基于 AHP 和 CRITIC 综合赋权的 K-means 算法①

丁晓琴, 张德生

(西安理工大学 理学院, 西安 710054)

摘 要: 传统的 K-means 算法认为被分析样本的各个属性在聚类中作用是相同, 针对这种不足, 提出一种基于 AHP 和 CRITIC 综合赋权的 K-means 聚类算法. 首先利用 CV-K-means 方法计算每个属性的权重, 从而两两进行比较得到判断矩阵. 然后, 根据层次分析法(AHP)确定各个属性的主观权重, 再利用 CRITIC 方法确定各个属性的客观权重. 采用差异系数法确定组合系数, 实验证明该算法的聚类精确度高于传统 K-means 算法.

关键词: K-means; 属性权重; AHP; CRITIC

K-Means Algorithm Based on Synthetic Weighting of AHP and CRITIC

DING Xiao-Qin, ZHANG De-Sheng

(College of Science, Xi'an University of Technology, Xi'an 710054, China)

Abstract: The traditional K-means algorithm is regarded that the attributes of swatches have the same effect on the clustering analysis. Based on AHP and CRITIC, comprehensive weighting of K-means clustering algorithm is proposed to solve the problem in this paper. First, each of attribute weight is calculated by CV-K-means method, thus judgment matrix is determined by comparing the two.Then, according to the analytic hierarchy process subjective weights of attributes is determined. And using the CRITIC method the objective weight of each attribute is determined, difference coefficient method is used to determine coefficient of combination. The experimental results show that the algorithm accuracy is higher than the traditional K-means algorithm.

Key words: K-means; weight; analytic hierarchy process; criteria importance through intercriteria correlation

聚类(Clustering)是数据挖掘的一种重要技术, 是分析数据并从中发现有用信息的一种有效的手段. 聚类属于一种无监督的学习方法. K-means 算法是较为常用的聚类方法之一, K-means 算法将欧氏距离作为相似性的评价指标, 即认为两个对象的距离越近, 其相似度就越大, 反之则相似性越小. 但是这种评价指标认为样本的各个属性对聚类结果的贡献相同, 忽略了不同属性特征对聚类结果可能造成的不同影响, 这导致算法难以获得稳定而精确的聚类结果. 一个有效的解决方法是为每一个属性加一个权值, 让不同的属性在聚类中起不同作用.

国内外对 K-means 加权算法主要研究集中在两个方面, 即对 K-means 算法单一赋权和综合赋权的研究. 倪少凯(2002)[1]为了更好融合各种赋权法的优点, 采用不同类的赋权方法进行组合. 上官延华等人(2010)[2]将客观权法中的熵值法和均方差法相结合, 得出综合赋权法 K-means 算法比传统 K-means 算法精度高的结论. 原福永(2011)[3]采用熵值法对数据对象的属性赋权, 充分利用各属性对聚类的作用, 使得聚类的结果更精确稳定. 李健森(2012)[4]用新的距离代替欧氏距离来确定属性的权重, 提高了新算法的执行效率. 综合前人研究结果可知, 综合赋权法比单一赋权法更科学、合理, 而对于结合主观赋权法和客观赋权法在 K-means 算法的研究较少, 值得探讨.

本文将主观层次分析法(AHP)法和 CRITIC 法相结合, 建立了一种综合赋权 K-means 算法, 实验结果

① 收稿时间:2015-11-24;收到修改稿时间:2016-01-07 [doi:10.15888/j.cnki.csa.005267]

表明该方法的聚类精度方面优于传统 K-means 算法.

