计算机科学 2000Vol. 27№. 12

面向数据挖掘的时间序列聚类方法研究*)

On the Data-Mining Oriented Methods for Clustering Time Series

李 斌 谭立湘 章劲松 庄镇泉
(中国科学技术大学电子科学与技术系 合肥 230026)

Abstract According to the characteristics of data mining, some improvements are made to the neural network fuzzy clustering algorithm FSART to make it more efficient and can perform incremental clustering. For the need of the clustering analysis of time series, a new fuzzy membership expression that can describe the modality similarity of vectors is proposed. The new fuzzy membership expression and the improved FSART algorithm are combined to implement the clustering analysis of the nonstationary time series.

Keywords Data mining, Clustering analysis, Time series, Neural network, Algorithm FSART

一、引言

自然界以及我们社会生活中的各种事物都在运动、变化和发展着，将它们按时间顺序记录下来，我们就可以得到各种各样的"时间序列"数据。对时间序列进行分析，可以揭示事物运动、变化和发展的内在规律，对于人们正确认识事物并据此作出科学的决策具有重要的现实意义。

数据挖掘，也称知识发现，研究从大量的历史数据中发现隐含的、事先未知的和有价值的知识的方法，是一种新兴的、面向决策支持的数据处理手段。针对时间序列的数据挖掘研究从大量时间序列历史数据中发掘有价值信息的算法及实现技术，也是一个新的、极具挑战性和有着重要应用前景的研究领域。

聚类又称无监督分类，是在没有任何导师信号的情况下，根据样本自身的分布对其进行分类的一种数据分析技术，也是数据挖掘中的关键技术之一。在数据挖掘应用中，一个重要的特点是数据驱动，即对于数据的分布没有任何先验知识，完全根据数据自身所提供的信息对其进行分类，因而要求面向数据挖掘的聚类算法应具有一定的自适应性，如应能够在聚类过程中自主地确定类的个数，而不需要事先人为设定。数据挖掘应用的另一个重要特点是待处理的数据量庞大，对数据集(或数据库)的一次遍历往往需要花费很长的时间，因而传统聚类算法中的块迭代方法将不再适用，而应努力研究合适的在线增量式聚类算法。此外，由于数据量庞大，对算法的计算效率也提出了很高的要求，目前数据挖掘研究的一个重要的任务就是寻找更加高效的计算方法[1]。

针对时间序列的聚类分析一般要求同一类中的时间序列片段应具有相似的变化形态，据此，本文提出了一个新的隶属度函数表达式，用该表达式可以根据矢量间形态的相似度对其进行分类。FSART(Fuzzy Simplified Adaptive Resonance Theory，模糊简化自适应谐振理论)算法是一种新颖的神经网络模糊聚类算法，它综合了多种神经网络无导师学习算法的优点，具有结构灵活、便于调节、聚类精度高、对噪声和初始状态不敏感等特点[2]。本文根据数据挖掘应用的要求对FSART算法做了进一步的改进，使其由一个离线块迭代学习算法变为一个可进行在线增量式学习的聚类算法，并对算法结构作了进一步的简化，使其计算效率得到进一步提高。利用本文提出的隶属度函数表达式和改进后的FSART算法对金融领域的时间序列数据进行了聚类分析，获得了良好的实验结果。

二、隶属度函数表达式

在模糊集理论中，隶属度被用来描述某一对象属于某一类的可能程度。本文采用隶属度函数来度量同类模式间的相似性。根据描述模糊对象时的要求不同，隶属度函数可分为两类：相对模糊隶属度 R_f 和绝对

*)973 国家重点基础研究发展规划项目资助(项目编号：G1998030413)。李 斌 博士生，研究方向为神经网络、数据挖掘、非线性信号处理；谭立湘 讲师，研究方向为数据库、数据通信、多媒体；章劲松 博士生，研究方向为混沌神经网络、人工生命；庄镇泉 教授，博士生导师，研究方向为智能信息处理。

