改进的 FCM 聚类算法在 Weka 平台的应用①

王 晶, 于威威

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 模糊 C-均值聚类算法是目前应用最广泛的聚类算法, 但其仍然存在对孤立点敏感及对初始中心点依赖等问题. 为此, 提出了一种改进的基于样本加权的模糊聚类算法, 该算法可以更加准确的获得初始中心点且去除噪声点. 同时, 针对 Weka 系统中聚类算法的薄弱性以及聚类问题在数据挖掘领域的广泛性, 本文对此平台进行二次开发并对传统 FCM 算法与改进算法进行研究. 研究发现, 改进算法使得聚类结果稳定, 且能准确获得聚类结果, 提高了算法准确率.

关键词: 模糊 C-均值聚类算法; 孤立点; 初始中心点; Weka

Application of an Improved FCM Clustering Algorithm on Weka Platform

WANG Jing, YU Wei-Wei

(Information Engineering College, Shanghai Maritime University, Shanghai 201306, China)

Abstract: The fuzzy C-means clustering algorithm is the most widely used clustering algorithm, but it still remains sensitive to outliers and dependent on initial centers and other issues. Therefore, this paper presents an improving fuzzy clustering algorithm based on sample weighting, the algorithm can get more accurate initial center points and remove noise. At the same time, to the weakness of the clustering algorithm in Weka system and the clustering problem is extensive in the field of data mining, this paper makes the platform the secondary development, researches the traditional FCM algorithm and improving algorithm. The study finds that the improving algorithm makes the clustering results stable, obtain the accurate clustering results and improve the accuracy of the algorithm.

Key words: fuzzy C-means clustering algorithm; outliers; initial center point; Weka

聚类[1]是数据挖掘的一个重要研究方向, 就是将数据对象分类, 同一类内的数据对象尽可能相似, 而不同类尽可能相异. 为了对数据对象进行聚类, 目前大量经典的算法涌现, 其中最受欢迎的是模糊 C-均值聚类算法(Fuzzy C-means, FCM), 该算法描述了一个样本对象可能同时属于几个簇, 这种属于不同簇的程度用模糊隶属度函数来描述[2,3]. 由于此模糊性质对现实世界的反映更加客观, 因此获得了很多学者对其的广泛研究.

但是模糊 C 均值算法还存在一定的问题, 该算法对数据集进行等级划分, 即对隶属度进行归一化约束, 这又与现实世界的数据分布不相符合. 显然, 此种均匀对称分布将会导致噪声敏感问题, 且当数据集密度程度相差较大时, 聚类结果准确率也会大大降低. 因此, 许多针对 FCM 的改进算法先后被设计出来解决 FCM 算法的不足. 其中基于密度加权的模糊 C 聚类(DWFCM)、基于样本加权的可能性模糊聚类等算法可以很好的解决对等级划分数据及噪声数据敏感的问题[4-7]. 但是这些算法都没有考虑初始簇中心的选择问题, 还有内存存储等综合性问题.

Weka, 全名为怀卡托智能分析环境(Waikato Environment for Knowledge Analysis), 是现今最完备的数据挖掘工具之一. Weka 是用 java 语言实现并提供了适用于任意数据集的数据预处理以及算法性能评估

① 基金项目:上海海事大学校基金(20120109)

收稿时间:2015-03-11;收到修改稿时间:2015-04-26

的方法,具有很强的扩展性和兼容性. 但是, 在 Weka 系统中聚类算法的薄弱性以及聚类问题在数据挖掘领域的广泛性, 因此聚类算法的研究对于数据挖掘领域具有一定的意义[8,9].

