文章编号:1009-2552(2019)12-0076-05 DOI:10.13274/j.cnki.hdzj.2019.12.016

基于K-means算法改进的短文本聚类研究与实现

王俊丰，贾晓霞，李志强

(华北计算技术研究所，北京 100083)

摘　要：文中讨论在文本类目数未知的情况下，如何对数据量过少的短文本进行有效聚类的问题。短文本的特点是每一份样本文章数据少，数据稀疏，用常规的聚类方法进行文本聚类不能取得很好的效果。文中提出了一种基于K-means的改进算法，提出一种简单降维方式和新的判别样本点距离的方法，经实验验证，文中改进算法比原K-means算法有更高的准确性。

关键词：短文本；K-means；聚类；无监督

中图分类号：TP391　　**文献标识码**：A

Research and implementation of short text clustering based on improved K-means algorithm

WANG Jun-feng，JIA Xiao-xia，LI Zhi-qiang

(North China Institute of Computer Technology，Beijing 100083，China)

Abstract: The problem of how to effectively cluster short texts with too little data in the case where the number of text categories is unknown is discussed in the article. The short text is characterized by less data in each sample article and sparse data. Conventional clustering methods for text clustering cannot achieve good results. In this paper，a new method based on K-means for discriminating sample point distance is proposed. Experiments show that the improved algorithm has higher accuracy than the original k-means algorithm.

Key words: hort text; K-means; clustering; unsupervised

0　引言

与大规模文本数据处理不同[1]，短文本聚类的特点是每一份样本中，各篇文章数据量过少，字数在100词左右；而整体需要聚类的样本中，文章数量也只有14到24篇。由此可见，需要处理的数据，若以单词为维度划分标准，十几篇文章将产生一个几百甚至上千维特征，而这么多特征下，只有十几条对应的数据，由此可见，这个初始数据构成是非常稀疏的。

在文本类目数未知的情况下，对数据量极少的短文本进行聚类，需要处理的问题有两点：第一点，是需要选择合理的聚类方法对数据进行聚类；第二点，是需要确定文本的类目数。

针对上述数据规模问题和需要解决的问题，在进行了多个试验比对后，发现K-means[2]算法在解决这个问题上表现最好，因此选择了K-means算法来对数据进行聚类；在确定文本类目数这个问题上，本文选择了斯坦福大学的Robert等教授提出的Gap Statistic方法[3]，结合K-means算法完成聚类数目的确定。其中，在K-means聚类算法上，文中设计了一种类似One-Hot[4]的编码方法来对文章进行编码，并提出了一种新的特征之间距离判断方法和降维方法，提升了聚类的准确性。

1　相关工作

1.1　K-means方法简介

K-means算法也叫K均值聚类算法，是一种通过迭代求解的聚类分析算法。其步骤是先确定需要

收稿日期：2019-08-26

作者简介：王俊丰(1993-)，男，硕士研究生，研究方向为自然语言处理。

聚类的数目 K，然后在特征空间中选择 K 个点作为初始的聚类中心点，接着计算每个特征到各个聚类中心点的距离，将每个特征分配给这个距离最短聚类中心点，最后得到每个聚类中心点对应的簇，从而完成第一轮聚类。

每结束一轮聚类后，需要结合各个簇所对应聚类点的特征坐标，根据设定方法更新聚类中心。下一步接着以更新后的中心重新进行聚类，选出新的簇，接着再次更新聚类中心。

这个过程将不断重复，直到满足某个终止条件，便完成聚类，最后得到一个以最新的聚类中心为中心点的 K 个类别。

1.2 Gap Statistic 方法介绍

Gap Statistic 方法是为了确定文本的聚类数目，在本质原理上是 Elbow 方法[5] 的提升改进，Elbow 方法也叫肘部判断法则。如图 1 所示，聚类数目 K 依次从 1 递增到文章段落数目量（图例为 10 个文章段落），当选择的 K 值小于真实聚类数目时，每次 K 增加 1，对应的 cost 值就会以极大速率减小；直到选择的 K 值大于真实值后，每次 K 值的增长变动，cost 值的变化就不会那么明显。随着聚类数目 K 不断提升，在某个 K 值范围处，K 值变化速率会大幅变小，处于拐点处位置的 K 值，即为聚类数目。