模糊隶属度 $A_{i,j}$。其中，i 为模式的索引顺序号，j 为类的索引顺序号。它们之间的关系如下式所示：

$$R_{i,j}=A_{i,j}/\sum_{k=1}^{c}A_{i,k}\quad i=1,\cdots,n,j=1,\cdots,c \tag{1}$$

其中，绝对隶属度函数 $A_{i,j}$ 的计算表达是我们研究的重点，它应满足条件：1）$A_{i,j}\in[0,1]$，$i=1,\cdots,n$，$j=1,\cdots,c$；2）$\max_i\{A_{i,j}\}>0$，$i=1,\cdots,n$，$j=1,\cdots,c$；3）$0<\sum_{i=1}^{c}A_{i,j}<n$，$i=1,\cdots,n$，$j=1,\cdots,c$。

目前文献中已有的绝对隶属度函数的表达式有以下几个[3~5]：

$$A_{i,j}=\begin{cases}1/(1+d_{i,j}^2/\eta_j)\in(0,1] & (2)\\ Gaussian_{i,j}=e^{-d_{i,j}^2/\sigma_j^2}\in(0,1] & (3)\\ 1/(d_{i,j})^{\rho_j}\in(0,\infty) & (4)\\ 1/(1-Gaussian_{i,j})^{\rho_j}\in(1,\infty) & (5)\end{cases}$$

其中，$d_{i,j}=d(X_i,W_j)$ 一般为输入模式 X_i 与第 j 类的接收域中心矢量 W_j 的欧氏距离，$\sigma_j,\eta_j,\rho_j\in(0,\infty)$。

这几个表达式都只是两点间欧氏距离的函数，不能充分反映两个矢量形态的相似程度。所谓形态相似，指的是两个矢量从一定宏观角度观察有相近的变化方向和模。例如，有两个矢量，保持其欧氏距离不变而改变它们的模，那么它们在不同模时所呈现的形态相似度是不一样的，模越大相似度越高。因而，要反映矢量间形态的相似性，隶属度函数表达式应同时包含距离和各矢量模的变化，而不仅仅是欧氏距离的函数。据此，本文提出了一个新的、反映矢量形态相似度的隶属度函数的计算表达式，如式(6)所示：

$$A_{i,j}=1-\frac{d_{i,j}^2}{d_i^2+d_j^2}=\frac{2\sum_{k=1}^{n}X_{i,k}W_{j,k}}{\sum_{k=1}^{n}X_{i,k}^2-\sum_{k=1}^{n}W_{j,k}^2}\in(0,1] \tag{6}$$

其中，$d_{i,j}$ 为矢量 X_i 与矢量 W_j 之间的欧氏距离，d_i，d_j 为两矢量的模。当两矢量完全相等时，$A_{i,j}=1$。

不难看出，上式展开后即可作为一个新的相似性距离度量，由于其变化满足上述绝对隶属度函数的条件，因而可以直接用来表示绝对隶属度。与 Dice 相似性距离度量不同之处在于，该相似性度量增强了矢量的模在判断两矢量是否相似时所起的作用。

此外，在国际标准数据集 IRIS 上进行的实验表明：该绝对隶属度函数表达式对样本空间呈球形分布的分类问题还具有一定的普适性，我们在 FSART 算法实现中采用该绝对隶属度函数表达式，对 IRIS 数据集做聚类分析，结果为：均方误差 MSE=0.0408，误分类数 $N_{mis}=12$，可与国内外同类工作相媲美（文献中已有的对 IRIS 数据集聚类的误分类数在 10—16 之间）[3~4]。在实验中我们还发现，相比其它表达式，采用本文所提出的隶属度表达式可以使算法更加稳定和易于调节。

三、模糊简化 ART(FSART)算法

FSART 算法是在 ART 算法的基础上演化而来的，它在保留 ART 算法基本框架的基础上，引入了软竞争学习策略、模拟退火学习策略、节点动态去除和分布式参数调整等优秀的算法思想，其主要步骤如下：

