基于特征加权的模糊聚类新算法

李 洁,高新波,焦李成
(西安电子科技大学电子工程学院,陕西西安 710071)

摘 要: 在聚类分析中,针对不同类型的数据,人们设计了模糊 k-均值、k-mode以及 k-原型算法以分别适合于数值型、类属型和混合型数据.但无论上述哪种方法都假定待分析样本的各维特征对分类的贡献相同.为了考虑样本矢量中各维特征对模式分类的不同影响,本文提出一种基于特征加权的模糊聚类新算法,通过 ReliefF算法对特征进行加权选择,不仅能够将模糊 k-均值、k-mode以及 k-原型算法合而为一,同时使样本的分类效果更好,而且还可以分析各维特征对分类的贡献程度.对各种实际数据集的测试实验结果均显示出新算法的优良性能.

关键词: 聚类分析; 模糊聚类; 数值特征; 类属特征; 特征加权
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2006) 01-0089-04

A New Feature Weighted Fuzzy Clustering Algorithm

LI Jie, GAO Xin-bo, JIAO Li-cheng
(School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China)

Abstract: In the field of cluster analysis, the fuzzy k-means, k-modes and k-prototypes algorithms were designed for numerical, categorical and mixed data sets respectively. However, all the above algorithms assume that each feature of the samples plays a uniform contribution for cluster analysis. To consider the particular contributions of different features, a novel feature weighted fuzzy clustering algorithm is proposed in this paper, in which the ReliefF algorithm is used to assign the weights for every feature. By weighting the features of samples, the above three clustering algorithms can be unified, and better classification results can be also achieved. The experimental results with various real data sets illustrate the effectiveness of the proposed algorithm.

Key words: cluster analysis; fuzzy clustering; numeric feature; categorical feature; feature weighte

1 引言

聚类分析是多元统计分析的方法之一,也是统计模式识别中非监督模式分类的一个重要分支[1],其任务是把一个未标记的样本集按某种准则划分成若干子集,要求相似的样本尽量归为一类,而不相似的样本应在不同的类.采用这种分析方法可定量地确定研究对象之间的亲疏关系,从而达到对其合理分类、分析等目的.

模糊 k-均值 (FKM)算法就是其中一种有效的聚类分析方法,在非监督模式识别、模糊控制等领域有着极为广泛的应用.然而,模糊 k-均值算法只能处理数值型数据,而在实际应用中我们遇到的数据既可能是数值型的,也可能是类属型的,还有可能是混合属性特征.而传统的将类属值转化为数值的方法不是总能得到有效的结果,这是因为类属域是无序的.因此,人们又相应设计出 k-mode算法和 k-原型算法以分别处理类属型数据和混合型数据[2,3].

然而,无论是传统的 k-均值算法, k-mode算法还是 k-原型算法都隐含假定待分析样本矢量的各维特征对分类的贡献均匀,不考虑各个特征对分类的不同影响,只是在 k-原型算法的聚类目标函数中,用一个权值来平衡类属特征和数值特征.而在实际应用中,由于构成样本特征矢量的各维特征来自不同的传感器,存在量纲差异和精度及可靠性的不同,另一方面,所选择的特征集未必适合于模式的分类.因此,传统的模糊 k-均值、k-mode算法和 k-原型算法在实际应用中都有一定的局限性.

为了考虑特征矢量中各维特征对模式分类的不同贡献,我们提出一种基于特征加权的模糊聚类新算法,新算法利用特征选择技术 ReliefF[4]对特征进行加权选择,可以将传统模糊 k-均值、k-mode算法与 k-原型聚类算法合而为一,不仅使分类效果优于传统聚类算法,同时可以分析各维

收稿日期: 2005-03-07;修回日期: 2005-09-06
基金项目:国家自然科学基金 (No. 60202004);中国博士后科学基金

特征对分类的贡献程度.

本文的安排如下:第 2节简单介绍特征选择技术 ReliefF算法,第 3节讨论基于特征加权的模糊聚类新算法,第 4节为实验结果,并将本文提出的新方法与传统的模糊 k均值算法、k-mode算法和 k原型算法分别进行了性能比较.最后总结全文,并指出进一步的研究方向.

2 ReliefF算法

特征选择在数据挖掘、图像处理、模式识别等领域有着广泛的应用,其基本任务是从诸多的特征中找出最有效的特征.基本的 Relief算法是 Kira和 Rendell在 1992年提出的[5],当时局限于解决两类的分类问题,1994年 Kononenko扩展了 Relief算法,使得 ReliefF可以解决多类问题.ReliefF算法是给特征集中每一特征赋予一定的权重.

