

文章编号:1001-9081(2011)06-1675-03

doi:10.3724/SP.J.1087.2011.01675

基于信息熵的精确属性赋权 K-means 聚类算法

原福永, 张晓彩, 罗思标

(燕山大学 信息科学与工程学院, 河北 秦皇岛 066004)

(callzxc@126.com)

摘要:为了进一步提高聚类的精确度,针对传统 K-means 算法的初始聚类中心产生方式和数据相似性判断依据,提出一种基于信息熵的精确属性赋权 K-means 聚类算法。首先利用熵值法对数据对象的属性赋权来修正对象间的欧氏距离,然后通过比较初聚类的赋权类别目标价值函数,选择高质量的初始聚类中心来进行更高精度和更加稳定的聚类,最后通过 Matlab 编程实现。实验证明该算法的聚类精确度和稳定性要明显高于传统 K-means 算法。

关键词:K-means; 精确度; 信息熵; 属性赋权; 初始聚类中心

中图分类号: TP181; TP301 **文献标志码:**A

Accurate property weighted K-means clustering algorithm based on information entropy

YUAN Fu-yong, ZHANG Xiao-cai, LUO Si-biao

(College of Information Science and Engineering, Yanshan University, Qinhuangdao Hebei 066004, China)

Abstract: Concerning the initial clustering center generation and the data similarity judgment basis of the traditional K-means algorithm, the paper proposed an accurate property weighted K-means clustering algorithm based on information entropy to further improve the clustering accuracy. First, property weights were determined by using entropy method to correct the Euclidean distance. And then, high-quality initial clustering center was chosen by comparing the empowering target cost function of the initial clusters for more accurate and more stable clustering. Finally, the algorithm was implemented in Matlab. The experimental results show that the algorithm accuracy and stability are significantly higher than the traditional K-means algorithm.

Key words: K-means; accuracy; information entropy; property weight; initial clustering center

0 引言

由于网络信息数据的急速膨胀,用户疲于从海量的检索结果中查看哪些结果符合自己的查询要求,查准率的呼声越来越高,聚类精确度的提高成为亟待解决的问题。Web 聚类应运而生,它将数据集划分为若干组或类,使得同一类内的数据对象具有较高的相似度,而不同类间的数据对象具有较大的差异。这使得用户可以先查找自己感兴趣的类,再仔细浏览所要查找的内容,大大节省了用户的搜索时间,提高了查找命中率。Web 聚类算法可以划分为基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法等^[1]。基于划分的方法是较为常用的聚类方法,其典型的代表是 K-means (K-平均)^[2] 算法。现今,国内外对 K-means 算法进行了大量的研究^[3-6],这些研究大都关注于数据对象集合的软聚类、初始聚类中心的选择方案和降低算法的时间复杂度,而对于提高 K-means 算法聚类的精确度却比较鲜见。

传统 K-means 算法随机选择初始聚类中心,同时忽略了数据对象各属性对聚类结果发挥的不同聚类作用,这些导致算法难以获得稳定而精确的聚类结果。信息熵是系统有序化程度的一个度量,一个系统越是有序,信息熵就越低;反之,一个系统越是混乱,信息熵就越高^[7]。针对传统 K-means 算法存在的问题,提出一种基于信息熵的精确属性赋权 K-means 聚类算法来提高聚类的精确度。该算法利用熵值法^[8-9]计算

数据对象各属性的权值,采用赋权欧氏距离作为相似性度量的依据,在确定初始聚类中心前预选聚类种子中心,取结果较好(赋权标准差小)的 k 个类的质心作为初始聚类中心继续来进行聚类,并用实验校验了该算法的性能。实验结果表明该算法能产生精确度更高的聚类结果,并且提高了聚类结果的稳定性。

1 相关定义

设待聚类的数据对象集为 $A = \{a_i | a_i \in \mathbb{R}^m, i = 1, 2, \dots, n\}$, k 个类别用 $T_i (i = 1, 2, \dots, k)$ 表示, k 个聚类的中心分别为 $c(T_1), c(T_2), \dots, c(T_k)$, 有如下定义。