1）初始化　为运行 FSART，需要事先指定警戒阈值 $\rho\in[0,1]$，并为每个节点必须维持的迭代周期数指定一个下限 $e_{min}\geq1$（该参数将影响学习率因子单调下降的速率，进而影响算法到达终点的时间）。当产生一个新的输出单元 E_c 时，其本地（基于处理单元的）时间计数器 e_c 被初始化为 0。FSART 为每个输出单元配备一个计数器，用于记录该输出单元的生存时间，计数器的值等于该单元经历的迭代周期数。该时间值将被用来计算该输出单元的学习率，采用基于神经元的时间计数器，使得分布式的、基于神经元的参数调整成为可能，使得算法对动态变化的环境具有更强的适应力。

2）处理单元的激活　输入矢量进入网络后，首先进行的是各处理单元的激活函数值的计算。激活函数采用式(1)所定义的相对隶属度函数。FSART 采用模糊隶属度函数作为其激活和匹配函数，因而可以处理模拟量。由于采用了相对隶属度函数，使得处理单元的激活是类间相关的。

3）更新邻域的确定　FSART 采用一种基于输出空间的相邻等级排序的软竞争学习机制。最高等级 $r_{j^*}^{(0)}=0$ 被赋予最佳匹配单元 E_{j^*}，其中，j^* 通过求解如下最大值得到：$j^*=\arg\max_{k=1,\cdots,c}(R_{ik}^{(0)})$。次高等级 $r_j^{(1)}=1$，如果激活函数值 $R_{ij}^{(1)}$ 为第二大，等等。

4）谐振域的产生　在单元激活并根据各单元的激活函数值大小对其进行相邻等级排序后，即进入"由上向下"的匹配过程，匹配函数 $A_{i,j}$ 取绝对隶属度函数（如本文所给出的绝对隶属度函数表达式）。所有满足条件：$A_{i,j}\geq\rho$（ρ 为警戒阈值）的处理单元均归入谐振域。谐振域是全部有资格参加权值调整的处理单元的集合。当谐振域中节点的个数大于或等于 1 时，表明输入矢量可归入某一个或多个已有的类中；当谐振域中节点的个数为 0 时，表明输入矢量与现有所有类的中心都相距太远，此时，系统自动产生一个新节点来匹配该输入矢量。由于在许多情况下谐振域中的节点个数是大于 1 的，因而一次有多个节点同时参加学习，这一点与传统 ART 算法不同，而与 SOM 算法非常相似，这也是软竞争学习的主要特征。与 SOM 采用几何邻域不同，FSART 采用相似性测度邻域，这更接近我们的分类要求。通过采用软竞争学习策略，使得算法相对不易于陷入局部最小，并且对输入模式的输入顺序相

对不敏感。

5)谐振——权值更新　对所有属于谐振域的处理单元采用同一Kohonen权值调整算法。

$$W_{kj}^{(t+1)}=W_{kj}^{(t)}-\beta_j^{(t)}\cdot(X_k^{(t)}-W_{kj}^{(t)}),$$

$$k=1,\cdots,d,\ r=1,\cdots,n,\ \forall\ j\in\{1,c\} \tag{7}$$

上式中，处理单元E_j(其本地时间计数器为e_j)的学习率$\beta_j^{(t)}$定义为：

$$\beta_j^{(t)}=\varepsilon_j^{(t)}\cdot h_j^{(t)},\ \forall\ j\in\{1,c$$

$$\beta_j^{(t)},\varepsilon_j^{(t)},h_j^{(t)}\in[0,1] \tag{8}$$

其中，因子$\varepsilon_j^{(t)}$随时间单调衰减，决定了学习率必然也随时间单调衰减；因子$h_j^{(t)}$随时间和单元的相邻等级(即更新域半径)的增加单调下降，反映了"软竞争"学习的特点以及算法结构由"软"到"硬"的演变过程。

