有混合数据输入的自适应模糊神经推理系统

张宇献 [1] 郭佳强 [2] 钱小毅 [1] 王建辉 [3]

摘 要 现有数据建模方法大多依赖于定量的数值信息, 而对于数值与分类混合输入的数据建模问题往往根据分类变量组合建立多个子模型, 当有多个分类变量输入时易出现子模型数据分布不均匀、训练耗时长等问题. 针对上述问题, 提出一种具有混合数据输入的自适应模糊神经推理系统模型, 在自适应模糊推理系统的基础上, 引入激励强度转移矩阵和结论影响矩阵, 采用基于高氏距离的减法聚类辨识模型结构, 通过混合学习算法训练模型参数, 使数值与分类混合数据对模糊规则的前后件参数同时产生作用, 共同影响模型输出. 仿真实验分析了分类数据对模型规则后件的作用以及结构辨识算法对模糊规则数的影响, 与其他几种混合数据建模方法对比表明本文所提出的模型具有较高的预测精度和计算效率.

关键词 自适应模糊推理系统, 结构辨识, 激励强度转移矩阵, 后件影响矩阵, 混合属性数据

引用格式 张宇献, 郭佳强, 钱小毅, 王建辉. 有混合数据输入的自适应模糊神经推理系统. 自动化学报, 2019, **45**(9): 1743−1755

DOI 10.16383/j.aas.2018.c170698

开放科学 (资源服务) 标识码 (OSID):

An Adaptive Network-based Fuzzy Inference System with Mixed Data Inputs

ZHANG Yu-Xian[1] GUO Jia-Qiang[2] QIAN Xiao-Yi[1] WANG Jian-Hui[3]

Abstract The available data modeling methods mostly depend on quantitative numerical information. But the data modeling with both numerical and categorical data input often has to build multiple sub-models on the basic of combination of categorical variables. It is likely to present unevenly data distribution of sub-models, time-consuming training process and other problems when the multiple categorical variables are input. For the above problems, an adaptive network-based fuzzy inference system with mixed data inputs is proposed. Based on the structure of the adaptive network-based fuzzy inference system, a firing-strength transform matrix and a consequent influence matrix are introduced. The subtractive clustering based on the Gaussian distance is adapted to identify structure of model, and a hybrid learning algorithm is used to train parameters of model. The numerical and categorical data play an important role on the antecedent and consequent parameters of fuzzy rules, and jointly affect the output of model. The simulation experiment analyzes the effect on categorical data to the consequent rules and structure identification to number of fuzzy rules. Comparing with others data modeling with mixed data inputs, the proposed model in this paper has higher prediction accuracy and computational efficiency.

Key words Adaptive network-based fuzzy inference system, structure identification, firing-strength transform matrix, consequent influence matrix, mixed attribute data

Citation Zhang Yu-Xian, Guo Jia-Qiang, Qian Xiao-Yi, Wang Jian-Hui. An adaptive network-based fuzzy inference system with mixed data inputs. *Acta Automatica Sinica*, 2019, **45**(9): 1743−1755

随着全球数据量出现爆炸式增长, 数据成了当今社会增长最快的资源之一. 如何对大量复杂数据进行分析和挖掘, 从中提取有价值的知识用于决策, 已经成为学术界和产业界广泛关注的问题[1−2]. 在各行业中数据分析与数据建模仍有许多核心技术问题有待解决[3−4]. 如商业金融领域, 电子商务企业通过对所销售产品的类别以及客户的浏览行为进行数据分析, 进而精准把握客户的购买意图, 其中数据信息中既包含数量、单价这样的数值信息, 又包含商品种类、属性这样的非数值信息. 银行金融企业对储户分类进行分析, 根据客户特点对其设计不同的金融管理方案, 其中数据信息中既包含账户余额、资金流

收稿日期 2017-12-11 录用日期 2018-02-26
Manuscript received December 11, 2017; accepted February 26, 2018
国家自然科学基金 (61102124), 辽宁省自然科学基金 (2015020064), 辽宁省教育厅项目 (LQGD2017035) 资助
Supported by National Natural Science Foundation of China (61102124), Natural Science Foundation of Liaoning Province (2015020064) and Educational Commission of Liaoning Province (LQGD2017035)
本文责任编委 刘艳军
Recommended by Associate Editor LIU Yan-Jun
1. 沈阳工业大学电气工程学院 沈阳 110870 2. 沈阳工业大学信息科学与工程学院 沈阳 110870 3. 东北大学信息科学与工程学院 沈阳 110819
1. School of Electrical Engineering, Shenyang University of Technology, Shenyang 110870 2. School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870 3. College of Information Science and Engineering, Northeastern University, Shenyang 110819

量这样的数值信息, 也包含储户年龄区间、职业、性别等非数值信息. 又如工业生产领域, 钢铁行业炼钢过程通过生产工艺参数建立生产过程数据模型对产量、产品质量、能耗等指标进行估计, 其中工艺参数中既包含氧气压力、流量等数值数据, 也包含造渣原料种类 (如石灰、白云石和萤石等) 带有分类性质的非数值数据. 汽车行业电动汽车电池管理系统通过电池组运行数据对电池组状态和汽车续航里程进行估计, 其中既包含电流、电压、内阻、温度等数值数据, 也包含电池结构、电池类型等非数值数据. 再如医疗领域中, 医疗辅助诊断借助医院综合管理信息系统数据进行大数据挖掘给出患者的诊断和治疗方案建议, 其中医学数据中既包含化验结果、基因数据等数值数据, 同时也包含波形信号、图像、文字等非数值数据. 上述领域中普遍存在一个共同特点, 即数据信息中同时包含数值/非数值两类数据 (这里我们将非数值数据统称为分类数据).

然而, 现有数据建模方法大多依赖于定量的数值信息, 难以加入定性的分类信息. 对此国内外学者进行了大胆的尝试, 并取得了一些进展. Jacobs 等[5] 利用多个独立网络子模型构建组合模型, 采用有监督学习对模型参数进行训练. 其子模型中仅包含数值变量, 各子模型由分类变量组合成完整数据模型. 但当分类属性值较多时, 不同分类变量的组合排序将呈几何倍数增长. Lee 等[6] 构建了多个参数的组合模型, 每个子模型输入仅有数值数据. 该方法采用 1-out-of-n 编码, 把子模型中分类数据编码为一个数值向量, 然后把该向量导入神经网络. 然而当训练数据分布不均匀时无法精确描述模型. Brouwer[7−8] 提出基于多层感知机 (Multi layer perception, MLP) 结构的改进神经网络模型, 该模型由多感知机和多输出编码器单元组成. 模型输入为数值变量, 由分类变量决定最终模型输出, 即每一个输出单元对应一个分类输入变量组合. 该方法适用于分类变量较少的数据建模问题, 当分类变量数量较大时该方法训练结构参数的时间较长. Reyes-del-Castillo 等[9] 提出一种模糊极小极大神经网络, 由模糊超立方体聚集形成的集合体定义模糊集, 模糊超立方体的极大点作为模糊操作算子, 并利用改进模糊极小极大神经网络模型结构实现数值/分类混合属性数据建模. 但由于神经网络的黑箱结构, 模型的输入输出映射关系难以解释. Hsu[10] 采用自组织神经网络 (Self-organizing map, SOM) 结构, 通过定义分类数据之间的距离把分类数据转化为数值数据. 张宇献等[11] 以自组织映射神经网络为框架, 采用基于样本概率的异构值差度量混合属性数据的相异性. 利用分类特征项在 Voronoi 集合中出现频率作为分类属性数据参考向量更新规则的基础, 通过混合更新规则实现数值属性和分类属性数据规则的更新.

尽管上述研究工作在数值/分类混合的数据建模中做出了积极贡献, 但对于数值/分类混合的数据建模研究中仍有一些难点问题尚未得到很好的解决, 具体体现在以下几方面: 1) 多个分类变量采用排列组合方式参与数值数据计算时, 不同分类变量的组合排序将呈几何倍数增长; 2) 按分类变量建立多个子模型, 各子模型训练数据分布不均匀; 3) 将分类变量转化为二进制数或定义成数值变量, 参与计算时易出现大数吃小数现象; 4) 分类数据转化为数值数据的过程, 忽略了各变量值之间内在的分类或约束关系.

针对上述问题, Liu 等[12] 提出带分类输入的自适应模糊推理系统 (Adaptive network-based fuzzy inference system with categorical inputs, C-ANFIS) 结构, 将激励强度转移矩阵 (Firing-strength transform matrix, FTM) 引入自适应模糊推理系统 (Adaptive network-based fuzzy inference system, ANFIS) 中, 把分类数据对规则的影响作用到规则前件的激励强度上. 该方法一定程度上取得了不错的效果, 但它却存在自身不足: C-ANFIS 只考虑分类数据对规则前件的影响, 而对规则后件并未做任何处理. 基于以上分析, 本文提出了一种具有混合数据输入的自适应模糊推理系统 (Adaptive network-based fuzzy inference system with mixed data inputs, MDI-ANFIS) 模型. 该模型在标准 ANFIS 结构基础上, 引入激励强度转移矩阵和后件影响矩阵 (Consequent influence matrix, CIM), 通过后件影响矩阵把分类数据对模糊规则后件的影响作用到 ANFIS 上, 使分类数据对整个模糊规则产生影响, 并提出适应 MDI-ANFIS 结构的参数学习算法. 同时, 针对 MDI-ANFIS 结构辨识问题, 给出了基于高氏距离的减法聚类算法, 通过在减法聚类中引入混合型数据的高氏距离来确定 MDI-ANFIS 的模糊规则数和规则前后件的初始参数.

