模糊 K-Harmonic Means 聚类算法

赵 恒，杨万海，张高煜
(西安电子科技大学 电子工程学院，陕西 西安 710071)

摘要： 对 K-Harmonic Means 算法进行扩展，考虑到数据点对不同类的隶属关系，将模糊的概念应用到聚类中，提出了模糊 K-Harmonic Means 算法，推导出聚类中心和模糊隶属度的迭代公式. 在中心迭代聚类算法统一框架的基础上，推导出 FKHM 算法聚类中心的条件概率表达式以及在迭代过程中的数据加权函数表达式. 最后，用 Folkes & Mallows 指标对聚类结果进行评价. 实验表明，模糊 K-Harmonic Means (KHM) 算法在聚类对于初值不敏感的同时提高了聚类结果的精确度，达到较好的聚类效果.

关键词： 模糊 K-Harmonic Means 聚类；聚类中心；条件概率；Folkes & Mallows 指标
中图分类号： TP18 **文献标识码：** A **文章编号：** 1001-2400(2005)04-0603-04

Fuzzy K-Harmonic Means clustering algorithm

ZHAO Heng, *YANG Wan-hai*, *ZHANG Gao-yu*
(School of Electronic Engineering, Xidian Univ., Xi'an 710071, China)

Abstract: Considering the fact that data belong to several clusters to some extent, we import the fuzzy membership of data to clustering analysis, and propose the fuzzy K-Harmonic Means clustering (FKHM) algorithm. The iterative expressions for cluster center and fuzzy membership are deduced respectively. We then describe a unified expression for the iteration of centers, and deduce the conditional probability expression for the centers and data weight functions for FKHM. Finally, the Folkes & Mallows index is used to evaluate the clustering result. Experiment indicates that the fuzzy K-Harmonic Means algorithm can not only overcome the sensitivity to the initial centers, but also improve the quality of clustering results, compared with the K-Harmonic Means.

Key Words: fuzzy K-Harmonic Means; cluster center; conditional probability; Folkes & Mallows index

在过去有关聚类问题的研究中，基于划分的聚类算法是聚类研究中一个重要领域，传统的基于划分的聚类对于初始值十分敏感，不同的初始化有可能导致完全不同的聚类结果. 目前解决初始化问题的方法有：一是进行多次聚类[1]，但计算量很大；二是研究单独的初始化算法，尽可能得到比较好的初始值，然后用得到的结果进行聚类[2~4]；三是研究对初始值不十分敏感的聚类算法，比如模糊 K-Means (FKM) 算法[5]，K-Harmonic Means (KHM) 算法[6] 以及利用遗传算法、神经网络进行聚类等.

笔者对 KHM 算法进行扩展，考虑数据点同时对不同类的隶属关系，将模糊的概念应用到 KHM 聚类中，提出了模糊 KHM 算法，解决某些数据点“在某种程度上既属于 A 也属于 B”的问题，既保留了 KHM 聚类对于初值不十分敏感的特性，同时也提高了聚类精度.

1 K-Harmonic Means 聚类算法

KHM 算法同 KM 算法类似，也是基于中心的迭代过程，不同的是它采用所有数据点到每个聚类中心的调和平均值的和作为目标函数. 假定 $X = \{X_1, X_2, \cdots, X_n\}$ 是一组数据元组，其中 $X_i = [x_{i1}, x_{i2}, \cdots, x_{im}]$ 表示具有 m 个属性的数据对象. 数据划分为 k 类. 目标函数为

收稿日期：2004-08-16
作者简介：赵 恒(1975-)，男，西安电子科技大学博士研究生.

$$E_{\mathrm{KHM}} = \sum_{i=1}^{n} \left(k \Big/ \left(\sum_{l=1}^{k} \left(1/d^2(X_i, C_l) \right) \right) \right) \quad , \tag{1}$$

这里 $C_l = [c_{l1}, c_{l2}, \cdots, c_{lm}]$ 是聚类中心,$d(X_i, C_l)$ 是距离测度,简写为 $d_{i,l}$,采用欧氏距离. 聚类中心的更新方法如下:

$$C_l = \sum_{i=1}^{n} \left(1 \Big/ \left(\sum_{j=1}^{k} \left(d_{i,l}^2 / d_{i,j}^2 \right) \right)^2 \right) X_i \Big/ \left(\sum_{i=1}^{n} \left(1 \Big/ \left(\sum_{j=1}^{k} \left(d_{i,l}^2 / d_{i,j}^2 \right) \right)^2 \right) \right) \quad . \tag{2}$$

它根据初始值不断迭代,使得式(1)不断减小直到稳定.