本文针对以上问题提出了一种改进的基于样本加权的FCM算法. 首先该算法利用样本间相异度性质去除孤立点, 获得准确的初始中心点[10]. 若初始中心点更接近于初始簇中心, 那么将会大大降低迭代的次数和聚类复杂性及算法稳定性. 再者通过簇中心的周围密度很大的特点可知每个样本对分类的贡献程度不同, 针对此特点本文提出基于样本权值的改进聚类算法IFCM(Improved Fuzzy c-means), 大大提高算法准确率. 另外, 本文将对 Weka 平台进行二次开发, 嵌入 FCM 及改进算法, 扩充了 Weka 聚类算法, 并将两种算法进行实验分析及比较.

1 FCM算法的基本思想

1969 年, Ruspini 率先在数据集合中提出了模糊划分概念[11]. Dunn 在 1974 利用 Ruspini 的定义将硬 C-均值聚类(HCM)算法引进到模糊集中,提出了 FCM 算法[12]. 模糊 C 聚类算法属于一种软聚类算法, 此类算法则是用隶属度约定每个数据点属于某个聚类的程度.

FCM 把 n 个向量 x_i (i=1,2,…,n)分为 k 个模糊组, 并求每组的聚类中心, 使得非相似性指标的价值函数达到最小. FCM 与 HCM 的主要区别在于 FCM 用模糊划分, 使得每个给定数据点用值在[0, 1]间的隶属度来确定其属于各个组的程度. 与引入模糊划分相适应, 隶属矩阵 U 允许有取值在[0, 1]间的元素. 不过, 加上归一化规定, 一个数据集的隶属度的和总等于 1:

$$\sum_{i=1}^{k} u_{ij} = 1, \forall j = 1,...,n \tag{1}$$

$$J(U,V) = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{2}$$

这里 u_{ij} 在 0 到 1 之间; k_i 为第 i 类的聚类中心, d_{ij} =‖v_i - x_j‖为第 I 个聚类中心与第 j 个数据点间的欧几里德距离; 且 m(m>0)是一个加权指数.

构造如下新的目标函数, 可求得使(2)式达到最小值的必要条件:

$$\begin{aligned}\bar{J}(U,V,\lambda_1,...,\lambda_n) &= J(U,V) + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{k} u_{ij} - 1\right) \\ &= \sum_{i=1}^{k}\sum_{j}^{n} u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{k} u_{ij} - 1\right)\end{aligned} \tag{3}$$

这里 λ_j , j=1…n, 是(1)式的 n 个约束式的拉格朗日乘子. 对所有输入参量求导, 使式(2)达到最小的必要条件为:

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{4}$$

和

$$u_{ij} = \frac{1}{\sum_{b=1}^{k}\left(\frac{d_{ij}}{d_{bj}}\right)^{2/(m-1)}} \tag{5}$$

由上述两个必要条件, 模糊 C 均值聚类算法是一个简单的迭代过程. 在批处理方式运行时, FCM 用下列步骤确定聚类中心 V 和隶属矩阵 U:

步骤 1: 用值在 0, 1 间的随机数初始化隶属矩阵 U, 使其满足式(1)中的约束条件

步骤 2: 用式(4)计算 k 个聚类中心 v_i , i=1,…,k.

步骤 3: 根据式(2)计算价值函数. 如果它小于某个确定的阀值, 或它相对上次价值函数值的改变量小于某个阈值, 则算法停止.

步骤 4: 用(5)计算新的 U 矩阵. 返回步骤 2.

上述算法也可以先初始化聚类中心, 然后再执行迭代过程. 该方法存在着缺点,对噪声敏感, 聚类结果依赖于初始中心点. 因此, 初始中心点选择不合适将会大大降低算法准确率, 而且降低了算法的性能.

2 改进的FCM算法

2.1 样本相异度

传统的FCM算法中, 我们通常是随机选取聚类中心, 而初始中心的正确选取对于聚类结果的影响较大, 如果能准确找到簇中心, 可以加快迭代. 另外, 若是提高聚类中心点的准确率, 则也会很大程度上提高本文改进算法的准确率.

