一种基于密度的加权模糊均值聚类算法

李翠霞 史苇杭 李占波

(郑州大学软件技术学院 郑州 450002)

摘 要 针对当数据集合中的数据属性差异不明显时,传统的均值聚类算法会收敛到局部最小值点,造成算法聚类结果不准、精度下降的问题,提出了一种基于密度的加权模糊均值聚类算法。该算法通过计算差异属性类中的相关密度,运用密度作为确定初始类中心的方法,得到了聚类效果更好的初始值。之后用加权模糊算法克服类划分中数据属性差异不明显带来的弊端,对类中差异属性进行归类划分。实验结果表明,该算法依然可以区分出不同属性的重要程度,而且其稳定性和聚类效果都有一定的提高。

关键词 聚类,模糊均值,属性加权,密度,误分类数

中图法分类号 TP391 **文献标识码** A

Density Based Weighted Fuzzy Clustering Algorithm

LI Cui-xia SHI Wei-hang LI Zhan-bo

(School of Software Technology, Zhengzhou University, Zhengzhou 450002, China)

Abstract The traditional clustering algorithm will converge to a local minimum point when the initial objects' attributes have no obvious difference, which can cause the decline of algorithms' accuracy and incorrectness of the results. In order to overcome these drawbacks, a density based weighted fuzzy c-mean clustering algorithm was proposed. It used the results of the calculation of the relative density differences attributes to determine the initial partition. After obtaining the better initial centers, a weighted fuzzy algorithm which can distinguish the importance of each attribute was implemented. Experimental results show that the algorithm not only can discriminate the attributes' contribution, but also can improve the stability and accuracy.

Keywords Clustering, Fuzzy C-means, Attribute weighted, Density, Number of misclassification

1 引言

聚类分析是一种无监督的学习,其目的是按照某种规则将包含大量数据点的集合划分为若干类,使得同一类中的数据尽可能相似,不同类中的数据尽可能相异。目前,聚类分析已广泛应用于机器学习、模式识别、信息检索、图像识别及市场营销等众多领域[1]。最经典的是C-means聚类算法,该算法是基于数据属性划分的聚类算法,此类算法从随机选择的初始类中心开始,求得能使目标函数取最小值的划分。因此,初始类中心对该类算法的影响较大。如果选择不当,将会使聚类收敛到局部极小值点。因此,准确建立初始聚类的中心聚类结果,是得到精准聚类结果的前提[2-4]。

在当前的实际应用中,传统均值聚类算法取得了不错的效果,但也反映出了一些问题。传统均值聚类过于依赖初始聚类的结果。后续的计算结果都和初始聚类结果有关,当需要聚类的数据属性过于一致,数据对象中各属性对聚类的贡献相同时,初始聚类划分结果就会出现偏差,导致初始聚类不准,进而影响聚类算法的精度。因此如何对数据属性同质性严重的数据进行分类,成为一个难题。

针对这一问题,提出一种基于密度的加权模糊均值聚类算法[5,6]。通过计算数据中的数据相关密度,克服属性一致带来的弊端,对初始类中心进行优化。将差异属性类中的相关密度计算结果作为确定初始类中心的依据,得到了聚类效果更好的初始聚类结果。运用加权模糊算法克服类划分中数据属性差异不明显带来的弊端,使其能够更好地对类中差异属性进行归类划分[7]。

2 传统聚类方法的原理及缺陷

利用聚类算法可以将需要识别的目标进行分类处理。按照事物近似特征,将目标与设置的衡量标准进行对比,可以将一组需要识别的目标分为不同的种类。在这个分类过程中,分类的结果是目标必须属于其中的某一类。但是,通常情况下,需要进行识别的目标并没有独特的属性用以与衡量标准进行对比,因此,这种需要进行识别的目标应该用聚类算法进行分类。

利用聚类算法进行分类处理的最为关键的一步是构造初

到稿日期:2011-06-15 返修日期:2011-08-25 本文受河南省重大科技攻关项目(102102210490)资助。

李翠霞(1977-),女,硕士,讲师,主要研究方向为机器学习、计算机软件开发技术,E-mail:qyliying@126.com;史苇杭(1980-),女,硕士,讲师,主要研究方向为嵌入式系统、计算机软件开发技术;李占波(1965-),男,硕士,教授,硕士生导师,主要研究方向为网络安全、计算机软件开发技术。

始聚类。得到的初始种类中，任意一个种类的目标相似度都比较高，并且不同类别中的目标相似度都比较低。算法利用需要进行识别的对象初始聚类结果的相关参数映射函数，将其函数值与衡量标准进行对比，最终获取所属类别。原理如下：

