

一种基于 D^2 权重的核 k-means 聚类算法

马翩翩^{1,2}, 苏一丹², 覃 华², 王晓帅²

(广西大学 计算机与电子信息学院, 广西 南宁 530004)

摘 要: 核 k-means 算法是标准 k-means 算法的扩展, 提高了 k-means 聚类中对非线性不可分数据的聚类效果. 传统核 k-means 算法的初始中心是随机选取的, 导致出现聚类时间较慢、聚类性能低等问题. 文中提出了一种基于 D^2 权重的核 k-means 算法, 它根据点对簇内距离的贡献, 选取对其贡献最大的点为簇中心, 然后在核空间内进行相应的聚类. 在 UCI 数据集上进行实验, 实验结果表明, 新算法相对于传统的核 k-means 算法, 可以有效地缩短聚类时间, 并提高聚类的质量, 新算法性能优于传统的核 K-means 算法.

关键词: 核 k-means; k-means; D^2 权重

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2012)07-0085-05

The D^2 Weighting Kernel K-Means Clustering Algorithm

MA Pian-pian², SU Yi-dan², QIN Hua², WANG Xiao-shuai²

(School of Computer and Electronics Information, Guangxi University, Nanning 530004, China)

Abstract: Kernel k-means is an extension of the standard k-means that improves the quality of nonlinearly separable clusters. In order to overcome the cluster initialization problem such as long computation time and low efficiency with this method, In this paper we propose the D^2 weight kernel k-means algorithm. Our method adds one cluster depending on the contribution to the sun of the distance that any point to the closest center that we have already chosen, and we choose the point as the cluster that has the most contribution. In order to verify the effectiveness of the proposed algorithm, we have done a large number of experiments on the UCI dataset. The experiment result shows that the proposed algorithm improves both the accuracy and the speed of kernel k-means. The new algorithm does better than the traditional k-means.

Key words: kernel k-means; k-means; D^2 weighting

1 引言

聚类(clustering)^[1]是一种重要的数据挖掘技术, 在机器学习、模式识别等领域具有重要的作用. K-means^[2]方法是聚类分析中应用最广泛的算法之一, 具有简单且收敛速度快的特点. 该算法在数据挖掘, 文本聚类, 图像处理等方面具有广泛的应用. 但是 k-means 算法有两个局限性: ① 聚类结果依赖于聚类初始中心的选择, 而初始中心是随机选取的; ② 它是线性可分的, 当样本的边界是线性不可分或是非高斯分布、椭圆分布时, 其聚类结果较差.

针对线性不可分的问题, 一般使用核 k-means 算法来解决. 核 k-means 算法是由 Scholkopf 等人^[3]提出的, 它是标准 k-means 算法的扩展. 核 k-means 算法通过非线性转换把原空间的数据映射到高维特征空间, 增加了数据点的线性可分概率, 减少了最小化聚类的误差.

但是核 k-means 仍无法克服初始中心随机选取导致聚类效果差, 聚类时间长的问题. 针对初始中心选择的问题, 在文献^[4]中使用全局的核 k-means 方法, 即采用渐进的方法通过全局搜索来确定 K 的初始中心, 但是该算法的时间复杂度较高. 在文献^[5]

中提出了一种基于局部密度的核 k-means 算法,该算法利用每个样本的局部相对密度来选择具有高密度且低相似性的样本来生成初始类中心点.在文献[6]中提出了一种基于免疫算法的模糊核 k-means 算法,该算法通过免疫遗传算法找到最优解,提高了数据中非线性不可分的聚类精度,但是该算法对初始点的随机选取仍是敏感的.文献[7]中提出了一种近似核 k-means 方法,该方法主要是对非线性的大规模数据集进行聚类,通过取全部数据集核矩阵中的子空间来进行聚类,以减少时间复杂度和空间复杂度,但是该聚类的性能有待提高.

针对核 k-means 中初始簇中心随机选取导致聚类性能低下,聚类结果差的问题,本文提出了一种基于 D^2 权重的核 k-means 算法.该算法首先将数据集映射到高维空间中,然后利用点对簇内距离的贡献来确定核 k-means 的初始中心,从而解决核 k-means 中初始中心随机选取导致聚类性能下降的问题,提高聚类质量.