在 Elbow 方法中，数据集的每个观测点用 x_i 表示，中心点为 c，用 $dist(x,c_i)^2$ 来表示观测点和对应的中心点的距离，则 Elbow Method 公式如下：

$$SSE = \sum_{i=1}^{K} \sum dist(x,c_i)^2 \quad (1)$$

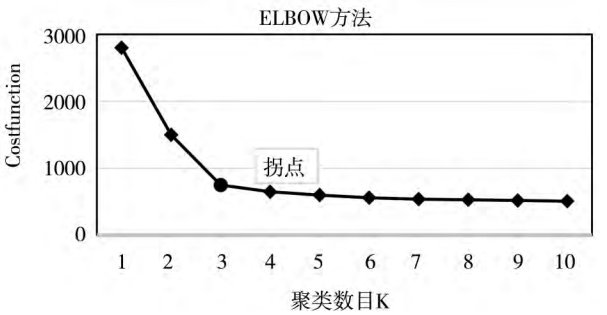

图 1 Elbow 方法

Elbow 方法缺点在于选择最适合的 K 值有时候并不是那么明显，而且大多时候需要人为判断，因此斯坦福大学的 Robert 等教授提出了 Gap Statistic 方法。Gap Statistic 需要引入参考的测值[6]，这个参考值可以由 Monte Carlo 采样[7] 的方法获得。

对于聚类数目为 K 的情况，将数据分为 k 类 $C_1, C_2, \cdots, C_k$，其中 C_r 表示观测点属于第 r 类的情况，定义 $n_r = |C_r|$ 为属于类 C_r 的个数。

定义 D_r 为第 r 类中任意两点距离和：

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \quad (2)$$

定义 W_k 如下：

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \quad (3)$$

最后定义 $Cap_n(k)$：

$$Gap_n(k) = E_n^*(\log(W_k)) - \log(W_k) \quad (4)$$

其中，E_n^* 表示参考数据期望值。

Gap Statistic 方法思想是通过对平均分布的参考数据集的期望值 $E_n^*(\log(W_k))$ 和观测数据集的 $\log(W_k)$ 进行对比，选择 $\log(W_k)$ 下降最快的 K 值为最优的聚类数目。

图 2 展示了随着聚类数目 K 的变化，参考数据和聚类数据的变化趋势。图 3 展示了参考数据和聚类数据差值变化，可以明显看到，在 K 值为 3 时候，他们的差值是最大的，这也表明了参考数据和观测数据下降最快的点，所以聚类数目 K 为 3。

