基于 K-means 的文本聚类算法①

Text Clustering Algorithm Based on K-Means

毛嘉莉 (西华师范大学 计算机学院 四川 南充 637002)

摘 要: 针对K—means 算法容易收敛到局部最优以及对初值的依赖性,基于多次采样一次预聚类搜索初始聚类中心的思想,提出了一种改进的 K—means 文本聚类方法。实验结果表明,改进的算法较原算法在准确率上有较大提高,并且具有更好的稳定性。

关键词: K—Means 算法 文本聚类 向量空间模型

1 引言

在机器学习领域中,文本聚类属于无指导的学习方法,不须依赖预先定义的类别和带有类别的训练集。它把一个文本集分成若干个类,并使同一类中的文本信息之间具有较高的相似度,而不同类之间的文本差别较大。文本聚类在大规模文本集的组织与浏览、文本集层次归类的自动生成等方面都具有重要的应用价值。

典型的文本聚类方法大致可分为层次凝聚法和平面划分法两种类型。层次凝聚法是一种自底向上的方法,它首先将每个文本看作一类,然后合并这些类成为较大的类,直到所有的文本都在一类中,或者用户指定的某个终止条件被满足。层次凝聚法可达到较高的精度,但时间复杂度较高。而以 K-means 算法为代表的划分方法,具有较好的可伸缩性和很高的效率,适合处理大文本集。

对于一个大小为 n 的文本集,K-means 算法首先随机地选取任意 k 个文本作为初始类中心,再根据每个文本与各个类中心的相似度,将它赋给最相似的类,然后重新计算每个类的中心。这个过程不断重复,直到准则函数收敛,算法的时间复杂度为 O(nkt),t 为迭代次数。K-means 算法中一般采用误差平方和准则函数作为聚类准则函数,其定义为:

$$J_c = \sum_{j=1}^{k} \sum_{X_i \in C_j} \left| X_i - Z_j \right|^2$$

其中,J_c 是所有文本对象平方误差的总和,X_i 是类 C_j 中的一个文本对象,Z_j 是类 C_j 的中心。误差平方和准则函数在各类的形状大小差异较大时,有可能出现将大的类分割的现象[1]。此外在运用误差平方和准则函数测度聚类效果时,最佳聚类结果对应于目标函数的极值点,由于目标函数存在着许多局部极小点[2],而算法是通过迭代重定位技术最小化目标函数,若初始化落在了一个局部极小点附近,就会造成算法在局部极小处收敛。因此初始类中心的随机选取可能会陷入局部最优解,而难以获得全局最优解。

针对使用误差平方和准则函数的聚类算法难以划分形状差异较大的类,文献[3]中采用了基于代表点的处理方法并取得了较优的聚类效果,而对于初值选取影响聚类效果的情况,通常希望找到相互之间距离较远的点作为初始中心点。但是在一般的基于贪心算法的初始中心点搜索过程中,由于仅仅基于距离因素,往往找到许多孤立点作为中心点,且初始中心点选择的随机性较强,导致聚类结果的随机性。有些算法亦采用了全局优化方法中的模拟退火技术以摆脱局部最小[4];还有的算法采用多次取样[5]数据集二次聚类以获取最优初值的思想(对多次提取的样本聚类产生新的多组类中心,并对这些类中心再次聚类,比较聚类结果从而得到最优的初值)。

本文提出了一种基于 K-means 算法的文本聚类新方法,该算法对文本进行聚类的主要过程是:首先

① 基金项目:四川省教育厅重点科研项目(07ZA121)
收稿时间:2009-02-06

运用向量空间模型(VSM)表示文本信息，对于给定的聚类数 K，基于多次取样一次聚类的思想搜索出最优的 K'(K'> K)个文本作为初始的类中心，再对该文本集执行 K-means 算法聚类，然后合并最为相似的类，直到类的数量减少到指定的 K 值为止。实验结果表明，该算法聚类效果优于原始算法并具有较好的稳定性。

2 改进的K-means文本聚类算法

对于给定的文本集 $D=\{d_1,d_2,d_3,\cdots,d_i,\cdots,d_n\}$，改进的 k-means 算法的流程如下：

1) 运用向量空间模型表示文档信息，采用常规方法去除无用词，并用 TF-IDF 法则约简特征词条得到文本特征集，再计算各个特征词条的权值，把文档表示为向量的形式。

2) 搜索 K'个文本向量 $S=(S_1,S_2,S_3,\cdots,S_j,\cdots,S_{k'})$作为初始的类中心，相似度采用文本向量夹角的余弦系数表示。在搜索过程中通过对文本集随机取样，尽量使得取样后的数据既不失真，又能体现数据的原始分布特征。为了尽可能减小取样对初始类中心选取所产生的影响，采取 J 次取样(每次提取的样本集大小应该能装入主存，并尽可能满足 J 次提取的样本集之和等于原始数据集)。对于每个样本集分别执行一次 K-means 算法，产生 J×K′ 个类中心；选择 min{Jc} 的一组为初始聚类中心。为避免采用准则函数容易将大的类分割的情况发生，算法设定初始聚类数为 K'(K'> K，根据质量要求和时间折中选择 K'值)，较大的 K'值可以扩大解空间的搜索范围，减少某些极值点附近无初值的现象。

