一种改进的基于特征赋权的K均值聚类算法*)

任江涛 施潇潇 孙婧昊 黄焕宇 印 鉴
(中山大学计算机科学系 广州 510275)

摘 要 聚类分析是数据挖掘及机器学习领域内的重点问题之一。近年来,为了提高聚类质量,借鉴和引入了分类领域特征选择及特征赋权思想,提出了一些基于特征赋权的聚类算法[1~3]。在这些研究基础上,本文提出了一种基于密度的初始中心点选择算法,并借鉴文[1]所提出的特征赋权方法,给出了一种改进的基于特征赋权的K均值算法。实验表明该算法能较为稳定地得到较高质量的聚类结果。

关键词 聚类,特征赋权,初始化

An Improved K-Means Clustering Algorithm Based on Feature Weighting

REN Jiang-Tao SHI Xiao-Xiao SUN Jin-Hao HUANG Huan-Yu YIN Jian
(Department of Computer Science, Zhongshan University, Guangzhou 510275)

Abstract Clustering analysis is one of the important problems in the data mining and machine learning areas. Recently, feature selection and feature weighting methods are introduced to clustering algorithms for improving the clustering quality[1~3]. Inspired by the research, an improved k-means clustering based on feature weighting is proposed, which proposes a density-based initial centers search algorithm. The experiments show that the proposed algorithm can result in high quality clustering steadily.

Keywords Clustering, Feature weighting, Initialization

1 引言

聚类分析在数据挖掘、模式识别、决策支持、机器学习及图像分割等领域有广泛的应用,是最重要的数据分析方法之一。它将数据对象分入不同的集合,使得同一集合中的数据对象相对来说有更大的相似性,而不同一集合里的对象相对来说有更大的差别。

在分类领域,通过特征分析,发现有些特征是与分类强相关的,有些特征与分类弱相关,还有许多特征与分类不相关,即不同的特征对于分类的贡献可能很不相同。因此在分类研究中很早就提出了许多特征选择及特征赋权方法,以提高分类精度。近年来,为了提高聚类质量,也借鉴和引入了分类领域的上述思想,提出了一些基于特征赋权的聚类算法。文[1,2]分别给出了一种基于特征赋权的类K均值聚类算法,文[3]给出了一种基于特征赋权的模糊C均值算法。但这些算法在初始点的选择上,由于采用随机选择方法,运行结果不稳定,且不易得到较好的聚类结果。

针对上述问题,本文提出了一种基于密度的初始中心点选择算法,并借鉴文[1]所提出的特征赋权方法,给出了一种改进的基于特征赋权的K均值算法,实验表明该算法能较为稳定地得到较高精度的聚类结果。

论文的第2部分对所提出的算法进行了描述,第3部分给出了实验研究结果,最后是对本文的总结。

2 算法描述

本算法主要解决两个问题,一是初始中心点的选择问题,另外就是特征权重的计算问题,下面分别针对这两个问题给出相应的算法并给出整个算法流程。

(1)初始中心点的选择

初始中心点的选择在K-Means算法中非常重要,通常希望找到散布较大的点作为初始中心点。但是在一般的基于贪心算法的初始中心点搜索过程中,由于仅仅基于距离因素,往往找到许多孤立点作为中心点,且初始中心点选择的随机性较强,导致聚类结果的随机性。实际上,对于初始中心点,除了希望分布得尽量散之外,还希望这些中心点具有一定的代表性,即具有较高的密度。因此,在初始点的选择中,除了考虑其散布程度外,还应考虑密度因素。因此,我们提出了一个基于密度及散布特征的初始中心点选择算法。

首先给出样本点 x 的密度定义如式(1)所示,其中 S 代表样本点集,$Dist(\cdot)$ 代表某种距离度量,如欧氏距离,ε 代表半径。式(1)表明点 x 的密度定义为以点 x 为中心,以 ε 为半径的超球体内所包含的样本的个数。

$$Density(x)=|\{p\in S|Dist(x,p)\leqslant\varepsilon\}| \quad (1)$$

其中,半径 ε 的确定是一个较为困难的问题,若 ε 取值过低或过高,都不能获得有效的密度估计。实际上,此问题的根源在于不了解样本间距离的概率分布。因此在本研究中,采用概率统计的方法来确定相似度阈值,具体方法如下:

