

一种优化的改进 k _means 算法张淑清^{1,2}, 黄震坤^{1,2}, 冯 铭^{1,2}

(1 武汉大学 国家多媒体软件工程技术研究中心, 湖北 武汉 430072;

2 武汉大学 计算机学院, 湖北 武汉 430072)

摘 要: 传统的 k _means 算法随机地选择初始中心, 导致最终聚类结果陷入局部最优且准确率低. $\min_max$ 算法针对初始中心随机选择的缺点提出了改进. 但原始的 k _means 和 $\min_max$ 算法都忽略了利用原空间欧式距离度量相似性的不合理性. 对此提出改进算法, 利用映射函数将输入向量转换到特征空间, 在 $\min_max$ 算法基础上确定初始中心后, 根据特征空间中的欧式距离来进行分类. 实验证明了改进算法的有效性, 在 iris 和 wine 数据集上获得了 92.86% 和 72.34% 的分类准确率.

关键词: 随机地; 欧式距离; 不合理; 映射函数; 特征空间

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2015)12-0036-04

An Optimized k _means AlgorithmZHANG Shu-qing^{1,2}, HUANG Zhen-kun^{1,2}, FENG Ming^{1,2}

(1 National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China;

2 College of Computer, Wuhan University, Wuhan 430072, China)

Abstract: The traditional k _means randomly selects initial cluster centers and divides the sample points by Euclidean distance in the original space, so the accuracy classification isn't enough. The $\min_max$ algorithm obtains more improvement than the traditional algorithm. However, the traditional algorithm and the $\min_max$ algorithm neglect the irrationality of the classification by Euclidean distance. The improved algorithm maps the input vectors into the feature space and determines the initial centers, at last clusters by the distance between two points in the feature space. The improved algorithm is evaluated on available datasets called iris and wine.

Key words: randomly; Euclidean distance; irrationality; mapping function; feature space

1 引言

k _means 算法原理可靠, 算法较为简单, 且能应用于大数据集, 从而成为应用最为广泛的划分聚类算法之一^[1]. 该算法的过程可简单分为两步: (1) 寻找初始聚类中心; (2) 根据原空间中样本点间的欧式距离进行聚类. 大多数人致力寻找最优的初始聚类中心来提高 k _means 分类准确率. 文献[2]中结合密度分布, 使用 kd_tree 选择初始聚类中心. 文献[3]中提出了密度函数法确定聚出的 DD 算法利用数据分布动态选取初始中心. 文献[4]中利用个体轮廓系数来选出优选样本, 最后确

定初始中心. 上述文献对 k _means 都做了合理的改进, 实验证明了聚类效果有了一定的提高. 但是这些算法忽略以下两点. (1) 度量样本点间相似性的方式的多样性. 原空间的欧氏距离只是度量样本点间相似性的方式之一, 例如向量空间中常利用余弦值度量相似性. (2) 利用欧氏距离来度量相似性的不合理性. 这种不合理性尤其是在高维非线性样本集下表现得尤为突出^[5].

2 $\min_max$ 算法选取初始聚类中心

将 N 个 M 维的输入集 $X = \{x_0, x_1, \dots, x_{N-1}\}$ 分为 K 个类. 传统的 k _means 算法的做法是首先随机选取 K

个样本点作为初始中心,对于剩下的其他样本点根据它们在原空间与初始点之间的欧式距离,将它们归为与之距离最小(相似度最大)的类中,然后重新计算聚类中心(该聚类中所有样本点的均值),重复这一过程,直到前后两次聚类中心不再变化为止.针对 k_means 随机选择初始中心的缺点, min_max 算法提出了新的方法用于更优的选择初始聚类中心.

min_max 算法选取初始中心的过程如下所示.

输入: N 个 M 维的输入集 $X = \{x_0, x_1, \dots, x_{N-1}\}$ 以及聚类个数 K .

输出: K 个初始聚类中心 $C = \{C_0, C_1, \dots, C_{k-1}\}$.

(1) 从输入集中随机选取一个样本点 x_i 作为第一个聚类中心即 $C_0 = x_i$.