定义 1 设两个向量 $a_i = (a_{i1}, a_{i2}, \dots, a_{im})$ 和 $a_j = (a_{j1}, a_{j2}, \dots, a_{jm})$ 分别表示两个对象,它们之间的欧氏距离定义如下:

$$d(a_i, a_j) = \sqrt{\sum_{d=1}^m (a_{id} - a_{jd})^2} \quad (1)$$

定义 2 同一类别数据对象的质心定义如下:

$$c(T_i) = \frac{1}{|T_i|} \sum_{a_j \in T_i} a_j \quad (2)$$

其中 $|T_i|$ 是 T_i 中数据对象的个数。

定义 3 同属于 T_j 组的 n_1 个数据对象 $a_i (i = 1, 2, \dots, n_1)$ 的标准差 σ 定义如下:

收稿日期:2010-12-22;修回日期:2011-03-01。

作者简介:原福永(1958-),男,黑龙江鸡西人,教授,主要研究方向:网络信息检索、数据库; 张晓彩(1985-),女,河北石家庄人,硕士研究生,主要研究方向:网络信息检索、数据库; 罗思标(1984-),男,江西吉安人,硕士研究生,主要研究方向:计算几何、机器人路径规划。

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n_1} (\mathbf{a}_i - c(T_j))^2}{n_1 - 1}} \quad (3)$$

2 本文方法

2.1 使用熵值法计算属性权重

在信息论中,信息熵是系统有序程度的度量。某个属性的变异程度越大,系统越是有序,该属性的信息熵越小,它提供的信息量越大,权重也就越大;反之,某个属性的变异程度越小,系统越是混乱,该属性的信息熵越大,它提供的信息量越小,权重也就越小。文中根据各属性的变异程度,利用熵值法计算各属性的权重,为无序数据对象集聚类提供依据。

使用熵值法确定属性权重值的步骤如下:

1) 设有 n 个待聚类数据对象,所有数据对象由 m 维属性来表示,根据实时数据构造属性值矩阵:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

2) 计算第 j 维属性对应的第 i 个数据对象的属性值比重。在实际问题中不同的数据往往有不同的量纲,为了使有不同量纲的数据能进行比较,进行数据标准化,即将数据压缩到区间 $[0,1]$ 上,其过程如式(4)所示:

$$M_{ij} = x_{ij} / \sum_{i=1}^n x_{ij} \quad (4)$$

其中: M_{ij} 为属性值比重, x_{ij} 为属性值, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$ 。

3) 计算第 j 维属性的熵值:

$$H_j = -p \sum_{i=1}^n M_{ij} \ln M_{ij} \quad (5)$$

其中: H_j 为属性熵值, $p = 1 / \ln n$ 。当 $M_{ij} = 0$ 时,有 $M_{ij} \ln M_{ij} = 0$ 。如果 x_{ij} 对于给定 j 的全部相等,那么 $M_{ij} = x_{ij} / \sum_{i=1}^n x_{ij} = 1/n$,此时 H_j 取极大值。

4) 计算第 j 维属性的差异系数:

$$q_j = 1 - H_j \quad (6)$$

其中 q_j 为差异系数。对于给定的 j ,当数据对象的属性值相差越小时, H_j 越大,该属性的聚类作用越小;当数据对象的属性值相差越大时, H_j 越小,该属性的聚类作用越大;当属性值全部相等时, $H_j = H_{\max} = 1$,此时属性的聚类作用为零。综上所述可知:当 q_j 越大时,属性越重要。

5) 计算第 j 维属性的权值:

$$\omega_j = \frac{q_j}{\sum_{j=1}^m q_j} \quad (7)$$

其中: $0 \leq \omega_j \leq 1$, $\sum_{j=1}^m \omega_j = 1$, $j = 1, 2, \dots, m$ 。

2.2 确定高质量的初始聚类中心

K -means 算法选择的相似性度量通常为欧氏距离,两个数据对象的欧氏距离越小,表示两者的相似度越大,反之,则相似度越小。文中算法选择的相似性度量为赋权欧氏距离。设第 j 维属性的权值为 ω_j ,根据定义 1 得到赋权后的欧氏距离为

$$d_w(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{\sum_{j=1}^m \omega_j (x_{aj} - x_{bj})^2} \text{。这相当于根据属性 } j \text{ 的权}$$

值对其对应的属性值进行了适当的放大与缩小,使权值大的属性聚类作用更大,权值小的属性聚类作用更小,真实反映了实际聚类时各属性发挥的作用。

K -means 算法一般采用标准差作为标准测度函数,在选择赋权欧氏距离作为相似性度量后,由定义 3 求得赋权类别目标价值函数表示为:

$$\sigma_i = \sqrt{\frac{\sum_{x_i \in T_j} d_w(x_i, c(T_j))}{|T_j| - 1}}$$

其中: σ_i 表示第 i 类的赋权标准差, $|T_j|$ 是 T_j 所含数据对象的个数。显而易见,赋权类别目标价值函数 σ_i 的值越小,说明类内数据对象相似度越大,数据对象越密集,其所在类的质心越能够体现分类决策面。