1 K-means算法

K-means 聚类算法[5]具有聚类速度快、效率高、适于处理大数据集等特点, 是一种具有较大影响力的无监督学习算法. K-means 聚类算法的原理是: 首先, 随机选取 k 个点作为初始聚类中心, 然后, 计算各个样本到聚类中心的距离, 把样本归到离它最近的那个聚类中心所在的类; 对调整后的新类计算新的聚类中心, 如果相邻两次的聚类中心没有发生任何变化, 说明样本调整结束, 这时某个误差平方和函数已经达到最小, 聚类准则函数已经收敛; 否则, 如果相邻两次的聚类中心不同, 需要继续调整全部样本来修改聚类中心, 再进入下一次的迭代过程.

K-means 算法也存在一些缺点: (1)算法对初始中心敏感; (2)算法不能处理非球形类、不同尺度和不同密度的类; (3)算法对孤立点数据和噪声数据较敏感; (4)算法对初始值 k 的选取依赖性较大; (5)算法经常陷入局部最优解, 无法得到全局最优解; (6)算法对每一个属性都同等看待, 难以产生高质量的聚类结果.

2 本文方法

2.1 AHP 方法确定属性的主观权重

层次分析法(AHP)[6-8]是一种解决多目标的复杂问题的定性与定量相结合的决策分析方法. 在主观赋权法中, AHP 法是目前使用最多、研究最多的方法, 用 AHP 法确定属性的权重, 其主观性体现在判断矩阵上, 本文将利用变异系数法先计算出各属性的权重, 然后两两比较每个属性值得出判断矩阵.

根据 AHP 法计算各属性的主观权重的步骤如下:

① 数据标准化

设有 m 个数据对象表示为 $S=\{S_1,S_2,\cdots,S_m\}$, 其属性表示为 $P=\{P_1,P_2,\cdots,P_n\}$,第 i 个数据 S_i 的第 j 个属性 P_j 的值记为 $x_{ij},i=1,2,\cdots,m,j=1,2,\cdots,n,X=\ (x_{ij})_{m\times n}$ 称为属性矩阵. 由于原属性中不同属性的数值大小差别很大, 为了便于采用各种多属性决策方法, 需要把属性值标准化, 即将数据变换到[0,1]区间上, 其过程如下式所示.

$$x'_{ij}=\frac{x_{ij}-x^{\min}_{j}}{x^{\max}_{j}-x^{\min}_{j}}$$

其中 x_{ij} 为属性值, $x^{\min}_{j}$ 是第 j 个属性的最小值, $x^{\max}_{j}$ 是第 j 个属性的最大值, $i=1,2,\cdots,m,\mathrm{j}=1,2,\cdots,n$.

② 估计判断矩阵 A

判断矩阵是表示本层所有因素针对上一层某个因素的相对重要性的比较, 判断矩阵的元素 a_{ij} 通过变异系数法来给出. 变异系数其含义是一组数据的变异指标与其平均指标之比, 即标准差与平均值的比值, 记为 CV .

$$CV=\frac{\sigma}{\mu}$$

其中 σ 是标准差, μ 是平均值.

设 $W=\{w_1,w_2,\cdots,w_n\}$ 是属性集 P 的 n 个属性的权重的集合, 则第 j 个属性的权重 w_j 为:

$$w_j=\frac{CV_j}{\sum_{j=1}^{n}CV_j}\qquad 0\le w_j\le 1$$

将各属性的权重值两两进行比较得到判断矩阵 A 为:

$$A=\begin{bmatrix}\frac{w_1}{w_1} & \frac{w_2}{w_1} & \cdots & \frac{w_n}{w_1}\\ \frac{w_1}{w_2} & \frac{w_2}{w_2} & \cdots & \frac{w_n}{w_2}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{w_1}{w_n} & \frac{w_2}{w_n} & \cdots & \frac{w_n}{w_n}\end{bmatrix}$$

其中 $\frac{w_t}{w_s}(t=1,2,\cdots,n,s=1,2,\cdots,n)$ 是第 t 个属性的权重值 w_t 与第 s 个属性的权重值 w_s 值的比值, 表示判断矩阵 A 中元素 a_{ts} 的取值.