(2) 分别计算各个样本点与中心点 C_0 之间的欧氏距离,将与之距离最远的样本点 x_j 作为第二个聚类中心即 $C_1 = x_j$. x_j 满足

$$\|x_j - C_0\| = \max(\|x_i - C_0\| \dots i = 0, 1, \dots, N-1).$$

(3) 分别计算各个样本点与两个初始中心 C_0 、 C_1 间的距离,分别记为

$$d_{i0} = \|x_i - C_0\|, i = 0, 1, \dots, N-1,$$

$$d_{i1} = \|x_i - C_1\|, i = 0, 1, \dots, N-1.$$

选择出较小的那个距离即 $\min(d_{i0}, d_{i1}), i = 0, 1, \dots, N-1$,再从这些最小的距离中选出最大的距离,与之所对应的样本点被选为第三个聚类中心.

if

$$\min(d_{i0}, d_{i1}) =$$

$$\max[\min(d_{i0}, d_{i1})], i = 0, 1, \dots, N-1.$$

那么 $C_2 = x_j$.

(4) 假如现在 $C = \{C_0, C_1, \dots, C_r\}, r < k-1$. 那么第 $r+1$ 个中心点 $C_{r+1} = x_j$. x_j 满足式(1). 一直寻找初始中心直到 $r = k-1$ 时停止.

$$\min(d_{j0}, d_{j1}, \dots, d_{jr}) =$$

$$\max[\min(d_{i0}, d_{i1} \dots d_{ir}), i = 0, 1, \dots, N-1]$$

(1)

该算法克服了传统 k_means 随机选取初始中心、容易陷入局部最优的缺点^[6],但此算法同样根据原空间中样本点间的欧氏距离来进行聚类.第三节的实验证明该算法在输入集维数较低的情况下,分类准确率较传统的 k_means 算法来说有了很大的提高,但当数据集维数较高时分类效果不单没有提高,反而降低了.在此算法基础上我们提出了自己的改进算法.

3 改进的 k_means 算法

3.1 改进算法的基本思想

1964 年 Aizerman 将核函数引入到机器学习中.核函数用于非线性的分类问题,它的基本思想是:输入集在原空间不可分时,可以将它们变换到新空间,在新空间中就可运用线性分类学习方法来进行分类或学习^[7].高斯核函数是最常用的核函数之一,它广泛应用于 SVM 分类和图像滤波等领域中.对改进算法的描述如下所示.

输入: N 个 M 维的输入样本集 $X = \{x_0, x_1, \dots, x_{N-1}\}$.

输出: K 个聚类中心.

① 将输入集通过非线性映射函数 φ 转换到特征空间,形成新的输入样本集 $R^x = \{\varphi(x_0), \varphi(x_1), \dots, \varphi(x_{N-1})\}$.

② 在特征空间中利用的 min_max 算法确定聚类初始中心的方法确定初始聚类中心集 $C = \{C_0, C_1, \dots, C_{k-1}\}$.两样本点的距离如式(2)所示,特征空间中两样本点的欧氏距离与难以确定的非线性映射函数 φ 无关,只与使用的核函数 $k(x, z) = \varphi(x) \cdot \varphi(z)$ 有关.

$$\varphi(z) = e^{-\frac{\|x-z\|^2}{2\sigma^2}}$$

$$\|\varphi(x_i) - \varphi(x_j)\|^2 =$$

$$[\varphi(x_i) - \varphi(x_j)] \cdot [\varphi(x_i) - \varphi(x_j)] =$$

$$\varphi(x_i) \cdot \varphi(x_i) + \varphi(x_j) \cdot \varphi(x_j) -$$

$$2 * \varphi(x_j) \cdot \varphi(x_i) =$$

$$k(x_i, x_i) - 2 * k(x_i, x_j) + k(x_j, x_j) \quad (2)$$

③ 计算各个输入样本点与各聚类中心在特征空间中的距离,将各个向量归入到与之距离最小的类中.

④ 更新聚类中心.其中 n 为所属 C_i 类的向量个数, $i = 0, 1 \dots k-1$.

$$C_i = \frac{\sum_{\varphi(x_m) \in C_i} \varphi(x_m)}{n} \quad (3)$$

⑤ 比较前后两次聚类中心是否变化,若是,则停止算法,否则重复 ③、④,直到聚类中心不再变化为止.

如上所述,映射函数 φ 是难以确定的.因此在完成第一次的聚类后,算法的 ④、⑤ 是很难实现的.因此我们做出了如下的改进.

在更新聚类中心后不再比较聚类中心是否变化,而是继续进行下一次的聚类.上接 ③,具体叙述如下所示.

