文章编号：1003-0077（2011）00-0000-00

高斯加权的重构性 K-NN 算法研究 *

刘作国，陈笑蓉

（贵州大学计算机科学与技术学院，贵州省 贵阳市 邮编：550025）

摘要：本文提出基于高斯加权距离以及聚类重构机制的 K-NN 文本聚类算法。文章提出 K-NN 邻近域的概念，通过高斯加权的邻近域算法实施 K-NN 聚类。利用高斯函数根据样本与聚类中心的距离为样本赋权，计算聚类距离。基于邻近域权重和聚类密度对形成的聚类实施重构，实现聚类数目的自适应调整。使用拆分算子拆分稀疏聚类并调整异常样本；使用合并算子合并相似聚类。实验显示聚类重构机制能够有效地提高聚类的准确率及召回率，增加聚类密度，使得形成的聚类结果更加合理。

关键词：文本聚类；K-NN 算法；高斯加权；邻近域规则；聚类重构

中图分类号：TP391.1　　**文献标识码：**A

Research on Gauss Weighed Reorganization K-NN

LIU Zuoguo, CHEN Xiaorong

(College of Computer Science & Technology, Guiyang,Guizhou, 550025, China)

Abstract: This paper illustrates a K-NN text clustering algorithm which uses Gauss Weighed Distance and Cluster Reorganization Mechanism. The concept of Nearest Domain is proposed and Nearest Domain Rules are elaborated. Then Gauss Weighing Algorithm is designed to Quantification samples' distance and weights. A text is weighed based on the distance from cluster kernel via Gauss function in order that distances of clusters can be calculated. What's more, Cluster Reorganization Mechanism will make a contribution to adjust amount of clusters self-adaptively. Splitting operator separates sparse clusters and adjusts abnormal texts while consolidating operator combines similar ones. Clustering experiment shows that reorganization process effectively improves the accuracy and recall rate and makes result more reasonable by increasing the inner density of clusters.

Key words: text clustering; K-NN; Gauss weighing; nearest domain rule; cluster reorganization

1 引言

K-NN 聚类算法简洁实用，是一类常见的文本聚类算法。K-NN 算法选定样本子集形成初始聚类分布，根据初始分布将测试样本划分入最近聚类。K-NN 算法初始聚类的选择直接影响聚类结果，聚类过程缺少对结果的检测和调整机制，难以实现聚类数目的自适应变更[1]。本文主要针对 K-NN 算法的距离判定策略和聚类重构机制进行了研究，通过高斯加权算法实施距离度量，判定样本归属。采用聚类重构机制对不合理聚类实施拆分及合并，实现聚类数目的自适应调整，同时保证形成的聚类更加紧密合理。

2 相关工作

2.1 文本表示

本文主要采用向量空间模型 VSM 进行文本描述，文本 t 表示为：

$$t = t(v_1, v_2, \cdots, v_n) \tag{1}$$

采用欧式距离描述文本 a,b 之间的关系：

* **收稿日期：2015-07-31**　　**定稿日期：2015-08-07**

基金项目：国家自然科学基金（61363028）

$$D(a,b)=\sqrt{\sum_{i=1}^{n}(a_{vi}-b_{vi})^2} \tag{2}$$

2.2 聚类密度

本文使用几何中心来定义聚类的中心，并通过聚类密度描述聚类内样本的相关性。

定义 1（聚类中心）：聚类C的中心定义为其几何中心：

$$K(C)=\frac{1}{|C|}\sum_{\forall t\in C}t \tag{3}$$

(3)式中，聚类中心$K(C)$的各维分量分别是各样本对应分量的向量均值。

定义 2（聚类密度）：聚类C的密度定义为：

$$Den(C)=\frac{1}{|C|}\sum_{\forall t\in C}\exp[-D(t,K(C))] \tag{4}$$

聚类内部样本与聚类中心越接近，聚类密度就越高，重构的必要性就越小。优先选择密度较低的聚类实施重构可以提高聚类效率。

2.3 簇间距离

文献[2]认为样本空间分布具有正态分布的性质，靠近聚类中心的样本权重较高，对聚类间距的影响较大。本文参考其中思想，设计一种高斯加权算法来计算聚类间距，使算法向聚类中心的高密度区域靠近。

