20, NG17-19, ohscal, k1b, hitech datasets

	NG20	NG17-19	ohscal	k1b	hitech
SPKmeans	0.935	0.663	0.863	0.730	0.726
soft-movMF	0.938	0.672	0.865	0.747	0.752
FDSC	0.889	0.677	0.865	0.746	0.732
RFDSC	0.924	0.674	0.867	0.779	0.757

表11 针对 la12、tr11、tr23、tr41、tr45 数据集的平均准确率

Table 11 AA results on la12, tr11, tr23, tr41, tr45 datasets

	la12	tr11	tr23	tr41	tr45
SPKmeans	0.830	0.840	0.680	0.853	0.886
soft-movMF	0.835	0.856	0.718	0.860	0.893
FDSC	0.805	0.859	0.732	0.861	0.877
RFDSC	0.821	0.864	0.702	0.865	0.902

从上述实验结果可以看出,RFDSC 算法的聚类效果在大多数情况下要好于其他算法,同时也发现了 FDSC 和 RFDSC 算法的互信息的标准差在大多数情况下要小于其他算法,说明 FDSC 和 RFDSC 算法的稳定性较好,能够更好地避免局部极值点. 当然,在部分数据集的试验中,RFDSC 算法与其他算法相比,还不是最优的,在下一步的工作中,将对 RFDSC 算法中权重指数 m、尺度参数 k 和模糊程度常数 η 的选择问题做更深入的研究,以提高算法的性能.

4 结束语

文中在方向相似性聚类算法 DSCM 基础之上,首先提出模糊方向相似性聚类算法 FDSC,继而从竞争学习角度,通过引入隶属度约束函数,并根据拉格朗日优化理论推导出鲁棒的模糊方向相似性聚类算法 RFDSC,最后将 RFDSC 算法很好的应用于文本方向性数据聚类中,从而解决了 DSCM 对高维大量数据的扩展问题.

附 录

定理1 证明

证明 使 FDSC 算法的目标函数即式(2)中的 J_{FDSC} 达到最大,即求在 $\sum_{i=1}^{c} u_{ij} = 1, 0 < \sum_{j=1}^{n} u_{ij} < n$, $u_{ij} \in [0,1]$ 以及 $w_i^{\mathrm{T}} w_i = 1$ 条件下的极值,为此引入 Lagrange 乘子 λ_j, b_i 并定义 Lagrange 目标函数 $L(w, \lambda, a, b)$:

$$L(w,\lambda,a,b) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}(e^{kx_j^{\mathrm{T}}w_i}) + \sum_{j=1}^{n}\lambda_j\Big(\sum_{i=1}^{c} u_{ij} - 1\Big) + \sum_{i=1}^{c} b_i(w_i^{\mathrm{T}}w_i - 1). \quad (9)$$

对 $L(w,\lambda,b)$ 关于 u_{ij} 求偏导,令其值为零得

$$u_{ij} = \Big[\frac{-\lambda_j}{m(e^{kx_j^{\mathrm{T}}w_i})}\Big]^{1/(m-1)}. \quad (10)$$

将式(10)代入 $\sum_{i=1}^{c} u_{ij} = 1$,消去 $(-\lambda_j)/m$ 得

$$u_{ij} = (e^{kx_j^{\mathrm{T}}w_i})^{-1/(m-1)} / \sum_{i=1}^{c}(e^{kx_j^{\mathrm{T}}w_i})^{-1/(m-1)}. \quad (11)$$

对 $L(w,\lambda,b)$ 关于 w_i 求偏导数,并令其值为零可得

$$\frac{\partial L}{\partial w_i} = \sum_{j=1}^{n} kx_j^{\mathrm{T}} u_{ij}^{m} e^{kx_j^{\mathrm{T}}w_i} + 2b_i w_i = 0. \quad (12)$$

则有

$$w_i = \sum_{j=1}^{n} kx_j^{\mathrm{T}} u_{ij}^{m} e^{kx_j^{\mathrm{T}}w_i} / (-2b_i). \quad (13)$$

由于 $w_i^{\mathrm{T}}w_i = 1$,由公式(13)可以进一步推出:

$$w_i = \frac{\sum_{j=1}^{n} x_j^{\mathrm{T}} u_{ij}^{m} e^{kx_j^{\mathrm{T}}w_i}}{\Big\| \sum_{j=1}^{n} x_j^{\mathrm{T}} u_{ij}^{m} e^{kx_j^{\mathrm{T}}w_i} \Big\|}.$$

证毕.