6)计时器更新与节点去除判断　当全部输入数据集被送入系统后，即完成一次训练周期，要进行下列操作：

· 各节点的时间计数器加1，$e_j=e_j+1$，$j=1,\cdots,c$；

· 执行节点去除机制。即，对于$\forall\ j\in\{1,c\}$，如果某处理单元E_j在刚才那一轮训练周期中对所有输入模式都不是最佳匹配单元，则将其去除，处理单元计数器c减1，$c=c-1$。

单元动态去除机制与动态生成机制相结合，使得算法对复杂环境呈现良好的"弹性"(动态适应能力)，也使算法对初始状态不那么敏感。

四、对FSART算法的改进

1. 算法结构的进一步简化

数据挖掘所面对的是海量数据，因此算法的计算复杂度成为评判算法优劣的一个重要的指标，在功能相同的情况下，计算复杂度越小，算法越好。FSART算法保留了ART算法的"由下向上"激活和"由上向下"匹配两步操作，之所以这样，是因为在ART算法中存在所谓"内星"和"外星"两个不同的参考矢量，并且所采用的激活和匹配函数为单向的非对称函数，即在函数表达式中输入矢量与参考矢量的位置不能互换。而在FSART算法中参考矢量只有一个，即代表聚类中心的权值矢量W_{kj}，如果所采用的绝对隶属度和相对隶属度函数均为双向的对称函数，且满足条件：$A_{r,1}>A_{r,2}\Rightarrow R_{r,1}>R_{r,2}$，我们就可以将两步操作合并为一步，使算法的计算复杂度得到明显降低。

具体做法是：取消"由下向上"的处理单元的激活计算，而直接计算各单元的匹配函数(即绝对隶属度函数)，在根据警戒阈值ρ确定谐振域的同时，按照绝对隶属度A对各处理单元进行相邻等级排序。如果谐振域中单元个数为0，则直接产生一个新的输出单元来匹配输入样本，不再需要有"搜索"过程。由上述条件可知，各单元的绝对隶属度与相对隶属度必然遵循相同的偏序排列。这样，原来的"由下向上"激活和"由上向下"匹配两个过程被简化为一个过程，网络结构也由原来的双向网络简化为单向网络，计算复杂度明显减小。改进后算法的计算步骤如下

1)初始化，同上节1)，

2)根据式(6)计算输入矢量对各输出节点的绝对隶属度函数值$A_{r,j}$，由绝对隶属度函数值和给定的警戒阈值确定谐振域。如果谐振域中的节点个数大于或等于1，则根据其隶属度函数值对输出节点进行相邻等级排序，并根据式(7)对谐振域中各节点作权值调整；如果谐振域中节点个数为0，则产生一个新的输出节点来匹配输入矢量。

3)当全部样本都已输入，各输出节点的计时器加1，同时进行节点去除判断，判断依据同上节6)。

本文所提出的绝对隶属度函数表达式满足上述偏序要求，因而在实现时对算法作了如上简化，实验表明，采用该简化措施对计算结果没有什么影响，而计算时间明显减少了。

2. 算法的在线化改造

数据挖掘的大数据量的特点要求其所采用的聚类算法应可以支持在线增量式学习。在线学习算法要求在保留已有知识的同时还能够对新知识进行学习。FSART算法是作为离线学习算法提出的，为了实现在线学习，本文对算法做了进一步的改造：

1)改变各单元计时器的计时步长。由原来每输入一批样本计时一次改为每输入n个($n\geq1$，n越大，算法对时间的变化越不敏感，本文中取$n=1$)样本计时一次。

2)将单元的动态去除判断改为每隔一定时间进行一次。当某单元的计时器值达到某一设定值时还未被选中一次，则认为该单元所代表的为噪声(outlier)类，将其去除。这样，在学习过程中零星出现的与其它点差异很大的点将被作为噪声滤除。