定义 3（簇间距离）：样本x相对于聚类C的高斯权重为：

$$W(x,C)=\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{5}$$

样本x与聚类C的距离为：

$$D(x,C)=\frac{1}{|C|}\sum_{\forall t\in C}W(t,C)D(x,t) \tag{6}$$

聚类C_i与C_j的距离为：

$$D(C_i,C_j)=\frac{1}{|C_i|\times|C_j|}\sum_{\forall t_i\in C_i}W(t_i,C_i)\sum_{\forall t_j\in C_j}W(t_j,C_j)D(t_i,t_j) \tag{7}$$

为便于计算，(5)式取$\mu=K(C),\sigma=1$。

3 加权 K-NN 聚类算法

K-NN 聚类是一类常用的文本处理算法。算法的思想是：如果一个样本 *S* 的 *K* 个最邻近样本更靠近聚类 *C*，就将 *S* 划分入聚类 *C*。

K-NN 聚类思想建立在理想假设下，要求初始状态的聚类划分是合理，并且已经形成的每个聚类内部联系紧密。但实际情况往往并非如此，即便初始聚类划分是理想的，聚类过程中大量测试样本的加入也将使所获得的聚类偏离理想状态 [3][4]。

3.1 邻近域

假设 K-NN 聚类算法取 *K=3*，图 1 中待测样本 *x*（以圆形表示）被划分入聚类 *b*（以三角形表示）。但实际上聚类 *b* 密度较低， *b* 类样本之间距离较大，两个邻近样本距离聚类 *b* 中心较远，因此样本 *x* 距离聚类 *a*（以方形表示）比较近。出现以上问题的原因在于 K-NN 聚类没有考虑两个聚类中的样本对待测样本 *x* 的影响力，换而言之没有考虑两个聚类中各个样本的权重。

图 1 邻近域权重的意义

文献[5]提出一种文本的权重量化思想，指出文本分布越密集的样本空间区域对聚类划分的影响越高。文献[6][7]提出距离聚类中心越近的样本对聚类的表征能力越强，因此权重也越高。

借鉴经典的 K-NN 聚类思想，本文提出加权邻近域的概念来处理待测样本的划分问题，并且认为越靠近聚类中心的样本对聚类的影响力越高，样本权重也越大 [8]。

定义 4（邻近域）：与样本x距离小于d的全部样本构成x的d邻近域，记为：

$$Domain(x,d) \tag{8}$$

其中d称为邻近域的半径。

定义 5（邻近域权重）：样本x的d邻近域为$Domain(x,d)$，聚类C_i与$Domain(x,d)$交集为$S_i = Domain(x,d) \cap C_i$，则：

$$\omega(C_i) = \sum_{\forall t \in S_i} W(t,C_i)\exp[-D(t,K(C_i))] \tag{9}$$

称为x在C_i上的d邻近域权重。

通常聚类中心附近的样本密度较大，由聚类中心向外密度逐渐降低，对于样本x，取适当的半径d求解$Domain(x,d)$，x应属于邻近域权重$\omega(C_i)$最高的聚类。

3.2 邻近域规则

邻近域规则：样本x的d邻近域为$Domain(x,d)$，邻近域上的最大权重为$\omega(C_k)$，样本x只有两种可能的聚类归属：①属于聚类C_k；②独立形成聚类。若$D(x,K(C_k)$距离不大于任意两个聚类的最大距离则将x划分入C_k，否则x独立形成聚类。

选取邻近域半径 d 内的 K 个（d 为确定值，K 为不确定数目）邻近对象进行聚类判定。采用邻近域规则判定待测样本 x 的类别划分：样本划分入 K 个邻近对象最接近的聚类。其中 d 为样本 S 到最近的聚类的距离。

