聚类模式下一种优化的K-means文本特征选择

刘海峰 刘守生 张学仁
(解放军理工大学理学院 南京 210007)

摘 要 文本特征降维是文本自动分类的核心技术。K-means方法是一种常用的基于划分的方法。针对该算法对类中心初始值及孤立点过于敏感的问题,提出了一种改进的K-means算法用于文本特征选择。通过优化初始类中心的选择模式及对孤立点的剔除,改善了文本特征聚类的效果。随后的文本分类试验表明,提出的改进K-means算法具有较好的特征选择能力,文本分类的效率较高。

关键词 特征选择,聚类,K均值,文本分类
中图法分类号 TP391 **文献标识码** A

Clustering-based Improved K-means Text Feature Selection

LIU Hai-feng LIU Shou-sheng ZHANG Xue-ren
(Institute of Sciences, PLA University of Science and Technology, Nanjing 210007, China)

Abstract Text feature reduction is the key technology in text categorization. In addition, K-means is an partitioning method which usually be used. With regards to this arithmetic excessively incentive to the initial centers and the isolated points, the improved K-means arithmetic was put forward which is used in text feature selection. Text feature clustering was improved by optimizing primitive class center's options and the elimination of isolated point. Following text classification test shows that the K-means arithmetic put forward in this paper has a good feature selection ability and high efficiency in text categorization.

Keywords Feature selection, Clustering, K-means, Text categorization

在网络时代的今天,文本信息是网络信息的主体。作为文本信息处理的核心技术之一,文本自动分类(Text Categorization,TC)技术的研究已经成为研究的热点[1-3]。文本自动分类技术日益广泛地应用在统计自然语言处理、模式识别、情报检索等领域,已经成为信息检索、信息过滤的有效手段,在提高信息利用的有效性、实时性和准确性方面具有重要的现实意义和广阔的应用前景。

文本自动分类是指在预先给定的类别标记集合下,根据待分类文本的内容对其类别归属进行判定的过程。文本特征向量的高维性及数据的稀疏性是文本分类的瓶颈,文本特征降维技术是文本自动分类的核心技术。目前,常用的特征降维方法有特征选择[4]和特征抽取[5]。其中基于选择的特征降维模式以其模型简洁、易于理解、效率较高而得到广泛的应用。

聚类(clustering)是将数据集按照其元素之间的相关性划分为若干簇或类的过程。聚类的目的是使同一簇内元素间的相关性最大而不同簇之间的相异性最大。作为一种无监督的机器学习模式,聚类技术已经成为对文本信息进行有效的组织和管理的重要手段[6]。文本特征聚类可以为文本特征选择提供前期准备,是文本特征降维的一种非常有效的方法。目前存在很多聚类算法。算法的选择取决于数据的类型、聚类的目的和应用。而在文本分类、聚类研究中,聚类的目的是将文本集划分成若干个簇,要求处于同一个簇内的文本之间的相似度最大而不同簇之间的文本的相似度最小。

1 基于划分的K-means聚类方法特点及其问题

聚类方法主要分为基于层次的聚类方法和基于划分的聚类方法。而在向量空间模型的文本表示模式下,基于划分的方法(partitioning method)是常用的聚类方法。基于划分的方法的基本思想是:给定一个有 n 个对象或元组的数据集,构建数据的 k 个划分,每个划分表示一个聚簇,产生的聚簇满足如下要求:

1)每个簇至少包括一个对象;

2)每个对象必须属于且只属于一个簇。

基于划分的方法首先规定簇数目 k,创建一个初始的划分后,采用一种迭代的重定位技术,尝试通过对象在划分间移动来改进划分。为了达到全局最优,基于划分的聚类会要求穷举所有可能的划分。而k-平均算法和k-中心点算法是在实际应用中经常采用的两个经典的基于划分的启发式方法。

1.1 K-means聚类方法及其特点

K-means算法最早是由MacQueen在1967年提出[7]的。该算法及其各种改进方法是数据挖掘及机器学习领域的一类

到稿日期:2010-02-22 返修日期:2010-04-28 本文受国家自然科学基金项目(编号:70571087)资助。
刘海峰(1962-),男,博士,副教授,CCF会员,主要研究方向为数据挖掘、文本挖掘,E-mail:liuhaifeng19620717@sina.com;刘守生(1965-),男,博士,副教授,主要研究方向为人工智能、遗传算法;张学仁(1955-),男,副教授,主要研究方向为人工智能、信息检索。