1 MDI-ANFIS 模型

1.1 MDI-ANFIS 的网络结构

学者 Jang 于 1993 年提出了 ANFIS[13], 它融合了神经网络的学习机制和模糊系统的语言推理能力等优点, 弥补各自不足, 属于神经模糊系统的一种. ANFIS 能够以任意精度逼近非线性函数, 具有便捷高效的特点, 并已在多个领域取得了成功应用[14−18].

然而, 标准的 ANFIS 结构只针对数值数据输

入, 当输入有分类数据时利用标准 ANFIS 建模将变得不再适合. 例如, 针对混合数据的自适应神经模糊推理建模问题, 假设 ANFIS 中的第 l 条规则有 2 个数值输入和 1 个分类输入, 其规则描述如下:

$$R_l\colon \text{If } x_1 \text{ is } A_1^l \text{ and } x_2 \text{ is } A_2^l \text{ and } x^C \text{ is } A_3^l$$
$$\text{Then } y_l = c_0^l + c_1^l \cdot x_1 + c_2^l \cdot x_2 + c_3^l \cdot x^C$$

其中, x_1 和 x_2 是数值数据输入, x^C 是分类数据输入, A_1^l、A_2^l 和 A_3^l 是第 l 条规则对应的模糊子集, y_l 是第 l 条规则的后件输出.

因为 x^C 是分类数据, 在规则中的 $\{x^C \text{ is } A_3^l\}$ 和 $\{c_3^l \cdot x^C\}$ 不能直接计算.

针对这个问题, 本文提出一种具有混合数据输入的自适应模糊推理系统 (MDI-ANFIS) 模型, 它在 C-ANFIS 的基础上, 引入后件影响矩阵, 使分类数据对规则前件和后件同时产生影响, 使得其对混合数据输入作用更加完善.

图 1 是一个多输入单输出的 MDI-ANFIS 结构图, 其对应的第 l 条模糊规则为:

$$R_l\colon \text{If } x_1 \text{ is } A_{1j}^l \text{ and } x_2 \text{ is } A_{2j}^l \text{ and } \cdots x_n \text{ is } A_{nj}^l$$
$$\text{and } x^C \text{ is } s$$
$$\text{Then } y_l = c_0^l + c_1^l x_1 + \cdots + c_n^l \cdot x_n + p_l$$

其中, $x^N = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ 为数值数据输入, $x^C = (x_1^C, x_2^C, \cdots, x_m^C)^{\mathrm{T}}$ 为分类数据输入, n 为数值输入变量个数, m 为分类输入变量个数. A_{ij}^l 为第 i 个数值输入对应的第 j 个模糊子集 (为了表述方便在图 1 中取 $j = 1, 2$), s 为分类数据的编码向量, $s \in \{s_1, s_2, \cdots, s_G\}$, $\{c_0^l, c_1^l, \cdots, c_n^l\}$ 为第 l 条规则的后件参数, p_l 为分类数据对第 l 条规则的后件影响, y_l 为第 l 条规则的后件输出, $l = 1, 2, \cdots, L$, L 为规则数.

MDI-ANFIS 网络结构分为 6 层: 输入层、规则层、正规化层、混合激励层、结论层和输出层, 具体各层的输出为:

第 1 层: 输入层, 该层的节点执行模糊化操作, 把数值输入转化为模糊子集的隶属度值. 各节点的输出可表示为

$$O_{ij}^1 = \mu_{A_{ij}}(x_i), \quad i = 1, 2, \cdots, n, \quad j = 1, 2 \tag{1}$$

其中, $\mu_{A_{ij}}(x_i)$ 为输入变量 x_i 的第 j 个模糊子集的隶属度函数, 当取高斯隶属度函数时, 其表达式为:

$$\mu_{A_{ij}}(x_i) = \exp\left[-\frac{(x_i - c_{ij})^2}{2\sigma_{ij}^2}\right], \quad i = 1, 2, \cdots, n, \quad j = 1, 2 \tag{2}$$

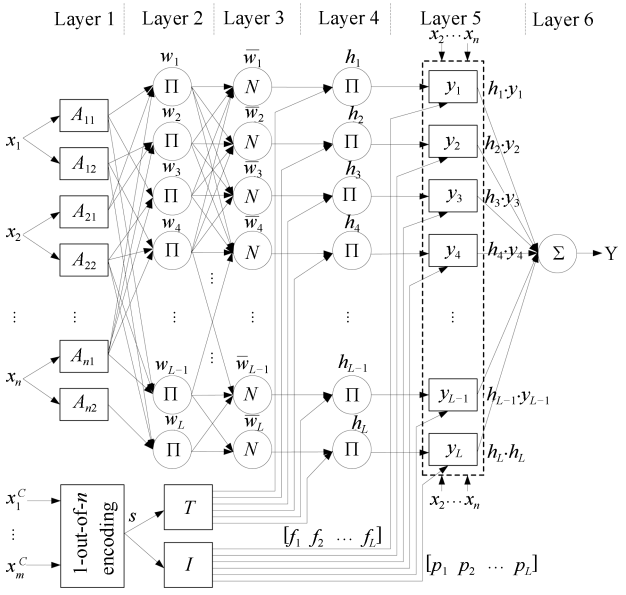


图 1　MDI-ANFIS 结构

Fig. 1　Structure of MDI-ANFIS

其中, $\{c_{ij}, \sigma_{ij}\}$ $(i = 1, 2, \cdots, n,\ j = 1, 2)$ 为模糊规则的前件参数集.

第 2 层: 规则层, 该层节点执行规则前件数值变量的模糊与运算, 计算出第 l 条规则数值变量的激励强度:

$$O_l^2 = w_l = \prod_{i=1}^{n} \mu_{A_{ij}^l}(x_i), \quad l = 1, 2, \cdots, L \tag{3}$$

其中, $\mu_{A_{ij}^l}(x_i)$ 是第 l 个规则层节点上的输入, 它由 $\mu_{A_{ij}}(x_i)$ 经结构辨识得到.

第 3 层: 正规化层, 正规化规则层的激励强度. 该层节点的输出为:

$$O_l^3 = \overline{w}_l = \frac{w_l}{\sum_{l=1}^{L} w_l}, \ l = 1, 2, \cdots, L \tag{4}$$

第 4 层: 混合激励层, 该层节点计算分类数据和数值数据对每条规则的激励强度. 各节点在该层的输出为:

$$O_l^4 = h_l = \overline{w}_l \cdot f_l, \quad l = 1, 2, \cdots, L \tag{5}$$

其中, f_l 为分类数据的编码向量经激励强度转移矩阵 T 得到的第 l 条规则上的分类激励值, $f_l = s \cdot T^l$, s 为分类数据的编码向量, T^l 为激励强度转移矩阵 T 的第 l 列.

第 5 层: 结论层, 该层计算每条规则的输出. 各

节点函数是一个线性函数, 各节点输出为:

$$O_l^5 = h_l \cdot y_l = h_l \cdot \left(c_0^l + c_1^l \cdot x_1 + \cdots + c_n^l \cdot x_n + p_l\right), \quad l = 1, 2, \cdots, L \tag{6}$$

其中, p_l 为分类数据通过后件影响矩阵 I 得到的分类数据对每条规则的后件影响值, $p_l = s \cdot I^l$, I^l 为后件影响矩阵 I 的第 l 列.

第 6 层: 输出层, 计算整个 MDI-ANFIS 的输出:

$$O^6 = Y = \sum_{l=1}^{L} O_l^5, \ l = 1, 2, \cdots, L \tag{7}$$

1.2 MDI-ANFIS 的参数学习

假设有 K 个训练样本点 x^k 和 s^k ($k = 1, 2, \cdots, K$) 分别为第 k 个样本点的数值输入向量和分类编码向量, 其中 $x^k = (x_{1k}, x_{2k}, \cdots, x_{ik}, \cdots, x_{nk})$, x_{ik} 为第 i 个数值变量在第 k 个样本上的取值. Y_k 为第 k 个样本点的训练输出值, $\overline{Y_k}$ 为第 k 个样本点的期望输出值. 对于单个样本点的 MDI-ANFIS 的输出误差为

$$e_k = \frac{1}{2}\left(Y_k - \overline{Y_k}\right)^2, \ k = 1, 2, \cdots, K \tag{8}$$

整个训练样本集的输出误差为:

$$E = \sum_{k=1}^{K} e_k = \frac{1}{2}\sum_{k=1}^{K}\left(Y_k - \overline{Y_k}\right)^2 \tag{9}$$

