

一种基于模糊加权的改进文本聚类方法

刘海峰, 姚泽清, 刘守生

(解放军理工大学 理学院, 江苏 南京 210007)

摘 要: 首先提出了一种优化初始中心点方法用以解决聚类的局部最优问题, 同时通过样本的模糊加权减少边缘噪音数据对聚类效率的影响. 文本聚类试验表明, 该模糊文本聚类算法取得较好的聚类效果.

关键词: k-means 算法; 模糊聚类; 文本聚类; 模糊加权

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)09-0039-04

An Improved Text Clustering Method Based on Fuzzy Weighting

LIU Hai-feng, YAO Ze-qing, LIU Shou-sheng

(Institute of Sciences, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: This paper proposes a new way that selects initial cluster center in order to solve the partially most superior phenomenon. By using the fuzzy weighting on the samples, this improved method decreases the influence that the k-means algorithm is very sensitive to the isolated point. Lastly, we have a test about text clustering and the result shows that this method obtains good clustering effect.

Key words: k-means algorithm; fuzzy clustering; text clustering; fuzzy weighting

1 引言

聚类是依据某种相似性将数据对象划分为多个簇,使得簇内高内聚,簇间低耦合,从而发现数据库中复杂数据之间的相似性,常常用于对大规模数据库的优化查询. 而文本聚类技术是文本信息处理的主要研究内容之一,文本聚类是将文本集里内容相似的文本聚为一簇,进而揭示隐藏在簇内各文本之间相同的类别属性. 作为一种无监督机器学习模式,聚类技术已经成为对文本信息进行有效的组织和管理的的重要手段^[1]. 目前存在很多聚类算法,总的说来聚类技术可分为:基于划分的方法、基于层次的方法、基于密度的方法以及基于网格的方法等. 其中基于划分的聚类技术在文本信息处理中成为一种主流技术^[2]. 而 k-means 聚类算法以其较快的聚类速度、较高的聚类效率以及良好的可伸缩性成为基于划分聚类算法的代表. 对其存在的不足进行改进,使

之更能适应日益增长的文本信息处理的技术需求.

2 基于划分的聚类方法特点及问题

在文本的向量表示前提下,基于划分的聚类方法(partitioning method)首先规定类簇数目 k , 在创建一个初始的划分后,采用一种迭代的重新定位技术,尝试通过对象在划分间移动来改进划分. 一个好的划分的准则是:在同一个簇中的对象之间尽可能“接近”或“相关”,而不同的簇的对象之间尽可能“远离”或“不同”,要求相应的准则函数收敛或满足一定的终止条件. 其中 k-平均算法(c-均值算法)和 k-中心点算法是经常采用两个经典的基于划分的启发式方法.

2.1 k-means 聚类算法

作为一种基于质心的聚类方法, k-means 聚类算法^[3]的主要步骤:

(1) 随机选择 k 个对象作为初始的类簇中心;

收稿日期: 2010-12-30; 修回日期: 2011-02-05

基金项目: 国家自然科学基金项目(71071161)

(2) 对于每个剩余对象, 计算其与各个簇中心之间距离, 将其分配到最近的簇中. 距离一般采用 Eudidean 距离或 Manhattan 距离进行计算;

(3) 更新各个簇的中心;

(4) 根据更新后的簇中心, 将每个对象重新赋给最近的簇;

(5) 如果每个样本的簇的归属不再发生变化或满足一定终止条件, 算法结束; 否则转入步骤(2).

k-means 聚类的准则函数通常使用平方误差准则函数;

$$J = \sum_{j=1}^k \sum_{x_i \in c_j} |x_i - \bar{c}_j| \quad (1)$$

其中, $\bar{c}_j = \frac{1}{|c_j|} \sum_{x_i \in c_j} x_i$ 表示类簇 c_i 内样本的几何中心.

k-means 算法试图找出使平方误差函数 J 值达到最小的 k 个划分. 算法复杂度是 $O(nk)$. 其中 n 是样本数, k 是簇数, t 是迭代的次数. 算法优点是具有较好的可伸缩性和较快的收敛速度, 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^[3].

但是, k-means 算法的缺点也是明显的.

(1) 该算法需要计算每个训练样本与簇中心的距离或相似度, 随着训练样本数量的增大, 计算开销也随之增大. 当训练样本数量过大时算法的性能随之降低.

(2) 由于 k-means 算法本质上是一种局部收敛的迭代寻优技术, 因而初始簇中心的选择对聚类结果影响很大: 不同的初始簇中心得到的聚类结果差异往往很大. 合理的初始簇中心选取能够改善算法、得到较为稳定的聚类效果.

