doi: 10.12068/j.issn.1005-3026.2017.08.001

一种基于黑洞算法的模糊C均值文本聚类方法

柳玉辉[1], 王伟超[1], 孟 磊[2]
(1. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110169; 2. 东网科技有限公司, 辽宁 沈阳 110169)

摘 要: FCM算法应用于文本聚类时,由于初始聚类中心点选择的随机性,以及容易陷入局部最优的问题,导致文本聚类效果较差. 为了提高FCM算法的聚类精度,提出了采用黑洞算法寻找FCM最优初始聚类中心的方法. 黑洞算法是一种启发式优化方法,在FCM初始聚类中心寻优的过程中,始终保持黑洞为全局最优解,最终发现FCM的最优初始聚类中心. 实验结果表明,基于黑洞算法的FCM文本聚类方法可以解决FCM算法对初始中心点敏感和容易陷入局部最优的问题,聚类精度明显提高.

关 键 词: 模糊C均值;黑洞算法;文本聚类;参数搜索;初始聚类中心
中图分类号: TP 391 **文献标志码**: A **文章编号**: 1005-3026(2017)08-1065-05

Document Clustering of Fuzzy C-Means Based on Black Hole Algorithm

LIU Yu-hui[1], *WANG Wei-chao*[1], *MENG Lei*[2]
(1. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China; 2. Neunn Technology Co., Ltd, Shenyang 110169, China. Corresponding author: LIU Yu-hui, E-mail: liuyh@neusoft.com)

Abstract: When fuzzy *c*-means (FCM) algorithm is applied to document clustering, the result is not ideal because of its initial cluster center points' random selection and falling into the local optimal solution easily. Aiming at improving the FCM's clustering accuracy, a method is proposed which uses the black hole algorithm (BHA), a heuristic algorithm, to find FCM's optimal initial clustering centers. During searching for the FCM's best initial clustering centers, the black hole is considered as the optimal option, and the FCM's best initial clustering centers can be found. The experiment's results show that the document clustering of FCM based on black hole algorithm can solve the problem that FCM is sensitive to initial centers and easy to fall into the local optimal solution, and finally, the clustering accuracy is improved significantly.

Key words: fuzzy *c*-means; black hole algorithm; document clustering; parameter searching; initial clustering center

文本聚类是用于自动话题提取、信息检索、信息推荐等领域的基本方法. 聚类的最终目标是将数据集的内部结构信息划分成为簇结构,使得同一个簇内部的数据集有较高的相似度,同时不同簇之间的数据集有较低的相似度.

模糊C均值(FCM)算法是划分聚类算法之一,在处理大数据集时具有较好的效果,但是需要提前确定簇数和随机初始化聚类中心,使聚类结果容易陷入局部最优解.

FCM算法初始值的随机选取容易使该算法陷入局部最优,针对这一问题,国内外学者给出很多解决方案. 文献[1]给出了FCM算法2000年至2014年的研究状况. 除此之外,Mekhmoukh等[2]提出采用粒子群算法寻找FCM最优初始聚类中心的方法,但是粒子群算法容易过早收敛,以至于陷入局部最优解[3]. Wang等[4]采用自适应粒子群算法解决FCM算法对中心点选择敏感的问题,一定程度上改进了粒子群算法,但算法性能

收稿日期: 2016-03-14
基金项目: 国家高技术研究发展计划项目(2015AA016005).
作者简介: 柳玉辉(1966-),男,山东烟台人,东北大学副教授.

仍有提高的空间. Ding 等[5]提出一种将遗传算法与基于核函数的 FCM 融合的方法进行聚类,但是遗传算法在种群进化的过程中,多样性容易丢失,存在收敛过早的现象[6]. Naik 等[7]提出用 TLBO 算法来优化 FCM 算法的聚类中心,但该算法解决高维复杂问题时,收敛速度慢且容易陷入局部最优[8].