设 $X=\{x_1, x_2, \cdots, x_n\}$是待聚类分析的对象的全体,其中 $x_i=[x_{i1}, x_{i2}, \cdots, x_{iN}]^{\mathrm{T}}$表示第 i个样本的 N个特征值,λ是 $N\times 1$的矩阵,表示各维特征的权值.对于任意的一个样本 x_i,首先找出 R个与 x_i同类的最近邻的样本 h_j, $j=1, 2, \cdots, R$,然后在每一个 x_i与不同类的子集中找出 R个最近邻的样本 m_{lj}, $j=1, 2, \cdots, R$, $l\neq \mathrm{class}(x_i)$,设 $diff_hit$是 $N\times 1$的矩阵,表示 h_j与 x_i在特征上的差异.

$$diff\_hit=\sum_{j=1}^{R}\frac{|x_i-h_j|}{\max(X)-\min(X)} \tag{1}$$

设 $diff_miss$是 $N\times 1$的矩阵,表示 m_{lj}与 x_i在特征上的差异.

$$diff\_miss=\sum_{l\neq \mathrm{class}(x_i)}\frac{P(l)}{1-P(\mathrm{class}(x_i))}\sum_{j=1}^{R}\frac{|x_i-m_{lj}|}{\max(X)-\min(X)} \tag{2}$$

$p(l)$为第 l类出现的概率,可以用第 l类的样本数比上数据集中样本的总数.ReliefF算法中 λ由下式更新

$$\lambda=\lambda-diff\_hit/R+diff\_miss/R \tag{3}$$

如此重复若干次,我们就可以得到特征集中每一个特征的权重.

3 基于特征加权的聚类新算法

在利用模糊 k均值算法、k-mode算法以及 k原型算法进行聚类分析时,总是假设特征提取相当完善,构成模式矢量的特征是独立且无冗余的,并且认为每维特征对分类的贡献是均匀的.而事实上,构成样本特征矢量的各维特征来自不同的传感器,存在量纲差异和精度及可靠性的不同,而且所选择的特征集也未必适合于模式的分类.鉴于此,本文将探讨一种新的模糊聚类算法,不仅将模糊 k均值算法、k-mode算法以及 k原型算法合而为一,而且考虑各维特征对模式分类的不同贡献,可以获得更有效的聚类分析结果.

下面在介绍基于特征加权的模糊聚类新算法时,我们以 k原型算法为例进行说明.设 $X=\{x_1, x_2, \cdots, x_n\}$是待聚类分析的对象的全体,其中 $x_i=[x_i^r, x_i^C]^{\mathrm{T}}$表示第 i个样本的 m个特征值,其中 $x_i^r=[x_{i1}^r, \cdots, x_{it}^r]$表示数值特征,$x_i^C=[x_{i,t+1}^C, \cdots, x_{im}^C]$表示类属特征,假设 $P=\{p_1, p_2, \cdots, p_k\}$, $p_i=[p_{i1}^r, \cdots, p_{it}^r, p_{i,t+1}^c, \cdots, p_{im}^c]^{\mathrm{T}}$,表示第 i类的聚类原型,则在 k原型算法中,我们定义聚类目标函数为

$$J(P)=\sum_{i=1}^{k}\left(\sum_{j=1}^{n}\sum_{l=1}^{t}|x_{jl}^r-p_{il}^r|^2+\lambda\sum_{j=1}^{n}\sum_{l=t+1}^{m}\delta(x_{jl}^c, p_{il}^c)\right) \tag{4}$$

式中第 1项是数值特征上的欧几里德距离平方,第 2项是类属特征上的简单的相异匹配测度,$\delta(\cdot)$定义为

$$\delta(a, b)=\begin{cases}0, & a=b\\ 1, & a\neq b\end{cases} \tag{5}$$

权值 λ用来调节两种特征在目标函数中的比例,从式(4)我们可以看出,在 k原型算法中,虽然有权值 λ来控制数值特征和类属特征的比例,但在数值特征集中,每一维数值特征对分类的贡献是均匀的;同样,每一维类属特征的贡献也是相同的.