参数学习的目的是通过训练 MDI-ANFIS 中的参数使总误差 E 达到最小. 假定系统输入为第 k 个样本点, 其输出为:

$$\begin{aligned} Y_k = & \sum_{l=1}^{L} \overline{w}_l(x^k) \cdot f_l(s^k) \cdot y_l(x^k, s^k) = \\ & \sum_{l=1}^{L} \overline{w}_l(x^k) \cdot f_l(s^k) \cdot (c_0^l + c_1^l \cdot x_{1k} + \cdots + \\ & c_n^l \cdot x_{nk} + p_l) = \\ & \sum_{l=1}^{L} \overline{w}_l(x^k) \cdot f_l(s^k) \cdot (\bar{x}^k \cdot C^l + s^k \cdot I^l) = \\ & \sum_{l=1}^{L} \overline{w}_l(x^k) \cdot f_l(s^k) \cdot z^k \cdot Q_l = \\ & z^k \cdot Q \cdot H(x^k, s^k), \quad k = 1, 2, \cdots, K \end{aligned} \tag{10}$$

其中, $\bar{x}^k = [\ 1 \quad x^k\ ]$, $C^l = [c_0^l, c_1^l, \cdots, c_n^l]^{\mathrm{T}}$, $z^k = [\ \bar{x}^k \quad s^k\ ]$, $Q_l = [\ C^l \quad I^l\ ]^{\mathrm{T}}$, $Q = [Q_1, Q_2, \cdots, Q_l, \cdots, Q_L]$, $H(x^k, s^k) = [\overline{w}_1(x^k) \cdot f_1(s^k), \cdots, \overline{w}_L(x^k) \cdot f_L(s^k)]^{\mathrm{T}}$, C^l 为第 l 条规则后件的参数向量, I^l 为后件影响矩阵 I 的第 l 列参数向量.

这里把所有由 C^l 元素组成的参数集合称为后件参数集 P^c, 所有由 I^l 元素组成的参数集合称为后件影响矩阵参数集 P^i.

将式 (10) 两边同时乘以 $[(z^k)^{\mathrm{T}} \cdot z^k]^{-1} \cdot (z^k)^{\mathrm{T}}$, 可以得到:

$$[(z^k)^{\mathrm{T}} \cdot z^k]^{-1} \cdot (z^k)^{\mathrm{T}} \cdot Y_k = Q \cdot H(x^k, s^k) \tag{11}$$

式 (11) 表明 $[(z^k)^{\mathrm{T}} \cdot z^k]^{-1} \cdot (z^k)^{\mathrm{T}} \cdot Y_k$ 通过 Q 与 $H(x^k, s^k)$ 成线性关系, 而 Q 包含后件参数和后件影响矩阵参数, 因此 P^c 和 P^i 可以通过最小二乘估计 (Least squares estimation, LSE) 得到.

同理, 在第 k 个样本点下的输出:

$$\begin{aligned} Y_k = & \sum_{l=1}^{L} \overline{w}_l\left(x^k\right) \cdot f_l\left(s^k\right) \cdot y_l\left(x^k, s^k\right) = \\ & \sum_{l=1}^{L} \left(s^k \cdot T^l\right) \cdot \overline{w}_l\left(x^k\right) \cdot y_l\left(x^k, s^k\right) = \\ & s^k \cdot T \cdot M\left(x^k, s^k\right), \quad k = 1, 2, \cdots, K \end{aligned} \tag{12}$$

其中, $M(x^k, s^k) = [\overline{w}_1(x^k) \cdot y_1(x^k, s^k), \cdots, \overline{w}_l(x^k) \cdot y_L(x^k, s^k)]^{\mathrm{T}}$, T^l 为激励强度转移矩阵 T 的第 l 列参数向量.

这里称由所有 T^l 元素组成的参数集合为激励强度转移矩阵参数集 P^t.

将式 (12) 两边同时乘以 $[(s^k)^{\mathrm{T}} \cdot s^k]^{-1} \cdot (s^k)^{\mathrm{T}}$, 可以得到:

$$[(s^k)^{\mathrm{T}} \cdot s^k]^{-1} \cdot (s^k)^{\mathrm{T}} \cdot Y_k = T \cdot M\left(x^k, s^k\right) \tag{13}$$

式 (13) 表明 $[(s^k)^{\mathrm{T}} \cdot s^k]^{-1} \cdot (s^k)^{\mathrm{T}} \cdot Y_k$ 通过 T 与 $M(x^k, s^k)$ 成线性关系, 因此激励强度转移矩阵参数集 P^t 也可由 LSE 得到.

以上有关 Y^k 的所有推导中 $\overline{w}_l(x^k)$ 为 MDI-ANFIS 正规化层的第 l 个输出, 这里把所有 $\overline{w}_l(x^k)$ 中的参数组成的集合称为前件参数集 P^p, 它可以通过预先固定参数集 P^t、P^c 和 P^i, 然后由反向传播算法 (Back propagation, BP) 求得. MDI-ANFIS 的参数学习步骤如表 1 所示.

表 1 MDI-ANFIS 混合学习算法

Table 1 Hybrid learning algorithm of MDI-ANFIS

参数集	算法
P^c, P^i	LSE
P^t	LSE
P^p	BP

1.3 MDI-ANFIS 的结构辨识

当输入 ANFIS 的维度不断增大时, 采用传统的网格划分会使规则数目呈指数增大, 这将不可避免导致维度灾难. 本文提出了基于高氏距离的减法聚类算法 (Gower distance-based subtractive cluster, GDSC) 对其进行结构辨识. 减法聚类 (Subtractive cluster, SC) 是一种无需预先确定聚类数和快速单次的聚类算法, 克服了其他聚类算法的计算量随着输入维数的增加而呈指数增长的不足. 然而, 减法聚类只适用于对数值数据进行聚类, 而对于混合数据就显得无能为力了, 同时仅对混合数据的数值部分聚类产生的模糊推理结构也不完备, 因为它没有考虑混合数据的分类部分对结构辨识产生的影响. 基于以上减法聚类的优缺点, GDSC 算法把高氏距离引入到减法聚类公式中. 这样 GDSC 算法既充分利用了 SC 的优势, 又充分考虑到分类数据对结构辨识的影响.

结合文献 [19] 和 [20], 本文定义样本点 X_k 和 X_r 的高氏距离为:

$$d_{kr}^{G}=\sqrt{\sum_{i=1}^{n}\left(X_{ik}-X_{ir}\right)^{2}}+\sum_{i=n+1}^{n+m}\delta\left(X_{ik},X_{ir}\right),$$
$$k=1,2,\cdots,K,\qquad r=1,2,\cdots,K \quad (14)$$

其中, $X_k=(x_{1k},\cdots,x_{nk},x_{1k}^{C},\cdots,x_{mk}^{C})$, X_{ik} 为 X_k 的第 i 个元素, $X_r=(x_{1r},\cdots,x_{nr},x_{1r}^{C},\cdots,x_{mr}^{C})$, X_{ir} 为 X_r 的第 i 个元素.

$$\delta\left(X_{ik},X_{rk}\right)=\begin{cases}0, & X_{ik}=X_{rk}\\ 1, & X_{ik}\neq X_{rk}\end{cases},\quad i\geq n+1$$

将式 (14) 代入到减法聚类公式中得到样本点的高氏密度

$$DD_{k}^{G}=\sum_{r=1}^{K}\exp\left[-\frac{\left(d_{kr}^{G}\right)^{2}}{\left(\frac{\beta_a}{2}\right)^{2}}\right],$$
$$k=1,2,\cdots,K \quad (15)$$

为排除已被选为聚类中心的附近数据作为下一个聚类中心的可能性, 将式 (15) 的减法聚类密度修正为:

$$D_{k}^{G}\left(v+1\right)=D_{k}^{G}\left(v\right)-D_{c_v}^{G}\exp\left[-\frac{\left(d_{kc_v}^{G}\right)^{2}}{\left(\frac{\beta_b}{2}\right)^{2}}\right],$$
$$k=1,2,\cdots,K \quad (16)$$

式 (15) 和 (16) 中, DD_k^G 表示混合属性样本点 X_k 的高氏密度值, $D_k^G\,(v{+}1)$ 表示第 v 次迭代修正后的样本点 X_k 的高氏密度值, $D_i^G\,(v=1)=DD_i^G$, $v=1,2,\cdots,V$, 其中 V 为最大迭代次数. $D_{c_v}^G$ 表示第 v 个聚类中心 c_v 的高氏密度值, β_a 定义了样本点的邻域半径, β_b 是一个密度显著减小的邻域半径, 一般取 $\beta_a=1.5\beta_b$. GDSC 算法根据迭代式 (16) 寻找具有高氏密度值最大的样本点并把它作为下一个聚类中心 c_{v+1}. 然后再次把 c_{v+1} 的高氏密度值带入到迭代公式中计算所有样本点的高氏密度修正值, 当聚类满足 $\left\|D_{c_{v+1}}^{G}\right\|/\left\|D_{c_1}^{G}\right\|<\varepsilon$ 或 $v>V$ 时聚类终止, 输出聚类中心矩阵 X_c. GDSC 算法的伪代码如下:

算法 1. GDSC 算法

Input: $X=\{X_1,X_2,\cdots,X_K\}$, neighborhood radius γ_a, threshold ε, maximum number of iterations V.