重要的数据处理方法。作为一种基于质心的聚类方法，该算法以其原理简单、收敛速度快以及适应性强而得到广泛的应用。K-means 分类算法描述如下[8]：

1)任意选择 K 个对象作为初始的簇中心；

2)repeat；

3)根据簇中对象的平均值，将每个对象(重新)赋给相似性最大的簇；

4)更新簇的平均值，即计算每个簇中对象的平均值；

5)until 不再发生变化。

该算法是一种基于质心的方法，它试图找出使平方误差函数值最小的 k 个划分。其优点是具有较好的可伸缩性和很快的收敛速度，复杂度为 $O(ntk)$。其中 n 是所有对象的数目，k 是簇的数目，t 是迭代的次数。该算法经常以局部最优结束，适合处理大数据集。当结果簇密集并且各簇之间的区别明显时，特别是当数据呈现球形分布时，采用 K-means 算法的效果较好[8]。

1.2 K-means 聚类算法主要问题及解决方法

K-means 算法是基于目标函数的算法。该算法不适合于发现非凸面形状的簇，或者大小差别很大的簇。主要问题有如下两个方面：一是该算法的初始簇种子的选择是随机的，难以代表真正的簇的中心甚至与簇中心相邻甚远，从而影响着最终的聚类效果；二是该算法对于数据集里孤立点非常敏感。较少的几个孤立对象却能对聚类结果产生较大的影响。这一缺陷是由该算法的簇中心计算方式决定的：因为在向量模型下进行 K-means 聚类计算时，进行下一次迭代前，将簇中各向量的几何中心作为新的簇中心，因而即使远离数据密集区域的数据个数较少，这种方式计算出的新的簇中心必然偏离真正的中心位置。

在利用 K-means 进行文本特征选择方面，主要考虑从下面两个方面入手对 K-means 算法进行改进：根据文本诸特征项对文本表示能力的大小来确定 K-means 算法的初始簇中心的选择标准，以避免簇中心选择上的随机性；根据文本特征之间的相似度确定删除噪声数据的依据，以解决噪声特征对簇中心定位的干扰。

2 一种改进的 K-means 聚类方法

2.1 对 K-means 聚类方法簇中心初始值的优化

K-means 算法首先要选择任意 K 个对象作为初始的簇中心，然后以此为基础进行迭代。选取的初始值不同，相应的聚类结果就可能不同。初始值的选取对 K-means 算法的聚类效果影响较大：初始值的选择是随机的，从理论上说这 k 个初始点最理想情况是应该选取 k 簇的中心点，这种情况下聚类的效果达到最佳；但是在较差的情况下也有可能将一簇中的 k 个元素选为 k 个簇的初始点，此时的聚类效果就较差。因此应该制定一个标准对初始值的选择进行控制，使其尽可能地接近各个簇中心。

设训练集的特征个数为 m，记 $w_i=(w_{i1},w_{i2},\cdots,w_{in})$ 为特征项 t_i 在训练集内的权重向量，以常用的 $tf\text{-}idf$ 因子法进行赋权，n 为训练集文本数。再记：

$$v_i=\max|w_{is}-w_{it}|;s,t=1,2,\cdots,n,s\neq t;i=1,2,\cdots,m;$$

相对于非类属特征词来说，由于类属特征词在其相应类别文本内出现的频数高而在其它类别文本内出现的频数较少，因此通常情况下其相应的 v_i 值大于非类属特征词相应的 v_i 值。所以，v_i 值较大的特征项对文本类属的标引能力更强。以 v_i 值的大小为标准将 m 个特征项预分为 k 组，方法如下：

记 $p=\frac{1}{k}[\max\{v_i\}-\min\{v_j\}],i\neq j,i,j=1,2,\cdots,m$，其中 k 为聚类的簇数。

则以 p 为单位长度，将区间 $d=[\min\{v_i\},\max\{v_i\}]$ 划分为 k 个小区间：

$$l_t=[\min\{v_i\}+(t-1)p,\min\{v_i\}+tp],t=1,2,\cdots,k$$

若特征项 t_i 的 v_i 值满足 $v_i\in l_r$，则将 t_i 划入相应的特征子集 $S_r,r=1,2,\cdots,k$ 内。

再记 $\bar{v}_r=\frac{1}{|S_r|}\sum_{t_i\in S_r}v_i,r=1,2,\cdots,k$；此时若

$$|v_s-\bar{v}_r|=\min\{|v_i-\bar{v}_r|,v_i\in l_r\} \tag{1}$$

则将特征项 t_s 作为第 s 簇的初始值。

这种簇初始值的选择方式具有以下优点：

1)初始值的选取基本上位于或接近位于各个聚类簇的几何中心，这更符合聚类要求；

2)由于本文研究的是文本特征选择，因此这种以 v_i 值的大小为标准的簇初始值选择模式通过聚类得到的位于各个簇的“核心区域”内的那些特征项更具有文本的类别标引能力。

2.2 对 K-means 聚类过程中孤立点的处理方法

如前所述，K-means 算法对孤立点过于敏感。一些远离数据分布主要区域的孤立点即使个数很少，K-means 聚类算法的类中心计算模式使得每次迭代前所计算出的新的簇中心常常偏离真正的中心位置，这样对聚类效果的影响很大。