相异度应用于去除孤立点挖掘算法之中, 从数据间属性的标准差来看, 其标准差越大, 则数据间的相

异度越大. 因此可以通过采用数据相异度来判断数据之间的差异度. 利用相异度去除孤立点以及寻找初始中心点, 可以减少与真正中心点的差异度[13,14].

相异度矩阵是对象结构的一种数据表达方式, 首先基于两个对象间的距离来计算相异度, 构造一种相异度矩阵.

设样本数据的集合: X={x_1, x_2, …, x_n}. K 个初始聚类中心: {m_1, m_2, … m_i}.

样本与样本间的相异度:

$$r_{ij} = \sum_{k=1}^{p} \frac{\sqrt{\left|x_{ik} - x_{jk}\right|^2}}{\max x_{R_k} - \min x_{R_k}} \tag{6}$$

x_{ik} 和 x_{jk} 是样本点 x_i 与 x_j 的第 k 个属性值, x_{R_k} 为第 k 个属性的所有取值. P 表示每个样本点 x_i 均有 p 个属性.

由于 $x_{ij} = x_{ji}$, 则样本间的相异度矩阵可以表示为三角矩阵:

$$\mathrm{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & & \\ & & \cdot & \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{7}$$

其中 $r_{ii} = r_{jj} = 0$.

$$AvgR_i = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} \tag{8}$$

计算样本的平均相异度, 其中 n 为样本总数, r_{ij} 为两样本间的相异度.

$$AvgX_i = \frac{1}{c} \sum_{j=1}^{c} r_{ik} \tag{9}$$

以 x_i 为中心, $AvgR_i$ 为半径, 计算此区域内 x_i 与其它所有样本点的相异度平均数. 其中 c 为此区域内其它样本的个数, r_{ik} 为 x_i 与其它样本间的相异度. 若是区域内平均相异度越小, 证明此区域内相似度的样本越高.

2.2 样本加权方法的改进

在传统的FCM算法中, 默认为每个样本对于聚类所做出的贡献是同等的, 而这样将会增大噪声点对聚类结果的影响. 考虑到文献[5]中基于样本加权的概念, 通过每个样本与周围样本间的接近程度为样本赋予权值或者样本对各个中心的接近程度, 但是其计算量较大以及对初始中心点的要求极高. 本文通过第 j 个样本点相对于第 i 个样本中心的加权系数作为样本的权值. 其权值计算公式如下:

$$\phi_{ij} = \exp\left(-\frac{\left\|x_j - \delta_i\right\|^2}{t_i^2}\right) \tag{10}$$

其中, t_i^2 是权系数参数, δ_i 表示第 i 个中心点, ϕ_{ij} 表示第 j 个样本对第 i 个聚类中心的权值. 改进的方法减少了计算量, 加上前期对初始中心点的处理, 使得权值更加合理. 此外, 对于 t_i 的选取方法如下:

$$t_i = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{j=1}^{k}\left\|x_j - v_i\right\|^2 & x_j \in N_{im} \\ \dfrac{1}{k^2}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{k}\left\|x_j - v_i\right\|^2 & other \end{cases} \tag{11}$$

其中, m 近邻数根据文献[13]概念[15], N_{im} 表示第 i 个聚类中心的 m 个近邻数.

对算法进行加权改进后, 其目标函数被重新定义为如下:

$$J(U, \delta_1, \ldots \delta_k, \phi) = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k}\sum_{j}^{n} \phi_i \mu_{ij}^m d_{ij}^2 \tag{12}$$

按照聚类有效性研究 m 可取值[1.1~2.5]. 以式(1)为约束条件构造拉格朗日函数求极值,

得到加权后 $u_{ij} = \dfrac{1}{\displaystyle\sum_{b=1}^{k}\left(\frac{d_{ij}}{d_{bj}}\right)^{2/(m-1)}}$ (13)

得到加权后 $v_i = \dfrac{\displaystyle\sum_{j=1}^{n}\phi_i \mu_{ij}^m x_j}{\displaystyle\sum_{j=1}^{n}\phi_i \mu_{ij}^m}$ (14)

2.3 IFCM 算法的一般描述

基于如上的介绍可知, 本文提出的算法主要分为两大步骤: 利用相异度去除孤立点及找到聚类中心, 计算基于样本加权的聚类算法.