(3) 区域边缘样本对聚类效果影响很大. 如何降低边缘样本对算法性能的影响, 成为改善算法效率的另一个因素.

2.2 基于样本分布的训练样本子集的选取

2.2.1 基于重复取样的 k-means 聚类算法性能的优化

上节所述 k-means 算法所存在的前两个问题可以通过对样本集的多次取样得以改善^[3]. 文献[3]提出了对训练集进行 m 取样, 每次取样后都进行 k-means 聚类, 共产生 $k \times m$ 个簇中心, 然后对这些候选簇中心进行基于层次的凝聚方法进行聚类以产生最后的 k 个中心作为聚类算法的初始簇中心; 文

献[4]的方法与文献[3]相似, 共进行了 m 次取样, 在 m 次 k-means 聚类产生的 m 组簇中心里, 选择使得误差准则函数最小的那一次聚类结果作为整个训练集的初始簇中心^[4]. 这种处理方式一方面减少了算法中训练样本的数量使得计算开销得以减少, 同时相对于随机选取 k 个样本作为聚类的初始簇中心的模式, 改进后的算法得到的 k 个初始簇中心优于经典的 k-means 算法.

但是上述方法仍然存在不足之处: 由于对样本的 m 次取样具有随机性, 每次得到的训练样本子集很难准确反映原始的训练样本集的分布状况, 从而对训练集的代表性较差. 可以考虑下述情况: 如果一次选取的样本基本上属于少数几个簇类甚至仅属于一个簇类, 那么通过聚类而得到的各个簇中心就不可能代表原始训练集的真正初始簇中心. 这种处理方式势必影响聚类的效率. 因此, 每次取样得到的样本子集能否与训练样本集里的样本分布尽可能的相似、并且所含样本数远远少于训练集的样本数就成为优化聚类算法性能的关键.

2.2.2 基于样本分布的训练样本子集构造以及初始簇中心选择

假设训练样本数为 N , 分 p 次取样. 文中的思路是基于训练样本相对于训练集的中心点之间的距离远近不同, 间隔地选取样本构造训练样本的子集用于每一次训练. 具体方法如下:

(1) 计算训练样本集的几何中心 $\bar{x}$:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (2)$$

其中, x_i 表示第 i 个训练文本向量.

(2) 计算 x_i 与 $\bar{x}$ 之间的距离:

$$d_i = |x_i - \bar{x}|, \quad i = 1, 2, \dots, N \quad (3)$$

根据 d_i 值大小按照从小到大的顺序对 x_i 进行排序, 依次记为 $x_1, x_2, \dots, x_n$.

(3) 记 $T = \lceil \frac{N}{p} \rceil$ 表示有理数取整, 将样本 x_j ,

$x_{j+p}, x_{j+2p}, \dots, x_{j+Tp}$ ($j = 1, 2, \dots, p$) 放入子集 S_j 中, 构成训练样本子集 S_j .

$$S_j = \{x \mid x = x_{j+tp}, t = 0, 1, \dots, T\} \quad (4)$$

对剩余的 $\alpha = N - T \times p < p$ 个样本(如果存在的话)依次放入训练集 $S_1, S_2, \dots, S_\alpha$ 里.

(4) 对每个训练子集 S_j 进行 k-means 聚类, 共产生 $k' \times p$ 个簇中心. 这里 k' 为 k-means 的类簇数.

(5) 对这 $k' \times p$ 个类簇中心再进行 k-means 聚

类,最后产生 k 个簇中心作为对整个训练集进行聚类的初始簇中心。

上述方式得到的 k 个类簇中心不再是训练集的 k 个样本点,但是一方面这种经过 p 次取样、从候选簇中心里选择出的 k 个簇中心的位置更接近训练集的实际簇中心,减少了初始簇中心选择的随意性;另一方面这种训练样本的选样方式使得训练子集里的样本分布更加均匀,能基本反映出整个训练集的分布状况,避免了选样的盲目性。

3 基于分布的训练样本模糊加权

3.1 文本聚类的特点及聚类模式的确定

在聚类分析理论中,根据样本是否以不同的隶属度分配到不同的类别,聚类模式相应的分为硬聚类以及软聚类。硬聚类的隶属度值只有 0 和 1,样本类别是唯一的。而现实生活里的数据常常没有严格的属性,尤其对于文本信息处理的数据而言,单一的类别归属与文本数据的内容常常不相符,软聚类的特性更符合现实世界对象与类的关系。这是因为文本聚类方面的数据有其鲜明的特点:随着交叉学科、边缘学科以及新兴学科的发展,现在一些文本特别是来自网络的文本涉及的内容越来越广泛,尤其是科技文本涉及的学科难以界定,文本属于几个类别的现象越来越普遍。这种文本可以归属于多个类别,或者更确切地说是以不同的隶属度属于不同的类别。而如果最终将其归属于某一个类别,是因为在该类别里其相应的隶属度最大。