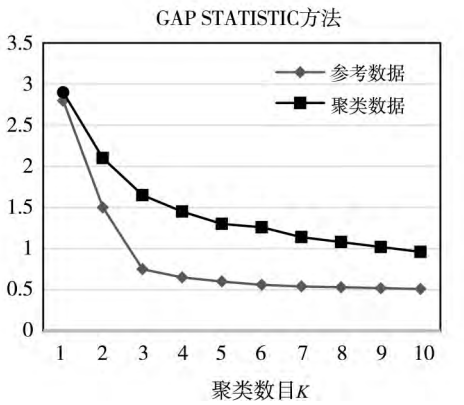

图 2 Gap Statistics 方法

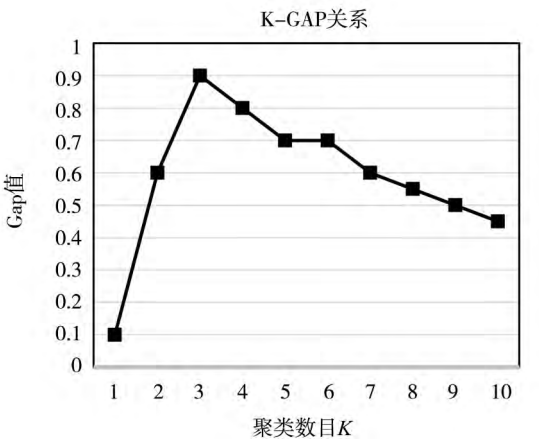

图 3 K-Gap 关系图

2 实验设计

实验分为数据预处理、特征建立、特征降维、K-

means 聚类，结果分析几个步骤，流程如图 4 所示。

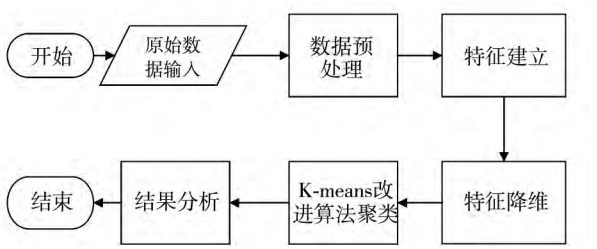

图 4　实验流程

2.1　数据说明

每份数据为从 3 篇以上的不同类型的文章中抽取出 1 个或多个简单段落，共 14 到 24 段，其中每个段落单词数量不超过 150 个。最后，将抽取到的这些段落存入一个 txt 文本中等待处理。

2.2　数据预处理及特征建立

在数据预处理阶段，对一份数据进行整体遍历，将所有单词提取出来组成一个列表，并去重，获得一个单词集合。接着根据停用词词表，去除集合中的停用词[8]；将剩下的单词映射为一个标签，如数字映射为 D_1、D_2、D_3、…、D_i，单词映射为 S_1、S_2、S_3、…、S_j 再利用字符串相似度判断，将相似的单词合并映射为同一个标签[9]。最后得到一个能够代表这一份数据的特征空间 $[D_1, D_2, D_3, \cdots, D_i, S_1, S_2, S_3, \cdots, S_j]$，空间维度即为处理后的标签的个数，$i+j$ 维。

特征建立阶段，对数据中的每一个段落，即每一篇文章去停用词以后，对照预处理阶段的映射表，若这一段中出现过映射表的单词，则在该维度位置记录出现次数，每段统计完成后，没有出现过的单词或数字位置，记为 0，最后得到个段落对应的向量，如 [1,2,0,4,…,3,0,0,0,5,3] 共 $i+j$ 维。

对每一段进行操作，最后得到一个 $n \times m$ 的矩阵，其中，n 为这一份数据中的段落数，m 为向量空间维度，即 $m = i + j$。

数据预处理及特征建立流程图。

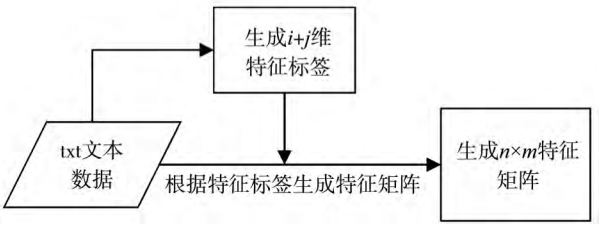

图 5　数据处理及特征建立

2.3　特征降维

除了在数据预处理中通过去停用词和将相似度归类为一个标签的降维方法以外，文中还提出了一种新的降维方式。

在特征矩阵中，由于整体的特征向量来自于整份数据，而不同的文章类别不同，其包含的单词也会有很大差异，因此每一段对应的向量有很多值为 0 的情况，最后导致的结果是特征矩阵过于稀疏，针对这个问题，提出下述降维方法。

由于每个段落都代表一个短文本，因此，在处理短文本聚类问题上个，不存在处理离散点这一步骤，若某个点过于偏离其他类别，则可以单独归为一类。依此设计了降维方法：对于 $n \times m$ 的特征矩阵，若在某一维度，即在某一列上，只有一个值，其他全为 0 时，则可以去除这个维度，以下为数学定义：

设特征矩阵 $X_{n\times m}$，若为 $X_{rj} \neq 0, X_{ij} = 0$，其中 $r, i \in [0, n], j \in [0, m]$，且 $i \neq r$；则 $del(col(X_{*j}))$。