我们可将硬 k划分扩展为模糊划分,这样对模糊聚类问题,目标函数进一步修正为:

$$J(W, P)=\sum_{i=1}^{k}\Big(\sum_{j=1}^{n}w_{ij}^2\sum_{l=1}^{t}|x_{jl}^r-p_{il}^r|^2+\lambda\sum_{j=1}^{n}w_{ij}^2\sum_{l=t+1}^{m}\delta(x_{jl}^c, p_{il}^c)\Big) \tag{6}$$

式中 $w_{ij}\in[0, 1]$是样本 x_j属于第 i类的隶属度,需要指出的是,我们给 w_{ij}加上幂指数 2,从而保证了硬划分向模糊划分的扩展是非平凡的[6].

我们在模糊 k原型算法的基础上进行修正,利用第 2节介绍的 ReliefF算法对每一维特征进行加权,设 $\lambda^r=[\lambda_1^r, \cdots, \lambda_t^r]^{\mathrm{T}}$表示数值特征的权值,$\lambda^C=[\lambda_{t+1}^C, \cdots, \lambda_m^C]^{\mathrm{T}}$表示类属特征的权值,这样,我们将聚类目标函数修正为:

$$J(W, P)=\sum_{i=1}^{k}\Big(\sum_{j=1}^{n}w_{ij}^2\sum_{l=1}^{t}\lambda_l^r|x_{jl}^r-p_{il}^r|^2+\sum_{j=1}^{n}w_{ij}^2\sum_{l=t+1}^{m}\lambda_l^c\delta(x_{jl}^c, p_{il}^c)\Big) \tag{7}$$

当目标函数 $J(W, P)$达到最小值时,就得到了最优的聚类结果.需要注意的是,在新算法中,特征权值包含两部分,一部分是数值特征的权值,一部分是类属特征权值,所以在利用 ReliefF算法更新权值时,也应该对此分别进行处理.数值特征权值的计算方法与上节介绍的相同,用下式表示.

$$\lambda^r=\lambda^r-diff\_hit^r/R+diff\_miss^r/R \tag{8}$$

式中 $diff_hit^r$和 $diff_miss^r$都是 $t\times 1$的矩阵,分别表示数值特征上的差异,由式(1)和式(2)定义.$diff_hit^C$和 $diff_miss^C$都是 $(m-t+1)\times 1$的矩阵,分别表示类属特征上的差异,由下式定义

$$diff\_hit^C=\sum_{j=1}^{R}\delta(h_j^C, x_i^C) \tag{9}$$

$$diff\_miss^C=\sum_{l\neq \mathrm{class}(x_i)}\frac{P(l)}{1-P(\mathrm{class}(x_i))}\sum_{j=1}^{R}\delta(m_{l_j}^C, x_i^C) \tag{10}$$

则类属特征的权值由下式计算

$$\lambda^C = \lambda^C - diff\_hit^C/R + diff\_miss^C/R \quad (11)$$

这样根据式(8)和式(11)就可以分别得到数值特征和类属特征的权值.

需要说明的是,ReliefF算法是针对分类技术的,在分类中,每一个样本的类别标记是确定的,而在聚类分析中,每一个样本的类别标记是未知的,这时,我们可以先对待分析的样本集进行一次聚类,选择隶属度较大的样本 x_i,并从划分矩阵中分别找出与 x_i 同类以及不同类的最近邻的 R 个样本,再按照上述方法计算特征权值. 当得到最终的权值后,利用权值对各维特征进行加权后,再进行聚类,就可以得到最后的聚类分析结果.

本文介绍的方法是基于 k-原型算法的,但事实上,新算法可以将模糊 k-均值算法和模糊 k-mode算法相结合,当权值 $\lambda^C = 0$ 时,新算法对应于加权的模糊 k-均值算法,当 $\lambda^r = 0$ 时,新算法为加权的模糊 k-mode算法,当 $\lambda^C \neq 0, \lambda^r \neq 0$ 时,是加权的模糊 k-原型算法.

4　实验结果

为测试基于特征加权的聚类新算法的有效性,我们对各种具有不同属性特征的数据集进行了测试实验,给出一些初步的实验结果,并将本文提出的新方法与传统的模糊 k-均值算法、k-mode算法和 k-原型算法进行了性能比较,显示出新算法的优良性能.