Output: cluster center matrix X_c

1) **Initialize:** $X_c\leftarrow[\ ]$ and $v\leftarrow 1$.
2) **for** $k\leftarrow 1$ to K
3) calculate Gower density DD_k^G of X_k according to equation (15);
4) **end for**
5) find the first cluster center X_{c_1} which has maximum Gower density in DD^G, then $X_c\leftarrow X_{c_1}$;
6) **while** $\|D_{c_{v+1}}^{G}\|/\|D_{c_1}^{G}\|<\varepsilon$ or $v>V$
7) update the Gower density $D_k^G\,(v)$ of X_k in X according to equation (16);
8) find next cluster center $X_{c_{v+1}}$ which has max-imum Gower density in $D^{G'}\,(v+1)$, then $X_c\leftarrow X_{c_{v+1}}$;
9) $v\leftarrow v+1$;
10) **end while**
11) Output cluster center X_c.

1.4 MDI-ANFIS 收敛性分析

本文针对基于 MDI-ANFIS 网络结构的 T-S 模糊系统, 给出收敛性证明[21−22]. 基于 MDI-ANFIS 的 T-S 模糊系统规则为:

R_l: If x_1 is A_{1j}^{l} and x_2 is A_{2j}^{l} and $\cdots x_n$ is A_{nj}^{l} and x^C is s
Then $y=c_0^l+c_1^lx_1+c_2^lx_2+\cdots+c_n^lx_n+p_l$

其中, $x^N=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为数值输入, $x^C=(x_1^C,x_2^C,\cdots,x_m^C)^{\mathrm{T}}$ 为分类输入, s 为分类数据的编码向量, $s\in\{s_1,s_2,\cdots,s_G\}$, $l=1,2,\cdots,L$, $j=1,2,\cdots,n_i$, 此处 n_i 为第 i 个变量的模糊子集个数, L 为规则数, n 为数值输入变量个数, m 为分类输入变量个数.

其模糊子系统的规则为:

$$R_{l_g}\text{: If } x_1 \text{ is } A_{1j}^{l_g} \text{ and } x_2 \text{ is } A_{2j}^{l_g} \text{ and } \cdots x_n \text{ is } A_{nj}^{l_g} \text{ and } x^C \text{ is } s_g$$

$$\text{Then } y = c_0^{l_g} + c_1^{l_g} x_1 + c_2^{l_g} x_2 + \cdots + c_n^{l_g} x_n + p_{l_g}$$

其中, R_{l_g} 为分类编码向量 $s = s_g$ 确定的模糊规则, $l_g = 1, 2, \cdots, L_g$.

由子系统的数值数据激励强度 $w_{l_g}(x^N) = \prod_{i=1}^{n} \mu_{A_{ij}^{l_g}}(x_i)$, 容易求得整个 T-S 模糊系统的输出为:

$$f_{TS}\left(x^N, x^C\right) = \sum_{g=1}^{G} f_{TS}\left(x^N, s_g\right) = \sum_{g=1}^{G} \left\{ \frac{\sum\limits_{l_g=1}^{L_g} \left[w_{l_g}(x^N) \cdot f_{l_g} \cdot \left(\sum\limits_{i=0}^{n} c_i^{l_g} x_i + p_{l_g} \right) \right]}{\sum\limits_{l_g=1}^{L_g} w_{l_g}(x^N)} \right\} \tag{17}$$

这里, f_{l_g} 为子系统第 l_g 条规则分类数据的激励强度, p_{l_g} 为子系统第 l_g 条规则分类数据的后件影响值, $c_i^{l_g}$ 为子系统第 l_g 条规则的后件参数.

定义 1. 数值数据取值 C^n [0, 1], 分类数据取值 $\{s_1, s_2, \cdots, s_G\}$ 的 ϕ 次 n+1 元多项式函数可以写为:

$$P_\phi(x^N, x^C) = \sum_{g=1}^{G} \left[\sum_{z_1=0}^{r_1} \sum_{z_2=0}^{r_2} \cdots \sum_{z_n=0}^{r_n} \beta_{z_1 z_2 \cdots z_n} (x_1)^{z_1} (x_2)^{z_2} \cdots (x_n)^{z_n} + s_g \cdot t \right] \tag{18}$$

其中, $\sum_{i=1}^{n} r_i = \phi$, $x^N = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, $x^C = \{x_1^C, x_2^C, \cdots, x_m^C\}^{\mathrm{T}}$, t 为 T 的某个列向量.

定义 2. 称论域 U 上的一组模糊集 A_{ij} $(j = 1, 2, \cdots, n_i)$ 是一致的, 如果对某些 $x_{i0} \in U$ 存在 A_{ij} $(j = 1, 2, \cdots, n_i)$ 使得 $\mu_{A_{ij}}(x_{i0}) = 1$, 且对任意 $\upsilon = 1, 2, \cdots, n_i$, 以及 $\upsilon \neq j$, 都有 $\mu_{A_{i\upsilon}}(x_{i0}) = 0$.

假设 1. 所研究的 T-S 模糊系统的每一个数值输入变量的模糊子集都是一致的.

假设 2. 所研究的 T-S 模糊系统采用的隶属度函数都是连续且分段可微的.

假设 3. 所研究的 T-S 模糊系统的每一个分类输入 s_g, $s_g \cdot T^{l_g} = 1$, 其余 $l \neq l_g$, $s_g \cdot T^l = 0$.

基于上述假设, 证明基于 MDI-ANFIS 网络结构的 T-S 模糊系统具有通用逼近性.

定理 1. 基于 MDI-ANFIS 网络结构的 T-S 模糊系统能够以任意精度一致逼近数值数据取值 C^n [0, 1] 上, 分类数据取值 $\{s_1, s_2, \cdots, s_G\}$ 的 ϕ 次 $n+1$ 元多项式函数 $P_\phi(x^N, x^C)$, 即 $\forall \gamma > 0$, 存在 T-S 模糊系统使得:

$$\left\| f_{TS}\left(x^N, x^C\right) - P_\phi\left(x^N, x^C\right) \right\|_\infty < \gamma \tag{19}$$

这里 $\|\cdot\|_\infty$ 定义为: 对任意定义在紧致集 $U \subset \mathbf{R}^n$ 的函数 $a(x)$, $\|a(x)\|_\infty = \sup_{x \in U} |a(x)|$.

证明. 假设 T-S 模糊子系统的每一条规则 R_{l_g}, $l_g = 1, 2, \cdots, L_g$, 它决定了一个特殊数值输入矢量 $x_{l_g} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, 每个分量 x_i $(i = 1, 2, \cdots, n)$ 的取值恰好等于对应的模糊子集 $A_{ij}^{l_g}$ 的中心点, 即

$$x_i = C_{ij}^i, \quad i = 1, 2, \cdots, n, \ \ j = 1, 2, \cdots, n_i \tag{20}$$

显然, 对应一个 T-S 模糊子系统, 全部输入矢量共有 L_g 个, 并且与子系统模糊规则一一对应的关系, 记它们的集合为

$$S_g = \left\{ x_{l_g} = \left(C_{1j}^1, C_{2j}^2, \cdots, C_{nj}^n\right) \right\}, \quad l_g = 1, 2, \cdots, L_g, \ i = 1, 2, \cdots, n \tag{21}$$

对于 $\forall x_{l_g} \in S_g$ $(l_g = 1, 2, \cdots, L_g)$, 根据假设 1 有, $\mu_{l_g}\left(x_{l_g}\right) = 1$, 而对任意 $l_{gg} \in \{1, 2, \cdots, L_g\}$, $l_{gg} \neq l_g$, 都有 $\mu_{l_{gg}}\left(x_{l_g}\right) = 0$, μ_{l_g} 和 $\mu_{l_{gg}}$ 为对应子系统规则的隶属度向量函数; 根据假设 3 有, $s_g \cdot T^{l_g} = 1$, 而对任意 $l \neq l_g$, 都有 $s_g \cdot T^l = 0$, 由公式 $f_{TS}(x^N, x^C)$ 得:

$$f_{TS}\left(x_{l_g}, s_g\right) = \sum_{i=0}^{n} c_i^{l_g} x_i + p_{l_g} = P_\phi\left(x_{l_g}, s_g\right) \tag{22}$$

在 $f_{TS}(x^N, x^C)$ 中 $x^N = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, 并且令 $x_0 \equiv 1$. 设数值输入变量 x_i $(i = 1, 2, \cdots, n)$ 的第 j 个模糊子集的中心点为 $C_{ij}^i, j = 1, 2, \cdots, n_i$, 归一化 x_i $(i = 1, 2, \cdots, n)$ 有 $0 \leq C_{i1}^i < C_{i2}^i < \cdots < C_{in_i}^i \leq 1$. 不失一般性, 设每个中心点处的隶属度为 1, 对每一个数值输入变量 x_i $(i = 1, 2, \cdots, n)$ 定义模糊分割间距:

$$\Omega_j^i = C_{ij}^i - C_{i,j-1}^i, j = 1, 2, \cdots, n_i + 1 \tag{23}$$

其中, $C_{i0}^i \equiv 0, C_{i,n_i+1}^i \equiv 1$.