虽然 K-means 聚类算法要求最终的聚类结果应满足每个数据必须属于且只属于一个簇，但是当该算法用于文本特征选择时，却不必满足这一基于划分的聚类算法的基本要求：文本特征聚类的目的是在原始文本特征集合里选择出少量的具有代表性的特征子集用于文本表示，而大量的特征项最终要被剔除。因此在特征聚类过程中删除一些对文本标引用处不大的特征项，特别是那些对聚类产生噪音作用的特征项，不仅能够提高聚类效率，还能提高特征选择的效率。

解决噪声数据的基本思想是：在计算类簇 C_i 的簇中心 X_i 时避开这些孤立点的影响，即：在进行第 k 轮聚类种子的计算时，将簇中那些与第 $k-1$ 轮聚类种子相似度明显小的数据剔除，使用剩余向量集合里的元素的均值点作为第 k 轮聚类的新种子。

具体地说，就是对于第 i 簇的第 $k-1$ 轮聚类获得的类簇 $C_{i,k-1}$，首先使用常用的向量夹角余弦计算簇内数据与簇中心向量 $X_{i,k-1}$ 的相似度：

$$\delta_{i,k-1,j}=\cos(X_{i,k-1},y_j)=\frac{(X_{i,k-1},y_j)}{|X_{i,k-1}||y_j|} \tag{2}$$

并给定阈值：

$$\varepsilon=\lambda\bar{\delta}_{i,k-1,j}=\lambda\frac{1}{|C_{i,k-1}|}\sum_{y_j\in C_{i,k-1}}\delta_{i,k-1,j},0<\lambda<1 \tag{3}$$

式中，$j=1,2,\cdots,|C_{i,k-1}|$，$X_{i,k-1}$ 为 $C_{i,k-1}$ 的几何中心，y_j 为 $C_{i,k-1}$ 的任意向量，λ 为较小的参数，一般可以限制在(0,0.2)之间，本文试验中取 $\lambda=0.1$。

最后将簇 $C_{i,k-1}$ 中与 X_{i-1} 相似度小于 ε 的 $p_{i,k-1}$ 个向量排除，取剩余的向量计算得第 k 轮聚类簇 $C_{i,k}$ 的中心 $X_{i,k}$ 的坐标；将这个过程迭代进行，直到每个元素的类属情况不再发生变化为止。实验表明该方法对孤立点的处理较 K-means 算

法的效果好。

2.3 一种改进的 K-means 聚类算法

改进后的 K-means 聚类算法描述如下：

算法 划分的 K-means 算法基于簇中对象的平均值

输入：簇的数目 k 和包含 m 个对象的数据集；

输出：k 个簇，使平方误差准则最小。

方法：

1)基于式(1)选择 K 个对象作为初始的簇中心；

2)repeat；

3)根据簇中对象的平均值，将每个对象(重新)赋给最类似的簇；

4)计算式(2)、式(3)(本文试验中取 $\lambda=0.1$)；

5)判断：对于 $C_{i,k-1}$ 内数据 y_j，若 $\delta_{i-1,j}<\varepsilon$，则在 $C_{i,k-1}$ 中删除 y_j，$j=1,2,\cdots,|C_{i,k-1}|$，否则 y_j 在 $C_{i,k-1}$ 中保留；

6) 更新簇的平均值，即计算每个簇中对象的平均值；

7) until 不再发生变化。

2.4 聚类模式下一种改进的 K-means 文本特征选择

下面提出一种基于聚类模式下的改进 K-means 文本特征选择模型：

1)使用文本分类中常用的 $tf\text{-}idf$ 因子[9]：

$$w_{ij}=\frac{tf_{ij}\times\log(\frac{n}{n_i}+0.01)}{\sum_{i=1}^{m}\left[tf_{ij}\times\log(\frac{n}{n_i}+0.01)\right]^2} \tag{4}$$

对第 i 个特征项 t_i 进行赋权，构造特征项 t_i 的赋权向量：$t_i=(w_{i1},w_{i2},\cdots,w_{in})$，$i=1,2,\cdots,m$，其中 n 表示文本数，n_i 为文本集里含有特征 t_i 的文本数，$tf_{ij}=\frac{df_{ij}}{\max(df_{ij})}$，$df_{ij}$ 表示文本 d_j 中特征项 t_i 的频数，$\max(df_{ij})$ 表示文本 d_j 中出现频率最高的词的词频；