定理 2 证明

证明: 1) 因为 $0 \leq u_i \leq 1, 0 < m < 1$, 所以 $\sum_{i=1}^{c} u_i(u_i^{m-1} - 1)$ 中每一项 $u_i(u_i^{m-1} - 1)$ 均大于等于 0, 因此 $f(u_1, u_2, \cdots, u_c) \geq 0$. $f(u_1, u_2, \cdots, u_c) = 0$ 当且仅当 $\sum_{i=1}^{c} u_i(u_i^{m-1} - 1)$ 中的每一项均为 0, 即 $u_i(u_i^{m-1} - 1) = 0 (0 \leq i \leq c)$, 那么必有 $u_i = 0$ 或 $u_i = 1$. 由于有 $\sum_{i=1}^{c} u_i = 1$ 的限制, 则 $f(u_1, u_2, \cdots, u_c) = 0$ 只能有某一个 $u_i = 1$, 其他 $u_k = 0 (k \neq i)$ 的情况.

2) 考虑 $f(u_1, u_2, \cdots, u_c) = \sum_{i=1}^{c} u_i(u_i^{m-1} - 1)$ $(0 \leq u_i \leq 1)$ 在约束条件 $\sum_{i=1}^{c} u_i = 1$ 下的极值. 应用拉格朗日乘数法构造函数: $G = \sum_{i=1}^{c} u_i(u_i^{m-1} - 1) - \lambda(\sum_{i=1}^{c} u_i - 1)$. 令 $\frac{\partial G}{\partial u_1}, \frac{\partial G}{\partial u_2}, \cdots, \frac{\partial G}{\partial u_c}$ 均为 0, 得 $u_1 = u_2 = \cdots u_c = \frac{1}{c}, \lambda = mc^{1-m} - 1$.

又因为 $f(u_1, u_2, \cdots, u_c)$ 在 $[0,1]^c$ 上是连续的, 故 $f(u_1, u_2, \cdots, u_c)$ 一定有最值, 而最值一定在驻点或边界点处取得.

下面分别求 $f(u_1, u_2, \cdots, u_c)$ 在驻点或边界点处的函数值.

1) 当 $u_1 = u_2 = \cdots u_c = \frac{1}{c}$ 时, $f(u_1, u_2, \cdots, u_c) = c^{1-m} - 1$.

2) 对于 $[0,1]^c$ 的边界点 $(u_1, u_2, \cdots, u_c)$, 设 $u_1, u_2, \cdots, u_c$ 中有 $p (1 \leq p \leq c)$ 个 u_i 为 0, $q (0 \leq q \leq 1)$ 个 u_i 为 1.

①若 $p + q = c$, 则 $\sum_{i=1}^{c} u_i(u_i^{m-1} - 1) = 0$.

②若 $p + q < c$, 将 $u_i = 0, 1$ 代入 $f(u_1, u_2, \cdots, u_c)$ 中, 不失一般性, 不防假设前 $p + q$ 个 u_i 为 0 或 1, 则 $f(u_1, u_2, \cdots, u_c)$ 转化为 $c - p - q$ 维的形如 $\sum_{i=p+q+1}^{c} u_i(u_i^{m-1} - 1)$ 的函数. 由反证法很容易证明 $\sum_{i=p+q+1}^{c} u_i(u_i^{m-1} - 1) < c^{1-m} - 1$.

综上所述, $f(u_1, u_2, \cdots, u_c)$ 在 $u_{1j} = u_{2j} = \cdots u_{cj} = \frac{1}{c}$ 时取最大值 $c^{1-m} - 1$. 证毕.

定理 3 证明

证明　使 RFDSC 算法的目标函数即式 (5) 中 J_{RFDSC} 达到最小, 即求在 $\sum_{i=1}^{c} u_{ij} = 1$, $0 < \sum_{j=1}^{n} u_{ij} < n, u_{ij} \in [0,1]$ 以及 $w_i^T w_i = 1$ 条件下的峰值, 为此引入 Lagrange 乘子 λ, b. 并定义 Lagrange 目标函数 $L(w, \lambda, a, b)$:

$$L(w,\lambda,a,b) = \sum_{i}^{c}\sum_{j}^{n} u_{ij}^{m}(e^{kx_j^T w_i}) - \sum_{j}^{n} a_j \sum_{i}^{c} u_{ij}(u_{ij}^{m-1} - 1) + \sum_{j}^{n} \lambda_j (\sum_{i}^{c} u_{ij} - 1) + \sum_{i}^{c} b_i (w_i^T w_i - 1). \quad (14)$$

对 $L(w, \lambda, a, b)$ 关于 u_{ij} 求偏导, 令其值为零得

$$u_{ij} = [\frac{-a_j - \lambda_j}{m(e^{kx_j^T w_i} - a_j)}]^{1/(m-1)}. \quad (15)$$

将式 (15) 代入 $\sum_{i=1}^{c} u_{ij} = 1$, 消去 $(-a_j - \lambda_j)/m$ 得

$$u_{ij} = (e^{kx_j^T w_i} - a_j)^{-1/(m-1)} / \sum_{i}^{c} (e^{kx_j^T w_i} - a_j)^{-1/(m-1)}. \quad (16)$$

当 $e^{kx_j^T w_i} - a_j < 0$, 出现负值, 隶属度 u_{ij} 也会为负. 故自适应提升 a_j, 使得 $e^{kx_j^T w_i} - a_j$ 为正.