黑洞算法[9]是根据自然界的黑洞现象生成的一种启发式优化方法,现阶段已被用于配电网潮流计算[10]、图像处理[11]、参数寻优[12]等领域,具有寻优精度高、容易达到全局最优等优点. Hatamlou 等[9]将黑洞算法与 k-Means, PSO, GSA, BB – BC 聚类算法做对比,证明了黑洞算法应用于数值型数据中具有良好的聚类效果. 本文将黑洞(BH)算法与 FCM 算法相结合,并引入到文本聚类领域中,提出了一种 BH – FCM 聚类方法.

1 基于 LDA 的文本特征提取

1.1 LDA 模型

LDA 模型[13]是一种常用的文本生成模型,它的图模型如图 1 所示.

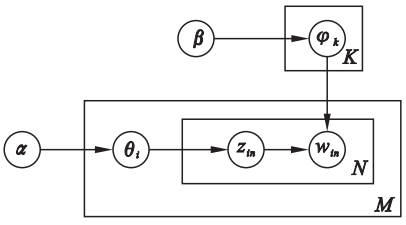

图 1 LDA 图模型

Fig. 1 Graphical model of LDA

图 1 中,w_{in}表示第 i 个文档的第 n 个词语;$\boldsymbol{\theta}_i$ 表示了第 i 篇文章的主题分布;$\boldsymbol{\varphi}_k$ 表示第 k 个主题的词分布;z_{in}表示第 i 个文档的第 n 个词的主题;α,β 分别是文档主题分布与主题词分布先验分布的超参数. LDA 模型的生成过程如下:

①对所有设定的主题 $k \in [1, K]$,随机初始化 $\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\beta)$;

②对所有存在的文档 $i \in [1, M]$,随机初始化 $\boldsymbol{\theta}_i \sim \mathrm{Dir}(\alpha)$ 和 $N_i \sim \mathrm{Poiss}(\xi)$;

③对步骤②中的每篇文档,选择其所有的词项 $n \in [1, N_i]$,采样 $z_{in} \sim \mathrm{Mult}(\boldsymbol{\theta}_i)$, $w_{in} \sim \mathrm{Mult}(\boldsymbol{\varphi}_{zin})$.

1.2 文本特征提取

吉布斯采样之后,可以得到矩阵 $\boldsymbol{\Theta}$,$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_i, \cdots, \boldsymbol{\theta}_M)^{\mathrm{T}}$,$M$ 是数据集中文本总数,$\boldsymbol{\theta}_i$ 为 K 维的向量,K 是预先设置的主题数目.

文本聚类中,不同的文本特征提取方法对文本数据挖掘的精度有较大的影响. 由 LDA 得到的模型 $\boldsymbol{\Theta}$ 不仅是每个文本在各个主题上的概率,也可以看成是每个文本在隐主题空间维度上的特征值,当处理的文本数据较长时,可以有效降低提取出的特征的维度. 因此,本文采用由 LDA 得到的文本主题矩阵 $\boldsymbol{\Theta}$ 作为文本特征提取的最终结果.

2 FCM 算法与黑洞算法

2.1 FCM 聚类算法

FCM 算法是基于模糊划分的聚类算法,即对于适应度函数 F,求使得 F 值收敛的模糊划分矩阵 $\boldsymbol{Y}$ 以及聚类中心 $\boldsymbol{V}$. F,$\boldsymbol{Y}$,$\boldsymbol{V}$ 的求解公式如下:

$$\boldsymbol{Y} = [y_{ij}]_{C \times N}, \tag{1}$$

$$F = \sum_{j=1}^{N} \sum_{i=1}^{C} (y_{ij})^{\alpha} \| \boldsymbol{p}_j - \boldsymbol{v}_i \|^2, \tag{2}$$

$$y_{ij} = \frac{1}{\sum_{t=1}^{C} \left(\frac{\| \boldsymbol{p}_j - \boldsymbol{v}_i \|}{\| \boldsymbol{p}_j - \boldsymbol{v}_t \|} \right)^{\frac{2}{\alpha - 1}}}, \tag{3}$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{N} (y_{ij})^{\alpha} \boldsymbol{p}_j}{\sum_{j=1}^{N} (y_{ij})^{\alpha}}. \tag{4}$$

式中:C 为簇数;N 为文本总数;$\boldsymbol{p}_j$ 表示第 j 个文本向量;y_{ij}表示文本向量 $\boldsymbol{p}_j$ 隶属于第 i 类簇的程度;$\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_i, \cdots, \boldsymbol{v}_C)$,$\boldsymbol{v}_i$ 表示第 i 个聚类中心;α 为加权指数.