初始簇中心生成步骤如下:

Step1. 按照式 $d_{ij} = \| x_i - x_j \|$ 计算样本间的距离, 并存储在距离矩阵 E 中, 方便后面步骤复用;

Step2. 按照(6)式计算样本间的相异度;

Step3. 按照(7)式构造相异度矩阵 R;

Step4. 按照(9)式计算 $AvgX_i$. 取最小的样本个数 k, 并以 k 中最小的为第一个初始中心点;

Step5. 将 x_i 样本与其区域内的样本删除, 剩下的样本重复步骤 1-5, 直到找到 k 个聚类中心为止.

利用第一步骤得到的聚类中心作为样本加权的 FCM 算法的初始中心点. 聚类步骤如下:

Step1. 利用第一步骤得到初始中心点;

Step2. 利用式(10)计算样本的权值系数 ϕ_{ij};

Step3. 利用式(13)更新隶属度矩阵 U;

Step4. 利用式(14)更新 V;

Step5. 利用式(12)计算目标函数 J,如果相对上次目标函数值的改变量小于某个阈值 ε 则算法停止, 否则返回步骤 3.

3 算法在Weka平台中的集成与实现

Weka 系统中集合了大量能承担数据挖掘任务的机器学习算法, 却只集成了少数的聚类算法. 针对此缺点, 本文对其进行二次开发, 并利用此平台比较两种聚类算法.

算法在 Weka 平台的实现是基于 Java 语言的, 采用面向对象的编程思想. 所以, 首先必须掌握接口文档, 然后通过接口开发新算法.

向 eclipse 中导入项目、开发算法程序及向 Weka 中添加算法的实现步骤如下:

① 我们可以从官网下载 Weka 安装软件进行安装, 然后通过 Weka 安装的位置找到 weka-src.jar 文件, 并将其解压为文件夹 weka-src.

② 打开 eclipse, 将此文件夹导入到 eclipse 工作区中.

③根据接口文档开发 FCM 算法的程序, 并将新建的的 FCM.java 放到 weka-src\src\main\java\weka \clusterer 的文件夹下, 另外更改 weka.gui.GenericObjectEditor.props, 在 #Lists the Clusterers I want to choose from 的 weka.clusterers.Clusterer=下加入 weka.clusterers.FCM.

然后, 将 IFCM 算法按照同样方法集成到平台中去.

IFCM 算法的添加的主要方法包括:

(1) Matrix solveE(Instances instances); //计算样本间的欧几里得矩阵 E

(2) Matrix solveR(Instances instances); //计算样本间相异度矩阵 R

(3) Matrix solveW(Instances instances); //获取样本权值矩阵 W

3.1 实验过程

在 eclipse 中运行 GUIChooser.java 文件, 出现 Weka 平台主页(如图 1). 点击 Explorer, 并加载数据, 应用 clusterers 的 FCM 算法与 IFCM 算法进行聚类, 并对聚类结果进行分析. 如图 2 所示, 则已成功将 FCM 算法及 IFCM 算法集成到平台中去.

