第33卷 第12期 计 算 机 工 程 2007年6月
Vol.33 No.12 Computer Engineering June 2007
·软件技术与数据库· 文章编号：1000—3428(2007)12—0063—02 文献标识码：A 中图分类号：TP301.6

一种基于模糊神经网络的数据挖掘算法

李良俊[1,2]，张 斌[1]，杨 明[2]
（1. 东北大学信息科学与工程学院，沈阳 110036；2. 鞍山师范学院计算中心，鞍山 114005）

摘 要：提出了一种基于模糊神经网络的数据挖掘算法，把模糊理论和神经网络结合起来构造、训练模糊神经网络，弥补了神经网络结构复杂、网络训练时间长、结果表示不易理解等不足。经过模糊神经网络的建立和训练达到精度要求，实现了运用模糊神经网络方法从数据库中提取知识的目标。

关键词：模糊理论；神经网络；数据挖掘

Data Mining Algorithm Based on Fuzzy Neural Network

LI Liangjun[1,2], ZHANG Bin[1], YANG Ming[2]
(1. School of Information Science and Eng., Northeast University, Shenyang 110036;
2. Computing Center, Anshan Normal University, Anshan 114005)

【Abstract】 This paper presents a data mining algorithm based on fuzzy neural network (FNN). Using fuzzy theory and neural network to structure and train fuzzy neural network, the algorithm overcomes the shortcomings of neural network such as complex structure, long training time and lack of understandable representation of results. Establishment and training of fuzzy neural network which meet the precision requests realize the utilization fuzzy neural network method to withdraw the knowledge from the database.

【Key words】 Fuzzy theory; Neural network; Data mining

1 概述

随着计算机应用的普及和数据库技术的不断发展，数据库管理系统的应用领域越来越广泛。日常生活、学习、工作等各方面涉及到的各种数据均成为数据库系统的信息来源，致使数据库规模越来越庞大。如何管理这样大规模的数据库，如何从大量的数据中获取需要的信息呢？为了解决这一问题，数据仓库与数据挖掘技术应运而生。

数据仓库是支持管理决策过程的、面向主题的、集成的、时变的、非易失的数据集合[1]。数据仓库早期主要用于产生报告和实现预先定义的查询，之后用于分析汇总的数据和细节的数据，并以报告和图表形式提供结果。再后来用于决策，进行多维分析和复杂的切片和切块操作，最后用于知识发现，并使用数据挖掘工具进行决策。

数据挖掘就是从大型数据库的数据中提取人们感兴趣的知识的过程。这些知识是隐含的，事先未知的潜在有用的信息，提取的知识表示为概念（Concept）、规则（Rule）、规律（Regularity）、模式（Pattern）等形式[2]。多数情况下，数据挖掘都要从数据仓库中取出数据存到数据仓库或数据集市中，如图1所示。

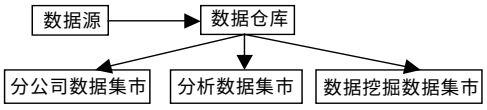


图1 从数据仓库中得出的数据挖掘库

基于模糊神经网络（fuzzy neural networks，FNN）的数据挖掘就是对数据进行处理，获得各种规则、条件等。神经网络模拟人脑内部结构，在模拟推理、自动学习等方面接近人脑的自组织和并行处理能力。其优点之一是不依赖于对象的数学模型，通过学习将输入与输出以权值的形式编码，把它们关联起来。它在数据挖掘和模式分类中有以下优势：

(1)分类精确，鲁棒性好；

(2)可用于各种算法进行规则提取；

(3)模式分类能力强[3]，具有在线自适应、非线性可分、刚柔相济决策、非参数分类等特点。

模糊理论在知识和规则获取中可发挥重要作用。人类语言和思维具有模糊性，模糊思维形式和语言表达具有广泛、完美和高效的特征。日常工作和学习中有很多知识是模糊的，这些模糊知识往往在控制和决策中发挥巨大的作用。在数据挖掘中运用模糊性处理方法可有效解决神经网络中存在的不足：训练时间长，规则提取困难，无法对不确切或模糊的数据进行数据挖掘。