FCM 聚类算法的执行过程如下:

①初始化聚类簇数 C 以及簇中心 $\boldsymbol{V}$;

②根据式(3),计算新的模糊划分矩阵 $\boldsymbol{Y}$;

③根据式(4),更新聚类中心 $\boldsymbol{V}$;

④如果某次迭代与上次迭代的 F 值的差值小于阈值,则结束循环,否则执行步骤②.

2.2 黑洞算法

黑洞算法通过适应度值的计算和星体与黑洞间的吸引吸收机制,在迭代的过程中不断地选择出具有最佳适应度值的星体作为黑洞,直到最终确定黑洞的位置与适应度值. 黑洞与星体间的吸引机制如式(5)所示.

$$\boldsymbol{l}_i(t+1) = \boldsymbol{l}_i(t) + \mathrm{rand} \cdot (\boldsymbol{l}_{\mathrm{BH}} - \boldsymbol{l}_i(t)). \tag{5}$$

式中:$\boldsymbol{l}_i(t)$ 表示第 i 个星体在第 t 次搜索时的位置;$\boldsymbol{l}_{\mathrm{BH}}$代表黑洞的位置;rand 是 0 到 1 之间的随机数.

在算法运行过程中,各个星体同时向黑洞靠

拢. 黑洞与星体间的吸收机制通过吸收半径 r 的计算实现：

$$r = \frac{f_{\mathrm{BH}}}{\sum_{i=1}^{N} f_i}. \tag{6}$$

式中：f_{BH}，f_i 分别代表黑洞的适应度值与第 i 个星体的适应度值；N 为星体的总数量.

黑洞算法的执行过程如下：

①在搜索空间中随机对几个星体初始化.

②计算各个星体的适应度值，并从中选择具有最好适应度值的星体作为黑洞.

③按照式(5)移动各个星体.

④计算全部星体的适应度值，如果某星体的适应度值优于黑洞的适应度值，则让该星体成为新的黑洞.

⑤根据式(6)计算半径，如果存在某个星体与黑洞的距离小于该半径，则删掉该星体，并随机产生一个新的星体.

⑥如果算法达到了最大迭代次数或已经收敛，结束执行过程，否则执行步骤③.

3　BH – FCM 文本聚类算法

3.1　聚类流程

首先对需要聚类的文本进行中文分词与去停用词处理，消除高频无用词对特征提取的影响. 其次，使用 LDA 模型对该文本进行主题建模，提取出文本的主题特征. 然后，对于不同的文本进行主题相似度的计算并执行 BH – FCM 聚类算法，得到最终的文本聚类结果.

BH – FCM 算法将黑洞算法与 FCM 算法相结合，该算法的设计如图 2 所示.

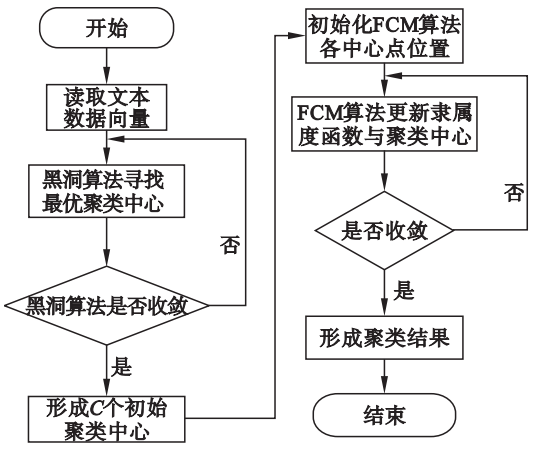

图 2　BH – FCM 聚类算法

Fig. 2　BH-FCM clustering algorithm

BH – FCM 算法首先使用黑洞算法进行全局最优初始中心点的寻找. 在黑洞算法收敛后，利用得到的全局最优解作为 FCM 算法的初始聚类中心，借以优化 FCM 聚类算法，解决 FCM 聚类算法对初始中心点的选择敏感以及容易陷入局部最优解的问题.