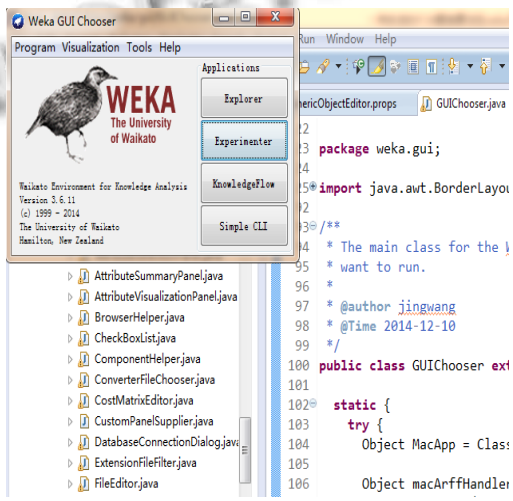

图 1　Weka 平台主页

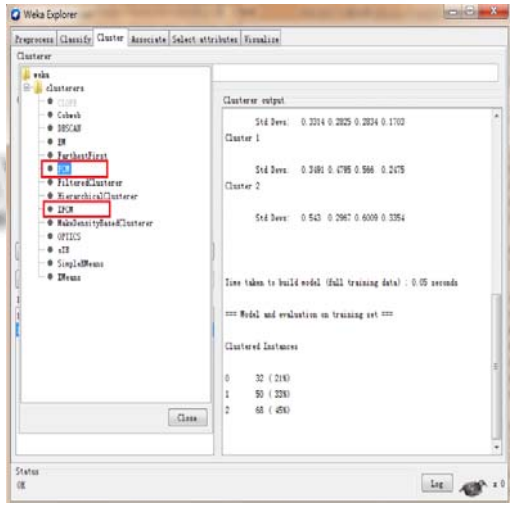

图 2　算法集成页面

4 实验数据与结果分析

UCI 数据库是一个专门用于测试机器学习、数据挖掘算法的数据库, 库中有专用于测试聚类算法性能的数据集, 因此可以用准确率来直观的表示聚类的质

量. 为了验证以上改进算法的有效性, 实验采用 UCI 数据库中的标准测试数据集 Iris、高维数据集 Wine 作为实验数据. 数据集的属性都列在表 1 中.

本实验环境为 WindowXP 系统, 实验工具为 Eclipse. 实验按步骤执行 GUIChooser.java 并操作界面后, 得出 FCM 聚类结果和 IFCM 聚类结果. 实验结果中 Sum of Squared Errors(SSE)表示均方误差和, Clusterd Instances 表示各个簇中实例的数目及百分比, Clusters centroids 表示各个簇中心的位置, 对于数值型的属性, 给出的还会是它在各个簇里的标准差.

表 1 数据集描述

数据集	类数目	属性数目	样本分布	描述
Iris	3	4	150(50Setosa,50Versicolour,50Virginica)	鸢尾花数据集
Wine	3	13	178(59 c1,71 c2,48 c3)	酒数据集

4.1 均方误差和比较

首先, 我们利用 Weka 环境自带的均方误差和来对算法性能进行评价.

SSE 属于内部度量, 其定义为

$$SSE = \sum_{j=1}^{k} \sum_{x_j \in c_j} \| x_i - \mu_j \|^2 \qquad (15)$$

其中, $\mu_j = \frac{1}{n_j} \sum_{x_j \in c_j} x_i$, n_j 代表簇 c_j 中的实例或样本的数目. SSE 越小, 说明簇内的各个样本点比较紧密, 类内耦合度较好, 说明聚类结果也越好.

实验中, FCM 算法由于其初始中心点的随机选取, 导致其实验结果不稳定, 本文多次实验, 选择结果最好的一次与改进算法作比较. 如表 2 可以看出, IFCM 算法的 SSE 远小于 FCM 的值.

表 2 FCM 与 IFCM 聚类结果的比较

DataSets	Measure	FCM	IFCM
Iris	SSE	428.070	219.827
Wine	SSE	38042.137	24501.065

4.2 Iris 与 Wine 数据集实验结果

为了验证算法的准确率, 本文引进了错分率的概念, 错分率即为每类分错实例数之和/实例总数. 错分率越小, 则准确率越高.