即删除矩阵 $X_{n\times m}$ 中的第 j 列，得到降了一维后的新特征矩阵 $X'_{n\times(m-1)}$。

如下例，设矩阵为 $X_{4\times5}$，则可以看到，在第二列，即 $j=2$ 时，当 $r=4$ 时，有 $X_{rj} \neq 0, X_{ij} = 0$，且 $i \neq r$，则可以删除第二列，得到新的特征矩阵 $X'_{4\times4}$。

在实际的一组实验中，通过上述的降维方法，实验成功从 437 维空间降到 98 维，一共降低了 339 维，并且经过前后对比，降维后的聚类效果明显优于降维前的聚类效果。

2.4　K-means 聚类

2.4.1　TF-IDF 词频加权

经过降维处理得到最新的特征矩阵 $X_{n\times m}$ 后，依据 TF-IDF 原理[10]，再对各段文章对应词进行处理，设 X_{ij} 为特征矩阵中第 i 行和第 j 列值，实际意义为第 i 段文章对应的 j 维度指代的词的数量，令 X'_{ij} 使用 TF-IDF 后得到的新值，则依据 TF-IDF 有如下公式：

$$X'_{ij} = \frac{X_{ij}}{\sum_{k=1}^{n} X_{ik}} \cdot \log \frac{n}{t+1} \tag{5}$$

其中，t 为整个数据集中个，包含 j 维度代表的词条段落数量。

依次对特征矩阵中的所有元素进行 TF-IDF 加权处理。最后，得到最新的特征矩阵 X'_{nm}，便可以开始进行 K-means 算法聚类。

2.4.2　新的距离计算方法

在特征点之间距离度量选取上，针对题设问题，实验先后采用了欧氏距离[11-12]、曼哈顿距离[13]、余弦距离[14] 作为聚类的度量距离，最后应用了新设计的距离计算方法，经过对比[15]，得到新的距离计算方法得到的效果优于前三者。

设需要衡量的两个点为 A、B，其中 A_i、B_i 分别为 A，B 两点第 i 维坐标值，新采用的计算 A、B 两点的距离公式为：

$$D(A,B) = \frac{same(A,B) \times 2}{A + B + 0.01} \tag{6}$$

其中：

$$A + B = \sum_{i=1}^{m}(A_i + B_i) \tag{7}$$

$$same(A,B) = \sum_{i=1}^{m} add(A_i,B_i) \tag{8}$$

$$add(A_i,B_i) = \begin{cases} A_i + B_i, & if\ A_i \times B_i \neq 0 \\ 0, & if\ A_i \times B_i = 0 \end{cases} \tag{9}$$

在 $D(A,B)$ 计算公式中，分母 0.01 为预防值，确保分母不会为 0。

2.4.3 初始聚类中心点选取及 K 值确定

对数据用层次聚类算法聚类[16]，选择聚类获得的簇中心点为 K-means 初始聚类中心点。

确定好初始聚类中心点的选择方法以后，K 值依次从 1 到 n 进行遍历，对特征矩阵进行类目数位 K 的聚类划分。

最后，应用 1.2 节提到的 Gap Statistic 方法，选择适合的 K 值作为聚类数目，得到最终的结果。

2.5 模型评价标准

模型评价从聚类数目和聚类准确性两个方面进行评价，聚类数目标准即为评价聚类后的类目数是否正确，若不正确，则判定该文本聚类效果为差，评价分数 3 分以下，与正确聚类数目相差越大，分数越低；若聚类正确，则给 6 分，再观测聚类准确性效果，即是否将同一类的文本聚为一类，准确性越高，分数越高，最高为 10 分，以下为具体公式描述。