目标种类的数量是 J，需要进行识别的样本集合是 $y=\{y_1,y_2,\cdots,y_N\}$。这个样本集合中所有的样本都属于不同的种类。V_{jk} 代表样本序号是 k、所属类别序号是 j 的样本，则该样本需符合下述要求：

(1)V_{jk} 的极大值是 1，极小值是 0；

(2)所有需要进行识别的对象属于目标种类的概率之和是 1；

(3)每一个类别中包括的需要进行识别的对象数目都大于 0 并且小于 N。

均值聚类算法是通过对目标进行优化处理获取的，需要利用下述公式进行初始聚类，才能获取其相似性系数。初始聚类方法如式(1)所示：

$$F(V,W)=\sum_{k=2}^{N} v_{jk}/e_n \tag{1}$$

式(1)需要符合下述条件：

$$\begin{cases} 1\leqslant p\leqslant\infty \\ v_{jk}>0 \\ 0<j<N \\ 0<k<N \end{cases} \tag{2}$$

式中，N 是需要进行识别的目标数目；V 是相似性函数矩阵；V_{jk} 是相似性函数中的第 k 行、第 j 列元素；W 是衡量标准参数，e_k 是第 k 个样本与衡量标准的区别参数，p 是均值系数，在指定区间中，均值系数的取值与关联程度成正比例关系。

通过算法的原理可知，传统的聚类算法过于依赖首次聚类结果。当数据中的属性同质化严重时，会造成初始的聚类结果特征不明显，从而引起后面的聚类精度变差，聚类效果不好。因此，对同质化数据进行划分聚类，就成为研究的重点。

3 密度加权模糊均值聚类算法

3.1 属性密度的计算

算法的第一步是求出聚类数据中的属性密度。由于初始类中心对后续计算的影响较大，因此较好的类中心应该具有较好的代表性，也就是类的内部有很高的密度，所以应尽量靠近类中心。为了更好地描述初始类中心的选择方法，运用属性密度的计算方式作为中心计算的基础，计算的具体方法如下：

对于数据集合中的任意数据点 x 和参考相似度 $Rsim$，以 x 为中心，半径为 $Rsim$ 的区域称为 x 的邻域。邻域中数据对象的个数称为 x 基于参考相似度 $Rsim$ 的密度，可记作：

$Den(x,Rsim)$

对于数据集合内的数据点 x、相似度 $Rsim$ 和阈值 $MinD$，如果满足 $Den(x,Rsim)\geqslant MinD$，则 x 为核心数据，$MinD$ 为密度阈值。若 x 在 y 的邻域中，且 y 是核心数据，则认为 x 从 y 直接密度可达。若 $x_1=x,x_n=y$，且满足 x_i 从 x_{i+1} 直接密度可达，则 x 从 y 密度可达。若存在 r 使得 x 和 y 都从 r 密度可达，则 x 和 y 是密度连接的。

邻接核心数据：给定相似度 $Rsim$ 和阈值 $MinD$，如果核心数据 x 满足 $sim(x,y)\geqslant 2Rsim$，即 x 和 y 之间的相似度大于或等于 2 倍的 $Rsim$，则 x、y 为邻接核心数据。利用 x，y 求出数据的密度特性，计算方法如下：

$$E_x=Rsim_x-\frac{MinD}{2}$$

$$E_y=-(Rsim_y-\frac{MinD}{2})$$

$$E(x,y)=(\frac{2E_x}{Rsim},\frac{2E_y}{MinD})=(\frac{2P_x}{Rsim}-1,1-\frac{2P_y}{MinD}) \tag{3}$$

式中，$E(x,y)$ 为计算出的密度特征。计算出密度后，可根据密度确定初始类的中心，克服传统方法的缺陷。

3.2 根据密度确定初始类中心

输入数据集合 $X=\{x_1,x_2,\cdots,x_n\}$，给定 $Rsim$，$MinD$，K（K 为小于等于 n 的任意正整数）。根据计算出的密度结果，计算初始类的中心，算法流程如下：

对于给定的数据集合 X 中的任意 x，根据输入参数求其邻域 $N_{minD}(x)$；若 $Z(Z\in X)$ 的邻域 $N_{minD}(x)$ 为 1，说明它的邻域内只有它自身，或者说它与集合中其它数据的相似度都小于 R_{sim}，则将求出的密度值放入参考点集合 CS；若 $L(L\in X)$ 的邻域 $N_{minD}(x)\geqslant MinD$，则选其邻域内所有对象的密度均值为参考点放入参考点集合 CS $(S_1,S_2,\cdots S_j)$，其中 $j<n$，如果 $j>K$，则将参考集合中相似度最大的 S_i 和 S_j 的密度均值作为参考点放入 CS，并删除这两个数据，直到参考集合中的数据点数目为 K；如果 $j<K$，重新输入密度值，返回第 1 步重新开始。如果参考集合中某个数据非原数据集合中的数据，则用数据集合中与它相似度最大的数据代替，从而得到最终参考点集合。