表 3 FCM 与 IFCM 算法在 Iris 数据集上的聚类结果

数据集	类别	聚类实例数	百分比(%)
FCM	Cluster0	32	21
	Cluster1	50	33
	Cluster2	68	45
IFCM	Cluster0	50	33
	Cluster1	47	31
	Cluster2	53	35

分析表 3 中可以看出 IFCM 中第一类十分准确, 第二类与第三类有少许的交叉, 刚好符合 Iris 数据本身性质, 而 FCM 则是第一类与第三类有较大的交叉现象. 再者, FCM 的错分率为(18+18)/150=24%,而 IFCM 的错分率为(3+3)/150=4%.

表 4 FCM 与 IFCM 算法在 Wine 数据集上的聚类结果

数据集	类别	聚类实例数	百分比(%)
FCM	Cluster0	47	26
	Cluster1	65	37
	Cluster2	66	37
IFCM	Cluster0	50	28
	Cluster1	68	38
	Cluster2	60	34

分析表 4 中可以看出 FCM 的错分率为(12+6+18)/178=20%,而 IFCM 的分错率为(9+3+12)/178=13.5%, IFCM 算法比传统的 FCM 算法相比聚类效率提高了.

5 总结

FCM 算法是一种被广泛应用的无监督聚类算法, 针对其对噪声点敏感, 加上未考虑到样本对聚类结果的贡献不相等, 本文对传统的 FCM 算法进行了改进. 除此之外, 本文针对开源数据挖掘平台 Weka 在聚类方面只集成了少数聚类算法的缺点, 对其进行二次开发, 扩充其聚类算法. 文中介绍了 weka 平台二次开发的主要步骤以及对实验结果的分析. 利用此平台实验证明, 本文改进算法能更加准确的找到初始中心点,并得到更加准确的聚类结果. 但是 Weka 平台的实验结果中没有指明每个实例隶属的类别. 为了更醒目的描述结果, 因此 Weka 平台还需作进一步改进.

参考文献

1 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48−61.

2 Tari L, Baral C, Kim S. Fuzzy c-means clustering with prior biological knowledge. Journal of Biomedical Informatics, 2009, 42(1): 74−81.

3 Kirindis S, Chatzis V. A robust fuzzy local information c-means clustering algorithm. IEEE Trans. on Image Process, 2010 , 19(5): 1328–1337.

4 孟海东,马娜娜,宋宇辰,徐贯东.基于密度函数加权的模糊 C 均值聚类.计算机工程与应用,2012,48(27).

5 刘兵,夏士雄等.基于样本加权的可能性模糊聚类算法.电子学报,2012,40(2).

6 王行甫,程用远,等.一种改进的密度加权的模糊 C 聚类算法.计算机系统应用,2012,21,(9).

7 Hathaway RJ, Hu YK. Density-weighted Fuzzy c-Means Clustering. IEEE Trans. on Fuzzy Systems, 2009, 17(1): 243–252.

8 郑世明,苗壮,宋自林,等.Weka 环境下基于模糊理论的聚类算法.解放军理工大学学报(自然科学版),2012,13(1):22–26.

9 陈慧萍,林莉莉,王建东,等.Weka 数据挖掘平台及其二次开发.计算机工程与应用,2008,44(19):76–79.

10 汪中,刘贵全,陈恩红.一种优化初始中心点的 K-means 算法.模式识别与人工智能,2009,22(2).

11 RusPini EH. A new approach to clustering. Information and Control, 1969, 15(1): 22–32.

12 Dunn JC. A graph theoretic analysis of pattern classification via Tamura's fuzzy relation. IEEE Trans. on Systems, Man and Cybernetics, 1974, 4 (3): 310–313.

13 涂丽红,仝海燕,杨丽萍.基于相异度的孤立点挖掘研究.计算机与数字工程,2008,36(1).

14 陆声链.基于距离的孤立点检测及其应用.计算机与数字工程,2004,32(5):94–97.

15 Zelnik-Manor L, Perona P. Self-tuning spectral clustering. Advancesin Neural Information Processing Systems, 2004.