对每组数据 $D = \{x_1, x_2, \cdots, x_n\}$，其标准结果为 $\{\lambda_1^*, \lambda_2^*, \cdots, \lambda_n^*\}$，其中 $\lambda_1 \in \{0,1,\cdots,m-1\}$ 为对应段落 x_i 输出值；模型测试结果为 $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$，其中 $\lambda_1 \in \{0,1,\cdots,k-1\}$ 为段落 x_i 对应的输出值，则定义：

$$a = |SS|,$$

$$SS = \{(x_i,x_j) \mid \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\} \tag{10}$$

$$b = |SD|,$$

$$SD = \{(x_i,x_j) \mid \lambda_i = \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\} \tag{11}$$

$$c = |DS|,$$

$$DS = \{(x_i,x_j) \mid \lambda_i \neq \lambda_j, \lambda_i^* = \lambda_j^*, i < j\} \tag{12}$$

其中，a 为预测结果和真实结果一致的真正例情况；b 为预测结果为一类，真实结果不为一类的假正例情况；c 为测试结果不为一类而真实结果是一类的假反例情况。

最终评分为：

$$score = \begin{cases} \dfrac{6}{e^{|k-m|}}, & if\ k \neq m \\ 6 + \dfrac{4a}{a+b+c}, & if\ k = m \end{cases} \tag{13}$$

由公式就能看出，分类的类别数目正确与否直接影响到分数结果，若分类类别数错误，则得分将低于 3 分；若分类类别数正确，则分数将为 6 分的基础上加上分类准确性分数。

3 实验结果

3.1 数据集及分数设定

实验中每组数据分别为 3 篇以上的不同文章中，抽取 1 段或多段组成短文本，每组数据共 14 ~ 24 段，抽取的每段词数不超过 150 个；这样的数据一共选择 10 组。

依据评分标准和选择的数据数量，10 组数据评分一共 100 分。

3.2 实验结果及评估

本文除了用 K-means 改进算法对文本聚类以外，还进行了分层次聚类、基于欧氏距离的 K-means 聚类方法和基于余弦距的 K-means 聚类方法。

实验得分结果如表 1 - 2 所示。

表 1 得分数据 I

聚类方法	第一组	第二组	第三组	第四组	第五组
分层聚类	7.548	2.207	7.864	8.202	0.812
欧氏距离 K-means	7.762	7.686	8.203	8.446	0.812
余弦距离 K-means	8.003	7.962	8.664	8.887	2.207
改进算法 K-means	8.458	8.354	8.571	10	8.224

表 2 得分数据 II

聚类方法	第六组	第七组	第八组	第九组	第十组	总分
分层聚类	7.004	2.207	8.332	7.686	7.221	59.083
欧氏距离 K-means	7.886	7.234	8.444	7.886	8.498	72.857
余弦距离 K-means	8.204	8.004	8.786	7.989	8.789	77.495
改进算法 K-means	9.242	8.645	10	8.643	9.324	89.461

得分数据变化趋势图，如图 6 所示。

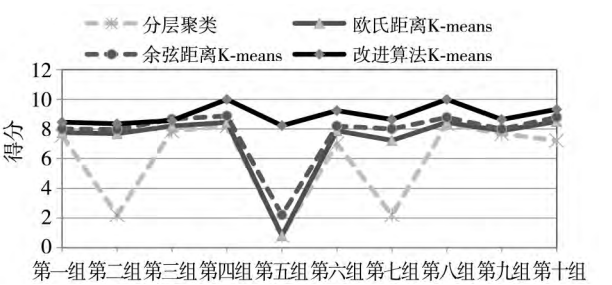

图 6 得分变化趋势图

由评分标准可知，每份数据评分满分是 10 分，只有在类目数确定情况下，得到分数才能超过 6 分，否则得到分数将只有 3 分以下。

由图 6 看出，分层聚类效果最差，聚类效果波动太大，无法准确地确定聚类数目，其中有三份数据聚类错误；基于余弦距离和基于欧式距离的 K-means 聚类算法效果相近，能够基本确定聚类数目，在同一份数据上出现聚类错误情况；效果最好的是改进算法，能够完全预测出正确聚类数目，并且聚类效果明显优于前面三种方法。