③ 判断矩阵 A 的一致性检验

判断矩阵 $A(a_{ij})_{m\times n}$ 满足一致性条件为: $a_{ij}=a_{ik}\cdot a_{kj}$, 由于判断矩阵数值是根据客观数据、专家意见和分析者的认识综合平衡后给出的, 因而难免会造成判断矩阵不满足一致性, 所以必须对判断矩阵进行一致性检验, 可以通过下式来检验: $CI=\frac{\lambda_{\max}-n}{n-1}$.

当判断矩阵的阶数比较大的时, 用平均随机一致性指标 RI 来修正 CI , 即 $CR=\frac{CI}{RI}$, 其中 $CI=\frac{\lambda_{\max}-n}{n-1}$, RI 是平均一致性指标, 是足够多个根据随机发生的判断矩阵计算的一致性指标的平均值, n 为判断矩阵的阶数. 一般而言 CR 越小, 判断矩阵的一致性越好, 通常认为当 CR 满足<0.1 时, 则可以认为该判断矩阵可以接受; 若 $CR\ge 0.1$, 则需要对判断矩阵进行相应修正, 直到 CR 满足条件为止.

④ 计算权重

在对判断矩阵 A 进行一致性检验之后，数据的属性通过(1)求得其权重:

$$w_j''' = \sqrt[n]{\prod_{i=1}^{m} a_{ij}} \quad j=1,2,\cdots,n \tag{1}$$

再按照式(2)归一化处理:

$$w_j' = \frac{w_j'''}{\sum_{j=1}^{n} w_j'''} \quad j=1,2,\cdots,n \tag{2}$$

得到数据各属性的主观权重向量 $w'=(w_1',w_2',\cdots,w_j')^T$.

2.2 CRITIC 方法确定属性的客观权重

CRITIC[9-11]是一种客观赋权法. 它的基本思路:确定指标的客观权重以两个基本概念为基础. 一是对比强度，它表示同一指标各个评价方案取值差距的大小，以标准差的形式来表现，即标准差的大小表明了在同一个指标内各个方案的取值差距大小，标准差越大各方案的取值差距越大. 二是评价指标之间的冲突性，指标之间的冲突性是以指标之间的相关性为基础，如两个指标之间具有较强的正相关，说明两个指标冲突性较低. 第 j 个指标与其它指标的冲突性的量化指标为 $\sum_{i=1}^{m}(1-r_{ij})$，其中 r_{ij} 是评价指标 i 和 j 之间的相关系数. 各个指标的客观权重确定就是以对比强度和冲突性综合衡量的，设 C_j 表示第 j 个指标所含的信息量，则 C_j 可表示为: $C_j=\sigma_j\sum_{i=1}^{m}(1-r_{ij}), j=1,2,\cdots,n, C_j$ 越大，第 j 评价指标所含的信息量越大，该指标的相对重要性也就越大，所以第 j 个指标的客观权重 w_j'' 应为:

$$w_j'' = \frac{C_j}{\sum_{j=1}^{n} C_j} \quad j=1,2,\cdots,n$$

2.3 确定属性的综合权重

由于 AHP 法主观性太强，忽略实际的数据信息，虽然 CRITIC 法能体现客观数据的信息，但它受原始数据质量的影响，且忽略了专家积累的经验信息和属性间的实际轻重关系，对样本区分度不佳. 为了既体现主观权重，又体现客观权重，建立确定组合权重中的系数 α 和 β 的最优化模型，令综合权重向量为 $w=\alpha w'+\beta w''$，其中 α,β 为主客观组合进行赋权的待定系数，满足: $\alpha,\beta\geq 0, \alpha+\beta=1$.