相对于 KM 算法,KHM 用数据点与所有聚类中心距离的调和平均替代了数据点与聚类中心的最小距离,从而引入了聚类中心对数据点的条件概率和每次迭代过程中数据点的动态加权,实际上起到了将“硬”聚类“软化”的作用.

2 模糊 K-Harmonic Means 聚类算法

FKHM 应用数据点对不同类的隶属度对目标函数(1)中的距离测度进行模糊加权. 其目标函数为

$$E_{\mathrm{FKHM}} = \sum_{i=1}^{n} \left(k \Big/ \left(\sum_{l=1}^{k} \left(1/(\omega_{l,i}^{\alpha} d^2(X_i, C_l)) \right) \right) \right) \quad , \tag{3}$$

其中 $\omega_{l,i} \in [0, 1]$ 是划分矩阵 $W_{k\times n}$ 的一个元素,它表示对象 X_i 从属于类 l 的程度,即隶属度,$\sum_{l=1}^{k} \omega_{l,i} = 1$,$\alpha \geqslant 0$ 是模糊因子. 从而可推导出模糊隶属度和聚类中心的更新公式为

$$\omega_{l,i} = \left(1/d_{i,l}^2 \right)^{1/(\alpha+1)} \Big/ \left(\sum_{j=1}^{k} \left(1/d_{i,l}^2 \right)^{1/(\alpha+1)} \right) \quad , \tag{4}$$

$$C_l = \sum_{i=1}^{n} \left(\omega_{l,i}^{\alpha} \Big/ \left(\sum_{j=1}^{k} \left(\omega_{l,i}^{\alpha} d_{i,l}^2 / (\omega_{j,i}^{\alpha} d_{i,j}^2) \right) \right)^2 \right) X_i \Big/ \left(\sum_{i=1}^{n} \left(\omega_{l,i}^{\alpha} \Big/ \left(\sum_{j=1}^{k} \left(\omega_{l,i}^{\alpha} d_{i,l}^2 / (\omega_{j,i}^{\alpha} d_{i,j}^2) \right) \right)^2 \right) \right) \quad . \tag{5}$$

算法描述为 初始化聚类中心,由式(4)和式(5)进行迭代,直到达到最大迭代次数或目标函数达到足够小,最后得到聚类中心和隶属度矩阵. 算法复杂度为 $O(n * k * m * t)$,其中 t 为迭代次数,m 为数据属性的个数. 在 FKHM 中,如果模糊因子 $\alpha = 0$,即不考虑模糊隶属度,算法 FKHM 变为 KHM.

3 FKHM 算法在中心迭代统一框架下的描述

基于中心迭代的聚类算法,如 KM,FKM,KHM,FKHM 等,都是通过迭代的方法得到聚类中心,同时使得目标函数达到局部最优. 由它们各自聚类中心的迭代公式,可以定义一个基于中心迭代的统一框架来表示它们[7].

定义 1 条件概率 $p(C_l \mid X_i)$ 表示数据点 X_i 属于以 C_l 为中心的类 l 的可能性,有 $P(C_l \mid X_i \geqslant 0$,$\sum_{l=1}^{k} p(C_l \mid X_i) = 1$. 这里,条件概率和模糊隶属度在数学形式上有相似之处,但它们是不同的两个概念. 条件概率表示的是可能性,数据点依概率属于不同的类,但同一时间某个数据点只属于某个类,即或者属于“A”或者属于“B”. 模糊隶属度表示的是一种模糊关系,数据点可以在同一时间属于不同的类,只是程度不同而已,即在某种程度上既属于“A”也属于“B”.

定义 2 权函数 $a(X_i)$ 表示每次迭代过程中,数据点对聚类中心的影响程度,$a(X_i) > 0$.

定义 3 基于中心迭代的聚类,其迭代公式为

$$C_l = \sum_{i=1}^{n} p(C_l \mid X_i)\ a(X_i)\ X_i \Big/ \left(\sum_{i=1}^{n} p(C_l \mid X_i)\ a(X_i) \right) \quad . \tag{6}$$

所以,有如下的聚类算法:(1)初始化聚类中心;(2)计算每个聚类中心对每个数据点的条件概率,以及迭代权函数;(3)由定义 3 计算新的聚类中心,如果是模糊聚类,还要根据隶属度函数计算隶属度;(4)重复(2),(3),直到收敛或达到最大迭代次数.