传统 K -means 算法随机选择初始聚类中心,致使初始聚类中心的质量良莠不齐。为了得到高质量的初始聚类中心,对初始聚类中心进行预处理,其思路为:首先将数据集均分为 k_1 ($k_1 > k$) 个子集,在每个子集里随机选择一个数据对象,再利用选择的 k_1 个数据对象作为聚类种子中心进行初聚类,计算各类别的赋权类别价值函数 σ_i 并将其从小到大进行排序,最后选择前 k 个类对应的质心作为初始聚类中心。

2.3 算法描述

基于信息熵的精确属性赋权 K -means 聚类算法步骤描述如下:

输入:待聚类数据集 A ,聚类种子中心个数 k_1 ,聚类个数 k

输出: k 个聚类,使每个数据对象到相应聚类中心的赋权欧氏距离之和最小。

- 1) 使用熵值法计算数据对象各属性的权值。
- 2) 将数据集均分为 k_1 ($k_1 > k$) 个子集,从每个子集中随机选择一个数据对象,把随机选择的 k_1 个数据对象作为聚类种子中心。
- 3) 扫描所有的数据对象,根据其与各聚类种子中心的相似度(赋权欧氏距离),将其归入与其最相似的(聚类种子中心代表的)聚类。
- 4) 计算每个类的质心。
- 5) 计算 k_1 个聚类的 σ_i ($i = 1, 2, \dots, k_1$),并按照 σ_i 值递增顺序排列,选前 k 个 σ_i 值对应的质心作为初始聚类中心。
- 6) 扫描所有数据对象,根据其与 k 个初始聚类中心的赋权欧氏距离,将其归入与其距离最近的聚类。
- 7) 计算 k 个类的质心。
- 8) 重复执行 6) 和 7),直到完成算法要求的迭代次数时,循环终止。
- 9) 计算出各类别的标准差来检测聚类的客观性,若标准差值存在非数值型数据,则重新聚类。
- 10) 扫描所有数据对象检测其聚类结果,将误判率降到最小,确保聚类结果的精确度。

传统 K -means 算法时间复杂度为 $O(tkn)$,其中 t 是迭代次数。文中算法含有四个处理操作:①计算每一属性的熵值;②选择高质量的初始聚类中心;③数据聚类;④最大限度减小数据误判率。分析可知:①操作的时间复杂度为 $O(nm)$;②操作的时间复杂度为 $O(k_1n)$;③操作的时间复杂度为 $O(t_1kn)$,其中 t_1 是迭代次数;④操作的时间为 $O(kn)$ 。由于传统 K -means 算法是随机选取初始聚类中心的,因此迭代次数比较大。文中算法优化了初始聚类中心的选择,可以大大减

少算法的迭代次数,从而降低了算法的时间复杂度。

3 实验及结果分析

为验证算法的效果,对传统K-means算法和文中算法进行对比实验。实验选用Matlab R2007a作为编程工具。实验环境如下,操作系统:Microsoft Windows XP;CPU:Intel Core2 2.93 GHz;内存:2 GB。采用UCI数据库^[10]的Iris和Statlog(Vehicle Silhouettes)数据集作为测试数据。Iris数据集取自3种鸢尾属植物的花朵样本,含150条数据,每条数据有4维属性。Statlog数据集取自4种交通工具轮廓样本,含846条数据,每条数据有18维属性。

1)计算得到Iris数据集各属性对应的权值如图1所示,Statlog数据集各属性对应的权值如图2所示。

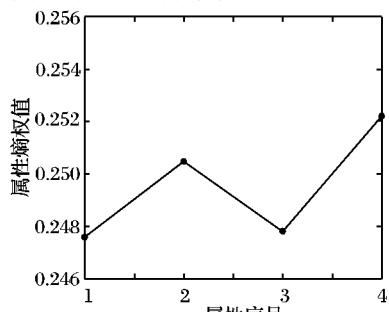


图1 Iris属性熵权值

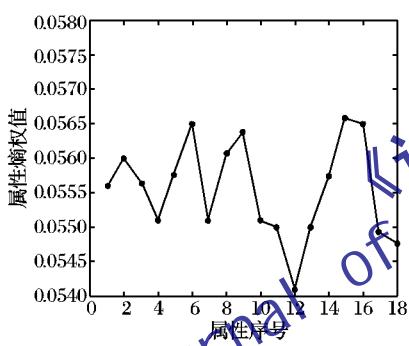


图2 Statlog属性熵权值

由图1和图2中的属性熵权值数据得知,在聚类过程中每个属性的聚类作用真实不同,应当加以区分对待。传统K-means算法忽略了属性对聚类作用的差异度,导致数据对象的误判情况屡屡发生,算法聚类结果和真实聚类结果之间存在差距。