3.2 模糊 k -means 聚类算法改进及特点

模糊 k -means 算法引入模糊集合理论,通过引入模糊权重指数改进收敛准则函数,采用 $[0, 1]$ 之间的权值刻画样本相对于类别的隶属程度,更能客观反映数据类属的模糊特性。在目前使用的模糊聚类算法中,由 Dunn 在 1973 年提出并经过 Bezdek 优化的模糊 C 均值(Fuzzy C -Means, FCM)算法最为常用。FCM 基本思想是使簇中的样本到簇中心的加权距离平方和达到最小,改变之处在于对样本引入相对于不同簇的不同隶属度,算法过程如下^[5]:

给定阈值 $\epsilon > 0$, 初始聚类中心 $C^{(0)}$

(1) 设定聚类簇数 k , 模糊系数 m , 初始化隶属度矩阵 $u^{(0)} = (u_{ij})_{KN}$ 初值, 令迭代次数 $\alpha = 0$;

(2) 计算模糊聚类中心 $C_i^{(\alpha)}, i = 1, 2, \dots, k$;

(3) 更新隶属度矩阵 $u^{(\alpha)} = (u_{ij}^{(\alpha)})_{KN}$

(4) 更新模糊聚类中心为 $C_i^{(\alpha+1)}$

(5) 更新聚类准则函数 $J^* = \sum_{j=1}^k \sum_{x_i \in c_j} |x_i - \bar{c}_j|$, 其中 $\bar{c}_j$ 为类簇 c_j 的几何中心;

(6) 如果 $|J - J^*| < \epsilon$, 算法终止; 否则转(3)。

这里隶属度矩阵 $u^{(\alpha)} = (u_{ij}^{(\alpha)})_{KN}$ 的元素 $u_{ij}^{(\alpha)}$ 表示在迭代过程中训练文本 d_j 相应于类簇 c_i 的隶属度, $j = 1, 2, \dots, N$ 。常用的一个计算公式为^[1]:

$$u_{ij}^{(\alpha)} = \frac{1}{\sum_{l=1}^k \left(\frac{d_{ij}^{(\alpha)}}{d_{lj}^{(\alpha)}} \right)^{\frac{2}{m-1}}} \quad (5)$$

其中, $d_{ij}^{(\alpha)}$ 表示迭代过程中文本 d_j 与簇 c_i 的簇中心之间的欧式距离。

考虑到比值 $d_{ij}^{(\alpha)} / d_{lj}^{(\alpha)}$ 可能大于 1 以及各个簇内样本数的差异,文中使用改进算式计算隶属度:

$$u_{ij}^{(\alpha)} = \frac{1}{\left[\frac{d_{ij}^{(\alpha)}}{\sum_{l=1}^k d_{lj}^{(\alpha)}} \right]^{\frac{2}{m-1}}} \bigg/ \sqrt{\left[\frac{1}{\left[\frac{d_{ij}^{(\alpha)}}{\sum_{l=1}^k d_{lj}^{(\alpha)}} \right]^{\frac{2}{m-1}}} \right]^2} \quad (6)$$

相比较于硬聚类的训练文本只属于一个类别的聚类算法,模糊加权方式对于每一个簇分配给样本一个不同的类别隶属度值,这与样本的自然属性更加贴近。样本与簇中心越近,分配到的隶属度越大;反之,与簇中心越远得到的隶属度值越小。这种处理方式不仅强调了距离簇中心较近的样本的作用,同时有效地降低了簇边缘样本特别是那些噪声数据对簇中心确定上的影响。虽然最终样本仍然归入一个固定的簇内,但是不同簇对样本的影响差异在聚类过程中得以体现:相比较于硬聚类的隶属度值只有 0 和 1 的模式,模糊加权方式优化了隶属度矩阵对训练样本在类别确定方面的表现能力,提高了聚类准则函数在迭代过程中的性能。

3.3 一种模糊加权的改进文本聚类算法

综上所述,基于对初始聚类中心选择的改进以及对训练样本的模糊加权,提出如下一种模糊加权的改进文本聚类方法如下:

首先,对文本集进行预处理,将文本表示为基于文本特征的向量形式,向量各分量使用文本分类中常用的 TF-IDF 公式进行计算^[6];

其次,通过对训练集进行划分、聚类产生初始的聚类簇中心:

(1) 使用式(2) 计算训练样本的几何中心;

(2) 使用式(3) 计算训练样本与样本集中心之

间距离并从小至大顺序对训练样本进行排序;