对于 FKHM 算法,由式(5),(6)可以推导出条件概率和权函数:

$$p_{\mathrm{FKHM}}(C_l \mid X_i) = (1/d_{i,l}^2)^{(\alpha+2)/(\alpha+1)} \Big/ \Big(\sum_{j=1}^{k} (1/d_{i,j}^2)^{(\alpha+2)/(\alpha+1)}\Big) \quad , \tag{7}$$

$$a_{\mathrm{FKHM}}(X_i) = \sum_{j=1}^{k} (1/d_{i,j}^2)^{(\alpha+2)/(\alpha+1)} \Big/ \Big(\sum_{j=1}^{k} (1/d_{i,j}^2)^{1/(\alpha+1)}\Big)^{(\alpha+2)} \quad , \tag{8}$$

其中 $\alpha \geqslant 0$ 是模糊因子. 可看到,FKHM 聚类中心的条件概率,随数据点与聚类中心的距离增大而减小,而权函数的引入使迭代过程中数据有一个动态的加权,动态加权取不同的值或函数,可得到 FKHM 的不同变形.

4　算法仿真

4.1　聚类结果评价方法

一般评价聚类结果用到的‘误分率’等统计方法是建立在聚类结果和输入样本的原始分类结构一一对应的基础上的,但聚类是无监督学习算法,其结果与输入样本原始分类结构并不一定有明显的对应关系,因此用 Folkes & Mallows(FM)指标[8] 来评价其结果.

假设原始数据分类是 $C = \{C_1, C_2, \cdots, C_k\}$,聚类算法所得到的结果是 $C' = \{C_1', C_2', \cdots, C_k'\}$,$C_l$ 和 C_l' $(l = 1,2,\cdots,k)$ 分别表示类 l 中的数据集. 元素 a_{ij} 表示 C_i 和 C_j' 中相同数据对象的个数.

对于原始数据集中的任何数据对 (x_i, x_j),$i, j = 1,2,\cdots,n$,一定具有以下 4 种情况之一:数据对 (x_i, x_j) 在 C 中属于同一类并且在 C' 中也属于同一类;数据对 (x_i, x_j) 在 C 中属于同一类但在 C' 中不属于同一类;数据对 (x_i, x_j) 在 C 中不属于同一类但在 C' 中属于同一类;数据对 (x_i, x_j) 在 C 中不属于同一类并且在 C' 中也不属于同一类.

假定以上 4 种情况数据对的个数分别是 ss, sd, ds, dd,它们与聚类结果表中的元素有如下关系:$ss = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}(a_{ij} - 1)/2$,$sd = \sum_{i=1}^{k}\sum_{j=1}^{k-1} a_{ij}\big(n_i - \sum_{l=1}^{j} a_{il}\big)$,$ds = \sum_{j=1}^{k}\sum_{i=1}^{k-1} a_{ij}\big(m_j - \sum_{l=1}^{i} a_{li}\big)$,$dd = N - ss - sd - ds$,其中 n_i, m_j 分别是 C_i 和 C_j' 中数据的个数,而总的数据对的个数是 $N = n(n-1)/2$,n 是数据的个数,$n = \sum_{i=1}^{k} n_i = \sum_{j=1}^{k} m_j$. FM 为 $ss((1/(ss + sd)) \cdot (1/(ss + ds)))^{1/2}$ 在 0 和 1 之间取值,其值越大表明划分 C' 和 C 越相似.

4.2　实验仿真

为了测试 FKHM 算法的聚类性能,采用 IRIS 实际数据作为测试样本进行实验仿真. IRIS 数据由四维空间中的 150 个数据点组成,每一个数据的 4 个分量表示 IRIS 的 4 个属性. 整个数据集包含 3 个 IRIS 种类,每类各有 50 个数据,其中一类与其他两类有较好的分离,而另外两类之间存在交迭. IRIS 数据的实际聚类中心位置分别为:(5.00　3.42　1.46　0.24),(5.93　2.77　4.26　1.32),(6.58　2.97　5.55　2.02).

(1) 分别用 FKM,KHM 和 FKHM 算法对 IRIS 数据进行聚类 100 次,每次 3 种算法采用相同的随机初始值,比较其性能. 实验过程中,模糊算法的模糊因子 $\alpha = 1.2$,迭代次数是 50 次.

聚类结果如图 1 所示,在随机初始值的 100 次聚类中,FKM 尽管采用了模糊的思想,但其结果对初值的选取依赖性依然相当大,有 20% 的 FM 指标非常低,而 KHM 尽管 FM 指标大部分时候要比 FKM 稍差,但它和 FKHM 对初值的选取不敏感,聚类结果非常稳定,几乎不随初值的不同而变化. 而且,FKHM 算法的平均 FM 指标要高于 FKM 和 KHM,这说明它所得到的聚类结果最接近于原始数据的类别分布,而且,从表 1 也可以看出,FKHM 算法所得到的聚类中心最接近于实际值.