3.3 基于模糊加权的数据分类

在确定了中心类后，可以对其进行聚类划分，方法如下：

设 $X=\{X_1,X_2,\cdots,X_n\}$ 是一个包含 n 个对象的数据集合。对象 $X_k=\{x_{k1},x_{k2},\cdots,x_{ks}\}$ 包含 s 个属性。设 $W=\{w_1,w_2,\cdots,w_s\}$ 是 s 个属性的权值集合，β 是属性权值 w_j 的参数，U 是划分矩阵，V 是 k 个类中心集合，该算法的目的是把 X 划分成 c 个类，并使得如下的目标函数最小：

$$J(u,v,w)=\sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{j=1}^{s}u_{ik}^{m}w_j^{\beta}(x_{kj}-v_{ij})^2 \tag{4}$$

式中，$\sum_{j=1}^{s}w_j=1$，$\sum_{i=1}^{c}u_{ik}=1$，$1\leqslant i\leqslant c$，$1\leqslant j\leqslant s$，c 是类数目，β 称为权重指数。

通过拉格朗日乘子法，可以得到最小化目标函数 $J(u,v,w)$ 的必要条件如下：

$$v_{ij}=\sum_{k=1}^{n}u_{ik}^{m}x_{kj}/\sum_{k=1}^{n}u_{ik}^{m}$$

$$u_{ik}=(\sum_{j=1}^{s}w_j^{\beta}\parallel x_{kj}-v_{ij}\parallel^2)^{1/(1-m)}\times (\sum_{i=1}^{c}(\sum_{j=1}^{s}w_j^{\beta}\parallel x_{kj}-v_{ij}\parallel^2)^{1/(1-m)})^{-1}$$

$$w_j=D_j^{1/(1-\beta)}(\sum_{j=1}^{s}D_j^{1/(1-\beta)})^{-1} \tag{5}$$

这里，$D_j=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}(x_{kj}-v_{ij})^2$

因此，本文算法的聚类过程可以描述如下。

Step1　选定聚类数目 c、模糊指数 m、最大迭代次数 $Tcount$、属性权值指数 β 和阈值 ε，初始化划分矩阵 U 和属性权值矩阵 W；

Step2　根据目前的划分和属性权值，计算目标函数式(1)的值，如果它大于迭代次数 $Tcount$，或它相对于上次目标函数值的改变量小于阈值 ε，则算法停止；

Step3　使用迭代式(4)修正聚类中心 v_{ij}；

Step4　根据新的类中心，使用迭代式(3)更新隶属度函数 u_{ik}；

Step5　使用迭代式(5)更新属性权值矩阵 W。返回Step2。

易于证明，给定 m 和 β 时，经过有限次迭代，AWFCM算法可以收敛到目标函数的局部最小值点或鞍点。

4　实验结果

以下通过数值实验检验本文所给算法的效果。聚类算法会受参数 m 和 β 的影响，可根据文献[5-7]中类似的方法，通过分析算法目标函数的海森矩阵，得出该算法的参数选择理论，可取 $m=2,\beta=3.2$。对算法进行比较时，采用平均误分类数作为评判标准。

实验1　本实验中采用包含600个数据点的人工合成数据集合Data。属性 x_1,x_2,x_5 服从正态分布且划分成3类，每一类中包含200个数据点，x_3,x_4 是服从均匀分布的随机干扰属性。

