# 基于改进的 K-means 算法在文本挖掘中的应用

杨　丹,朱世玲,卞正宇
(南京邮电大学 计算机学院,江苏 南京 210003)

**摘　要**:K-means 算法具有简单易于理解的特征,广泛运用于聚类过程中,但是其初始聚类中心是随机确定的,这样极容易导致聚类结果的稳定性很差。针对传统 K-means 算法对于初始聚类中心选择的敏感性及最大最小距离法容易选取离散点的不足,提出了一种新的聚类中心选择评判函数,依次考察每个点的函数值,选取当前函数值最大的点作为新的聚类中心,直到满足事先确定的聚类中心数。新聚类中心评判函数既可以保证新中心点周围是紧凑的,又可以保证远离其他中心点。最后将该算法运用于文本聚类之中,根据准确率、召回率及 $F$ 度量值来衡量算法的聚类质量。实验结果表明,该算法相对于传统算法和最大最小距离算法,准确率更高,聚类质量更好,较适合于文本聚类。

**关键词**:K-means 算法;聚类中心;文本聚类;文本距离;稀疏度
**中图分类号**:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2019)04-0068-04
doi:10.3969/j.issn.1673-629X.2019.04.014

# Application of Improved K-means Algorithm in Text Mining

YANG Dan,ZHU Shi-ling,BIAN Zheng-yu
(School of Computer Science and Technology,Nanjing University of Posts and Telecommunications,
Nanjing 210003,China)

**Abstract**:The K-means algorithm is simple and easy to understand,widely used in the clustering process. However,the initial cluster centers are randomly determined,which can easily lead to poor stability of the clustering results. In view of the sensitivity of the traditional K-means algorithm to the selection of the initial clustering center and the shortcoming of the maximum and minimum distance method in the selection of discrete points,we propose a new evaluation function for the selection of the clustering center. The function value of each point is examined successively,and the point with the largest current function value is selected as the new clustering center until the predetermined number of clustering centers is satisfied. The new clustering center evaluation function can not only ensure the compactness around the new center point,but also keep it away from other centers. In the last,the improved algorithm is applied to text clustering,and its clustering quality is measured according to the accuracy rate,recall rate and F metric. The experiment shows that the proposed algorithm has higher accuracy,better clustering quality,which is more suitable for text clustering than the traditional algorithm and the maximum and minimum distance algorithm.

**Key words**:K-means algorithm;clustering center;text clustering;text distance;sparseness

## 0　引　言

K-means 算法作为一种无监督的机器学习算法[1],具有简单易理解、聚类速度快等特点。目前聚类算法大致可以分为基于划分的、密度的、分层的、网格的、模型的等类型[2-6]。虽然 K-means 算法在众多领域应用广泛,但是其聚类质量的优劣则取决于多个因素。针对 K-means 算法对于初始聚类中心选择的敏感性,文献[7-8]提出了一种基于最大最小距离法选取初始聚类中心的方法,该算法基于距离最远的数据对象最不可能分到一个簇的事实,获取若干个距离彼此相隔较远的点作为初始中心[9]。虽然该算法的聚类效果相对于传统算法有了一定的提高,但是并没有考虑到距离较远的数据对象可能是离群点的情况,离群点作为初始聚类中心会影响到聚类的最终结果[10]并且也会增加迭代的次数。

在上述算法的基础上,文中依据稀疏度能够度量点周围紧凑程度的特点[11-12],结合最大最小距离法的原理,提出了一种新的结合稀疏度和距离的方式来选

**收稿日期**:2018-05-15　　**修回日期**:2018-09-25　　**网络出版时间**:2018-12-20
**基金项目**:国家“863”高技术发展计划项目(2006AA01Z201)
**作者简介**:杨　丹(1992-),女,硕士研究生,研究方向为机器学习、文本挖掘。