2)对文本集进行切词，统计得到特征集 S，对 S 使用 2.3 节方法进行聚类，得到相应的类簇 $C_1,C_2,\cdots,C_k$；

3)计算类簇 $C_1,C_2,\cdots,C_k$ 的类别中心向量(簇内向量的算术平均向量)$X_1,X_2,\cdots,X_k$；

4)计算簇内各特征项与类别中心向量的相似度，分别取其最大前 f 个特征值对应的特征，将这 $f\times k$ 个特征项构成特征集 S 的子集用于文本表示。

3 文本分类试验结果及其分析

本文对上述方法的文本分类效果进行了试验。试验数据为从新浪、百度下载的 1300 篇文本，其中分为经济(210 篇)、生活(180 篇)、体育(230 篇)、卫生(220 篇)、工业(120 篇)、文学(235 篇)以及农业(105 篇)共 7 类。试验时采用 4 分交叉试验法，将 1300 篇文本平均分为 4 份，3 份为训练集，1 份为测试集；每份轮流作为测试集循环测试共 4 次，取平均值为测试结果。对 Web 页面进行清洗获得纯文本文档，使用东北大学自然语言处理实验室提供的分词软件进行分词，使用禁用词表过滤停用词，人工删除冷僻的低频词并剔除虚词、助词、人称代词、特高频词等预处理后建立候选特征项集合得到特征个数 4276 个，分类效果评估函数使用常用的查准率、查全率和 F_1 测试值：

查准率＝分类的正确文本数/实际分类文本数；查全率＝分类的正确文本数/应有文本数；

$$F_1=\frac{\text{查准率}\times\text{查全率}\times 2}{\text{查准率}+\text{查全率}}$$

使用 2.3 节改进的 K-means 方法聚类后(取 $k=7$)得到 2846 个特征项，取 $f=25$ 对特征项簇进行筛选，以 175 个特征进行文本标注，分类器使用常用的 KNN 分类器[10]，并与常见的 k-means 聚类方法(类别数与簇内特征选取数均与改进的方法相同)在文本分类效果上进行了比较(相应的数据以下标 old 记)，实验结果统计如表 1 所列。

表 1 文本分类试验结果统计

	经济	生活	体育	卫生	工业	文学	农业	均值
查准率	0.8762	0.9126	0.8913	0.9047	0.8468	0.8845	0.7762	0.8703
查全率	0.8675	0.8847	0.8752	0.8791	0.8572	0.8681	0.7913	0.8604
F1 值	0.8718	0.8989	0.8832	0.8917	0.8520	0.8762	0.7836	0.8653
查准率 old	0.8218	0.8429	0.8126	0.8279	0.8531	0.8176	0.7413	0.8167
查全率 old	0.8157	0.8071	0.8247	0.8416	0.8613	0.8231	0.7382	0.8160
F1 值 old	0.8187	0.8246	0.8136	0.8347	0.8572	0.8203	0.7397	0.8155

从分类精度可以看出，本文提出的改进方法与经典的 k-means 方法在分类效果上相比较还是令人满意的。其中经济、生活、体育类别的文本分类效果改进幅度较大，而工业与农业两类效果略微差一些，这可能与这两类的文本数略少一些有关。总的 F1 值提高了 6.1%，效果还是明显的。

结束语 文本特征降维问题是文本信息处理所必须面对的主要问题之一，高维特征是制约文本分类效率的瓶颈。本文讨论了一种 K-means 聚类的改进模型在文本分类中的应用。总的说来，目前基于聚类的方法在文本特征降维应用上主要是用于特征选择。而特征降维的另外一种主要方式特征抽取与特征聚类的结合研究不足。特征选择和特征抽取从不同的立足点出发对文本特征进行压缩，两种模式各有其优点和不足。在聚类模式下如何将两种模型进行合理的结合，构造组合型的特征降维模型应该是提高特征降维效率的思路，这也是我们下一步要进行研究的方向。

参 考 文 献

[1] Makrehchi M, Kamel M S. Text classification using small number of features[C]// Perner P, Imiya A, eds. Proc. of the 4th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition: (MLDM 2005). 2005: 580-589

[2] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

[3] 刘海峰，姚泽清，刘守生，等. 文本分类中基于核的非线性判别[J]. 应用科学学报，2008，26(6)：627-631

[4] 刘海峰，王元元，姚泽清，等. 文本分类中一种混合型特征降维方法[J]. 计算机工程，2009，35(2)：194-196

[5] 刘海峰，王元元，张学仁，等. 文本分类中一种基于正交变换的特征降维方法[J]. 计算机科学，2008，35(5)：125-126

[6] 东波，白硕，李国杰. 文本聚类中权重计算的对偶性策略[J]. 软件学报，2002，13(11)：2083-2089

[7] MacQueen J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297

[8] Han Jiawei. Micheline Kamber. Data Mining Concepts and Techniques[M]. 范明，孟小峰，译. 北京：机械工业出版社，2001

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

[10] 赵万磊，王永吉，张学杰. 一种优化初始中心点的 K 平均文本聚类算法[J]. 计算机应用，2005，25(9)：2037-2040