3.3 聚类重构

为解决初始聚类对聚类结果的影响，采取聚类重构策略对获得的聚类实施重构。

聚类重构机制根据聚类的密度及各样本的距离拆分稀疏的聚类，合并相近聚类从而实现聚类的数目及空间分布的自适应调整。重构机制需要考虑以下情形：

1）异常样本调整。若聚类内少数样本与簇内其他样本联系较弱，应当将这些“另类”样本调整到其他聚类中；

2）稀疏聚类拆分。若聚类密度过低，说明簇内样本分布稀疏，应当将稀疏聚类拆分为多个密集聚类；

3）相似聚类合并。若多个聚类联系紧密，考虑将它们合并为一个聚类，合并后可能需要考虑 1）、2）类问题。

1）类问题采用邻近域算法处理；2）、3）两类问题分别采用拆分算子和合并算子进行处理。聚类过于稀疏不利于判断聚类间距，会影响聚类合并，因此聚类重构应当先拆分后合并，并优先处理密度低的聚类 [9]。本文参照文献[10]阐述的聚类改进策略，设置密度阈值来限定拆分算子的作用范围，聚类拆分的算法如下：

聚类拆分算法：

Step1：在密度低于阈值的聚类中选择密度最低的聚类C_i；

Step2：获取簇内任意未处理成员t；

Step3：寻找t最近聚类C_j；

Step4：若$C_i = C_j$转 Step6，否则继续：

①若$\exp[-D(t, C_j)] \geq Den(C_j)$，$t$归入$C_j$；

②若$\exp[-D(t, C_j)] < Den(C_j)$，新建聚类容纳$t$；

Step5：更新聚类中心及聚类密度；

Step6：迭代处理聚类C_i内所有样本。

算法 Step4 中，样本t最近聚类为C_j，若$\exp[-D(t, C_j)] \geq Den(C_j)$，说明$t$比$C_j$中大多数样本都更接近聚类中心，允许将$t$归入$C_j$；反之说明$t$距离$C_j$中心较远，进而断定没有与$t$相近的聚类，需要新建聚类来容纳样本$t$。

设样本规模为n，理论上拆分算子完成所有计算的平均复杂度为$O(n^2)$，由于聚类中心、聚类密度、高斯权重等复杂计算在聚类过程中已经完成，拆分算子实际时间开销为$O(n \times \log n)$。

聚类合并的算法如下：

聚类合并算法：

Step1：整个聚类集添加到未处理聚类集合C_u；

Step2：获取任意未处理聚类C_i；

Step3：寻找C_i最近聚类C_j；

Step4：分析C_i与C_j关系：

若$\exp[-D(C_i, C_j)] \geq Den(C_i)$或$\exp[-D(C_i, C_j)] \geq Den(C_j)$，合并聚类$C_i$与$C_j$，更新聚类中心及密度，将新聚类添加到$C_u$。否则不予以合并；

Step5：C_u中删除已处理聚类；

Step6：迭代处理C_u中所有聚类。

算法 Step4 中，若$\exp[-D(C_i, C_j)]$大于等于C_i或C_j任意一个的聚类密度，说明两个聚类存在较大交集，二者具有包含或较大的重叠关系，考虑将两个聚类合并。合并产生的新聚类仍作为未处理聚类参与迭代过程。

理论上合并算子复杂度为$O(n^2)$，实际为$O(n \times \log n)$。

重构机制示例：假设样本空间共包括 3 类 16 个文本，用三种图形各代表一类文本。初始状态文本集被分为 4 类，星形表示各聚类几何中心，箭头指向文本的最近聚类。理想状态重构过程如图 2：