3.2　文本相似度计算

在 LDA 模型中，因为所有文本在各个隐主题上的分布都是由同一分布中采样得到[13]，所以本文使用 KL 散度的对称平滑变形 Jensen – Shannon 距离来衡量两个文本向量的相似度. Jensen – Shannon 距离越小，说明两文本越相似. 对于向量 $\boldsymbol{v}$ 和 $\boldsymbol{e}$ 来说，它们的 KL 散度定义如式(7)所示.

$$D_{\mathrm{KL}}(\boldsymbol{v} \parallel \boldsymbol{e}) = \sum_{i=1}^{N} v_i \ln(v_i / e_i). \tag{7}$$

Jensen – Shannon 距离的定义如式(8)所示.

$$D_{\mathrm{JS}}(\boldsymbol{v} \parallel \boldsymbol{e}) = \frac{1}{2}[D_{\mathrm{KL}}(\boldsymbol{v} \parallel \boldsymbol{a}) + D_{\mathrm{KL}}(\boldsymbol{e} \parallel \boldsymbol{a})]. \tag{8}$$

式中：$\boldsymbol{a} = \frac{1}{2}(\boldsymbol{v} + \boldsymbol{e})$.

3.3　黑洞算法的编码结构

黑洞算法用于 FCM 初始聚类中心点寻优时，关键是如何表示每一个星体及确定适应度函数. 由于黑洞算法中每一个星体都是聚类结果的候选解，所以每一个星体(候选解)的表示如图 3 所示.

$Z_{11}\ Z_{12}\ Z_{13}..Z_{1k}$	$Z_{21}\ Z_{22}\ Z_{23}..Z_{2k}$	…	$Z_{C1}\ Z_{C2}\ Z_{C3}..Z_{Ck}$

图 3　星体的表示

Fig. 3　Representation of a star

图 3 中：算法初始化时，Z_{ij} 为星体随机赋值的第 i 个聚类中心的第 j 维值，算法执行结束之后，Z_{ij} 表示最优初始聚类中心中第 i 个中心的第 j 维值；C 值为簇的数目；k 值为文本特征提取向量的维度.

黑洞算法的适应度函数计算如下：

$$\mathrm{fitness}(\boldsymbol{Z}, \boldsymbol{X}) = \sum_{i=1}^{T} \sum_{j=1}^{C} w_{ij} \cdot | D_{\mathrm{JS}}(\boldsymbol{x}_i \parallel z_j) |. \tag{9}$$

式中：$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_T)$，$\boldsymbol{x}_i$ 为第 i 个文本向量；$\boldsymbol{Z} = (z_1, z_2, \cdots, z_j, \cdots, z_{\mathrm{C}})$，$z_j$ 为第 j 个聚类中心；T 为文本向量的总数；C 为聚类的簇数. 若第 i 个文本向量属于第 j 个簇，则 w_{ij} 为 1，否则为 0.

4 实验与分析

4.1 文本数据选择

为了测试该算法的稳定性与有效性,本文采用复旦大学语料作为聚类测试样本,随机选择150 篇艺术类文本、210 篇历史类文本、240 篇电脑类文本作为数据集 Data1,150 篇运动类文本、210 篇政治类文本、240 篇经济类文本、260 篇农业类文本、240 篇环境类文本作为数据集 Data2,100 篇艺术类文本、140 篇历史类文本、160 篇电脑类文本、240 篇环境类文本、160 篇农业类文本、100 篇经济类文本、100 篇政治类文本作为数据集 Data3. 数据集的具体描述如表 1 所示. 本文采用 ICTCLAS[14] 中文分词系统,在对文本分词处理后去停用词,使用 LDA 模型进行文本特征提取,LDA 的超参数 α 通常取 0.05,β 为 0.01. 使用内部簇距离之和作为适应度函数值.