图1示出本数据集合在不同的三维子空间中的分布。

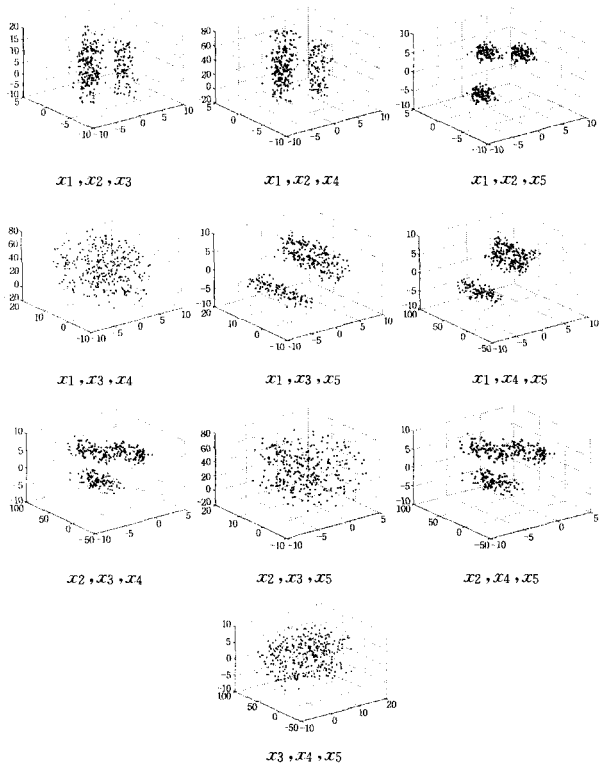

图1　Data数据集合三维子空间分布图

由于本文算法需要确定输入参数 $MinD$ 和 $Rsim$，通过多次实验得到，$MinD=200$，$Rsim=0.35$ 时，聚类效果较好。因为算法受参数 m 和 β 的影响，参照文献[5-7]中的参数选择理论和实验方法，能够得出 $m=2.5$，β 从2变化到8时，各属性的权值如图2所示。

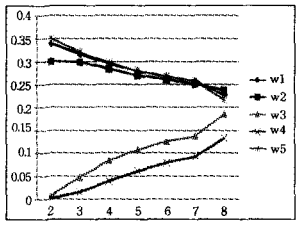

图2　属性权值图

从图中可以看出，属性 x_1,x_2,x_5 对聚类贡献较大，x_3,x_4 属性的重要程度略低一些，这与实际数据集合的特征是相吻合的。

实验2　本实验将采用UCI数据库中的鸢尾花数据IRIS来验证文中所提出聚类方法的有效性。在对比各算法的有效性时，采用了最大误分类个数、最小误分类个数以及平均误分类个数共3个指标。表1列出了FCM算法、随机选择的加权模糊均值算法，以及文中所提算法的比较结果。有文献指出，$m=2$ 时，FCM算法在IRIS数据上有效，因此采用 $m=2$，执行FCM算法100次；加权模糊均值算法在对IRIS数据集合进行聚类时，$m=2,\beta=2$ 时效果较好，同样取 $m=2,\beta=2$，执行算法100次；取 $m=2,\beta=2$，通过改变输入参数执行本文算法100次实验。所得结果如表1所列。

表1　FCM、WFCM和本文算法的稳定性比较表

数据集合	算法	迭代次数
Data	FCM	37
	WFCM	24
	本文算法	18
Iris	FCM	21
	WFCM	12
	本文算法	10

通过实验数据比较可以看出，本文得到的结果无论在哪种数据集合上，其迭代次数和平均误分类数都优于其他的分类算法，取得了明显的优化效果。

结束语　本文采用密度的思想提出一种基于密度的加权模糊均值聚类算法，实验证明了该方法的有效性。基于密度的方法在进行初始类中心选择时，充分考虑了类中心的分散和代表性，又很好地结合了区别对待数据对象各属性的属性加权模糊均值聚类算法，提高了算法的效率和稳定性。该算法受参数 $Rsim$ 和 $MinD$ 的影响，文中只是通过实验给出了参考选择方法，如果能有理论方法给出这些参数的选择规则，将会增强改进算法的效果。

参 考 文 献

[1]　Han I, Kamber M. Data Mining: Concepts and Techniques[M]. Berlin: Morgan Kaufmann Publishers, 2000: 335-389

[2]　张文明，吴江，袁小蛟. 基于密度和最近邻的K-means文本聚类算法[J]. 计算机应用，2010，30(7)：1933-1935

[3]　Faber V. Clustering and the continuous K-means algorithm[EB/OL]. http://library. lanl. gov/cgi-bin/get-filefi00412967. pdf, 2009-10-03

[4]　关云鸿. 改进K-均值聚类算法在电信客户分类中的应用[J]. 计算机仿真，2011

[5]　于剑. 论模糊均值算法的模糊指标[J]. 计算机学报，2003，26(8)：968-973

[6]　Yu Jian, Cheng Qian-sheng, Huang Hou-kuan. Analysis of the weighting exponent in the FCM[J]. IEEE Transactions on Systems, Man and Cybernetics-part B: Cybernetics, 2004, 34(1): 634-639

[7]　Jian Yu, Yang Miin-shen. Optimality Test for Generalized FCM and Its Application to Parameter Selection[J]. IEEE Transactions on Fuzzy Systems, 2005, 13(1): 164-176

[8]　刘宴雄，刘宴兵，罗来明. 一种基于密度和网格 的高效聚类算法[J]. 重庆邮电大学学报：自然科学版，2010，22(2)：242-247