本文采用差异系数法[12]来确定，方法如下:

$\alpha=\frac{n}{n-1}\cdot T'$，其中 T' 为 w' 各分量的差异系数，$T'=\frac{2}{n}(1p_1+2p_2+\cdots+np_n)-\frac{n+1}{n}$，$p_1,p_2,\cdots,p_n$ 是主观权重向量 w' 中各分量从小到大的重新排列，n 为属性的个数. 则 $\beta=1-\alpha$, 代入 $w=\alpha w'+\beta w''$ 得到各属性的综合权重向量 w .

2.4 算法描述

基于 AHP 和 CRITIC 综合赋权的 K-means 算法步骤如下:

输入: 待聚类数据集 X，聚类个数 k

输出: k 个聚类，聚类的准确率 f

① 按照公式 $x_{ij}'=\frac{x_{ij}-x^{\min}_j}{x_j^{\max}-x_j^{\min}}$ 对数据集进行标准化;

② 随机选取 k 个初始中心;

③ 根据 $CV=\frac{\sigma}{\mu}$ 计算各属性的权重，然后进行两两比较得到判断矩阵 A;

④ 利用层次分析法 AHP 计算数据各属性的主观权重值 w';

⑤ 采用 CRITIC 法计算数据各属性的客观权重值 w'';

⑥ 建立组合系数 α,β 的优化模型，得到综合权重值 w;

⑦ 扫描所有的数据对象，根据其与聚类中心的相似度，将其归入与其最相似的类中;

⑧ 重新计算每个类的中心;

⑨ 计算出各类的准则函数值，若不满足要求，则需重新聚类;

⑩ 扫描所有的数据的聚类结果，计算聚类结果的准确度 f.

3 实验及结果分析

3.1 实验描述

为了验证算法改进的效果，对传统 K-means 算法、基于 CRITIC 法加权 K-means 算法和本文算法进行对比实验. 实验选用 Matlab R2012a 作为编程工具，实验环境如下，操作系统 Windows 7; CPU:Inter® core(TM)i5-3470@3.20GHZ; 内存:4GB. 采用 UCI 数据库中的 4 个常用数据集作为测试数据, UCI 数据库是一个专门用于测试机器学习、数据挖掘算法的公共数据库，库中的数据都有确定的分类，因此可以用准确率

直观地表示聚类的质量. 为评价数据聚类的效果, 本文采用准确率和聚类熵(Entropy)[13]作为衡量聚类效果的度量标准. 各数据的特征如表 1 所示.

表 1 数据集

数据集	样本个数	属性个数	聚类个数
Iris	150	4	3
Wine	178	13	3
Haberman	306	4	2
Contraceptive	1473	9	3

3.2 实验结果

实验结果如表 2 和图 1 所示. 表 2 是聚类精确度测试结果, 图 1 是聚类熵测试结果.

表 2 聚类结果精确度

数据集	传统 K-means 算法	基于 CRITIC 法加权 K-means 算法	本文改进算法
Iris	88%	89.33%	95.33%
Wine	74.72%	94.38%	95.51%
Haberman	51.96%	54.90%	55.23%
Contraceptive	45.01%	45.42%	45.42%