2 利用模糊神经网络进行的网络构造和训练

针对一个各个属性为连续变量的多属性数据库，若希望提取出其中n个属性和另外m个属性之间关系的规则，则设有n个输入，用$x_1,x_2,x_3,\ldots,x_n$表示，其对应的m个输出用$y_1,y_2,y_3,\ldots,y_m$表示。利用模糊神经网络通过隶属度函数把它们化为相应的模糊输入和模糊期望输出的隶属度值，以这些隶属度值作为训练样本，最终希望得到如下形式规则：

if$(x_1\theta\upsilon_1)$ and $(x_2\theta\upsilon_2)$ and…and $(x_n\theta\upsilon_n)$ then $(y_1\theta\zeta_1)$ and $(y_2\theta\zeta_2)$ and…and$(y_m\theta\zeta_m)$

其中，θ为关系运算符（=、<、>、≥、≤）；υ_i、ζ_i为模糊属性（大、中、小）。

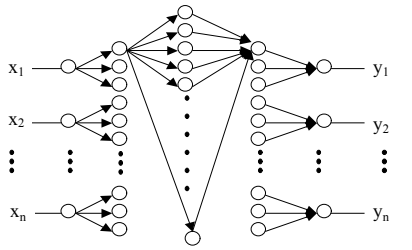
作者简介：李良俊(1967－)，男，博士生，主研方向：数据挖掘，神经网络；张 斌，教授、博导；杨 明，副教授

收稿日期：2006-07-04 **E-mail：**llj@mail.asnc.edu.cn

2.1 模糊神经网络的结构

模糊神经网络根据模糊逻辑与神经网络的结合方式，分为 2 大类：(1)仍采用普通神经网络的结构，但将普通非线性神经元用模糊运算神经元代替；(2)采用普通神经网络的结构和神经元作为信息处理工具，而网络的输入量、输出量等则采用输入输出信息的模糊隶属度。

通过对第 2 种类型进行研究和改进，在 Takagi 等人提出的神经网络驱动的模糊推理基础上，构造一个能够实现模糊矢量分类分析的 5 层前向网络解决这一问题。第 1 层为输入层；第 2 层计算输入数据的隶属度，通过每个属性各自模糊隶属度函数化为 3 个分属于大、中、小语言变量的隶属度值。然后使隶属度最大的单元输出为 1，其余为 0，作为第 3 层的输入；第 3 层完成模糊控制规则的前件处理，按乘积原则计算；第 4 层完成模糊控制规则的后件处理，按求和原则进行，完成所有规则前件部的隶属度求和计算；第 5 层为期望输出层，通过隶属度函数化为隶属度值，最大取 1，其余为 0。模糊神经网络如图 2 所示。



(a)第1层 (b)第2层 (c)第3层 (d)第4层 (e)第5层

图 2 模糊神经网络的结构

2.2 模糊神经网络的学习算法

(1)样本数据的标准化

设有k个样本$x_1,x_2,\ldots,x_k$，每个样本x_i有n个样本指标$z_1,z_2,\ldots,z_n$；x_{ij}表示第i个样本的第j个指标，k个样本的n个指标表示如表 1 所示。

表 1 k 个样本的 n 个指标

指标	z_1	z_2	z_3	…	z_n
x_1	x_{11}	x_{12}	x_{13}	…	x_{1n}
x_2	x_{21}	x_{22}	x_{23}	…	x_{2n}
…	…	…	…	…	…
x_k	x_{k1}	x_{k2}	x_{k3}	…	x_{kn}

k 个样本第 j 个指标的平均值及标准差分别为

$x_j=\frac{1}{k}\sum_{i=1}^{k}x_{ij}$，$s_j=\sqrt{\sum_{i=1}^{k}(x_{ij}-x_j)^2}$

原始数据标准化为

$x_{ij}'=\frac{x_{ij}-x_j}{s_j}$

运用极值标准化公式，将标准化数据压缩在[0,1]内，即

$$x_{ij}=\frac{x_{ij}'-x_{jmin}'}{x_{jmax}'-x_{jmin}'} \quad (1)$$

其中，x_{jmin}'和x_{jmax}'分别表示$x_{1j}',x_{2j}',\ldots,x_{kj}'$中的最大值和最小值；$x_{ij}$为标准化后的指标。

(2)对于隐层((b)层~(d)层)数据，利用统计的方法[4]进行标准化处理，得到模糊隶属度函数，完成模糊化处理，使其特征值映射到[0,1]区间上。隶属度函数形式如下：

$$S=\frac{1}{1+e^{-wg_1(u-wc_1)}},\quad M=\frac{1}{1+e^{-wg_2(u-wc_1)}}-\frac{1}{1+e^{-wg_3(u-wc_2)}},\quad B=\frac{1}{1+e^{-wg_4(u-wc_2)}} \quad (2)$$