在此基础上可以对每个数值输入变量 x_i $(i = 1, 2, \cdots, n)$ 定义最大模糊分割间距:

$$\Omega_{\max}^i = \max_{j=1}^{n_i+1} \Omega_j^i \tag{24}$$

注意 x^N 的任意分量 x_i $(i = 1, 2, \cdots, n)$, 总可以找到下标 $j \in \{1, 2, \cdots, n_i+1\}$, 使得 $C_{i,j-1}^i \leq x_i \leq C_{ij}^i$, 从而 $\left|x_i - C_{ij}^i\right| \leq \Omega_{\max}^i, i = 1, 2, \cdots, n$.

根据假设 2 可知 $f_{TS}\left(x_{l_g}, s_g\right) = P_\phi\left(x_{l_g}, s_g\right)$ 连续且分段可微, 由 $f_{TS}\left(x_{l_g}, s_g\right) = P_\phi\left(x_{l_g}, s_g\right)$, 利用多元泰勒公式有:

$$
\begin{aligned}
&\left|f_{TS}\left(x^N, x^C\right) - P_\phi\left(x^N, x^C\right)\right| = \\
&\quad\left|\sum_{g=1}^{G} f_{TS}\left(x^N, s_g\right) - \sum_{g=1}^{G} P_\phi\left(x^N, s_g\right)\right| \le \\
&\quad\sum_{g=1}^{G}\left|f_{TS}\left(x^N, s_g\right) - P_\phi\left(x^N, s_g\right)\right| = \\
&\quad\sum_{g=1}^{G}\left|f_{TS}\left(x^N, s_g\right) - f_{TS}\left(x_{l_g}, s_g\right) + \right. \\
&\quad\left. P_\phi\left(x_{l_g}, s_g\right) - P_\phi\left(x^N, s_g\right)\right| \le \\
&\quad\sum_{g=1}^{G}\left[\left|f_{TS}\left(x^N, s_g\right) - f_{TS}\left(x_{l_g}, s_g\right)\right| + \right. \\
&\quad\left.\left|P_\phi\left(x_{l_g}, s_g\right) - P_\phi\left(x^N, s_g\right)\right|\right] \le \\
&\quad\sum_{g=1}^{G}\left\{\sum_{i=1}^{n}\left[\left(\left\|\frac{\partial f_{TS}\left(x^N, s_g\right)}{\partial x_i}\right\|_\infty + \right.\right.\right. \\
&\quad\left.\left.\left.\left\|\frac{\partial P_\phi\left(x^N, s_g\right)}{\partial x_i}\right\|_\infty\right) \cdot \left|x_i - C_{ij}^i\right|\right]\right\} \le \\
&\quad\sum_{g=1}^{G}\left\{\sum_{i=1}^{n}\left[\left(\left\|\frac{\partial f_{TS}\left(x^N, s_g\right)}{\partial x_i}\right\|_\infty + \right.\right.\right. \\
&\quad\left.\left.\left.\left\|\frac{\partial P_\phi\left(x^N, s_g\right)}{\partial x_i}\right\|_\infty\right) \cdot \Omega_{\max}^i\right]\right\} \le \\
&\quad\sum_{g=1}^{G}\left\{\max_{i=1}^{n}\Omega_{\max}^i \cdot \sum_{i=1}^{n}\left(\left\|\frac{\partial f_{TS}\left(x^N, s_g\right)}{\partial x_i}\right\|_\infty + \right.\right. \\
&\quad\left.\left.\left\|\frac{\partial P_\phi\left(x^N, s_g\right)}{\partial x_i}\right\|_\infty\right)\right\} < \sum_{g=1}^{G}\gamma_g = \gamma
\end{aligned}
$$

上式 $\left\|\partial P_\phi\left(x^N, s_g\right)/\partial x_i\right\|_\infty$ 是有限数, 函数 $\left\|\partial f_{TS}\left(x^N, s_g\right)/\partial x_i\right\|_\infty$ 连续且分段可微, 因此 $\left\|\partial f_{TS}\left(x^N, s_g\right)/\partial x_i\right\|_\infty$ 也是有限数, 当 $\max\limits_{i=1}^{n}\Omega_{\max}^i$ 充分小时, 存在一个极小的正数 γ 使得基于 MDI-ANFIS 网络结构的 T-S 模糊系统无限逼近于一个多项式函数. □

引理 1. $\forall\zeta > 0$, 存在多项式 $P(x)$, 使得对一切 $x \in [a, b]$ 的 $f(x)$ 成立:

$$
\left|P\left(x\right) - f\left(x\right)\right| < \zeta \tag{25}
$$

定理 2. 基于 MDI-ANFIS 网络结构的 T-S 模糊系统能够以任意精度一致逼近数值输入在紧致集 $U \subset \mathbf{R}^n$ 上的任意实函数 $\Psi\left(x^N, x^C\right)$, 即 $\forall\delta > 0$, 存在基于 MDI-ANFIS 网络结构的 T-S 模糊系统使得:

$$
\left\|f_{TS}\left(x^N, x^C\right) - \Psi\left(x^N, x^C\right)\right\|_\infty < \delta \tag{26}
$$

证明. 根据引理, 在 $U \subset \mathbf{R}^n$ 上存在 ϕ 次多项式函数 $P_\phi(x^N, s_g)$, 一致逼近任意连续实函数 $\psi(x^N, s_g)$, 即 $\forall\zeta_g > 0$ 存在 $P_\phi(x^N, s_g)$ 使得 $\|P_\phi(x^N, s_g) - \psi(x^N, s_g)\|_\infty < \zeta_g$, 另一方面, $\forall\gamma_g > 0$, 根据定理 1, 存在 $\|f_{TS}(x^N, s_g) - P_\phi(x^N, s_g)\|_\infty < \gamma_g$.

$$
\begin{aligned}
&\left\|f_{TS}\left(x^N, x^C\right) - \Psi\left(x^N, x^C\right)\right\|_\infty = \\
&\quad\sum_{g=1}^{G}\left(\left\|f_{TS}\left(x^N, s_g\right) - \psi\left(x^N, s_g\right)\right\|_\infty\right) = \\
&\quad\sum_{g=1}^{G}(\|f_{TS}(x^N, s_g) - P_\phi(x^N, s_g) + \\
&\quad P_\phi(x^N, s_g) - \psi(x^N, s_g)\|_\infty) \le \\
&\quad\sum_{g=1}^{G}(\|f_{TS}(x^N, s_g) - P_\phi(x^N, s_g)\|_\infty + \\
&\quad\|P_\phi(x^N, s_g) - \psi(x^N, s_g)\|_\infty) < \\
&\quad\sum_{g=1}^{G}(\gamma_g + \zeta_g) = \sum_{g=1}^{G}\delta_g = \delta
\end{aligned}
$$

即 $\left\|f_{TS}\left(x^N, x^C\right) - \Psi\left(x^N, x^C\right)\right\|_\infty < \delta$. □

2　仿真实验及结果分析

为了验证所建模型的性能, 我们将从规则后件参数影响分析、结构辨识方法比较以及几种混合数据建模方法预测精度对比几方面来说明本文所提出的 MDI-ANFIS 的优越性.

实验操作系统为 Windows 8.1, 仿真软件为 MATLAB 2009b. 硬件条件: CPU 为 Intel Core I5 2.5 GHz, 内存为 4 GB.

2.1　后件参数影响分析

对于参数预测问题, 文献 [12] 提出的 C-ANFIS 算法把分类数据对规则的影响作用到规则前件上, 但并未考虑其对后件的影响. 本文在 C-ANFIS 结构上做了改进, 提出适用于混合数据参数预测的算法 MDI-ANFIS, 使混合数据中的分类数据对规则的前后件均产生影响.

这里采用 UCI 机器学习库中的 Abalone 数据集来训练 C-ANFIS 和 MDI-ANFIS 参数, 然后预测鲍鱼的年龄. Abalone 数据集包含 4 177 个样本点, 分别记录了鲍鱼的性别、长度、直径、高度、整体重量、脱皮重量、内脏重量、壳重量和年龄属性

值, 其中鲍鱼的性别是分类属性数据, 其他变量是数值属性数据.

表 2 给出了两种算法对比结果, 其中平均规则后件值反映了 C-ANFIS 与 MDI-ANFIS 对规则后件结论的影响大小, 预测误差选取均方根误差作为误差指标. 为了更加体现分类数据对规则后件的影响, 我们选取表 2 第 1 组实验产生的平均规则后件值数据制作对比柱状图见图 2, 横坐标表示本组实验一共产生 9 条规则, 纵坐标记录了每条规则的平均输出值. 图 2 非常直观地显示出考虑分类数据对规则后件的影响将极大地改变规则后件大小. 表 2 同时体现出 MDI-ANFIS 相较 C-ANFIS 能够有一个更好的预测精度. 图 3 是两种算法的训练误差对比, 从图上可以看出随着训练周期的增加, 两者的误差距离正在逐渐拉大. 图 4 是训练后的 C-ANFIS 模型和 MDI-ANFIS 模型对测试样本点做预测的结果. 对比结果显示相对于 C-ANFIS 模型, 本文所提出的 MDI-ANFIS 模型在后件参数的影响和预测精度上更具优势.