择初始聚类中心,并通过实验对该算法进行验证。

## 1　稀疏度

在数据集 $D\{x_1,x_2,\cdots,x_n\}$ 中,对象 $F$ 的稀疏度定义为:

$$S(x_i)=\frac{1}{K}\sum_{x_i\in N_i}d(x_i,x_j) \tag{1}$$

其中, $N_i$ 表示 $x_i$ 的 $K$ 个近邻对象组成的对象集合; $d(x_i,x_j)$ 表示 $x_i,x_j$ 之间的距离。

## 2　K-means 算法

### 2.1　基本思想

给定一个数据集 $D\{x_1,x_2,\cdots,x_n\}$ ,每一个数据对象都是 $m$ 维的,将数据集 $D$ 划分为 $K$ 个簇 $\{S_1,S_2,\cdots,S_K\}$ ,使用 K-means 算法过程需要使用的定义如下:

定义 1:两个数据对象之间的距离(欧氏距离)。

$$d(x_i,x_j)=\sqrt{(x_i,x_j)(x_i,x_j)} \tag{2}$$

其中, $x_i,x_j$ 为数据集中的两个数据对象。

定义 2:聚类中心。

$$Z_c=\frac{1}{n_c}\sum_{i=1}^{n_c}x_i \tag{3}$$

其中, $n_c$ 表示簇 $C$ 包含数据对象的个数; $x_i$ 表示簇 $C$ 的第 $i$ 个数据对象。

定义 3:聚类的终止条件。

$$E=\sum_{i=1}^{K}\sum_{j=1}^{n_j}d(x_j,Z_j) \tag{4}$$

其中, $K$ 表示簇的个数; $Z_j$ 表示第 $j$ 个簇的中心; $n_j$ 表示第 $j$ 个簇包含的数据对象的个数。

定义 4:Minps 值。

$$\text{Minpts}=\text{beta}\times\frac{N}{K_{\max}} \tag{5}$$

$$K_{\max}=\sqrt{N} \tag{6}$$

上述公式都是一种经验的规则, $x_i$ 邻近的 Minpts 个点组成 $N_i$ , $K_{\max}$ 表示最大的聚类簇数,beta 为用户自定义的参数。

算法 1:K-means 算法。

输入:数据集 $D$ ,簇的个数 $K$ ,聚类终止条件 $E$ ,迭代次数 $T$

输出: $K$ 个满足终止条件和迭代次数的簇

Step1:在数据集 $D$ 中任选 $K$ 个初始聚类中心;

Step2:循环 Step3 ~ Step5,判断是否满足终止条件 $E$ 和迭代次数;

Step3:按照定义 1 计算剩余的数据对象到每个聚类中心的距离,将其归并到距离最小的簇中;

Step4:按照定义 2,重新计算每个簇的聚类中心;

Step5:按照定义 3,计算聚类结果 $E$ 。



### 2.2　改进的 K-means 算法描述

对于初始聚类中心的选择,不仅要考虑到初始聚类中心周围的紧密程度,而且还要保证初始聚类中心尽量的离散。因此文中在采用稀疏度评判数据对象周围稀疏度的同时,为了保证聚类中心点尽量离散,采用了一种新的结合稀疏度和距离的度量方式。首先通过稀疏度来判断数据周围的紧密程度,然后结合最大最小距离的原则构建一个新的评价函数。该评价函数可以有效避免文献[10]必须人为确定参数的缺点,使得初始聚类中心的选择更加稳定。评价函数如下:

$$S_i=\frac{D(x_i)-S(x_i)}{\max[D(x_i),S(x_i)]} \tag{7}$$

其中, $S(x_i)$ 表示数据对象 $x_i$ 的稀疏度; $D(x_i)$ 表示数据对象 $x_i$ 与其他已选初始聚类中心距离的最小值; $S_i$ 的值是介于-1 和 1 之间。 $S_i$ 接近 1 时,表示数据对象 $x_i$ 的周围是紧凑的,且远离其他的聚类中心。

改进的算法流程如下:

输入:簇的数目 $K$ ,数据集 $D$ ,最大迭代次数 $T$ 及终止条件 $E$

输出: $K$ 个满足终止条件和迭代次数的簇

Step1:按照式 1 ~ 5,计算每个数据对象由 Minpts 个点组成的对象集合的稀疏度,选择稀疏度最小点作为第一个初始聚类中心;

Step2:按照式 7 计算剩余的数据对象与已选取的初始聚类中心值,选取值最大的点作为剩余初始聚类中心,依次循环下去,直到满足 $K$ 个初始聚类中心;