其中，S、M、B分别代表属于输入属性的隶属度值；参数wg为控制 3 个隶属度函数交点处斜率参数；参数wc为隶属度函数交点，通过数列的均值μ和方差δ计算：$wc_0=e^{\mu}$，$wc_1=e^{\mu-\delta}$，$wc_2=e^{\mu+\delta}$。其中，$\mu=\frac{1}{k}\sum_{i=1}^{k}\log u_i$；$\delta=\sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\log u_i-\mu)^2}$；k为输入属性x的样本总数；$u_i$为输入属性x的第i个样本，i=1, 2, 3,…, k。

在网络的计算中从模糊输入层到模糊输出层之间基本上是一个前馈反传网络。模糊输入层与隐层节点之间为全连接，隐层节点与模糊输出层节点之间也是全连接。各层内部节点之间没有连接。在计算隐层各节点的输出时，为有利于隐层各节点激活值的聚类需求，激活函数应扩大值域。在这里采用值域在[-1, 1]的双曲正切函数δ，$\delta(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$。

(3)利用梯度优化法进行网络训练

定义误差函数：

$$E=\frac{1}{2}\sum_{i=1}^{p}\sum_{\substack{j=1\\j\neq i}}^{p}y_{ij}^{\ 2}+\frac{1}{2}\sum_{i=1}^{p}(1-y_i)^2=\frac{1}{2}\sum_{i=1}^{p}[\sum_{\substack{j=1\\j\neq i}}^{p}y_{ij}^{\ 2}+(1-y_i)^2]$$

经梯度优化，得到权值修正公式为

$$\omega_{ij}^{new}=\omega_{ij}^{old}+a_t\left|\sum_{j\neq i}^{p}y_{ij}+1-y_i\right| \quad (3)$$

其中，y_i为输出层第i神经元输出；y_{ij}为在第i神经元抑制之下其余神经元的输出；a_t为学习系数，$a_t=a_0[1-t/t_{max}]$(a_0为初始学习系数，t为学习次数，t_{max}为最大学习次数)。

综上所述，利用模糊神经网络对输入层、隐含层数据进行训练、学习后，最终得到网络输出，整个过程如下：

(1)给定原始数据集$X=\{x_1,x_2,\ldots,x_n\}$，利用式(1)对输入数据((a)层)进行标准化处理；

(2)设定误差常数ε，初始化学习系数a_0；

(3)对于隐层((b)层~(d)层)数据，利用式(2)完成模糊化处理，并计算系统总误差 ERR；

(4)判定 ERR$\leq\varepsilon$，如果是，转步骤(7)；否则，转步骤(5)；

(5)利用式(3)进行权值调节；

(6)转步骤(3)；

(7)学习完毕，输出计算结果((e)层)。

3 基于模糊神经网络进行数据挖掘的步骤

模糊神经网络可以对一些实体的不确定性用模糊数学的方法来表示，对数据库中大量数据样本通过网络学习使数据所含知识压缩在结点之间的权值中。利用模糊神经网络方法进行数据挖掘大致可经历以下 5 个步骤：

(1)数据选择。从数据库中提取所需数据及其相关属性。

(2)数据预处理。对在数据选择阶段产生的数据，根据需要进行再加工，保证数据的完整性和一致性，对缺失、失真等噪声数据应用数据平滑技术进行处理。针对数据特点，可选取分箱、聚类、回归、计算机与人工检查结合等多种方法进行处理。最后将经过处理的数据存入一个数据库中。

(3)构造训练数据。将经过预处理的数据库中数据根据实体的特性构造其特性集合$\{A_1,A_2,\ldots,A_m\}$，用$val(A_k)$表示k实体属性值的值域；设C为将实体分归n个类$C_1,C_2,\ldots,C_n$的集合，构造出m+1 元组训练数据，其表达形式为（$a_1,a_2,\ldots,a_m,C_i$），其中，$a_j\in val(A_j)$，($1\leq j\leq n$)；$C_i\in C$，($1\leq i\leq n$)。把已知的数据分为一个训练数据组T和另一个检验数据组。已知数据中的N个数据构成训练数据组T，对于T中的每一个(m+1)元组，

(下转第 67 页)