2)对Iris和Statlog数据集应用两种聚类算法进行聚类,实验结果见表1和表2,采用聚类精确度对聚类的效果进行评价。所谓聚类精确度,即聚类结果与预定义类别一致的个体数与参与聚类的全部个体数的比,该指标容易理解,是对聚类效果进行评价的一个常用指标^[11]。

表1 两种算法对Iris数据集的聚类结果

类别	传统K-means算法		本文算法	
	实例个数	正确个数	实例个数	正确个数
setosa	50	50	50	50
versicolor	50	48	50	48
virginica	50	36	50	46
精确度/%	89.33		96.00	

1)实验结果是在多次实验后统计聚类情况计算平均值得到的,Iris聚类精确度由89.33%提高到96.00%,Statlog聚类精确度由64.42%提高到70.45%,充分证实了文中算法的有效性。

2)图1中Iris数据集的属性熵权最大值与最小值之差为0.0064,图2中Statlog数据集的属性熵权最大值与最小值之差为0.0025。从表1和表2可以看出Iris数据集的聚类精确度提高效果略优于Statlog数据集。

表2 两种算法对Statlog数据集的聚类结果

类别	传统K-means算法		本文算法	
	实例个数	正确个数	实例个数	正确个数
saab	217	124	217	143
opel	212	136	212	156
bus	218	185	218	185
van	199	100	199	112
精确度/%		64.42		70.45

综上分析可知,统计属性熵权值方法对于提高聚类精确度的幅度通常依赖于具体的数据集和属性,属性差异度越大,对提高聚类精确度的效果越显著。

4 结语

聚类搜索引擎是当前搜索引擎发展的一个趋势,提高聚类的精确度成为一个迫在眉睫必须解决的问题。本文针对传统K-means算法聚类结果的精确度问题,提出一种基于信息熵的精确属性赋权K-means聚类算法。新算法首先使用熵值法计算数据属性的权值,其次使用赋权欧氏距离对数据进行初聚类,然后通过比较各类别赋权标准差值选择高质量的初始聚类中心继续算法直到符合聚类要求。与传统K-means算法相比,新算法克服了对随机初始聚类中心的依赖性,通过简单预选聚类种子中心处理使得选出的初始聚类中心更能够体现实际分类的决策面,同时充分利用数据对象各属性的聚类作用差异度,使聚类结果更加接近数据对象的实际分类。实验结果表明利用该算法所得到的聚类结果更加精确稳定。

参考文献:

- MUATA K, BRYSO O. Towards supporting expert evaluation of clustering results using a data mining process model[J]. Information Sciences, 2010, 180(3): 414–431.
- CAO F Y, LIANG J Y, JIANG G. An initialization method for the K-means algorithm using neighborhood model [J]. Computers and Mathematics with Applications, 2009, 58(3): 474–483.
- ALIK K R. An efficient K-means clustering algorithm[J]. Pattern Recognition Letters, 2008, 29(9): 1385–1391.
- 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.
- LAI J Z C, HUANG T J, LIAW Y C. A fast K-means clustering algorithm using cluster center displacement[J]. Pattern Recognition, 2009, 42(11): 2551–2556.
- HAN J W, KAMBER M. Data mining concepts and techniques [M]. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006: 383–461.
- 孟洋,赵方. 基于信息熵理论的动态规划特征选取算法[J]. 计算机工程与设计,2010,31(17):3879–3881.
- 陈雷,王延章. 熵权法对融合网络服务质量效率保障研究[J]. 计算机工程与应用,2005, 41(23): 1–3.
- 高孝伟. 熵权法在教学评优中的应用研究[J]. 中国地质教育, 2008, 17(4): 100–104.
- UCI Machine Learning Repository [DB/OL]. [2010-12-20]. <http://archive.ics.uci.edu/ml/>.
- AHMAD A, DEY L. A k-mean clustering algorithm for mixed numeric and categorical data[J]. Data & Knowledge Engineering, 2007, 63(2): 503–527.