2 核 k-means 聚类算法

核 k-means 聚类算法^[8]的思想是首先利用非线性映射函数 $\phi: R^n \rightarrow F, x \rightarrow \phi(x)$,将原始空间 R^n 中的样本 x 映射到高维特征空间中,然后在这个高维的核空间中进行 k-means 聚类.在核空间中,待分类的样本变为: $(\phi(x_1), \dots, \phi(x_n))$,进行核 k-means 聚类就是最小化式(1)中的目标函数:

$$\phi = \sum_{k=1}^K \sum_{i=1}^{N_k} \|\phi(x_i) - m_k^{\phi}\|^2 \quad (1)$$

其中,均值: $m_k^{\phi} = \frac{1}{N_k} \sum_{i=1}^{N_k} \phi(x_i)$, $i = 1, \dots, N, k = 1, \dots, K$, N_k 是第 k 类的样本数目.

在核空间中,任意一个样本 $\phi(x)$ 和某类均值 m_k^{ϕ} 之间的距离如下所示:

$$\begin{aligned} \|\phi(x) - m_k^{\phi}\|^2 &= \left\| \phi(x) - \sum_{i=1}^{N_k} \phi(x_i) \right\|^2 \\ &= k(x, x) - 2 \sum_{i=1}^{N_k} k(x, x_i) + \sum_{i,j=1}^{N_k} k(x_i, x_j) \end{aligned} \quad (2)$$

其中, $K(x, y)$ 是核函数.

具体算法流程如下:

输入:簇的数目 m ,核函数 $K(x, y)$ 和包含 n 个对象的数据集,距离参数 DIS.

输出: m 个簇,使目标函数(1)最小.

Step1: 利用核函数 $K(x, y)$ 将数据集 X 对象 x_i 从原空间映射到高维核空间,映射后的资源对象

为 X' ,其相应的对象为 x'_i ;

Step2: 从数据集中随机选定 k 个对象作为初始聚类簇中心集合 $C = (m_1^{\phi}, \dots, m_k^{\phi})$;

Step3: 逐个将资源对象 x'_i 按核距离公式(2)分配给距离最近的一个聚类簇中心 m_i^{ϕ} ;

Step4: 重新计算聚类中心 $(m_1^{\phi}, \dots, m_k^{\phi})$ 和 J^{ϕ} 值.由于在核空间类中心不能明确地进行计算,因此只能在每个类别中选择一个样本来代替类中心,具体方法是:根据步骤 Step3 所得出的分类,分别以每个样本为类中心,计算类内其它各样本点到类中心的距离,并算出距离之和,距离之和最小的类中心就是该类的类中心.

Step5: 重复 Step3 和 Step4,直到连续 n 次迭代 J^{ϕ} 值不变(或变化很小)为止.

其中的 $K(x, y)$ 为核函数,本文选取高斯核函数,其高斯核函数表达式如下所示:

$$K(x, y) = \exp \left[-\|x - y\|^2 / 2\sigma^2 \right] \quad (3)$$

其中参数 σ 是由经验值所得.

3 基于 D^2 权重的核 k-means 算法

核 k-means 算法对初始聚类中心比较敏感,随机选取簇中心导致容易陷入局部最优.本文的算法利用 D^2 权重的思想来确定初始中心点,从而避免聚类结果限于局部最优. D^2 的思想是来自文献[9], D^2 是选择初始中心的一种方法, D^2 方法是首先随机选取第一个簇中心点,剩余的 $k-1$ 个簇中心点是根据任一点到最近簇中心的距离与数据集中任意点到最近簇中心的距离之和的比来选择.即每次选择的是所有点中对簇内距离贡献最大的点,这样就可以使簇内距离单调递减,即满足簇内距离越小,聚类结果越好.