④ 在完成前一次的聚类后,记录下各个样本点所属的类别构成类别表.

⑤ 更新聚类中心如式(3)所示,回到③对样本点进行聚类.聚类过程中中心点与样本点间的距离计算方式如式(4)所示.同式(2)类似,计算样本点与中心点之间的欧氏距离只与高斯核函数有关.

⑥ 记录该次聚类后各个样本点所属的类别构成新的类别表,比较前后两次类别表是否变化.

```
if(类别表无变化)
{
    停止算法,聚类结束.否则回到⑤继续进行聚类
}
```

从描述中可以看出改进算法不仅改变了原算法中度量相似性的方式,而且将停止聚类过程的准则由前后两次聚类中心不再变化改变为前后两次聚类结束时类别表不再变化为止.

$$\begin{aligned} & \| \varphi(x_i) - C_i \|^2 \\ &= \left\| \varphi(x_i) - \frac{\sum_{\varphi(x_m) \in C_i} \varphi(x_m)}{n} \right\|^2 \\ &= (x_i) \cdot \varphi(x_i) - 2 * \varphi(x_i) \cdot \frac{\sum_{\varphi(x_m) \in C_i} \varphi(x_m)}{n} + \\ & \quad \frac{\sum_{\varphi(x_m) \in C_i} \varphi(x_m)}{n} \cdot \frac{\sum_{\varphi(x_j) \in C_i} \varphi(x_j)}{n} = k(x_i, x_i) - \\ & \quad 2 * \frac{\sum_{\varphi(x_m) \in C_i} k(x_i, x_m)}{n} + \frac{\sum_{\varphi(x_m) \in C_i} \sum_{\varphi(x_j) \in C_i} k(x_j, x_m)}{n^2} \end{aligned} \quad (4)$$

3.2 实验结果与分析

实验采用 UCI^[8] 中的 iris 和 wine 数据集.

3.2.1 原始 k_means 算法、min_max 算法和改进算法在两种数据集上的比较

将原始 k_means 算法、min_max 算法、改进算法分别在 iris 和 wine 数据集上做了 30 次实验,对 30 次实验结果进行了平均得到如表 1、图 1、图 2 所示的结果.表 1 中,min_max 算法在低维(维数为 4)输入样本集 iris 上通过选取最优的初始中心提高了分类准确率,但在高维非线性输入样本集(维数为 13)wine 上的准确率,较原始的 k_means 来说,非但没有提高,反而降低了.而改进算法正是因为改变了度量相似性的方式,利用特征空间的欧氏距离来进行分类.无论是较原始 k_means 算法还是 min_max 算法来说,在低维和高维样本集上都大大提高了分类准确率.

表 1 三种算法在 iris 和 wine 数据上平均准确率的比较 %

数据集	原始 k_means 算法	文献[6]	本文算法
iris	77.93	89.33	92.86
wine	63.81	70.22	72.34

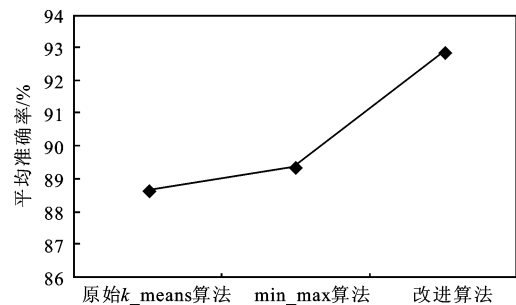


图 1 三种算法在 iris 数据集上平均准确率的比较

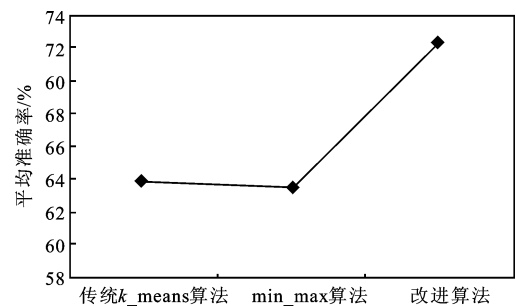


图 2 三种算法在 wine 数据集上平均准确率的比较

3.2.2 改进算法和其他算法聚类准确率的比较

从表 2 可以看出较文献[9]、文献[10]、文献[11]、文献[4]来说,改进算法在两个数据集上的聚类准确率也大大提高.文中实验的参数是通过大量的实验选取而来,其值设定如下:在 iris 上 δ 取 0.45,在 wine 上 δ 取 3.4.