Step3:利用 $K$ 个初始聚类中心进行聚类。

## 3　文本聚类

### 3.1　TF-IDF 算法

文本表示常用的模型有布尔逻辑模型(BM)、向量空间模型(VSM)[13-15]、潜在语义索引(LSI)[16]和概率模型(PM)等。文中采用的是向量空间模型,对于文档 $D$ ,采用 $(\text{TF}-\text{IDF}_1,\text{TF}-\text{IDF}_2,\cdots,\text{TF}-\text{IDF}_n)$ 来表示[17-19]。IF - IDF 算法公式如下:

$$\text{TF}-\text{IDF}(t_i,d)=\frac{\lg[\text{tf}(t_i,d)+1]\times\lg\left(\frac{N}{n_i+1}\right)}{\sqrt{\left\{\lg[\text{tf}(t_i,d)+1]\times\lg\left(\frac{N}{n_i+1}\right)\right\}^2}} \tag{8}$$

其中, $n_i$ 为含词条 $t_i$ 的文档数; $N$ 为文档总数。

### 3.2　改进的文本距离计算

K-means 算法通常是根据两个数据对象之间的距离来归类的[20-22]。但在对文本进行聚类之前,通常采用余弦公式来计算两个文本间的相似度。文中根据相似度值总是处于 0 和 1 之间的特点,采取一种简单

的数学变化将文本相似度转化为文本距离,避免文献[8]采用 lg 方法需设计参数的问题,距离公式如下:

余弦计算公式:

$$\cos\theta = \frac{v_i \times v_j}{\sqrt{v_i^2} \times \sqrt{v_j^2}} \tag{9}$$

距离计算公式:

$$\text{Distance}_{i,j} = 1 - \cos\theta \tag{10}$$

其中,$v_i$ 和 $v_j$ 分别表示两个文本特征向量,在距离公式中,当两个文本的相似度为 1 时,距离为 0;当两个文本之间相似度为 0 时,两者之间的距离最大为 1。

## 4 实验结果与分析

### 4.1 实验一

#### 4.1.1 实验结果

实验环境:Window7 操作系统,1T 硬盘,MyEclipse 软件,Java 编程软件。

实验数据:为了验证改进后的算法优于传统算法和文献[4]算法,采用了 UCI 数据库中的 3 个数据集:Iris、Wine 及 Balance-Scale。其中 Iris 数据集包含 4 个属性,150 个数据对象,分为 3 类;Wine 数据集包含 13 个属性,178 个数据对象,分为 3 类;Balance 数据集包含 4 个属性,625 个数据对象,分为 3 类。

实验结果:实验通过准确率、召回率及 $F$ 度量值对传统算法、最大最小距离法及改进后的算法进行了比较,结果见表 1。

表 1 算法比较(1) %

| 算法 | Iris 数据集 | | | Wine 数据集 | | | Balance-Scale 数据集 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确率 | 召回率 | $F$ 度量值 | 准确率 | 召回率 | $F$ 度量值 | 准确率 | 召回率 | $F$ 度量值 |
| 传统算法 | 90.72 | 89.33 | 90.01 | 61.73 | 52.94 | 57.00 | 49.40 | 43.32 | 46.16 |
| | 53.74 | 50.00 | 51.80 | 72.37 | 69.60 | 70.96 | 50.82 | 45.82 | 48.19 |
| | 89.79 | 88.67 | 89.23 | 61.33 | 52.57 | 56.61 | 41.23 | 40.73 | 40.98 |
| | 55.63 | 54.00 | 54.80 | 61.73 | 52.94 | 57.00 | 46.81 | 41.69 | 44.10 |
| | 92.04 | 92.00 | 92.02 | 60.49 | 52.02 | 55.94 | 52.50 | 48.14 | 50.23 |
| | 53.74 | 50.00 | 51.80 | 64.78 | 58.60 | 61.54 | 55.77 | 51.60 | 53.60 |
| | 89.79 | 88.67 | 89.23 | 61.64 | 51.46 | 57.40 | 47.64 | 43.77 | 45.62 |
| | 53.96 | 51.33 | 52.61 | 60.35 | 52.25 | 55.55 | 51.51 | 47.94 | 49.66 |
| | 90.82 | 90.67 | 90.74 | 61.46 | 52.47 | 56.61 | 46.81 | 41.69 | 44.10 |
| | 90.72 | 89.33 | 90.01 | 61.00 | 52.25 | 56.29 | 43.68 | 40.63 | 42.10 |
| 平均 | 76.10 | 74.40 | 75.23 | 62.69 | 54.71 | 58.49 | 48.62 | 44.53 | 46.47 |
| 最大最小 | 90.72 | 89.33 | 90.02 | 65.73 | 56.94 | 61.02 | 50.76 | 45.81 | 48.16 |
| 改进算法 | 89.79 | 88.67 | 89.23 | 72.37 | 69.60 | 70.96 | 65.33 | 60.56 | 62.85 |