3.1 初始点选取定理

定理 1 D_{Opt} 代表当簇内距离最小,聚类结果最优时所选择的簇中心的集合, A 代表最佳簇中心的子集, C 代表从 A 中随机选取任一个簇中心所对应的簇的集合, ϕ 代表簇 C 所对应的簇内距离,则有:

$$E[\phi(A)] = 2\phi_{Opt}(A) \quad (4)$$

定理 2 C_{Opt} 代表当簇内距离最小,聚类结果最优时所选择的簇中心的集合, A 代表最佳簇中心的子集, C 代表从 A 中使用 D^2 选择的一个簇中心所对应的簇的集合, ϕ 代表簇 C 所对应的簇内距离,则有:

$$E[\phi(A)] \leq 8\phi_{Opt}(A) \quad (5)$$

定理 3 C 代表任意一个簇的集合, ϕ 代表对应簇内的距离. 从数据集中随机选择非最优簇中心 u 个簇中心, X_u 代表以 u 个数据点为簇中心所对应的簇的集合. 规定 $X_C = X/X_U$. 使用 D^2 权重方法向簇集合 C 中添加 t 个簇中心, 且满足 $t \leq u$, 则 C' 代表在使用 D^2 添加簇中心后最终的聚类结果, ϕ' 代表簇 C' 相对应的簇的距离. 则结果满足下式:

$$E[\phi'] \leq (\phi(X_C) + 8\phi_{Opt}(X_U)) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \phi(X_u) \quad (6)$$

综上所述, 使用 D^2 选择初始化簇中心的方法来进行 k-means 聚类使得其目标函数 $\phi = \sum_{x \in X} \min_{c \in C} \|x - c\|^2$ 以 $O(\log K)$ 的优势于随机初始化簇中心进行 k-means 所得到的目标函数.

$$E(\phi) \leq 8(Lnk + 2)\phi_{OPT} \quad (7)$$

3.2 D^2 初始中心实现

使用 D^2 选取簇中心的过程是从数据集 $X = (x_1, x_2, \dots, x_n)$ 中任意选取一点作为簇中心 c_1 , 在 0 到 $D(x_1)^2 + D(x_2)^2 + \dots + D(x_n)^2$ 中选择一个实数 y , 选择下一个簇中心坐标正整数 i , 使其满足 $D(x_1)^2 + D(x_2)^2 + \dots + D(x_i)^2 \geq y \geq D(x_1)^2 + \dots + D(x_{i-1})^2$, 其中 $D(x_i)$ 代表任意一点到已经生成的簇中心的最短距离.

新改进的算法步骤如下:

输入: 簇的数目 k , 核函数 $K(x, y)$ 和数据集 X .

输出: k 个簇, 使目标函数最小.

Step1: 利用核函数 $K(x, y)$ 将数据集 X 对象 x_i 从原空间映射到高维核空间, 映射后的资源对象为 X' , 其相应的对象为 x'_i .

Step2: 从数据集中选定 k 个对象作为初始聚类簇中心集合 $C = (m_1^*, \dots, m_k^*)$. 其步骤为:

① 从数据集 X' 随机选取第一个初始簇中心 m_1^* ;

② 依据正态分步, 选取下一个质心 $m_i^* = j' \in X'$, 使其相对概率为

$$\frac{D(j')^2}{\sum_{j' \in X'} D(j')^2}$$

③ 重复步骤②, 直到选取 k 个质心.

Step3: 逐个将资源对象 x'_i 按核距离公式(2)依据分配给距离最近的一个聚类核中心 m_i^* .

Step4: 重新计算聚类中心 $(m_1^*, \dots, m_k^*)$ 和 J^* 值. 由于在核空间类中心不能明确地进行计算, 因此只能在每个类别中选择一个样本来代替类中心, 具

体方法是: 根据 Step3 所得出的分类, 分别以每个样本为类中心, 计算类内其它各样本点到类中心的距离, 并算出距离之和, 距离之和最小的类中心就是该类的类中心.

Step5: 重复 Step3 和 Step4, 直到连续 n 次迭代 J^* 值不变(或变化很小)为止.

4 实验

4.1 实验设计

实验中算法代码采用 C++ 编写, 编译平台为 Visual studio 2008. 仿真代码运行在一台 PC 机上 (CPU 2.78GHz, 512MB 内存, WindowsXP 操作系统). 为了检验算法的有效性, 实验的数据集采用 UCI 公共数据库提供的四个数据集^[10]: Iris、glass、Yeast、Soybean, 这些数据集有原始分类, 可用于聚类性能评价. 在实验中统一使用高斯核函数, 核函数参数由 σ 经验值得到. 如表 1 所示.