1)计算两两样本距离的均值 μ_{Dist};

2)由用户根据经验给定一个系数 θ;

3)令 $\varepsilon=\theta*\mu$,作为计算密度的半径。

实验表明,该方法能较为有效地减少密度半径 ε 选择时的盲目性,提高实验效率。

根据上述分析,首先从原样本集 S 中找出一个高密度点集 S_d,再从 S_d 中应用贪心算法搜索出散布较大的初始中心点集。高密度点集 S_d 定义为密度不小于平均密度 $\mu_{density}$ 的 β 倍的数据点的集合。下面给出初始中心点集 M 的搜索算法如下。

算法1 $Initial(S,k,\beta)$

*)本研究得到国家自然科学基金项目(60374059)及广东省自然科学基金项目(04300462)资助。任江涛 博士,讲师。

输入:待处理数据集 S,聚类个数 k,β

输出:初始中心点集 M

步骤:

1)根据式(1)计算每个点的密度 $Density(x)$,并计算平均密度 $\mu_{density}$;

2)计算点集 $S_d=\{x|Density(x)\geqslant\beta*\mu_{density},x\in S\}$;

3)初始中心点集 M 初始化为空集,即 $M=\{\}$;

4)选择密度最大的点 m_1 为第 1 个初始中心点,即

$Density(m_1)=\mathrm{Max}\{Desity(x)|x\in S_d\}$;

$M=M\cup\{m_1\}$;

5)寻找满足如下条件的点 m_i:

$Dist(x)=\mathrm{Min}\{Dist(x,q)|q\in M\},x\in S_d\backslash M$

$m_i\in S_d\backslash M$

$Dist(m_i)=\mathrm{Max}\{Dist(x)|x\in\partial S_d\backslash M\}$;

6)将点 m_i 加入中心点集 M,即 $M=M\cup\{m_i\}$;

7)重复步骤 5)、6),直至找到 k 个中心点,即 $|M|=k$;

8)输出中心点集 M,算法结束。

从上述算法可以看出,由于初始中心点集的第一点为确定的(最大密度点),在基于距离最远的其它中心点搜索过程中,得到的中心点也基本上是确定的,消除了初始中心点选择的随机性,同时保证了获得较高质量的初始中心点。

(2)权重学习

权重学习是该算法中关键的一环,本文采用文[1]所给出的权重方法。该方法的基本原理是对类内分布一致性好的特征赋予较大的权重,而且在不同的类内相同的特征赋予不同的权重。类内分布的一致性主要通过类内该特征的方差大小度量。式(3)给出了特征 j 在第 i 类内的分布方差 X_{ij} 的定义。

令 S 代表整个数据集,S_i 代表第 i 类数据集,x 代表样本点,X_{ij} 代表特征 j 在第 i 类内的分布方差,W_{ij} 代表特征 j 在第 i 类内的权重;c_l 代表第 l 类的类中心向量。

$$S_i=\{x|i=\arg\min_l Dist_w(c_l,x)\}$$

$$Dist_w(c_l,x)=\left(\sum_{j=1}^{d}w_{lj}(c_{lj}-x_j)^2\right)^{1/2} \tag{2}$$

$$X_{ij}=\sum_{x\in S_i}(C_{ij}-x_j)^2)/|S_i| \tag{3}$$

其中 c_{ij} 是第 i 类中心的第 j 特征,x_j 则是点 x 的第 j 个特征,$|S_i|$ 是第 i 类的样本数。第 i 类的第 j 个特征的权重 w_{ij} 定义如式(4)所示。

$$w_{ij}=\exp(-h\times X_{ij})/\left(\sum_{o=1}^{d}\exp(-h\times 2\times X_{io})\right)^{1/2} \tag{4}$$