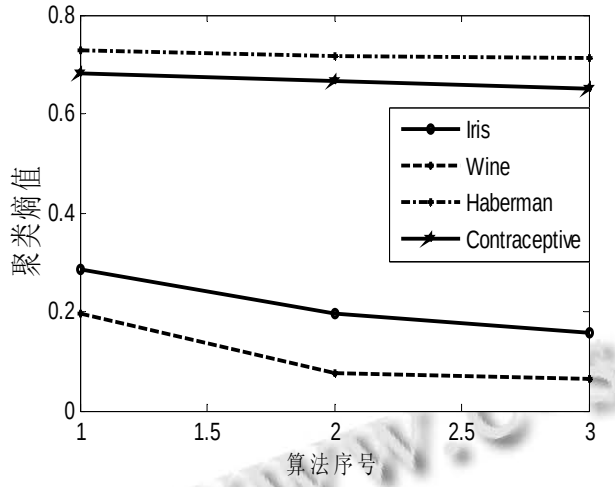

图 1 聚类熵

3.3 实验结果分析

从以上的实验结果来看, 随着选取的聚类算法不同, 产生的聚类效果也不相同. 从表 2 中可以看出, 在 Iris 数据集中, 聚类精度由 88%分别提高到 89.33%和 95.33%. Wine 数据集有 14 个属性, 各个属性的取值范围差距较大. 所以在传统 K-means 算法中精确度只有 74.72%, 在基于 CRITIC 加权 K-means 算法中, 在计算距离时, 通过 CRITIC 法给每个属性赋予不同的权重, 这样各属性的重要程度就可以表现出来, 所以精确度达到 94.38%. 本文的算法综合实际数据的信息和属性实际的轻重程度, 赋予各属性综合权重, 精确度达到 95.51%, 明显优于之前两种算法. 应用改进后的算法, 由于给每个属性赋一个综合权值, 所以在计算距离过程中能得到精确的样本的相似度, 在第 3 和 4 组数据集中, 加权后的准确率略有提高.

从图 1 可以发现, 各数据集在传统 K-means 算法中的聚类熵都比较大, 随着应用基于 CRITIC 法加权 K-means 算法和本文改进算法各数据集的聚类熵都趋于下降的趋势, 而本文算法的聚类熵指标远远优于其它算法, 熵值越小, 说明聚类效果越好.

4 结论

由于传统 K-means 算法没有考虑不同属性对聚类结果可能造成的不同影响, 本文建立了一种基于 AHP 和 CRITIC 综合赋权的 K-means 算法. 引入属性权重以提高聚类的准确率, 将 AHP 法和 CRITIC 法权重计算结果根据差异系数法对数据属性进行综合赋权处理. 实验结果表明:基于 AHP 和 CRITIC 综合赋权的 K-means 算法能够充分体现各个属性在聚类中的重要程度, 采用主观赋权法和客观赋权法结合, 综合赋权结果更科学, 弥补了单一赋权的缺点, 有较好的聚类效果.

参考文献

1 倪少凯.7 种确定评估指标权重方法的比较.华南预防医学, 2002,28(6):54−55.

2 上官延华,冯荣耀,刘宏川.一种基于熵和均方差综合赋权的 K-means 算法.计算机与现代化,2010,(4):34−36.

3 原福远,张晓彩,罗思标.基于信息熵的精确属性赋权 K-means 聚类算法.计算机应用,2011,31(6):1675−1677.

4 李健森,百万民.一种改进的距离度量的聚类算法.电子设计工程,2012,20(20):86−88.

5 Tan PN, Steinbach M, Kumar Vipin. 范明,范宏建等译.数据挖掘导论.北京:人民邮电出版社,2011.

6 T Saaty. The Analytic Hierarch Porcess. New York: McGraw-Hillinc, 1980.

7 袁能文.上市公司经营业绩综合评价模型优化研究[学位论文].长沙:湖南大学,2010.

8 麻小娟,张继荣.基于层次分析法和熵理论的网络选择算法. 陕西科技大学学报,2014,32(3):163−164.

9 Diakoulaki D, Mavrotas G, Papayannakis L. Determining objective weights in multiple criteria problems: The CRITIC mehtod. Computer Ops Res, 1995, (22): 763−770.

10 袁和才,辛艳辉.基于 AHP 和 CRITIC 方法的水资源综合效益模型.安微农业科学,2011,39(4):2225−2229.

11 张玉,魏华波.基于CRITIC的多属性决策赋权方法.方法应用,2012(16):75−77.

12 席荣宾,黄鹏,赖雪梅,郑巧凤.组合赋权法确定权重的方法探讨.中国集体经济·学术探讨,2010,(7):75−76.

13 Tsai CY, Chiu CC. Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm. Computational Statistics & Data Analysis, 2008, 52(10): 4658−4672.