表 1 实验数据基本情况

数据集	类别数	属性个数	样本总数
Iris	3	4	150
glass	6	10	214
Yeast	10	9	1484
Soybean	19	35	307

为了验证算法的有效性, 从以下三个方面进行

评价: 核 k-means 的目标函数值 $\phi = \sum_{k=1}^K \sum_{i=1}^n \| \phi(x_i) - m_k^* \|^2$, 即聚类后类内距离之和的总和, 其中 m_k^* 为核空间中簇中心, $\phi(x_i)$ 为映射到核空间中的点, 目标函数值越小, 就说明这个资源类越紧凑, 类内的资源越相似, 聚类效果越好; 运行该算法所使用的时间, 单位为 s, 运行时间越短, 表明效率越高; 聚类结

果的精确度, 其表达式为 $\sum_{i=1}^k N_i/n$, 其中 N_i 是在第 i 个聚类结果簇中能够被正确识别的数据对象的数目, n 是测试数据集所有对象的数目, 聚类精确度的数值越大, 表明被划分到正确所属类的数据点越多, 聚类效果越好. 提升度 $\text{Improve} = 1 - D^2$ 核 k-means/核 k-means.

4.2 实验结果与分析

分别用核 k-means 算法与 D^2 核 k-means 算法对 UCI 数据集 Iris、Glass、Yeast、Soybean 进行聚类, 每个算法运行 10 次, 取其平均值来进行算法比

较. 实验的结果如表 2、表 3 和表 4 所示.

表 2 为核 k-means 算法与 D^2 核 k-means 算法在 UCI 数据集 Iris, Glass, Yeast, Soybean 运行 10 次最终聚类后簇内距离之和的平均值的对比表.

表 2 类内距离之和

数据集	参数 σ	核 k-means (s)	D^2 核 k-means (s)	Improve (%)
Iris	0.2	0.657357	0.526284	19.939
Glass	5	1.65413	0.66658	59.701
Yeast	17	0.327909	0.318783	2.783
Soybean	10	5.73097	5.51691	3.735

表 3 为核 k-means 算法与 D^2 核 k-means 算法在 UCI 数据集 Iris, Glass, Yeast, Soybean 运行 10

次所使用时间的平均值的对比表. 由于初始中心选择的不同,会导致聚类中迭代次数的变化.

表 3 运行时间

数据集	参数 σ	核 k-means (s)	D^2 核 k-means (s)	Improve (%)
Iris	0.2	0.5424	0.403274	25.65
Glass	5	1.00618	0.885842	11.959
Yeast	17	18.6422	15.784	15.332
Soybean	10	32.0037	31.2827	2.252

表 4 为核 k-means 算法与 D^2 核 k-means 算法在 UCI 数据集 Iris, Glass, Yeast, Soybean 运行 10 次聚类精确度的平均值的对比表.

表 4 聚类精确度

数据集	参数 σ	核 k-means(s) 错分样本	核 k-means(s) 正确率	D^2 核 k-means 错分样本	D^2 核 k-means(s) 正确率
Iris	0.2	13	0.91333	10	0.9333
Glass	5	46	0.78505	33	0.84579
Yeast	17	410	0.72372	365	0.75404
Soybean	10	63	0.79479	34	0.88925

综上所述,在选取初始化中心时,依据点对簇中心的距离的贡献来选取,从而导致簇中心距离下降,即聚类效果越好,簇之间越紧凑. 由表 2 可知,在使用参数相同的情况下, D^2 核 k-means 的簇内距离比核 k-means 簇内距离小,都有一定的幅度的提高;由于初始中心选择的不同,会导致聚类中迭代次数的变化,即聚类效果达到稳定时所使用的时间. 由表 3 表明,利用 D^2 进行初始化中心时,选取初始中心不再随机化,从而使其在相对较少的时间内就可以达到稳定的聚类效果. 表 4 进一步验证了当簇内的距离的越紧凑,聚类的效果越好,即聚类的正确性得到了提高.

