DOI:10.13382/j.jemi.2007.03.020
第 21卷 第 3期 电子测量与仪器学报 Vol.21 No.3
2007年 6月 JOURNAL OF ELECTRONIC MEASUREMENT AND INSTRUMENT

一种新的基于粗集编码的模糊聚类数据处理方法

曾黄麟[1] 师奕兵[2] 曾晓辉[3]

(1. 四川理工学院,自贡 643000; 2. 电子科技大学,成都 610054; 3. 成都信息工程学院,成都 610225)

摘 要:本文提出一种新的基于粗集编码的模糊聚类数据处理方法。该方法对电子测量信息处理中的数据,根据粗集理论进行编码、特征属性简化,然后利用模糊隶属度函数将输入精确信息映射为模糊变量信息,提出把数据特征的重要性因子结合在模糊聚类的分类隶属度函数中以提高数据聚类处理的能力,并利用最小化目标函数离线学习来搜索测量数据聚类的聚类中心,该方法可以通过人工神经网络实现。

关键词:模糊,粗集,神经网络,聚类

中图分类号:TP18 文献标识码:A

New Method of Data Processing Based on Fuzzy Clustering of Rough Set Encoding

Zeng Huanglin Shi Yibing Zeng Xiaohui

(Sichuan University of Science and Engineering, 643033, P. R. China;
University of Electronic Science and Technology of China, 610054, P. R. China;
Chengdu University of Information Technology, 610225, P. R. China)

Abstract: A new approach of data processing based on fuzzy clustering of rough set encoding is presented in this paper. An equivalence class encoding input data is defined to eliminate insignificant feature attributes in data sets of electronic measurement data processing by means of rough sets. Fuzzy representation of precise input data is used to deal with either incomplete or imprecise even ill-defined database. A class membership function incorporated the significant factor of the input feature attribute is made to enhance data processing characteristic corresponding to consequent class in the fuzzy clustering output space. One kind of supervised algorithms with batch expression is suggested in searching for the cluster center of data classification by way of LMS rule. The method proposed here can be realized by way of an artificial neural network.

Keywords: fuzzy set; rough set; neural network; clustering

众所周知,电子测量的大量数据要通过处理才能挖掘知识,才能发现规律,才能应用到物理系统中去,所以,数据处理技术在电子测量中越来越显示出它的重要性。近十年迅速发展起来以粗集理论、模糊逻辑、人工神经网络为代表的智能计算[1-4],在数据处理研究领域中十分活跃。

在数据处理中,聚类是从数据中发现知识的关键技术。本文研究一种新的基于粗集编码的模糊聚类数据处理方法。根据粗集理论的方法,首先讨论基于粗集概念的等价类知识编码,从而进行输入数据矢量属性和数据模块的化简,然后,利用模糊隶属度函数将输入精确信息映射为模糊变量信息,以解决病态数据、增强系统分类的非线性映射能力等问题;为了提高数据聚类处理对于输入数据矢量中属性(特征量)的依赖性,我们将数据特征量的重要性因子结合在数据聚类中构建出分类隶属度函数,以提高数据聚类处理的能力。最后,我们将提出一种新的模糊聚类的递推学习方法,该学习方法可用能进行并行信息处理的神经网络实现。

本文研究工作得到四川省教育厅基础应用研究课题基金的部分资助(编号:2005A140)。

本文于 2006年 4月收到。曾黄麟:教授,博士生导师;师奕兵:教授,博士生导师;曾晓辉:助教,研究生。

1 测量数据的粗集编码与简化

对于一个被测系统，测量的数据特征量中可能包含一些相关的因素，是信息的重复和浪费，这就需要进行冗余信息的删除与简化，我们将通过一个数据聚类系统来完成，测量的原始数据是系统的输入，系统通过聚类产生输出，并简化测量数据的结构及模块。