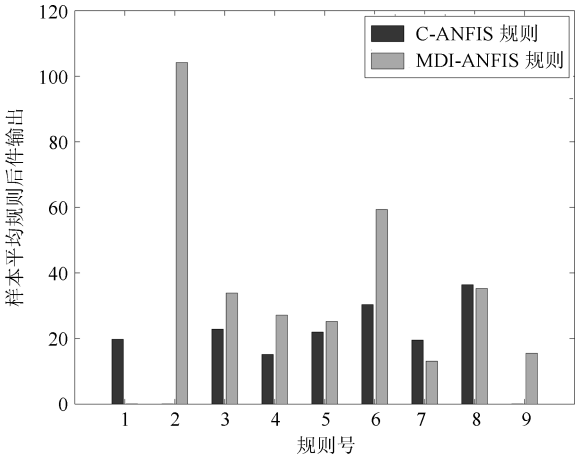


图 2 样本平均规则后件输出

Fig. 2 Average consequent output of samples

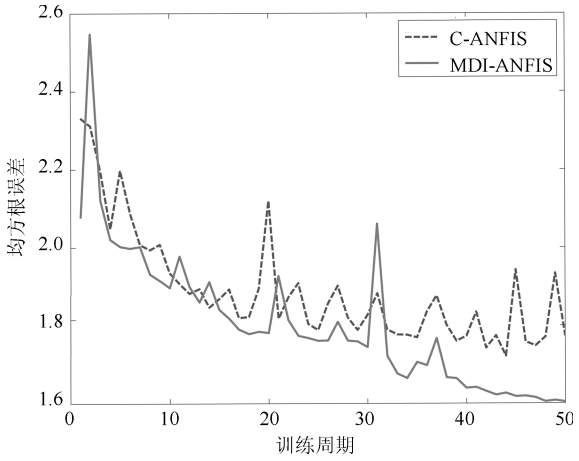


图 3 模型训练误差对比

Fig. 3 Comparison of model training error

2.2 结构辨识对比分析

MDI-ANFIS 的结构辨识问题对具有高维输入数据的网络性能具有重要影响, 本文提出的 GDSC 算法, 将高氏距离引入到减法聚类中, 实现数值数据和分类数据同时对初始规则产生影响, 从而完成混合属性数据的 ANFIS 结构辨识. 实验采用 UCI 中的 Boston Housing 数据集, 它包含 506 个样本点, 其中 11 个数值属性和 2 个分类属性, 这里把数值属性记为 NA1 ∼ NA11, 分类属性记为 CA1 和 CA2. 实验首先利用 SC 算法和 GDSC 算法对 Boston Housing 数据集聚类, 然后利用聚类结果产

表 2 两种算法的平均规则后件影响和误差

Table 2 Average consequent influences and errors of two algorithms

组号	样本点个数	平均规则后件值		预测误差	
		C-ANFIS	MDI-ANFIS	C-ANFIS	MDI-ANFIS
1	200	19.659	17.735	2.040	1.519
2	200	18.905	36.270	1.690	1.463
3	200	21.323	27.297	2.980	1.881
4	400	34.202	66.760	3.230	1.604
5	400	14.050	39.905	2.330	2.145
6	400	16.385	35.070	2.510	2.002
7	500	18.901	17.804	3.680	2.194
8	600	21.659	30.857	2.290	2.395
9	600	16.299	22.267	2.800	2.242
10	600	18.426	34.818	3.730	2.187
平均值	410	19.981	32.878	2.728	1.963

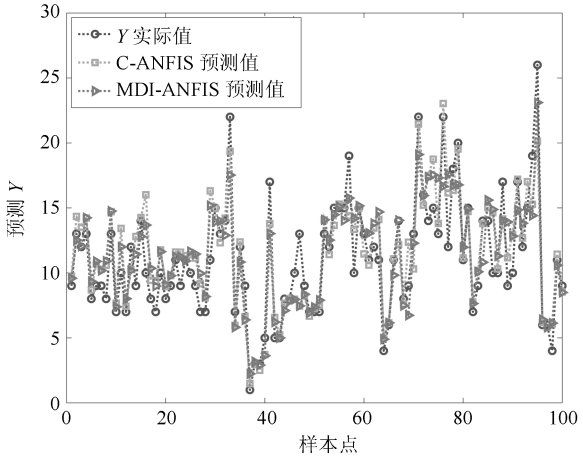
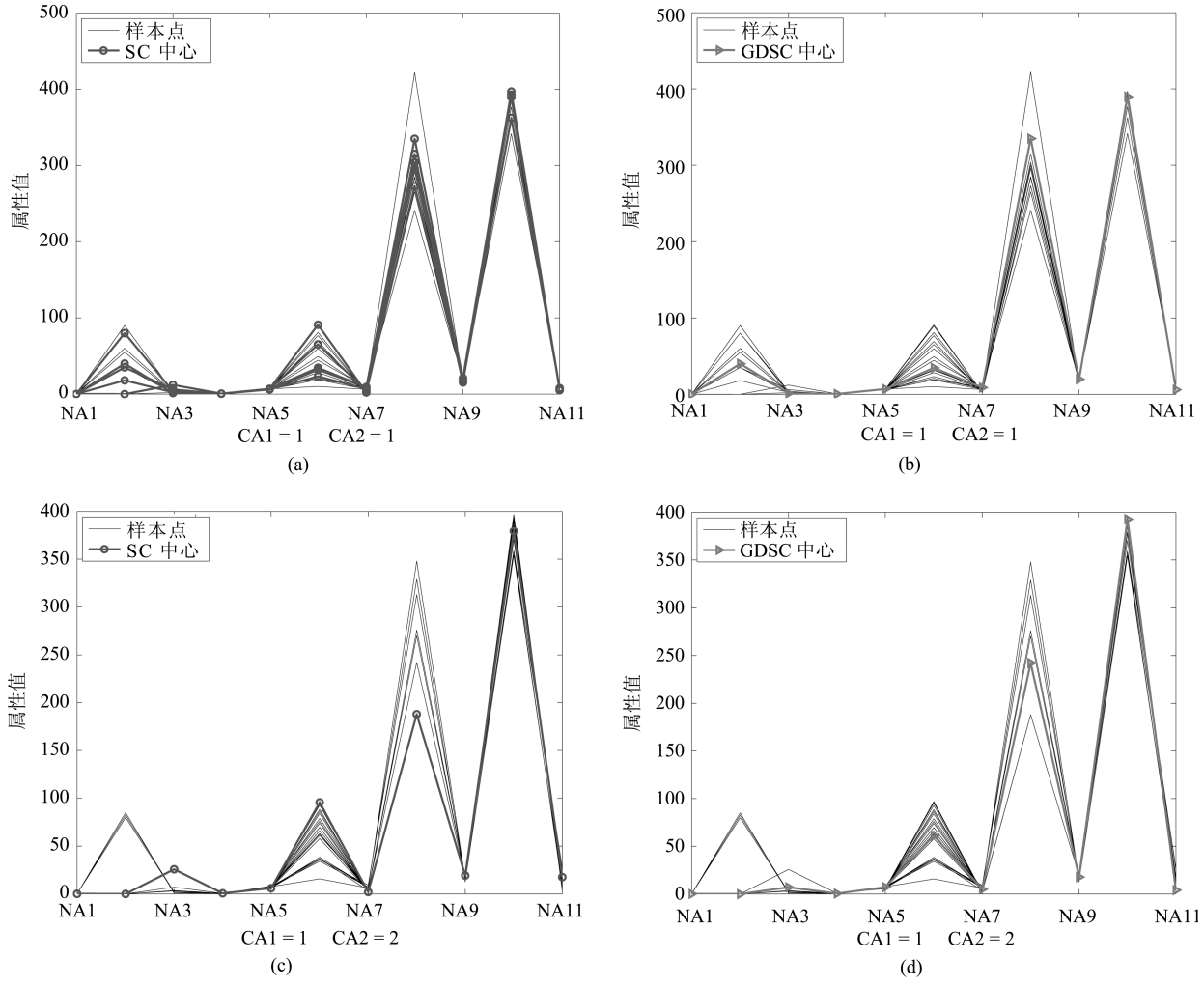


图 4　模型预测结果对比

Fig. 4　Comparison of model prediction

生的规则作为 MDI-ANFIS 的网络结构, 再通过对 MDI-ANFIS 进行训练得出模型来预测波士顿的房价. 为了可视化方便, 我们选取 CA1 = 1, CA2 = {1, 2, 3} 的样本点且使用平行坐标系显示 (其结果见图 5), 图 5 通过平行坐标系实现高维混合属性数据的可视化, 从图中我们可以看出减法聚类得到的聚类中心数是 12, 聚类中心相对集中, 存在一致性的问题. 而基于高氏距离减法聚类得到的聚类中心数是 4, 且聚类中心位置分布相对比较合理. 我们可以发现, GDSC 算法得出的聚类中心数比 SC 算法得到的聚类中心数显著减小, 且 GDSC 算法得到的聚类中心更具代表性.

表 3 从 Boston Housing 数据集中随机选取 10 组样本集作训练对结构辨识性能对比, 其中规则数反映出利用两种算法做辨识得到的规则数目多少, 预测误差反映采用两种算法作结构辨识时模型的预测精度. 通过 10 组样本预测结果比较可以看出, 两种辨识算法的预测误差平均值较为接近, 但 GDSC 算法在结构辨识中产生的规则较少, 降低了需要训练的规则参数个数, 因此模型的参数辨识速度相对较快. 图 6 是 MDI-ANFIS 在第 1 组数据下采用两种辨识算法做训练的模型训练误差, 图 7 是 MDI-ANFIS 模型预测波士顿房价的结果.