#### 4.1.2 实验分析

根据表 1 的实验结果可以看出,传统算法在 Iris 数据集的 10 组实验结果中,准确率、召回率及 $F$ 度量值中最高值与最低值都相差 30 个百分点以上,在 Wine 数据集上,三个度量值中最高值与最低值相差也在 15 个百分点以上。在相差值较小的 Balance-Scale 数据集中三个度量值的最高值与最低值相差也达到了 10 个百分点左右。因此传统算法的聚类结果的随机性很大,很不稳定。

最大最小距离法相对于传统算法在 Iris 数据集上的准确率、召回率及 $F$ 度量值都优于传统算法,但是在 Wine、Balance-Scale 数据集上聚类效果略低于传统算法。首先对 Wine 数据集进行分析,可以发现该数据集的最后一个属性值的范围跨度较大,最大值为 1 680,最小值为 278,采用最大最小距离法选取初始聚类中心会受到最后一条属性的影响,因此所选取的聚类中心不能很好地代表数据的实际分布。其次再对 Balance-Scale 数据集进行分析,可以发现它的分布极度不均匀,最大的簇包含 288 个数据对象,而最小的簇仅仅包含 49 个数据对象,如果仅仅采用最大最小距离法选取初始聚类中心,可能会导致所选取的中心来源于某一个或其中的某几个簇。

文中改进算法是在最大最小距离乘积法的基础上,采用稀疏度来评判所选取的聚类中心,既考虑到了数据的分布情况,又使所选取的聚类中心较远。从实验结果来看,该算法无论是在分布均匀的数据集中,还是在分布不均匀的数据集中,聚类的准确率、召回率以及 $F$ 度量值都优于传统算法和最大最小距离算法。

### 4.2 实验二

#### 4.2.1 实验数据

为了证明改进后的算法聚类结果的优越性,实验数据采用了搜狗语料库中的财经文档 150 篇,教育文档 500 篇和军事文档 350 篇。通过准确率、召回率及 $F$ 值对实验结果进行分析,比较算法的聚类效果。其

中传统算法的值为 5 次实验结果的平均值。实验结果见表 2。

4.2.2　实验分析

由于文本具有高维稀疏的特点,所以文本聚类的效果并不像普通数据集那样产生较好的结果。针对文本聚类过程中不可以采用簇的均值替代新的聚类中心的原理,采用 K-中心算法中计算新的中心点的方法来获取迭代中的中心点。通过表 2 可以看出,改进算法相较于传统算法和最大最小距离法在聚类质量上都有了一定的提高,因此,该算法相较于传统算法和最大最小距离法较适合用于进行文本聚类。

表 2　算法比较(2)

| 特征提取率 | 传统算法 | | | 文献[4]算法 | | | 改进算法 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确率 | 召回率 | *F* 度量值 | 准确率 | 召回率 | *F* 度量值 | 准确率 | 召回率 | *F* 度量值 |
| 0.015 | 44.00 | 43.80 | 43.90 | 46.01 | 45.86 | 45.93 | 48.52 | 44.42 | 46.38 |
| 0.020 | 44.61 | 43.80 | 44.20 | 43.41 | 43.16 | 43.28 | 47.43 | 46.53 | 46.98 |
| 0.025 | 54.21 | 53.78 | 53.99 | 45.81 | 44.03 | 44.90 | 58.41 | 58.29 | 58.35 |
| 0.030 | 42.83 | 41.57 | 42.19 | 37.49 | 37.34 | 37.41 | 45.85 | 43.46 | 44.62 |
| 0.035 | 45.69 | 44.43 | 45.04 | 47.69 | 45.91 | 46.78 | 51.32 | 55.00 | 53.10 |