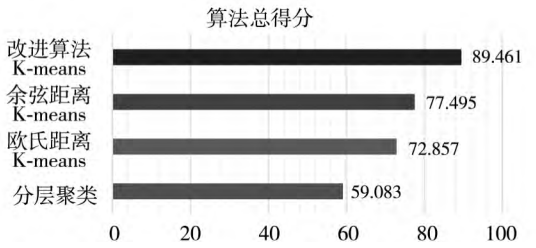

图 7　算法总得分对比图

由图 7 可以明显看到，基于 K-means 改进的算法效果是最好的，接近于 90 分；而基于余弦距离和基于欧式距离的 K-means 效果相近，表现一般分数在 70 到 80 分之间；表现最差的是分层聚类算法，分数仅有 59 分。

实验结果表明，本文使用的基于 K-means 改进算法的短文本聚类算法相较于传统 K-means 算法有更高的准确率。

4　结束语

本文对数据量少、数据规模小的短文本聚类问题进行了进一步探讨。在小规模和小数据量的条件下，大型模型并不能够对这样的数据进行很好的分析，极易出现过拟合现象。因此只能选择传统的机器学习方法进行聚类研究，本文就此设计了基于 K-means 的改进算法。

相较于传统 K-means 算法，本文算法在模型特征构建、特征降维以及设计新的算法距离度量方法上进行了一些创新改进；本文提出的特征降维算法实现简单且高效，新的距离度量方法能够更明显地刻画这类特征之间的关系。在解决这类短文本聚类问题上，本文提出的基于 K-means 改进算法能够提升聚类的准确率。

参 考 文 献：

[1] 刘澎，陆介平. 基于 MapReduce 的改进 k-means 文本聚类算法 [J]. 信息技术，2016，40(11)：201－205.

[2] Hartigan J，Wong M. Algorithm as 136：a K-means clustering algorithm [J]. Journal of the Royal Statistical Society. Series C(Applied Statistics)，1979，28(1)：100－108.

[3] Tibshirani R，Walther G，Hastie T. Estimating the number of clusters in a data set via the gap statistic [J]. Journal of the Royal Statistical Society B，2001，63(2)：411－423.

[4] Hanzawa S，Sakata T，Kajigaya K，et al. A dynamic CAM-based on a one-hot-spot block code-for millions-entry lookup [C]. Symposium on Vlsi Circuits. IEEE，2004：382－385.

[5] 王千，王成，冯振元，等. K-means 聚类算法研究综述 [J]. 电子设计工程，2012，20(7)：21－24.

[6] 肖宇，于剑. Gap statistic 与 K-means 算法 [J]. 计算机研究与发展，2007，44(Z2)：176－180.

[7] Hastings W K. Monte carlo sampling methods using markov chains and their applications [J]. Biometrika，1970，57(1)：97－109.

[8] Silva C，Ribeiro B. The importance of stop word removal on recall values in text categorization [C]. International Joint Conference on Neural Networks. IEEE，2003：1661－1666.

[9] 李星毅，曾路平，施化吉. 基于单词相似度的文本聚类 [J]. 计算机工程与设计，2009，30(8)：1966－1968.

[10] Salton G，Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management，1988，24(5)：513－523.

[11] Deza M M，Deza E. Encyclopedia of distances [J]. Reference Reviews，2009，24(6)：1－583.

[12] 李云. 快速语音识别算法的研究 [J]. 信息技术，2017，41(7)：118－120.

[13] Krause E F. Taxicab geometry [J]. Mathematics Teacher，1973，66(8)：695－706.

[14] Liao H，Xu Z. Approaches to manage hesitant fuzzy linguistic information based on the cosine distance and similarity measures for HFLTSs and their application in qualitative decision making [J]. Expert Systems with Applications，2015，42(12)：5328－5336.

[15] Gang Q. Similarity between euclidean and cosine angle distance for nearest neighbor queries [C]. Acm Symposium on Applied Computing，2004：1232－1237.

[16] 尉景辉，何丕廉，孙越恒. 基于 K-Means 的文本层次聚类算法研究 [J]. 计算机应用，2005，25(10)：2323－2324.

责任编辑：梁毅菲