因此有以下结论,FKHM 聚类算法对初值选取不太敏感,而且相对于 FKM 和 KHM,它提高了算法的精确度,能够得到与数据的自身类别分布更加相似的聚类结果.

(2) 改变模糊因子的值,令 α 分别等于 0.2,0.6,1.2,2,4,8,16,用 FKHM 对 IRIS 数据进行聚类,测试 α 的变化对 FKHM 算法性能的影响.

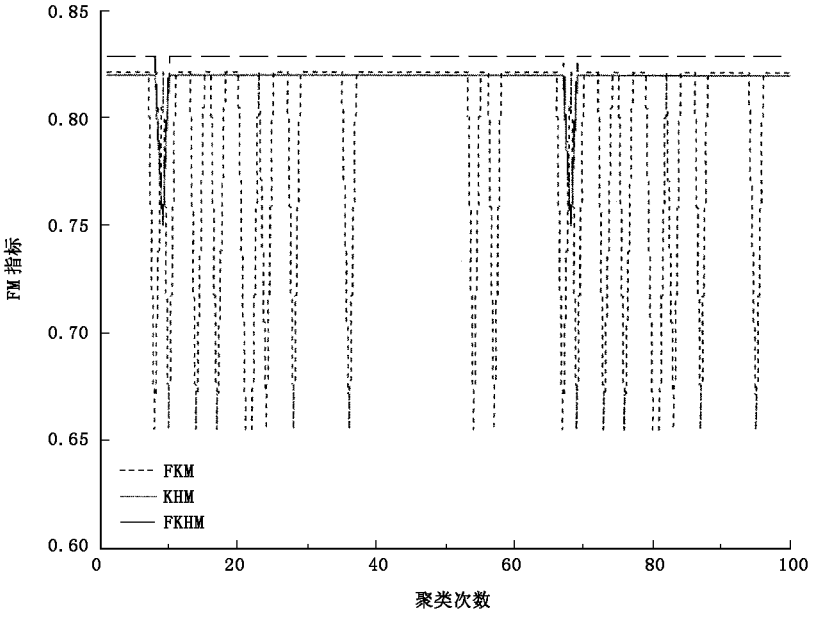

图 1 3 种算法各聚类 100 次的 FM 指标

表 1 3 种算法对 IRIS 数据聚类结果

聚类算法	FM	聚类中心				与实际中心的误差平方和
FKM	0.82081	5.0061	3.4157	1.4678	0.2460	0.1459
		5.8922	2.7449	4.3904	1.4272	
		6.8468	3.0742	5.7250	2.0705	
KHM	0.81967	5.0036	3.4030	1.4850	0.2515	0.0750
		5.8892	2.7612	4.3643	1.3974	
		6.7751	3.0524	5.6469	2.0536	
FKHM	0.82847	5.0014	3.3879	1.4953	0.2518	0.0399
		5.9152	2.7964	4.3829	1.3982	
		6.6839	3.0351	5.5370	2.0313	

从表 2 可看出,FKHM 算法的目标函数值随着模糊因子 α 的增大而减小,FM 指标逐渐增大而趋于稳定. 聚类中心与实际中心的误差平方和随着 α 的增大先减小,当 α 增大到 1.2 附近时,误差平方和达到极小点,而后随着 α 的增大逐渐增大. 由于当 α 的值非常大时,聚类结果太模糊而失去划分特性,继续增大 α 已不能改善算法的性能,因此 α 的取值应在 0 ~ 4 之间.

表 2 α 的变化对 FKHM 算法性能的影响

α	0.2	0.6	1.2	2.0	4.0	8.0	16.0
E_{FKHM}	161.2003	121.2135	73.2955	34.6385	4.5113	0.0623	0.0000
FM	0.8197	0.8285	0.8285	0.8376	0.8266	0.8285	0.8279
聚类中心与实际中心的误差平方和	0.062571	0.04742	0.039887	0.04336	0.077525	0.14854	0.21208

5 结 论

将模糊的概念应用到 KHM 聚类中,考虑到数据点同时对不同类的隶属关系,提出了 FKHM 算法,解决某些数据点‘ 在某种程度上既属于 A 也属于 B ”的问题. 实验表明,模糊 K-Harmonic Means 算法在聚类对于初值不敏感的同时提高了聚类结果的精确度,达到较好的聚类效果.

(下转第 638 页)