3)保留模拟退火学习策略，并为学习率因子的衰减设置下限。模拟退火策略在学习初期取较大的学习率因子，使得算法可以快速地收敛到稳定状态附近。随着时间的推移，学习率因子单调减小，新的输入样本对结果的影响越来越小。但是，在线(增量式)学习算法要求在保留过去学习成果的同时，应能根据新的输入样本作增量式调整。因而，学习率因子应有一个下限，当学习率因子单调下降到该下限时就不再减小，以后算法将以恒定的学习率进行增量式学习，此时，学习策略便退化为普通的"爬山式"学习策略。

4)保留软竞争学习策略。软竞争学习策略在学习

初期有助于降低算法对初始状态的敏感性，使系统不易陷入局部最小。随着时间的推移，算法结构逐渐由软竞争学习过渡到硬竞争学习，算法的模糊性逐渐减小，直至完全失去模糊性。

五、实验结果

在对时间序列数据作数据挖掘分析时，往往首先需要将以数值数据表示的时间序列转换成以相对抽象的符号表示的符号序列，其过程一般是：首先，将长的时间序列依据其变化特征分割成若干短的时间序列片段；然后，对这些时间序列片段进行聚类，并为每个类分配一个类标识符；最后，以该类标识符表示所有属于该类的时间序列片段，进而得到一个由各个类标识符所构成的符号序列。

本文的实验使用美国股市 Nasdaq 指数（反映美国股市高科技类股变化的综合指数）500 天的数据。在做时间序列分割时，我们希望分割后获得的每一个时间序列片段都具有相对独立的变化模式。为此我们选择线性化分段方法对时间序列进行分割，因为该方法具有很好的形态表达和分割能力。本文采用的线性化分段算法的基本思想是：先将整个时间序列分为许多小的分段（如 3 个点为一段），然后通过相邻分段之间不断的融合来减少分段的数量，直至达到要求的分辨率。处理后的结果如图 1 所示。

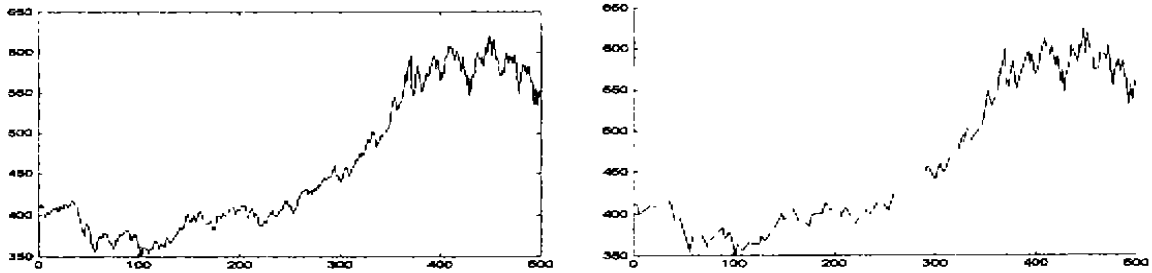


图 1　时间序列的线性化分段表示

然后，利用本文所提出的隶属度函数表达式和改进后的 FSART 算法对分割后的时间序列片段进行聚类分析。为简化问题，我们将这些分段全部映射到同一个二维坐标平面，且使得所有线段都从原点出发，这样，我们就得到了一组二维点集。接下来，我们就对该二维点集作聚类分析，图 2 给出了聚类的结果（警戒阈值 $\rho=0.8$，不同的点型表示该点属于不同的类，大的圆点表示类的中心），由图可见所产生的聚类中心较好地代表了具有相似形态的点在空间的分布，实验还表明，聚类算法具有很好的收敛性和稳定性。