4.1　具有数值型特征的数据集聚类性能检验

为了测试本文提出的新算法的对数值型数据的分类性能,我们采用著名的 IRIS实际数据作为测试样本集[7]. IRIS数据由四维空间中的 150个样本点组成,每一个样本的 4个分量分别表示 IRIS的 Petal Length, Petal Width, Sepal Length和 Sepal Width. 整个样本集包含了 3个 IRIS种类 Setosa, Versicolor和 Virginica,每类各有 50个样本. 其中 Setosa与其他两类间较好地分离,而 Versicolor和 Virginica之间存在交迭. 我们分别采用传统的 FKM 算法和加权的 FKM算法对 IRIS样本进行分类,测试两种算法的误分样本数,以比较它们的性能.

IRIS数据经常被用作检验聚类算法性能的标准数据, Hathaway在 1995年给出了这组测试数据的实际类中心位置分别为[8]: $p_1 = (5.00, 3.42, 1.46, 0.24)$, $p_2 = (5.93, 2.77, 4.26, 1.32)$, $p_3 = (6.58, 2.97, 5.55, 2.02)$. 我们分别用传统的 FKM算法和加权的 FKM算法对 IRIS样本进行分类,得到的分类效果如表 1所示.

从表中可知,基于特征加权的 FKM算法不仅误分率小,而且得到的聚类原型模式也更接近实际的类中心位置. 特征加权的 FKM 算法得到的加权矩阵为 $\lambda^r = [3.9720, 2.6880, 9.4350, 14.4810]^T$,从得到的加权矩阵来看,第四维特征的贡献最大,而第二维特征的贡献最小.

表 1　特征加权的 FKM和传统 FKM算法性能比较

聚类算法	误分数	误分率	聚类原型矢量	误差平方和
传统 FKM	16	10.67%	$p_1 = (5.0062, 3.4242, 1.4684, 0.2492)$ $p_2 = (5.8946, 2.7460, 4.4154, 1.4273)$ $p_3 = (6.8484, 3.0750, 5.7283, 2.0741)$	0.1554
特征加权的 FKM	6	4%	$p_1 = (5.00603, 3.4276, 1.4626, 0.2463)$ $p_2 = (5.9082, 2.7490, 4.2671, 1.3313)$ $p_3 = (6.6459, 3.0050, 5.5923, 2.0462)$	0.0125

4.2　具有类属型特征的数据集聚类性能检验

为了测试本文提出的新算法的对类属型数据的分类性能,本实验中我们使用的数据是大豆疾病的实际数据[9]. 大豆疾病数据共有 47个记录,每个记录由 35个特征描述. 每个记录都被标记为四种疾病中的一种: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot和 Phytophthora Rot除了 Phytophthora Rot有 17个记录外,其他的每种疾病都有 10个记录.

我们分别用传统的 k-mode算法和加权的模糊 k-mode算法对大豆疾病数据进行分类,得到的分类效果如表 2所示,表 2中大写字母 D、C、R、P分别表示每一种大豆疾病,我们看到利用新的聚类算法,所有的样本都被正确分类,说明我们提出的新算法是有效的.

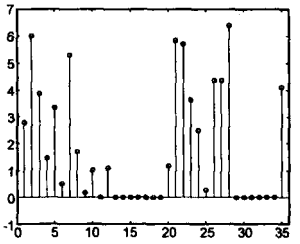

图 1　大豆疾病数据的各维权值

表 2　大豆疾病数据聚类结果

聚类算法	D	C	R	P	误分数	误分率
特征加权的 k-mode算法	10	10	10	17	0	0%
传统的 k-mode算法	13	10	10	14	9	20%

图 1所示为新算法得到的各维特征的权值,从图中可以看出,第 11维特征与第 13～19以及 29～34维特征的权值为 0,说明这 14维特征对分类不起作用. 对照原始数据,我们发现在这 14维特征上,所有样本的特征均相同,这也证明了新算法不仅提高了聚类的性能,而且还可以用于模式识别中的特征提取和优选.

4.3　具有混合属性特征的数据集分类性能检验

我们知道在数据挖掘中,经常会遇到既具有数值特征,也具有类属特征的混合属性数据. 为了测试新算法对混合型数据的分类性能,本实验中我们使用的是动物园的实际数据[10]. 动物园数据共有 101个记录,每个记录由 15个类属特征和 1个数值特征描述.

表 3 动物园数据聚类结果

动物标准分类 (每类样本数)	类 1 (31)	类 2 (20)	类 3 (14)	类 4 (10)	类 5 (8)	类 6 (8)	类 7 (10)
哺乳类 (41)	31						10 *
鸟类 (20)		20					
鱼类 (13)			13				
昆虫类 (8)				8			
软体类 (10)				2 *	8		
爬行类 (5)			1 *			4	
两栖类 (4)						4	

我们分别用传统的 k原型算法和本文的新算法对动物园数据进行分类,得到的分类效果如表 3所示. 表中带“ * ”的为错分的样本数. 由于哺乳动物的数目比其他动物多,所以被错分为两类 (第 1类和第 7类),爬行类和两栖类合为一类,其他类基本正确,错分数目 (不包括错分的哺乳动物)为 3个. 而传统的 k原型算法错分了 19个样本.