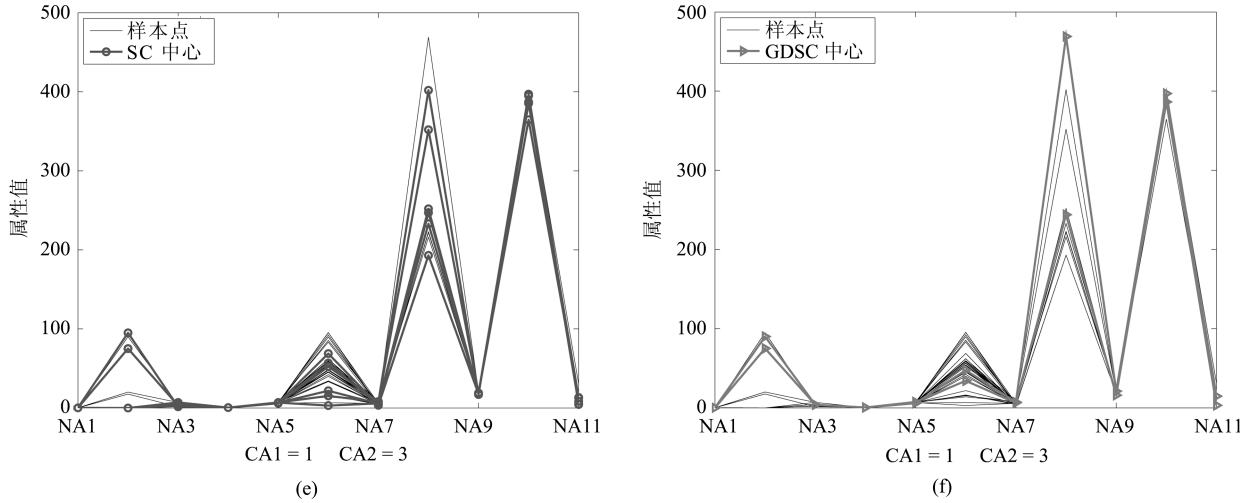


图 5 聚类结果对比图

Fig. 5 Comparison of clustering results

表 3 结构辨识性能对比

Table 3 Performance comparison of structure identification

组号	样本点个数	规则数		预测误差	
		SC	GDSC	SC	GDSC
1	100	50	13	0.452	0.356
2	100	32	14	0.575	0.466
3	200	37	21	0.517	0.709
4	200	25	14	0.908	0.613
5	300	40	18	0.586	0.690
6	300	34	16	0.661	0.705
7	400	31	16	0.642	0.459
8	400	32	14	0.630	0.747
9	500	30	13	0.788	0.836
10	506	30	14	0.726	0.827
平均值	300	34	15	0.648	0.641

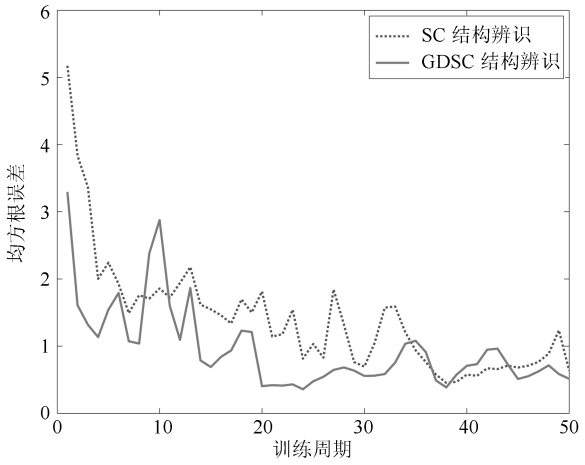


图 6 模型训练误差

Fig. 6 Model training error

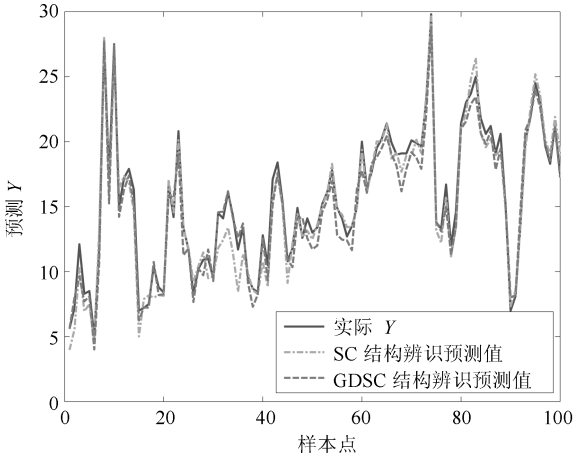


图 7 MDI-ANFIS 模型预测对比

Fig. 7 Prediction results comparison of MDI-ANFIS

2.3 模型误差对比分析

为了比较分析 MDI-ANFIS 模型在混合属性数据建模上的性能, 现选取几种已有混合属性数据建模方法与之对比, 对比建模方法说明如下:

1) ANFIS 模型: 采用标准的 ANFIS 算法, 其中混合属性数据只考虑数值输入, 而不考虑分类输入.

2) 带有数值转化的自适应模糊推理系统 (Adaptive network-based fuzzy inference system with numeric conversion, N-ANFIS) 模型: 将分类数据转化为数值数据 (如 1, 2, 3, $\cdots$), 然后和数值输入一起导入标准的 ANFIS 网络中.

3) 带有频率转化的自适应模糊推理系统 (Adaptive network-based fuzzy inference system with frequency conversion, F-ANFIS) 模型: 通过频率给分类数据赋值, 之后与数值输入导入 ANFIS

网络.

4) 分离多层感知机 (Multi-layer perception with separation method, S-MLP) 模型: 是由 Brouwer 提出的混合属性数据预测模型, 分类数据经编码后与以数值数据做输入的 MLP 的输出作点乘, 产生预测输出.

5) C-ANFIS 模型: 是由 Liu 等提出的 C-ANFIS 混合属性数据预测模型, 其分类数据经激励强度转移矩阵作用到 ANFIS 结构上.

6) MDI-ANFIS 模型: 本文所提出的混合属性数据预测模型, 分类数据经激励强度转移矩阵和后件影响矩阵作用到 ANFIS 上.

对比实验选取 UCI 数据库中的 Abalone、Boston Housing、Auto MPG、Servo、TAE、Zoo 和 Heart Disease 数据集, 验证本文提出的算法对不同数据集的性能.

这里对 ANFIS、N-ANFIS、F-ANFIS 和 C-ANFIS 模型的结构辨识采用 SC 算法; 对 MDI-ANFIS 模型的结构辨识采用 GDSC 算法, 其初始参数设置为: 邻域半径 $\gamma_a = (1/2)\min_k\{\max_r\{\|X_k - X_r\|\}\}$, $\|X_k - X_r\|$ 表示样本点 X_k 和 X_r 的高氏距离, 阈值 $\varepsilon = 0.06$, 最大迭代次数 $L = 100$, 训练周期 epoch = 50, 初始化步长 step = 0.01, 惯性因子 gamma = 0.75, 激励强度转移矩阵 FTM 和后件影响矩阵 CIM 初始化为 0～1 区间的随机矩阵. 而 S-MLP 模型设置学习率 deta = 0.001, 训练周期 epoch = 1 000, 权值矩阵初始化为 0～1 区间的随机矩阵.

对比实验采用十折交叉验证, 选取均方根误差 (Root mean squared error, RMSE) 为模型预测误差的评价指标.

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(Y_k - \overline{Y}_k\right)^2} \tag{27}$$

其中, Y_k 为第 k 个样本点的预测输出值, $\overline{Y}_k$ 为第 k 个样本点的期望输出值, K 为样本点总数.

实验过程, 记录每次测试集的 *RMSE*, 然后对十次测试得到的 *RMSE* 求其平均值, 以此来判断模型对一种数据集的预测精度.

同时, 本文还通过误差降低率 (Error reduction rate, ERR) 对各模型进行比较, 其反映了各模型的 *RMSE* 相对原有模型下降的大小, 误差降低率定义为:

$$ERR = \frac{RMSE_{\text{other}} - RMSE_{\text{M}}}{RMSE_{\text{other}}} \tag{28}$$

其中, $RMSE_{\text{M}}$ 是 MDI-ANFIS 模型的预测误差, $RMSE_{\text{other}}$ 是本文对比的其他混合属性数据建模方法的预测误差.

表 4 给出 UCI 数据库中 7 个数据集对应不同建模方法的预测误差以及误差降低率. 通过几种建模方法的预测误差和误差降低率结果对比可以看出, 对于不同数据集本文所提出的 MDI-ANFIS 相对 ANFIS、F-ANFIS、S-MLP 和 C-ANFIS 具有相对较高的预测精度, 仅相对于 N-ANFIS 误差降低率较小. 当对比 7 个数据集的误差降低率平均值时, N-ANFIS 相对 MDI-ANFIS 高出 0.203.