设由 N个输入数据矢量构成的测量数据集，每一个输入数据矢量有 M个特征量（被测数据特征量）。即 $U=\{X_j\}$，

$X_j=(x_{j1}, x_{j2}, \cdots, x_{jn})\in R^n \quad j=1, 2, \cdots, N$

定义 1：被测数据矢量中的特征量 x_i的大小归一化为 L个等价类，测量数据集输入中第 j（$j=1, 2, \cdots, N$）个输入数据矢量的第 i（$i=1, 2, \cdots, M$）个被测数据特征量 x_i等价类归一化为 $[(x_i)]_R=$

$$\left\{x_i \frac{(x_{max}-x_{min})(k-1)}{L}\le x_k\le\frac{(x_{max}-x_{min})k}{L}\right\}$$

$$k=1, 2, \cdots, L \qquad (1)$$

定义 2：测量数据集 N个输入数据矢量的第 i（$i=1, 2, \cdots, M$）个被测数据特征量 x_i对于输出数据分类 W的重要性因子[5]为：

$$\alpha_{x_i}(W)=\frac{card(POS_X(W)-POS_{X-x_i}(W))}{card(U)}$$

$$i=1, 2, \cdots, n \qquad (2)$$

这里 U由训练集 N个输入模式矢量（对象）组成，card表示集合的基数，$POS_X(W)-POS_{X-x_i}(W)$表示系统去掉第 i（$i=1, 2, \cdots, n$）个被测数据特征量 x_i时，对于输出模式分类的正域的减小[4]。

显然，若某些被测数据特征量 x_i对于输出模式分类 W的重要性因子 $\alpha_{x_i}(W)$大，则表明在进行聚类时它的作用大，若 $\alpha_{x_i}(W)$很小，则表明在进行聚类时它不重要，若 $\alpha_{x_i}(W)=0$，我们就可以忽略该被测数据特征量，简化数据结构的知识表达。对于控制系统，可以达到简化系统控制的目的；对于神经网络系统，可以达到缩短网络训练时间的目的；对于模式识别系统，可以达到简化识别计算量等目的。

2 精确信息映射为模糊变量信息

在电子测量数据处理中，我们常常遇到由于环境、测试、传输以及处理过程中引起数据的不精确、不完整，甚至产生病态数据等问题[6-8]，为此，根据模糊理论方法，我们利用模糊隶属度函数将输入精确信息映射为模糊变量信息，解决分类中病态定义的数据问题。

定义 3：模糊化的模糊隶属函数为高斯函数形隶属函数，表达为

$$\mu_A(x)=\exp\left(-\frac{1}{2}\left(\frac{x-c_i}{\sigma_i}\right)^2\right) \qquad (3)$$

这里，高斯函数形隶属函数由二个参数 $\{c_i, \sigma_i\}$确定；c_i确定函数的中心，σ_i确定函数的宽度。

定义 4：设输入数据矢量的一个特征映射为 C个模糊变量，若一个特征的精确值为 x_i，输入矢量特征数据的值域为 (x_{min}, x_{max})，定义各隶属函数的中心为：

$$c_p=\frac{x_{max}-x_{min}}{C-1}(p-1)+x_{min}$$

$$p=1, 2, \cdots, C \qquad (4)$$

由 N个输入数据矢量组成的数据表，每一个输入数据矢量有 M个被测数据特征量，假设通过粗集方法简化后成为 s个特征变量，$s<M$，s维输入模式矢量通过高斯函数形隶属函数模糊化后，得到 $C\times s$维输入模糊变量的隶属度 $\mu(X_j)$。例如，$C=3$，$\mu(X_j)=(\mu_{高}(x_{j1}), \mu_{中}(x_{j1}), \mu_{低}(x_{j1}), \cdots, \mu_{高}(x_{js}), \mu_{中}(x_{js}), \mu_{低}(x_{js}))$为 $3\times s$维输入模糊变量的隶属度。

通过把输入矢量的特征变量映射为模糊变量，把模糊变量$_{高}(x_i)$，模糊变量$_{中}(x_i)$，模糊变量$_{低}(x_i)$再进行等价类划分，不仅使输入数据模糊化，而且使模糊变量$_{高}(x_i)$，模糊变量$_{中}(x_i)$，模糊变量$_{低}(x_i)$数据全部归一为 L个等价类的值。

定义 5：设数据集信息 S=(U, A)可划分为 Q类，根据输出的每一类，由输入数据构成的等价类族 U_l构成对 S=(U, A)的 Q个划分，即 $S=(U, A)_l$，$l=1, 2, \cdots, Q$

其中，A_l是第 l个划分时的输入条件属性和聚类决策属性。

由输入数据矢量构成的等价类族 U_l是由 n_k个输入数据矢量组成。则根据输出的第 l类，由输入及输出重新构成的数据表就是一个 $\sum_{k=1}^{Q} n_{lk}=N$行，$3s+|d_l|$列的数据聚类决策表，其中 $|d_l|$是数据聚类的属性（特征量）的个数。