图 2所示为新算法得到的各维类属特征的权值,我们看到第 4维特征的权值最大,表明该维特征对分类的贡献最大,特征空间中第 4维特征描述了动物是否哺乳,正好对应区分最大类 - 哺乳动物,而第 14维特征描述了动物是否可以被驯化,对应的权值最小. 数值特征 $\lambda^r = 46.1842$.

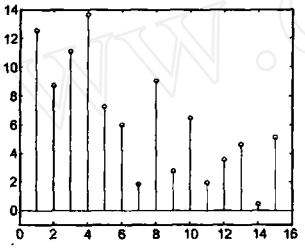

图 2 动物园数据的各维类属特征权值

对各种类型数据集的测试实验结果均表明:本文提出的基于特征加权的聚类新算法是合理有效的.

5 结论

本文提出一种基于特征加权的模糊聚类新算法. 实验结果表明该方法不仅可以将模糊 k均值、k-mode算法与 k-原型聚类算法合而为一,使聚类效果优于传统聚类算法,同时可以分析各维特征对分类的贡献程度,有效地进行特征提取和优选,这在实际应用中是非常方便的.

参考文献:

[1] 何清. 模糊聚类分析理论与应用研究进展 [J]. 模糊系统与数学, 1998, 12(2): 89 - 94.
He Qing. Advance in fuzzy clustering theory and application[J]. Fuzzy Systems and Mathematics, 1998, 12(2): 89 - 94.

[2] Zhexue Huang, Michael K Ng. A fuzzy k-modes algorithm for clustering categorical data[J]. IEEE Trans on Fuzzy Systems, August, 1999, 7(4): 446 - 452.

[3] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining[A]. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery[C]. USA: ACM Press, 1997. 1 - 8.

[4] Kononenko I. Estimating attributes: Analysis and extensions of Relief[A]. Proceedings of the 7th European Conference on Machine Learning[C]. Berlin: Springer, 1994. 171 - 182.

[5] Kira K, Rendell L A. A practical approach to feature selection[A]. Proceedings of the 9th International Workshop on Machine Leaning[C]. San Francisco, CA: Morgan Kaufmann, 1992. 249 - 256.

[6] 李洁, 高新波, 焦李成. 一种基于 CSA的混和属性特征大数据集聚类算法 [J]. 电子学报, 2004, 32(3): 357 - 362.
Li Jie, Gao Xinbo, Jiao Licheng. A CSA-based clustering algorithm for large data sets with mixed numeric and categorical values[J]. Acta Electronica Sinica, 2004, 32(3): 357 - 362. (in Chinese)

[7] Duda R O, Hart P E. Pattern classification and scene analysis[M]. New York: John Wiley & Sons, 1973. 89 - 91.

[8] Hathaway R J, Bezdek J C. Nerf C-means: Non-Euclidean relation fuzzy clustering[J]. Pattern recognition, 1994, 27(3): 429 - 437.

[9] Michalski R S, Stepp R E. Automated construction of classifications: Conceptual clustering versus numerical taxonomy[J]. IEEE PAMI, 1983, 5: 396 - 410.

[10] Jollois F X, Nadif M. Clustering large categorical data [A]. Advances in Knowledge Discovery and Data Mining [C]. Heidelberg: Springer-Verlag, 2002. 257 - 263.

作者简介:

李 洁 女, 1972年出生于陕西西安,工学博士,西安电子科技大学副教授. 主要从事人工智能、模式识别、数据挖掘等方面的研究.
E-mail: leejie@mail.xidian.edu.cn

高新波 男, 1972年出生于山东莱芜,工学博士,西安电子科技大学教授,博士生导师, IEEE会员,中国电子学会高级会员. 主要从事智能信息处理、计算机视觉、基于内容的图像与视频信息检索等领域的研究.

焦李成 男, 1959年 10月出生于陕西白水,工学博士, IEEE高级会员,现为西安电子科技大学教授,博士生导师. 主要从事非线性科学和智能信号处理以及神经网络与大规模并行处理等领域的研究.