其中,h 是用户输入的一个参数,在本算法中,h 定义为 15。另外,为了避免由于 X_{ij} 过大而导致 $\exp(-h\times X_{ij})$ 由于计算精度问题而趋于 0,有必要对 x 的取值进行一定的预处理。经过实验,发现令 $x_j=x_j/\mu_{x_j}$ 可以取得较好的效果。其中,μ_{x_j} 为 x_j 的均值。

$$x_j=x_j/\mu_{x_j} \tag{5}$$

基于上述方法对特征权重进行计算后,基于调整后的特征权重重新计算各样本点与各聚类中心的距离,并在此基础上重新分配给样本点,并通过式(6)得到新的聚类中心 c_i 向量。

$$c_i=\left(\sum_{x\in S_i}x\right)/|S_i| \tag{6}$$

(3)算法流程

根据(1)、(2)节的讨论结果,与 K 均值算法相结合,得到一种新的经过改良的基于特征赋权的 K 均值聚类算法,算法具体描述如下:

算法 2 $WeightKmeans(S,K,\beta)$

输入:待处理数据集 S,聚类个数 k,β

输出:聚类结果

步骤:

1)调用函数 $Initial(S,k,\beta)$,获得选择 k 个初始聚类中心;

2)初始设置权重 $w_{ij}=1/d$;

3)根据式(2)将样本分配至各类 S_i;

4)根据式(3)、(4)计算新的权重系数 w_{ij};

5)根据式(2),重新计算各样本与当前各聚类中心 c_i 的距离,将样本分配至各类 S_i;

6)根据式(6),重新计算各聚类中心 c_i;

7)重复步骤 3)、4)、5)、6)直至算法收敛或达到指定的迭代次数。

3 实验研究

为了评估算法的有效性,采用了三个真实高维数据集进行测试,包括雷达数据集 Ionosphereh 及两个基因表达数据集 MLL_Leukemia 和 Lung_Cancer,采用分类正确率度量聚类的质量。为了验证本算法的优越性,采用文[1]的 LAC(Locally Adaptive Clustering)算法与本算法进行性能比较。数据集的说明及实验结果说明分别在表 1 及表 2 中给出。

表 1 选用数据集及说明

数据集	特征维数	类别数	样本数
Ionosphere	34	2	351
MLL_Leukemia	12582	3	57
Lung_Cancer	12533	2	32

表 2 聚类实验结果比较

算法	Ionosphere	MLL_Leukemia	Lung_Cancer
LAC	75%	65%	69%
WeightKmeans	81%	94.7%	97%

从上述实验结果可看出,经过改进,算法的聚类精度得到了较大提高,取得了较为令人满意的结果。

结论 本文主要针对当前基于特征赋权的 K 均值算法结果不稳定这一问题,提出了一种改进算法。主要通过基于密度的方法进行初始中心点选择,排除了传统选点算法中随机性对结果的影响。另外,在对数据的预处理方面也进行了一些改进,使权值的计算更为稳定可靠。实验结果表明,该算法对高维数据聚类的质量较好。

参考文献

1 Domeniconi C, Papadopoulos D, Gunopulos D, Ma S. Subspace Clustering of High Dimensional Data. In: Proc. of the Fourth SIAM Intl. Conf. on Data Mining, 2004. 517～521

2 Chan E Y, Ching W K, Ng M K, Huang J Z. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognition, 2004, 37: 943～952

3 Wang Xizhao, Wang Yadong, Wang Lijuan. Improving fuzzy c-means clustering based on feature-weight learning. Pattern Recognition Letters, 2004, 25: 1123～1132

4 Aggarwal C C, Procopiuc C, Wolf J L, et al. Fast Algorithms for Projected Clustering. In: Proc. of ACM SIGMOD Conference 99, 1999. 61～72

5 Kaufman L, Rousseeuw P. Finding Groups in Data - An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics, 1990

6 Huang Zhexue. Extensions to the K-Means Algorithm for Clustering LargeData Sets with Categorical Values. Data Mining and Knowledge Discovery, 1998. 283～304