同样,如果我们定义一个模糊变量的隶属度门限函数为 T,若隶属度小于门限函数的模糊变量的等价类消去,则还可以使由输入数据及聚类决策属性构成的知识分类表进一步简化。

3　一种新的模糊聚类学习方法

假定通过上述粗集方法简化后,被测数据特征量由 M个减少到 s个,模糊聚类决策分类的输入为 s维矢量 $X_j=(x_1, x_2, \cdots, x_s)\in R^s$,通过输入精确特征值转化为模糊变量,一个输入为 s维矢量转换为 C× s维输入模糊变量,例如,

$$F_j=(\mu_{高}(x_{ij}), \mu_{中}(x_{ij}), \mu_{低}(x_{ij})),$$
$$i=1, 2, \cdots, s \quad j=1, 2, \cdots, N$$

设训练集信息 $S=(U, A)$ $j=1, 2, \cdots, N$可划分为 Q类,根据输出的每一类,假设 $n_k(\sum_{k=1}^{Q} n_k=N)$个输入数据矢量映射为第 $l(l=1, 2, \cdots, Q)$类输出分类。

定理 1: $n_k(\sum_{k=1}^{Q} n_k=N)$个输入数据矢量的第 i个特征分量,映射到输出分类属于第 l类时的中心的第 i个分量为:

$$u_{il}=\frac{\sum_{j=1}^{n_k} f_l(X_j)\, x_{ij}}{\sum_{j=1}^{n_k} f_l(X_j)}, \quad i=1, 2, \cdots, s$$
$$l=1, 2, \cdots, Q \tag{5}$$

搜索测量数据聚类的聚类中心对于第 l个分类的中心的各分量的更新为:

$$u_{il}(t+1)=u_{il}(t)+\eta(t)(x_{ij}-u_{il}) f_l^2(X_j)$$
$$i=1, 2, \cdots, s \tag{6}$$

这里 $\eta(t)=\frac{1}{|t|}$表示学习衰减因子。

证明:定义 n_k个输入数据矢量中,输出分类属于第 $l(l=1, 2, \cdots, Q)$聚类时,第 j个训练数据矢量 X_j与输出分类属于第 l类时的中心的加权距离为:

$$D_{jl}=\frac{\|\mu(X_j)-\mu(U_l)\|}{\sum_{i=1}^{s}\alpha_{xi}(W)\, x_{ij}}$$
$$=\frac{\sum_{i=1}^{s}\sum_{p=1}^{3}[\mu_p(x_{ij})-\mu_p(u_{il})]^2}{\sum_{i=1}^{s}\sum_{p}^{3}\alpha_{xi}(W)\mu_p(x_{ij})}$$
$$l=1, 2, \cdots, Q \quad j=1, 2, \cdots, n_k \tag{7}$$

定义在 n_k个输入数据矢量中,第 j个训练数据矢量 X_j属于第 $l(l=1, 2\cdots, Q)$类输出分类时的隶属函数定义为:

$$f(X_j)=(1+D_{lj})^{-1}$$
$$=\frac{\sum_{i=1}^{s}\alpha_{xi}(W)\, x_{ij}}{\sum_{i=1}^{s}\alpha_{xi}(W)\, x_{ij}+\|\mu(X_j)-\mu(U_l)\|},$$
$$j=1, 2, \cdots, n_k \quad l=1, 2, \cdots, Q \tag{8}$$

定义一个网络的目标函数:

$$E=\frac{1}{2}\sum_{l=1}^{Q}\sum_{j=1}^{n_k} f_l(X_j)\|X_j-U_l\| \tag{9}$$

这里

$$\|X_j-U_l\|=\sum_{i=1}^{s}(x_{ij}-u_{il})^2$$
$$=\sum_{i=1}^{s}\sum_{p}^{3}[\mu_p(x_{ij})-\mu_p(u_{il})]^2$$