## 5　结束语

针对 K-means 算法相较于其他聚类算法具有对初始中心敏感的特点,结合最大最小距离的方法提出了一种基于稀疏度选取初始聚类中心的算法。实验结果表明,该算法优于传统的 K-means 算法和最大最小距离算法,并得到较好的聚类效果。最后将该算法应用于文本的聚类处理中,在聚类的过程中通过一个简单的变换将相似度转化为对象之间的距离值。实验表明该算法相较于传统算法和最大最小距离法更加适用于文本聚类处理,准确率和召回率和 *F* 度量值都有了一定的提高。

**参考文献:**


[1] HAN J W, KAMBER M. Data mining: concepts and techniques[M]. 2nd ed. Beijing: China Machine Press, 2007: 263-266.

[2] 雷小锋,谢昆青,林　帆,等. 一种基于 K-Means 局部最优性的高效聚类算法[J]. 软件学报,2008,19(7):1683-1692.

[3] 张玉芳,毛嘉莉,熊忠阳. 一种改进的 K-means 算法[J]. 计算机应用,2003,23(8):31-33.

[4] 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.

[5] 刘立平,孟志青. 一种选取初始聚类中心的方法[J]. 计算机工程与应用,2004,40(8):179-180.

[6] 汤九斌,陆建峰,唐振民,等. 基于层次的 K-means 初始化算法[J]. 中国工程科学,2007,9(11):74-79.

[7] 周　涓,熊忠阳,张玉芳,等. 基于最大最小距离法的多中心聚类算法[J]. 计算机应用,2006,26(6):1425-1427.

[8] 杨大鑫,王荣波,黄孝喜, 等. 基于最小方差的 K-means 用户聚类推荐算法[J]. 计算机技术与发展,2018,28(1):104-107. DOI:10.3969/j. issn. 1673-629X. 2018. 01. 022.

[9] 翟东海,鱼　江,高　飞,等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J]. 计算机应用研究,2014,31(3):713-715.

[10] 袁　方,周志勇,宋　鑫. 初始聚类中心优化的 k-means 算法[J]. 计算机工程,2007,33(3):65-66.

[11] 张科泽,杨鹤标,沈项军,等. 基于节点数据密度的分布式 K-means 聚类算法研究[J]. 计算机应用研究,2011,28(10):3643-3645.

[12] 戚后林,顾磊. 基于密度与最小距离的 K-means 算法初始中心方法[J]. 计算机技术与发展,2017,27(9):60-63,69. DOI:10.3969/j. issn. 1673-629X. 2017. 09. 013.

[13] 石晓敬,韩　燮. 文本聚类算法的设计与实现[J]. 计算机工程与设计,2010,31(9):2013-2015.

[14] 彭　京,杨冬青,唐世渭,等. 一种基于语义内积空间模型的文本聚类算法[J]. 计算机学报,2007,30(8):1354-1363.

[15] 姚清耘,刘功申,李　翔. 基于向量空间模型的文本聚类算法[J]. 计算机工程,2008,34(18):39-41.

[16] 胡　燕,吴虎子,钟　珞. 中文文本分类中基于词性的特征提取方法研究[J]. 武汉理工大学学报,2007,29(4):132-135.

[17] 路永和,李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作,2013,57(3):90-95.

[18] 周丽杰,于伟海,郭　成. 基于改进的 TF-IDF 方法的文本相似度算法研究[J]. 泰山学院学报,2015,37(3):18-22.

[19] 黄承慧,印　鉴,侯　昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011,34(5):856-864.

[20] 万小军,杨建武,陈晓鸥. 文档聚类中 k-means 算法的一种改进算法[J]. 计算机工程,2003,29(2):102-103.

[21] 安计勇,高贵阁,史志强,等. 一种改进的 K 均值文本聚类算法[J]. 传感器与微系统,2015,34(5):130-133.

[22] 黄建宇,周爱武,肖云, 等. 基于特征空间的文本聚类[J]. 计算机技术与发展,2017,27(9):75-77,81.