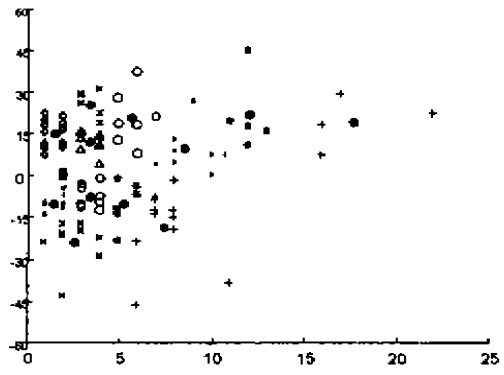


图 2　FSART 算法的聚类结果

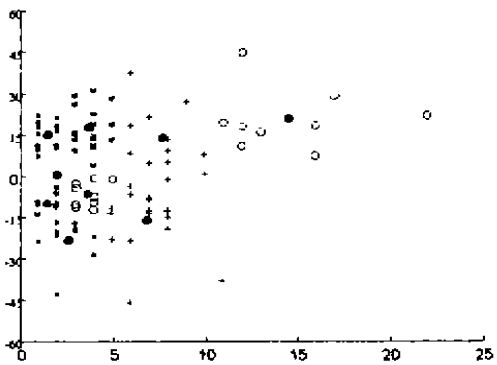
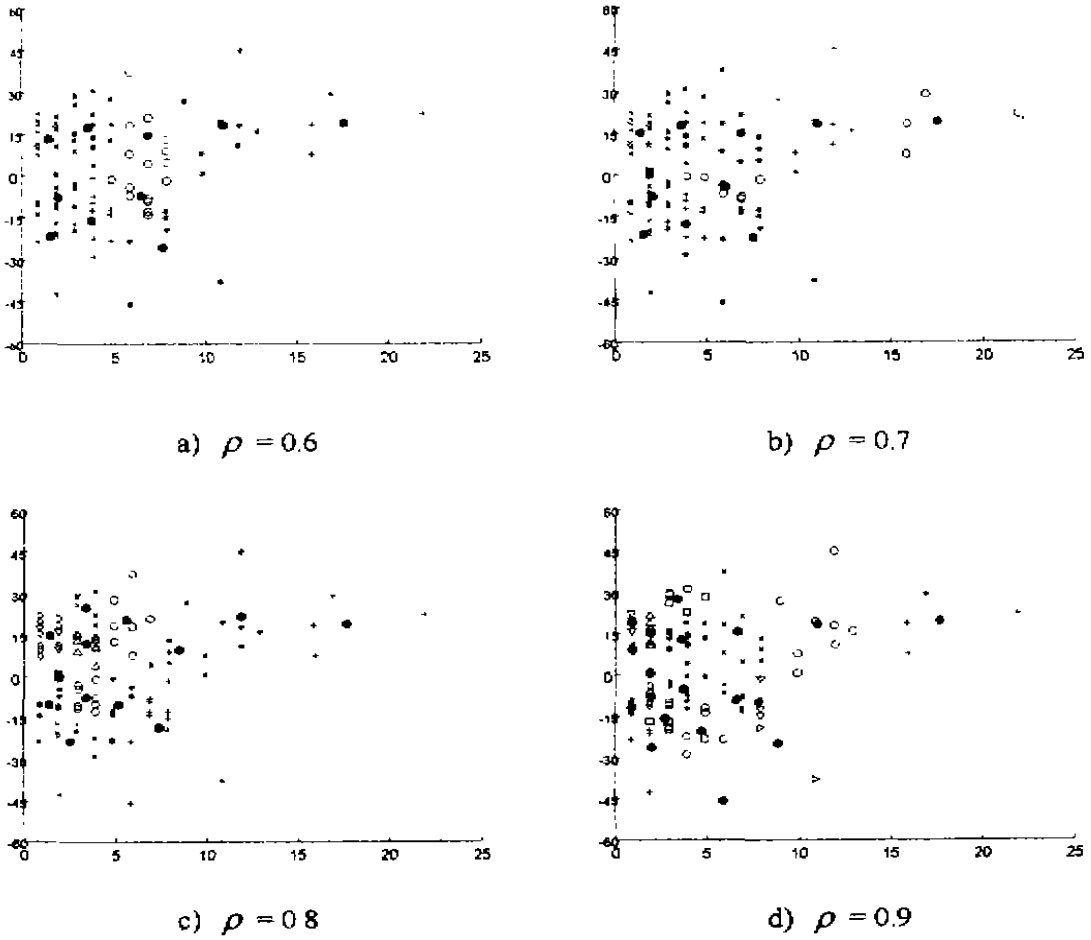