我们通过最小化网络的目标函数来求网络的分类的中心[8,9]。对于分类的中心的各分量,我们求

$$\frac{\partial E}{\partial u_{il}}=\frac{1}{2} f_l(X_j)(\|X_j-U_l\|)'+\frac{1}{2}\|X_j-U_l\| f_l'(X_j)$$
$$=f_l(X_j)(-(x_{ij}-u_{il}))-\frac{1}{2}\|X_j-U_l\|(1+D_{lj})^{-2} D_{lj}'$$
$$=(-(x_{ij}-u_{il})) f_l(X_j)+(x_{ij}-u_{il}) f_l(X_j)$$
$$\frac{\|X_j-U_l\|}{(1+D_{lj})\sum_{i=1}^{s}\alpha_{xi}(W)\, x_{ij}}$$
$$=-(x_{ij}-u_{il}) f_l(X_j)\left(1-\frac{\|X_j-U_l\|}{(1+D_{lj})\sum_{i=1}^{s}\alpha_{xi}(W)\, x_{ij}}\right)$$
$$=-(x_{ij}-u_{il}) f_l(X_j)\left(1-\frac{\|X_j-U_l\|}{\|X_j-U_l\|+\sum_{i=1}^{s}\alpha_{xi}(W)\, x_{ij}}\right)$$
$$=-(x_{ij}-u_{il}) f_l^2(X_j) \tag{10}$$

根据梯度下降原理,有

$$\frac{du_{il}}{dt}=-\eta(t)\frac{\partial E}{\partial u_{il}}$$
$$l=1, 2, \cdots, Q \quad i=1, 2, \cdots, s \tag{11}$$

故对于第 l个分类的中心的各分量的更新为

$$u_{il}(t+1)=u_{il}(t)+\eta(t)(x_{ij}-u_{il}) f_l^2(X_j)$$
$$i=1, 2, \cdots, s$$

批训练学习后的网络,我们就得到一组确定的

分类中心 u_{ij}, $i=1,2,\cdots,s$ $j=1,2,\cdots,Q$然后计算任一输入数据矢量 X_j与输出分类属于第 i类时的中心的加权距离 D_{ij}, $i=1,2,\cdots,Q$ $j=1,2,\cdots,n$,再计算任一输入数据矢量 X_j属于第 i类输出分类时的隶属函数 $f_i(X_j)=(1+D_{ij})^{-1}$, $j=1,2,\cdots,n$, $i=1,2,\cdots,Q$ 从而确定输入数据矢量 X_j属于第 i类输出分类。

上述模糊聚类方法,可以利用具有并行数据处理能力的神经网络方法完成学习和数据的聚类处理[7]。

4 结 论

在数据处理中,聚类是从数据中发现知识的关键技术。在这篇文章中,我们利用粗集理论的方法,对知识进行编码、属性简化,从而达到简化数据表达等目的。为了提高聚类数据的能力,我们把测量数据矢量的属性重要性因子结合在学习和分类中构建了一个分类隶属度函数,提出了有导师的批处理学习方法,该方法可以由完全并行和分布的神经网络来实现和搜索数据模式分类及聚类中心。

本文提出利用模糊隶属度函数将输入精确信息映射为模糊变量信息,能将输入空间混叠特征矢量通过非线性映射到输出分类矢量空间,能实现病态定义数据的分类,该学习方法不仅可以应用于电子测量数据的聚类处理,也可以应用于模式识别等领域。

参考文献:

[1] Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Nowowiejska: Warsaw, Poland, 1990: 2-25.

[2] A. Kandel. Fuzzy expert systems[M]. CRC Press, Inc., Boca Raton, FL, 1992: 68-97.

[3] B. Kosko. Neural networks and fuzzy systems: a dynamical systems approach[M]. Prentice Hall: Upper Saddle River, NJ, 1991: 45-64.

[4] 曾黄麟. 粗集理论及其应用(修定版)[M]. 重庆大学出版社, 1998: 40-55.

[5] 曾黄麟. 智能计算(关于粗集理论、模糊逻辑、神经网络的理论及其应用)[M]. 重庆大学出版社, 2004: 20-40.

[6] M. Banerjee, S. K. Pal. Rough fuzzy MLP: knowledge encoding and classification[J]. IEEE Transactions on neural networks, 1998, 9: 1203-1216.

[7] R. Swiniarski, L. Hargis. Rough sets as a front end of neural-networks texture classifiers[J]. Inter. Journal Neurocomputing, 2001, 36: 85-102.

[8] W. Skarbek. Rough sets for enhancements of local subspace classifier[J]. Inter. Journal Neurocomputing, 2001, 36: 67-84.

作者简介:

曾黄麟:四川理工学院教授、院长,电子科技大学博士生导师。主要研究方向为智能信息处理、模式识别等。