3) K-means(D, K')，用搜索到的初始聚类中心对原文本集执行 K-means 算法。为了使整个聚类结果的类内平均相似度尽量大，算法采用渐变中心的优化方法[6,7]，即在算法的每轮迭代中，一旦将某个文本归入到某个类中心所在的类，立即根据这个文本向量修改类中心，而不是等到迭代结束时取类内各文本向量的平均值作为类中心。

4) Merging(K'→K)，用凝聚的层次聚类算法(采用平均距离 Average Link)对 K'个类再次聚类，直到类的数量减少到指定的 K 值为止。

在改进的 K-means 算法中，对提取的文本样本子集搜索初始聚类中心的过程，由于文本数较少，迭代次数很小，速度很快。对于文本集合非常大，提取样本以及搜索初始聚类中心的过程所耗费的时间在整个算法中可以忽略不计，而利用搜索出的初始聚类中心对原数据集聚类这个过程所需时间为 O(nk't)，最后运用层次聚类算法合并聚类的过程所需时间 $O((k')^2 \log k')$，因此改进的 k-means 算法所需总的时间为 $O(nk't+(k')^2 \log k')$。

3 实验结果及分析

为验证改进 k-means 算法的有效性，本文选择了 http://www.re-search.att.com/~lewis 下的 Reuters-21578 中 4 个文本集作为测试数据集，首先对其进行预处理，用 Porter 切词器对文本切分词语并排除停用词，统计词频，将每个文本生成该文本的特征向量空间，再对每个向量空间进行特征提取，提取出能代表该文本的特征向量。经过预处理后，各个文本集的统计信息如表 1 所示。

表 1　文本数据集描述

文本集	文本数	类数	词数
DS1	1504	13	2886
DS2	1657	25	2758
DS3	2730	28	3475
DS4	1287	12	2369

为评价聚类结果，采用了常用的 F-measure 来衡量，F-measure 综合了信息检索领域中的查准率(Precision)和查全率(Recall)的概念。对于一个聚类 i 和一个主题 j 的 P(Precision)与 R(Recall)定义如下：

$$P(i,j)=\frac{N1}{N2};\quad R(i,j)=\frac{N1}{N3},$$

其中 N1 为在聚类 i 中且主题类别为 j 的文本数，N2 为聚类 i 中所有的文本数，N3 为属于主题类别 j 的文本数，则主题 j 的 F(F-measure)定义为

$$F(j)=\frac{2PR}{(P+R)}$$

最终计算出所有类别的 F-mea sure 加权平均值：，其中，|j| 表示分类 j 中所有对象的数目。F 值越大，表明算法的整体聚类精度越高。

在实验中，分别运行随机选择初始聚类中心的传统 k-means 算法和本文提出的改进 k-means 算法，所得结果如表 2 所示。

表 2 传统 k-means 算法同改进 k-means 算法 F 值的比较

文本集	k-means	改进 k-means(K'= 2.5K)
DS1	0.637	0.783
DS2	0.584	0.632
DS3	0.645	0.696
DS4	0.695	0.724

从表 2 中我们可以看出：对于 Reuters-21578 中的 4 个测试文本集，使用改进 K-means 算法比传统的 K-means 算法聚类的 F-measure 值都高，这主要是由于初始聚类中心和聚类个数的选择能够很大程度地影响 K-means 聚类结果的有效性。改进的 K-means 算法虽也采用聚类准则函数为标准，但初始选取 $k'(k'=2.5k)$ 个聚类中心，最后合并聚类结果，避免了因使用聚类准则函数、随机选取初始聚类中心所造成的陷入局部解的情况。

4 结语

k-means 算法是一种广泛应用的文本聚类算法，但初始聚类中心选择的随意性造成了聚类结果的不稳定，本文提出的改进 k-means 算法，通过对样本而不是整个文本集预聚类搜索初始聚类中心，可以避免孤立点的影响，提高了结果初始中心的代表性，采取预先设定大的聚类数，最后结合层次聚类算法合并聚类的方法，进一步消除了算法对初始聚类中心及聚类个数的敏感性，并能获得较高的聚类精度。在标准数据集 Reuters-21578 上进行的实验证明了它的有效性和可行性。

参考文献

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

2 唐立新,杨自厚,王梦光.用遗传算法改进聚类分析中的 K-平均算法.数理统计与应用概率, 1997,12(4):350－356.

3 陈恩红,王上飞,宁岩,等.一种利用代表点的有效聚类算法设计与实现.模式识别与人工智能, 2001,14(4).

4 Selim SZ, Alsultan K. A Simulated Annealing Algorithm for the Clustering Problem. Pattern Recognition.1991,24(10):1003－1008.

5 Fayyad U, Reina C, Bradley PS. Initialization of Iterative Refinement Clustering Algorithms. Microsoft Research Technical Report MSR-TR-98－38. June 1998.

6 Faber V. Clustering and the Continuous k-Means Algorithm.http://library.lan.lgov/cgi-bin/getfile?00412967. pdf.1994.

7 Larsen B, Aone C. Fast and Effective Text Mining Using Lineartime Document Clustering. Proc. of the fifth ACM SIGKDD International Conf. on Knowledge Discovery and Datamining. San Diego, 1999:16－22.

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