表 2 改进算法与其他算法聚类准确率的比较 %

数据集	文献[9]	文献[10]	文献[11]	文献[4]	本文算法
iris	83.33	88.67	89.33	90.67	92.86
wine	57.30	70.22	70.22	70.22	72.34

4 结束语

传统的 k_means 算法随机选择初始中心,且利用欧氏距离来度量样本点间的相似来进行聚类,聚类效果容易陷入局部最优且准确率也较低.改进算法利用 min_max 算法选择初始中心,继承了 min_max 算法能避免聚类结果陷入局部最优的优点,但改进算法不同于 min_max 算法的是它利用特征空间中的欧氏距离来度量相似性.相比于其他算法和

min_max 算法,改进算法的聚类准确率在低维和高维输入样本集上都有了很大的提高.但是算法采用的高斯核函数的参数值的设定目前还没有简单可行的方法,这是我们未来研究的方向.

参考文献:

- [1] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm[J]. Expert Systems with Applications, 2013, 40(1): 200-210.
- [2] Redmond S J, Heneghan C. A method for initialising the k-means clustering algorithm using kd-trees[J]. Pattern Recognition Letters, 2007, 28(8): 965-973.
- [3] 毛韶阳,李肯立. 优化 K-means 初始聚类中心研究[J]. 计算机工程与应用, 2007, 43(22): 179-181.
- [4] 张靖,段富. 优化初始聚类中心的改进 k-means 算法[J]. 计算机工程与设计, 2013, 34(5): 1691-1694.
- [5] Dhillon I S, Guan Y, Kulis B. Kernel k-means: spectral clustering and normalized cuts[C]// Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. USA, Seattle, ACM, 2004: 551-556.
- [6] Li Y G. A clustering method based on k-means algorithm[J]. Physics Procedia, 2012(25): 1104-1109.
- [7] Filippone M, Camastra F, Masulli F, et al. A survey of kernel and spectral methods for clustering[J]. Pattern recognition, 2008, 41(1): 176-190.
- [8] Bache K, Lichman M. UCI machine learning repository [D]. Irvine, CA: University of California, School of Information and Computer Science, 2013.
- [9] 傅德胜,周辰. 基于密度的改进 K 均值算法及实现[J]. 计算机应用, 2011, 31(2): 432-434.
- [10] 谢娟英,郭文娟,谢维信,等. 基于样本空间分布密度的初始聚类中心优化 K-均值算法[J]. 计算机应用研究, 2012, 29(3): 888-892.
- [11] 陈光平,王文鹏,黄俊. 一种改进初始聚类中心选择的 k-means 算法[J]. 小型微型计算机系统, 2012, 35(6): 1320-1323.

作者简介:

张淑清 女, (1990-), 硕士研究生. 研究方向为图像聚类. E-mail: 15527179027@163.com.

黄震坤 男, (1983-), 博士研究生. 研究方向为视频编码、视频分析.

冯 铭 男, (1988-), 硕士研究生. 研究方向为视频分析.

(上接第 35 页)

- [2] Heinzelman W B, Chandrakasan A P, Balakrishnan H. Energy-efficient communication protocol for wireless microsensor networks [C]// Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS'00). Maui, Hawaii, USA, 2000: 10-19.
- [3] Hoang D C, Kumar R, Panda S K. Realisation of a cluster-based protocol using fuzzy C-means algorithm for wireless sensor networks [J]. IET Wireless Sensor Systems, 2013, 3(3): 163-171.
- [4] Singh A K, Goutele S, Verma S, et al. An energy efficient approach for clustering in WSN using fuzzy logic [J]. International Journal of Computer Applications, 2012, 44(18): 8-12.
- [5] Chiu S L. Fuzzy model identification based on cluster estimation [J]. Journal of Intelligent and Fuzzy Systems, 1994, 2(3): 267-278.
- [6] Xie X L, Beni G. A validity measure for fuzzy clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847.
- [7] 黄小青. 大规模 WSN 中基于分环分簇的能量高效路由协议研究[D]. 长沙: 中南大学, 2011: 48-49.

作者简介:

朱 攀 男, (1990-), 硕士研究生. 研究方向为无线传感器网络. E-mail: eduzhupan7807@163.com.

吴多龙 男, (1964-), 博士, 副研究员. 研究方向为射频与天线、无线传感器网络、无线通信等.