5 结束语

文中提出了一种基于 D^2 权重的核 k-means 思想,该算法使用点对簇内距离的贡献的大小来确定初始中心,有效地提高了核 k-means 算法的精确度. 实验表明,该算法通过减少簇之间的距离和运行时间,来提高聚类质量,并通过实验验证了该算法的有

效性. 未来的研究重点是将该算法应用到具体的案例中,这对促进核 k-means 算法的发展与应用具有积极的意义.

参考文献:

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2008.

[2] Xu Rui, Wunsch D. Survey of clustering algorithms [J]. IEEE Trans on Neural Network, 2005, 16(3): 645-678.

[3] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10(5): 1299-1399.

[4] Tzorzis, G, Likas, A. The global kernel K-means clustering algorithm [J]. Neural Networks, 2008: 1977-1984.

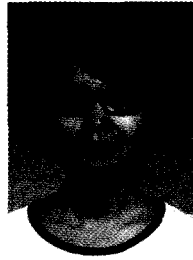
[5] 李米娜, 朱玉全, 郝洪星, 等. 一种基于局部密度的核 k-means 算法[J]. 计算机应用研究, 2011, 28(1): 79-90.

[6] Chengjie Gu, Shunyi Zhang, Kai Liu, et al. Fuzzy kernel K-means clustering method based on immune genetic algorithm[J]. Journal of Computational Infor-

mation System, 2011(7):221-231.

- [7] Radha Chitta, Rong Jin, Timothy C Havens, et al. Approximate kernel k-means: solution to large scale kernel clustering[C]// Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, USA, 2011: 895-903.
- [8] Muller K, Mika S, Ratsch G, et al. An introduction to kerner-based learning algorithms[J]. IEEE Trans. Neural Net, 2001,12(2): 181-201.
- [9] David Arthur. Sergei vassilvitskii, k-means++: the advantages of careful Seeding [C]// SODA. USA, New Orleans, 2007.
- [10] UCIDatSet[EB/OL]. [2011-01-15]. <http://archive.ics.uci.edu/ml/datasets.htm>.

作者简介:



马翩翩 女,(1986—),硕士研究生. 研究方向为个性化推荐、数据挖掘.

苏一丹 男,(1962—),博士,教授. 研究方向为网络安全、电子商务.

覃华 男,(1972—),博士,副教授. 研究方向为电子商务数据挖掘与最优化技术.

王晓帅 女,(1985—),硕士研究生. 研究方向为个性化推荐、数据挖掘.

(上接第84页)

- [3] Paul R, Patra A. A temperature compensated bandgap voltage reference circuit for high precision applications [C]//Proceeding of the IEEE INDICON. India, 2004: 553-556.
- [4] Lee I, Kim G, Kim W. Exponential curvature-compensated BICOMS bandgap references[J]. IEEE Journal of Solid-State Circuits, 1994, SC-29(11):1396-1403.
- [5] Rincon-Mora G A, Allen P E. A 1.1V current mode and piecewise linear curvature corrected bandgap reference[J]. IEEE Journal of Solid-State Circuits, 1998, SC-33(10):1551-1554.
- [6] Allen P E, Douglas R H. CMOS analog circuit design [M]. Beijing: Publishing House of Electronics Industry, 2002.
- [7] 肖明, 吴玉广, 何凤琴. 一种低压 I^2PTAT 曲率校正 CMOS 带隙基准源的设计[J]. 微电子学与计算机,

2007, 24(3):85-87.

- [8] 廖敏, 周玮. 一种二阶曲率补偿的带隙电压基准[J]. 现代电子技术, 2009(22):10-12.
- [9] Guan X, Wang X, Wang A, et al. A 3V 110uW 3.1ppm/ $^{\circ}\text{C}$ curvature-compensated CMOS bandgap reference[C]//IEEE ISCAS. USA, 2006: 2861-2864.

作者简介:

宋宇 男,(1986—),硕士研究生. 研究方向为集成电路设计.

吴春瑜 男,(1953—),博士,教授,硕士生导师. 研究方向为半导体器件和集成电路设计.

王绩伟 男,(1968—),博士,副教授,硕士生导师. 研究方向为半导体物理与器件.