定义 $a_j = \min\{e^{kx_j^T w_i} \mid i \in \{1, \cdots, c\}\} - \eta, (\eta > 0)$, 代入式 (10) 从而得到

$$u_{ij} = (e^{kx_j^T w_i} - \min e^{kx_j^T w^*} + \eta)^{-1/(m-1)} / \sum_{i}^{c} (e^{kx_j^T w_i} - \min e^{kx_j^T w^*} + \eta)^{-1/(m-1)}. \quad (17)$$

对 $L(w, \lambda, a, b)$ 关于 w_i 求偏导数, 并令其值为零得

$$\frac{\partial L}{\partial w_i} = \sum_{j}^{n} k x_j^T u_{ij}^m e^{kx_j^T w_i} + 2 b_i w_i = 0. \quad (18)$$

则有:

$$w_i = \sum_{j}^{n} k x_j^T u_{ij}^m e^{kx_j^T w_i} / (-2 b_i). \quad (19)$$

由于 $w_i^T w_i = 1$, 由式 (19) 可以进一步推出:

$$w_i = \frac{\sum_{j}^{n} x_j^T u_{ij}^m e^{kx_j^T w_i}}{\| \sum_{j}^{n} x_j^T u_{ij}^m e^{kx_j^T w_i} \|}.$$

证毕.

参考文献:

[1] DHILLON I S, MODHA D S. Concept decompositions for large sparse text data using clustering [J]. Machine Learning, 2001, 42(1): 143-175.

[2] BANERJEE A, DHILLON I S, GHOST J, et al. Generative model based clustering of directional data [C]// Conference on Knowledge Discovery in Data. Washington, DC, 2003.

[3] LI H X, WANG S T, XIU Y. Applying robust directional similarity based clustering approach RDSC to clas-

sification of gene expression data [J]. J Bioinformatics and Computational Biology, 2006, 4(3):745-768.

[4]ZHANG Y J, LIU Z Q. Self-splitting competitive learning: a new on-line clustering paradigm [J]. IEEE Trans on Neural Network, 2002, 13(2):369-380.

[5]WU S H, LIEW W C, YAN H, et al. Cluster analysis of gene expression data based on self-splitting and merging competitive learning [J]. IEEE Trans on Information Technology in Biomedicine, 2004, 8(1):5-15.

[6]XU L, KRZYAK A, OJA E. Rival penalized competitive learning for clustering analysis, RBF net and curve detection [J]. IEEE Trans on Neural Network, 1993, 4(4):636-649.

[7]魏立梅, 谢维信. 对手抑制式模糊C均值算法[J]. 电子学报, 2000, 28(7):63-66.
WEI Limei, XIE Weixin. Rival checked fuzzy C-means algorithm [J]. Acta Electronica Sinica, 2000, 28(7): 63-66.

[8]TAN P N. MICHAEL S, KUMAR V. Introduction to data mining [M]. Boston: Addison Wesley, 2005.

[9]姜 宁, 宫秀军, 史忠植. 高维特征空间中文本聚类研究[J]. 计算机工程与应用, 2002, 38(10):63-67.
JIANG Ning, GONG Xiujun, SHI Zhongzhi. Text clustering in high-dimension feature space[J]. Computer Engineering and Applications, 2002, 38(10): 63-67.

[10]ALEXANDER S, JOYDEEP G. Cluster ensembles-a knowledge reuse framework for combining partitions [J]. Journal of Machine Learning Research, 2002, 3(3):583-617.

[11]MAKOTO I, TAKENOBU T. Hierarchical Bayesian clustering for automatic text classification[R]. Department of Computer Science, Tokyo Institute of Technology, 1995.

[12]RAND W. Objective criteria for the evaluation of clustering methods[J]. Journal of the American Statistical Association, 1971, 66(336):846-850.

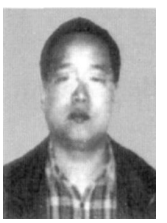
[13]Available on http://kdd.ics.uci.edu./databases/20newsgroups/20newsgroups.html.

[14]Available on ftp://www.cs.umn.edu/~karypis/CLUTO/flies/datasets.tar.gz.

[15]Mow: A toolkit for statistical language modeling, text retrieval, classification and clustering Available on http://www.cs.cmu.edu/mccallum/bow.

作者简介:

朱 林,男,1983年生,硕士研究生,主要研究方向为图像处理、模式识别.



王士同,男,1964年生,教授,博士生导师,中国计算机学会高级会员,主要研究方向为人工智能、模式识别、数据挖掘、神经网络及生物信息学.

修 宇,男,1976年生,硕士研究生,主要研究方向为模式识别、数据挖掘.