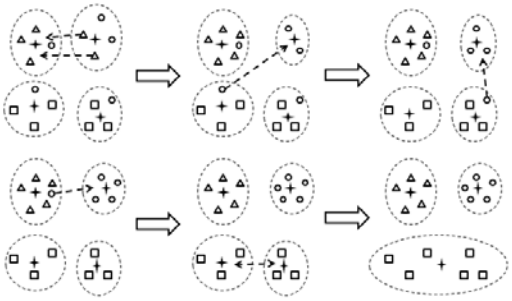

图 2 聚类重构示例

经过聚类重构，稀疏聚类得到优化，初始状态对聚类结果的影响也被削弱。聚类的拆分及合并使得聚类数目动态调整，无需用户干预，更符合聚类处理的实际应用需求。

4 实验分析

本文从复旦大学中文语料库分别随机选取 500 和 1000 个样本进行聚类实验。采用 K-NN 算法和加权重构 K-NN 模型分别进行聚类。统计各类别的准确率、召回率、*F-Score* 值并计算获得的聚类密度：

表 1 - 500 样本 K-NN 聚类及加权重构 K-NN 结果对比

分类	文本数	K-NN				加权重构 K-NN			
		准确率	召回率	F-Score	聚类密度	准确率	召回率	F-Score	聚类密度
艺术	39	77.1%	69.2%	73.0%	0.29	94.6%	89.7%	92.1%	0.35
哲学	46	84.4%	82.6%	83.5%	0.30	97.8%	95.7%	96.7%	0.36
经济	87	90.8%	79.3%	84.7%	0.31	96.6%	96.6%	96.6%	0.36
政治	64	88.4%	95.3%	91.7%	0.33	93.7%	92.2%	92.9%	0.34
军事	50	87.0%	94.0%	90.4%	0.32	94.1%	96.0%	95.0%	0.35
农业	46	78.8%	89.1%	83.7%	0.30	93.2%	89.1%	91.1%	0.33
通信	52	91.5%	82.7%	86.9%	0.31	90.9%	96.2%	93.5%	0.34
运输	39	86.1%	79.5%	82.7%	0.30	94.7%	92.3%	93.5%	0.33
法律	30	82.4%	93.3%	87.5%	0.31	90.6%	96.7%	93.5%	0.35
医药	47	80.8%	89.4%	84.8%	0.31	93.8%	95.7%	94.7%	0.32
均值		84.7%	85.4%	84.9%	0.31	94.0%	94.0%	94.0%	0.34

表 2 - 1000 样本 K-NN 聚类及加权重构结果 K-NN 对比

分类	应有文本	K-NN				加权重构 K-NN			
		准确率	召回率	F-Score	聚类密度	准确率	召回率	F-Score	聚类密度
历史	93	89.4%	90.3%	89.8%	0.31	96.8%	96.8%	96.8%	0.35
矿业	123	88.2%	91.1%	89.6%	0.29	96.0%	97.6%	96.8%	0.34
教育	69	91.0%	88.4%	89.7%	0.30	92.9%	94.2%	93.5%	0.33
运输	112	95.2%	88.4%	91.7%	0.33	98.2%	96.4%	97.3%	0.36
环境	85	89.8%	92.9%	91.3%	0.32	97.6%	97.6%	97.6%	0.37
通信	141	94.8%	90.1%	92.4%	0.34	98.6%	97.2%	97.9%	0.37
环境	97	89.1%	92.8%	90.9%	0.32	96.9%	95.9%	96.4%	0.35
政治	106	88.3%	92.5%	90.3%	0.31	95.4%	97.2%	96.3%	0.33
经济	78	97.3%	91.0%	94.0%	0.36	97.5%	98.7%	98.1%	0.37
体育	96	87.1%	91.7%	89.3%	0.30	95.8%	94.8%	95.3%	0.33
均值		91.0%	90.9%	90.9%	0.32	96.4%	96.5%	96.5%	0.35

实验结果显示邻近域算法和聚类重构机制对文本聚类的处理是有效的。经过重构处理后各类文本准确率、召回率均有显著提升，聚类密度有所提高，说明重构之后聚类内部样本关联性更强。