图 3　无软竞争学习时的聚类结果

ART 类算法的一个重要的特点是具有警戒阈值判断功能，警戒阈值 ρ 反映了外界对类内相似度的一种要求（或期望），算法根据 ρ 值大小，通过竞争学习，自主地确定类的个数及类中心的位置。FSART 算法保留了 ART 类算法的这一特性，所不同的是，由于引入了软竞争学习策略，警戒判断所产生的可参加权值调整的节点不再只有一个，而可能有多个，这一方面降低了初始状态和输入顺序对算法的影响，另一方面也提高了算法对警戒阈值 ρ 的敏感程度。图 3 给出了从

FSART 算法中去除了软竞争学习后所得出的聚类结果，与图 2 相比较可见，在其它条件不变的情况下取消软竞争学习使得聚类所产生的类的个数明显减少。图 4 给出了 ρ 取不同值时聚类的结果，可以看出，ρ 值取得越大，对类内相似度要求越高，相应的分出的类的个数也越多。



a) $\rho = 0.6$　　b) $\rho = 0.7$

c) $\rho = 0.8$　　d) $\rho = 0.9$

图 4　警戒阈值 ρ 取不同值时的聚类结果

结束语　针对时间序列的数据挖掘是一个新兴的研究领域，聚类分析是其中的一个重要的技术环节，面向数据挖掘的聚类要求算法具有自适应性、高效性和在线学习等特性，目前国际上广泛采用的是基于精确描述的归纳学习算法，本文通过对 FSART 算法进行改进，用神经网络模糊聚类算法实现了对时间序列数据的在线聚类分析，同时将软竞争学习、模拟退火等先进思想引入在线聚类分析，取得了较好的实验结果，拓展了神经网络的应用领域，同时也丰富了数据挖掘的研究方法。

参 考 文 献

1　Mingsyan C, Jiawei H, Philip SY. Datamining: An Overview From a Database Perspective. IEEE Trans. Knowledge and Data Engineering, 1996, 18(6): 833～866

2　Baraldi A, Alpaydm E. Simplified ART: A new class of ART algorithms. [Technique Report of INTERNATIONAL COMPUTER SCIENCE INSTITUTE]. Berkeley, California, 1998

3　Krishnapuram R, Keller J M. A possibilistic approach to clustering. IEEE Transactions on Fuzzy Sytems, 1993, 1(2): 98～110

4　Williamson J R, A constructive, incremental-learning network for misture modeling and classification. Neural Computation, 1997, 9: 1517～1543

5　Bezdek J C, Pal N R. Generalized clustering networks and Kohonen's self-organizing scheme. IEEE Trans. On Neural Networks, 1993, 4(4): 549～557

6　Baraldi A, Parmiggiani F. Novel neural network model combining radial basis function, competitive hebbian learning rule, and fuzzy simplified adaptive resonance theory. In: Proc. SPIE's Optical Science, Engineering and Instrumentation'97: Applications of Fuzzy Logic Technology IV, San Diego, CA, 1997, 3165: 98～112

7　Bezdek J C, Pal N R. Two soft relatives of learning vector quantization. Neural Networks, 1995, 8(5): 729～743

8　Tsao E C, Bezdek J C, Pal N R. Fuzzy Kohonen clustering network. Pattern Recognition, 1994, 27(5): 757～764