表 1 文本数据集

Table 1 Document data sets

数据集	文档数量	隐主题数量	簇的数量
Data1	600(150,210,240)	200	3
Data2	1 100(150,210,240,260,240)	350	5
Data3	1 000(100,140,160,240,160,100,100)	300	7

4.2 文本聚类对比

在 PSO – FCM 聚类算法中,设置粒子数为 50,设置初始惯性权重 w 为 0.72,加速系数 c_1 与 c_2 设置为 1.49. 在 GA – FCM 算法中,设置个体数为 50,采用单点交叉、交换突变的方法,交叉概率设置为 0.4. BH – FCM 聚类算法中,设置星体个数为 50. FCM 算法、PSO – FCM 算法、GA – FCM 算法、BH – FCM 算法的文本聚类平均准确率对比如图 4 所示.

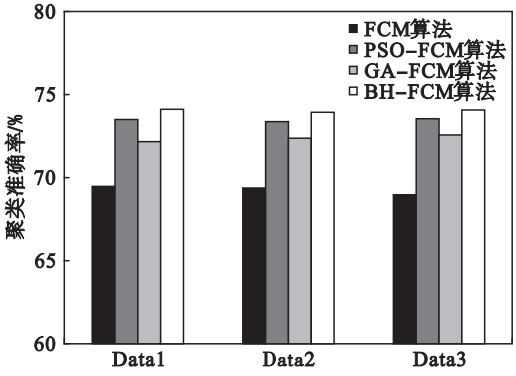

图 4 各算法的聚类效果对比

Fig. 4 Comparison of all algorithms for clustering effect

由图 4 可知,在三类数据集上,BH – FCM 算法的聚类准确率最高,PSO – FCM 算法次之,GA – FCM 算法再次之,FCM 算法准确率最低.

遗传算法、离子群算法、黑洞算法在 Data1 上寻找聚类中心点的过程中,适应度值的变化趋势如图 5 所示.

由图 5 可知,在聚类过程中,黑洞算法比遗传算法与粒子群算法的收敛速度最快,且能收敛到一个最优的适应度值,从而发现最优的初始聚类中心.

综上所述,BH – FCM 算法较 FCM 算法、PSO – FCM 算法、GA – FCM 算法有更优的效果,证明了该方法的有效性.

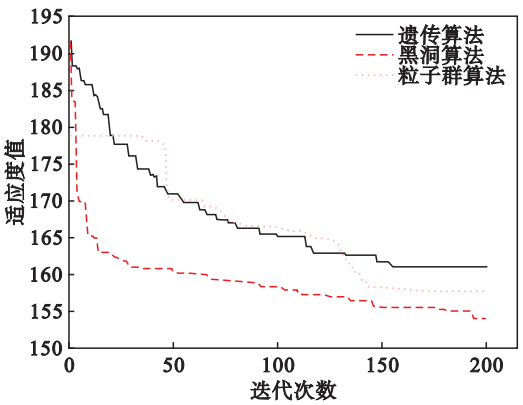

图 5 三种算法聚类过程对比

Fig. 5 Comparison of three algorithms for clustering process

5 结 语

本文提出一种基于黑洞算法的模糊 C 均值文本聚类方法 BH – FCM. 通过对黑洞算法在启发式搜索与聚类算法的研究,提出使用黑洞算法搜寻 FCM 算法的最优初始聚类中心,解决 FCM 算法容易陷入局部最优的问题. 结合 LDA 模型提取出的文本特征值,将 BH – FCM 算法应用于文本聚类领域中,通过仿真实验,将该算法与 FCM 算法、PSO – FCM 算法、GA – FCM 算法对比. 实验结果表明,黑洞算法能发现最佳的聚类中心,基于黑洞算法的 FCM 文本聚类方法具有最高的聚类准确率,证明了文本提出的方法是有效的. 在未

(下转第 1074 页)