(3) 使用式(4)对训练样本集进行分割构造训练样本子集 $S_1, S_2, \dots, S_p$, 文中实验中取 $p = 10$;

(4) 对每个 S_j 进行 k -means 聚类, 共产生 $k' \times p$ 个簇中心, 文中实验取 $k' = 6$;

(5) 对这 $k' \times p$ 个簇中心再进行聚类, 最后得到 k 个簇中心 $\bar{c}_i (i = 1, 2, \dots, k)$ 作为聚类的初始簇中心, 其中 $\bar{c}_i$ 为类簇 c_i 的几何中心。

最后, 使用 3.2 节的 FCM 算法进行模糊加权聚类。但是在聚类的迭代过程中隶属度矩阵的更新使用式(6)进行计算, 实验中取模糊系数 $m = 2$ 。

4 聚类试验结果及其分析

文中对上述方法的聚类效果进行了试验。试验数据为从新浪、百度下载的 1850 篇文本, 其中分为经济(280 篇)、艺术(140 篇)、文学(210 篇)、工业(270 篇)、卫生(320 篇)、体育(310 篇)以及农业(320 篇)共 7 类。试验时采用 4 分交叉试验法, 将 1850 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。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; 剔除虚词、助词、人称代词、特高频词等预处理后建立特征项集合, 使用 TF-IDF 公式进行特征赋权, 取权重最大的前 175 个特征项构造文本向量。考虑到文本类别数和聚类簇数之间的关系, 文中取聚类簇数 $k = 7$, 这样聚类簇数与样本类别数相同, 一方面与实际问题的需求相符, 同时也便于聚类效果评估。

传统聚类分析的聚类效率很难度量。但是文本聚类由于训练样本的类别属性确定, 聚类效果的评估函数可以使用信息检索领域里常用的查准率、查全率和 F_1 测试值:

查准率 = 类内的正确文本数 / 实际类内文本数;

查全率 = 类内的正确文本数 / 类内应有文本数;

$F_1 = 2 \times \text{查准率} \times \text{查全率} / (\text{查准率} + \text{查全率})$

并与 2.1 节中常见的 k -means 聚类方法的聚类效果上进行了比较(取类簇数 $k = 7$, 相应的数据以下标 old 记), 实验结果统计如表 1 所示。

表 1 聚类算法实验结果统计

	经济	艺术	文学	工业	卫生	体育	农业	均值
查准率	0.864 7	0.812 5	0.805 7	0.897 2	0.827 6	0.882 3	0.901 4	0.855 9
查全率	0.881 3	0.797 4	0.826 3	0.858 9	0.875 9	0.849 1	0.897 5	0.855 2
F_1 值	0.872 9	0.804 9	0.815 9	0.877 6	0.851 1	0.865 4	0.899 4	0.855 3
查准率 old	0.812 6	0.781 6	0.791 3	0.832 7	0.801 7	0.832 5	0.869 1	0.817 4
查全率 old	0.821 3	0.776 9	0.762 6	0.818 6	0.825 3	0.807 2	0.830 7	0.802 2
F_1 值 old	0.816 9	0.779 2	0.776 7	0.825 6	0.813 3	0.819 7	0.849 5	0.811 6

从表 1 的聚类结果可以看出, 文中提出的改进方法较经典的 k -means 方法在聚类效果上还是令人满意的。其中经济、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。总的查准率提高约 0.45%, 查全率提高约 6.2%, F_1 值提高了约 5.1%, 改进的效果还是明显的。

参考文献:

- [1] 刘志勇, 耿新青. 基于模糊聚类的文本挖掘算法[J]. 计算机工程, 2009, 35(5): 44-49.
- [2] 焦远锋, 李万龙, 郑山红, 等. 一种新的两阶段 FCM 聚类算法[J]. 微电子学与计算机, 2009, 26(4): 40-42.
- [3] 龚静, 李安民. 一种改进的 k -means 中文文本聚类算法[J]. 湖南工业大学学报, 2008, 22(2): 52-54.

[4] 毛嘉莉. 基于 k -means 的文本聚类算法[J]. 计算机系统应用, 2009, 18(10): 85-87.

[5] 吉翔华, 陈超, 邵正荣, 等. 基于主题子空间的文本模糊 C 均值聚类算法[J]. 小型微型计算机系统, 2007, 28(12): 2225-2228.

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

作者简介:

刘海峰 男, (1962-), 博士, 副教授, 计算机学会会员。研究方向为模式识别、文本挖掘。

姚泽清 男, (1960-), 教授。研究方向为统计分析、数据挖掘。

刘守生 男, (1965-), 博士, 副教授。研究方向为神经网络。