表 4　UCI 数据集模型误差对比

Table 4　Model error comparison on UCI dataset

数据集	样本	混合属性	预测误差						误差降低率				
	个数	(N, C)	ANFIS	N-ANFIS	F-ANFIS	S-MLP	C-ANFIS	MDI-ANFIS	ANFIS	N-ANFIS	F-ANFIS	S-MLP	C-ANFIS
Abalone	4 177	7, 1	2.608	1.842	1.997	3.985	2.632	1.951	0.336	−0.056	0.023	1.04	0.349
Boston Housing	506	11, 2	0.779	0.631	0.657	7.096	0.824	0.638	0.221	−0.011	0.029	10.1	0.291
Auto MPG	398	4, 3	2.072	0.912	0.871	6.969	0.963	0.605	2.42	0.507	0.439	10.5	0.591
Servo	167	2, 2	1.012	0.060	0.051	3.119	0.362	0.025	39.4	1.40	1.04	123	13.4
TAE	151	1, 4	2.972	0.196	0.385	0.849	0.192	0.225	12.2	−0.128	0.711	2.77	−0.146
Zoo	101	1, 15	1.276	0.062	0.059	2.542	0.126	0.072	16.7	−0.138	−0.181	34.3	0.750
Heart Disease	303	6, 7	0.255	0.073	0.062	1.483	0.108	0.086	1.96	−0.151	−0.279	16.2	0.255
平均值	–	–	1.568	0.539	0.583	3.720	0.744	0.515	10.462	0.203	0.255	28.273	2.213

进一步我们对比 N-ANFIS 和 MDI-ANFIS 的计算时间复杂度, 这里我们假设 W 为训练周期, K 为样本点个数, n 为数值属性个数, m 为分类属性个数, L 为规则数, 则 N-ANFIS 和 MDI-ANFIS 的时间复杂度分别为 $\mathrm{O}\left(W \times K \times (n+m) \times L^3\right)$ 和 $\mathrm{O}(W \times K \times n \times L^3)$, 因此, 在输入是高维混合属性数据时, MDI-ANFIS 的程序运行效率要高于 N-ANFIS.

3 结论

本文针对已有混合数据模型存在的模型组合随分类变量呈几何增长以及子模型训练数据分布不均匀问题, 提出一种具有混合数据输入的自适应模糊神经推理系统模型. 该模型引入激励强度转移矩阵和后件影响矩阵, 构建新型模糊神经网络结构, 使混合属性数据对模糊规则的前后件同时产生影响. 在模型的结构辨识中, 将高氏混合距离引入减法聚类, 计算混合型样本点的密度值, 克服了经典 ANFIS 网络仅适用于数值数据不适用分类数据的缺陷. 在模型的参数学习中, 使用 BP 和 LSE 混合学习算法来训练前件参数、激励强度转移矩阵、后件参数以及后件影响矩阵. 仿真实验验证了后件规则对模型的影响作用, 并验证了结构辨识中采用 GDSC 算法能够以更少的规则数达到模型精度要求. 最后, 选取 UCI 数据库中 7 组数据进行对比实验, 结果表明所提出的具有混合数据输入的自适应模糊神经推理系统模型相比其他模型具有更高的预测精度.

References

1 Alexander F J, Hoisie A, Szalay A. Big data. *Computing in Science & Engineering*, 2011, **13**(6): 10−13

2 Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. *Big Data*, 2013, **1**(1): 51−59

3 Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(1): 97−107

4 Chen Jin-Yin, He Hui-Hao. Research on density-based clustering algorithm for mixed data with determine cluster centers automatically. *Acta Automatica Sinica*, 2015, **41**(10): 1798−1813
(陈晋音, 何辉豪. 基于密度的聚类中心自动确定的混合属性数据聚类算法研究. 自动化学报, 2015, **41**(10): 1798−1813)

5 Jacobs R A, Jordan M I, Nowlan S J, Hinton G E. Adaptive mixtures of local experts. *Neural Computation*, 1991, **3**(1): 79−87

6 Lee K W, Lee T. Design of neural networks for multi-value regression. In: Proceedings of the 2001 International Joint Conference on Neural Networks. Washington DC, USA: IEEE, 2001. 93−98

7 Brouwer R K. A feed-forward network for input that is both categorical and quantitative. *Neural Networks*, 2002, **15**(7): 881−890

8 Brouwer R K. A hybrid neural network for input that is both categorical and quantitative. *International Journal of Intelligent Systems*, 2004, **19**(10): 979−1001

9 Rey-del-Castillo P, Cardeñosa J. Fuzzy min-max neural networks for categorical data: application to missing data imputation. *Neural Computing and Applications*, 2012, **21**(6): 1349−1362

10 Hsu C C. Generalizing self-organizing map for categorical data. *IEEE Transactions on Neural Networks*, 2006, **17**(2): 294−304

11 Zhang Yu-Xian, Peng Hui-Deng, Wang Jian-Hui. Self-organizing mapping clustering algorithm based on heterogeneous value difference metric for mixed attribute data. *Chinese Journal of Scientific Instrument*, 2016, **37**(11): 2555−2562
(张宇献, 彭辉灯, 王建辉. 基于异构值差度量的 SOM 混合属性数据聚类算法. 仪器仪表学报, 2016, **37**(11): 2555−2562)

12 Liu M, Dong M Y, Wu C. A new ANFIS for parameter prediction with numeric and categorical inputs. *IEEE Transactions on Automation Science and Engineering*, 2010, **7**(3): 645−653

13 Jang J S R. ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 1993, **23**(3): 665−685

14 Abdelrahim E M, Yahagi T. A new transformed input-domain ANFIS for highly nonlinear system modeling and prediction. In: Proceedings of the 2001 Canadian Conference on Electrical and Computer Engineering. Toronto, Canada: IEEE, 2001. 655−660

15 Mar J, Lin F J. An ANFIS controller for the car-following collision prevention system. *IEEE Transactions on Vehicular Technology*, 2001, **50**(4): 1106−1113

16 Lima C A M, Coelho A L V, Von Zuben F J. Fuzzy systems design via ensembles of ANFIS. In: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems. Honolulu, USA: IEEE, 2002. 506−511

17 Paramasivam S, Arumugan R, Umamaheswari B, Vijayan S, Balamurugan S, Venkatesan G. Accurate rotor position estimation for switched reluctance motor using ANFIS. In: Proceedings of the 2001 Conference on Convergent Technologies for the Asia-Pacific Region. Bangalore, India: IEEE, 2003. 1493−1497

18 Lih W C, Bukkapatnam S T S, Rao P, Chandrasekharan N, Komanduri R. Adaptive neuro-fuzzy inference system modeling of MRR and WIWNU in CMP process with sparse experimental data. *IEEE Transactions on Automation Science and Engineering*, 2008, **5**(1): 71−83

19 Chiu S L. Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology*, 1994, **2**(3): 267−278

20 Tuerhong G, Kim S B. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Systems with Applications*, 2014, **41**(4): 1701−1707

21 Zeng Ke, Zhang Nai-Yao, Xu Wen-Li. Sufficient condition for linear T-S fuzzy systems as universal approximators. *Acta Automatica Sinica*, 2001, **27**(5): 606−612
(曾珂, 张乃尧, 徐文立. 线性 T-S 模糊系统作为通用逼近器的充分条件. 自动化学报, 2001, **27**(5): 606−612)

22 Liu Hui-Lin, Feng Ru-Peng, Hu Rui-Dong, Liu Chun-Hua. Decennary development of fuzzy systems as universal approximators. *Control and Decision*, 2004, **19**(4): 367−371
(刘慧林, 冯汝鹏, 胡瑞栋, 刘春华. 模糊系统作为通用逼近器的 10 年历程. 控制与决策, 2004, **19**(4): 367−371)

张宇献 沈阳工业大学电气工程学院副教授. 2007 年获得东北大学控制理论与控制工程专业博士学位. 主要研究方向为智能控制, 复杂系统建模, 智能优化. 本文通信作者.
E-mail: yuxian524524@163.com
(**ZHANG Yu-Xian** Associate professor at the School of Electrical Engineering, Shenyang University of Technology. He received his Ph. D. degree from Northeastern University in 2007. His research interest covers intelligent control, complex system modeling and intelligent optimization. Corresponding author of this paper.)

郭佳强 沈阳工业大学信息科学与工程学院硕士研究生. 主要研究方向为智能控制, 复杂系统建模.
E-mail: guo_dataworld@163.com
(**GUO Jia-Qiang** Master student at the School of Information Science and Engineering, Shenyang University of Technology. His research interest covers intelligent control and complex system modeling.)

钱小毅 沈阳工业大学电气工程学院博士研究生. 主要研究方向为智能优化, 复杂机电装备的故障诊断.
E-mail: qianxiaoyi123@163.com
(**QIAN Xiao-Yi** Ph. D. candidate at the School of Electrical Engineering, Shenyang University of Technology. His research interest covers intelligent optimization and fault diagnosis for complex mechanical and electrical equipment.)

王建辉 博士, 东北大学信息科学与工程学院教授. 主要研究方向为智能控制, 复杂系统建模, 康复机器人.
E-mail: wangjianhui@ise.neu.ediu.cn
(**WANG Jian-Hui** Ph. D., professor at the College of Information Science and Engineering, Northeastern University. Her research interest covers intelligent control, complex system modeling and rehabilitation robot.)