从表 1 及表 2 可见，艺术类准确率、召回率及聚类密度较低，这是由于语料库对文本的人工标注不够细致。语料库艺术类包括音乐、书画、舞蹈、美学等多个领域的文章，虽然这些领域都属于“艺术”范畴，但文本的词汇特征相差甚远。通过聚类重构，“艺术”类被划分为 4 个子类，如表 3 所示，每个子类密度仍然是可接受的：

表 3 “艺术”子类

分类	聚类中心距		聚类密度
	最大值	最小值	
艺术类	6.25	2.63	0.28
聚类均值	3.87	2.61	0.34
艺术子类 1	3.85	2.86	0.33
艺术子类 2	3.23	2.63	0.34
艺术子类 3	3.13	2.70	0.36
艺术子类 4	3.03	2.78	0.34

表 1 与表 2 的对比结果显示，不同样本规模下准确率、召回率有一定差别，但重构后聚类密度却相差无几，这说明聚类算法对样本规模是敏感的，但重构机制不受到样本规模的影响。

5 总结

本文提出一种高斯加权的 K-NN 文本聚类算法。采用高斯函数对初始聚类中各个样本的影响力进行评估。文章引入聚类重构机制调整稀疏聚类，能够有效提高聚类密度并实现聚类数目的自适应调整。实验表明，重构机制不受到样本规模和初始划分的影响，能够有效地提高聚类精度，保证聚类的紧密性，其算法时间开销在可接受范围。

本文在邻近域的加权规则和距离度量方面还存在改进和优化的空间。更合理的邻近域加权规则可以使得 K-NN 聚类所获得的聚类更加合理，同时也有助于对稀疏聚类的判定，减小聚类重构的代价。

参考文献：

[1] Hyeong-Il Kim and Jae-Woo Chang. K-Nearest Neighbor Query Processing Algorithms for a Query Region in Road Networks[J]. Journal of Computer Science & Technology, 2013, 28(4): 585-596.

[2] 刘金岭,冯万利,张亚红.初始化簇类中心和重构标度函数的文本聚类[J].计算机应用研究,2011,28(11): 4115-4117.

[3] 王灿田,孙玉宝,刘青山.基于稀疏重构的超图谱聚类方法[J].计算机科学,2014,41(2): 145-148,156.

[4] 曾依灵,许洪波,吴高巍,等.一种基于空间映射及尺度变换的聚类框架[J].中文信息学报,2010,24(3): 81-88.

[5] Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, etal. A Study of Density-Grid based Clustering Algorithms on Data Streams[C] //Ding Yongsheng. FSKD 2011. Shanghai China. 2011: 1652-1656.

[6] 陈建超,胡桂武,杨志华,等.基于全局性确定聚类中心的文本聚类[J].计算机工程与应用,2011,47(10): 147-150.

[7] 季铎,王智超,蔡东风,等.基于全局性确定聚类中心的文本聚类[J].中文信息学报,2008,22(3): 50-55.

[8] 王骏,王士同,邓赵红. 特征加权距离与软子空间学习相结合的文本聚类新方法[J].计算机学报,2012,35(8): 1655-1665.

[9] M. Shahriar Hossain, Praveen Kumar Reddy Ojili, Cindy Grimm, etal. Scatter/Gather Clustering: Flexibly Incorporating User Feedback to Steer Clustering Results[J]. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2012, 18(12): 2829-2838.

[10] Nisha.M.N, Mohanavalli.S, Swathika.R. Improving the quality of Clustering using Cluster Ensembles[C]// 2013 IEEE Conference on Information and Communication Technologies. 2013: 88-92.

作者简介：

刘作国（1987——），男，博士研究生，主要研究领域为中文信息处理。Email:412769371@qq.com。

陈笑蓉（1954——），女，教授，主要研究领域为中文信息处理。Email:xrchengz@163.com。