# Research on a Method about Fuzzy-Rules Automatic Generation

**Huimin Zhao, Dongyan Li**
*Software Technology Institute, Dalian Jiaotong University, Dalian, Liaoning, China*
*Email: hm_zhao1977@126.com*

**Abstract:** This paper puts forward an approach about fuzzy rules automatic generating from metrical data. Firstly, fill the vacant values with sample data in the collected fuzzy system，applied the idea of the spline interpolation. Then classify the input-output sample data, which have been interpolated with K- average clustering method. Finally, discretize the data based on the projections of sample space clustering on every dimension, offer an objective basis for the defining of fuzzy subclass of each variable in fuzzy system, and make a good base for the further automatic establishment of fuzzy rules and the adaptive regulation of fuzzy membership function. Then define the occupancy, present a theoretical basis for the reduction of fuzzy rules and the generation of consistency fuzzy model. With the thinking above, proceeded the extraction of fuzzy rules.

**Keywords:** automatic generation of fuzzy-rules; data pretreatment; clustering; occupancy ratio

# 一种模糊规则的自动生成方法研究

**赵慧敏，李东艳**
大连交通大学软件学院，大连，辽宁，中国，116028
Email: hm_zhao1977@126.com

**摘　要：**本文提出了一种从测量数据中自动提取模糊规则的方法，该方法首先利用样条插值思想对所采集的模糊系统样本数据进行空缺值的填充，然后运用基于K—均值的聚类方法对插值后的输入—输出样本数据进行聚类；最后根据样本空间聚类在各维上的投影，对数据进行离散化处理，为模糊系统各变量的模糊子集定义提供客观依据，为下一步模糊规则的自动确定，模糊隶属度函数自适应调整等策略，打下基础。之后给出了占有率的定义，为模糊规则的化简和生成一致性模糊模型提供理论依据。并用以上的思想方法进行了模糊规则的提取。

**关键词：**模糊规则自动提取；数据预处理；聚类；占有率

## 1 引言

30 多年来，模糊控制在各个领域中得到了长足的发展，但是，模糊控制规则的获取一直是困扰模糊控制发展的主要问题之一。在模糊控制系统中，模糊规则库的构建是至关重要的。然而从大量文献[1-2]中可以看出：模糊规则的数量基本上是人为决定的。尽管用各种方法进行优化计算，人工设定模糊规则还是很困难的，甚至是不可能的。因此研究模糊规则的自动生成有着重要的理论和应用价值。

本文提出了一种从测量数据中自动提取模糊规则的方法。文中首先介绍了数据预处理的基本思想，接着对输入输出空间划分方法和数据离散原理进行介绍，最后介绍了模糊规则的自动提取方法。

资助信息：国家自然科学基金项目支持（资助号：60870009）

## 2 数据预处理

### 2.1 空缺数据的插值填充

在所采集的样本数据中，数据缺损是不可避免的，尤其是实时控制过程采集数据情况下，样本数据在输入输出空间中的分布是非均匀和不完备的。为了能够给模糊模型的建立提供一个比较好的样本数据集，必须对所采集的原始样本数据进行数据清理。

使用最可能的值填充空缺值是最常用的方法，与


978-1-935068-40-2 

其它方法相比，它使用现存数据的多数信息来推测空缺值，有更大的机会使插入的填充值和原值保持一定的属性联系。

### 2.2 样条插值基本原理

采用样条插值的方法对空缺值进行填补的基本思想是：首先，假设原始采样数据已进行了噪声与一致性处理，利用这些采样数据进行样条插值(或最小二乘法拟合)，得到一个光滑的曲线或超平面；然后，在这个插值或拟合出的曲线或超平面上再进行插值计算，得到空缺点处的空缺值(插值函数值)。

插值法的定义为：设实变量 x 的实函数 f (x)具有直到某阶的导数，它在 n+1 个不同的点 x0，x1，…，xn 处的值分别为 f0 ， f1 ，…， fn ，使用这些值来求 f (x)在其它 x 处的值 f 的方法，称为插值法。通常通过 n+1 个点(xi，fi)作 n 次多项式(称为插值多项式)来进行插值，这种方法称为 Lagrange 插值法。

## 3 输入输出空间划分[3]

为了能够对输入输出空间进行模糊自动分区、隶属函数自动生成和模糊规则的自动提取，对样本空间进行划分和数值规约处理是必要的，在此采用 K－均值聚类方法。通过 K－均值方法，将 N 个样本数据点划分为 K 个簇，各簇在样本数据集 D 中构成了 K 个样本子集 $C_i$ ( i =1，…，K )。各子集中样本点的平均值可以作为自动定义各变量模糊子集的代表值(隶属函数为 1 的点)；根据子集中元素对应的各变量的变化区间形成输入输出空间的划分；各子集标号 i 即为数据规约值。

具体的实现过程说明如下：设在输入输出空间给定一个具有 N 个样本点构成的数据集 D，且每个样本点具有 m 个输入参数和 n 个输出参数，即每一个样本点可以用一个 m+n 维的向量表示。首先，在数据集 D 中使用相应的聚类中心赋初值方法选择 K 个样本点，每个样本点代表一个簇的初始平均值(或中心)；然后，对剩余的 N - K 个样本点，通过计算其与各个簇中心的距离，将它赋给最近的簇；最后，在重新计算每个簇的平均值。不断地重复该过程，直到准则函数收敛。在此，采用平方误差准则，即

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - \overline{m}_i |^2 \quad (1)$$

式中，E 为数据集 D 中所有点的平方误差的总和，$p \in D$ 是输入输出空间中的点表示给定的样本点，$\overline{m}_i$ (i=1，2，…，K)表示簇 $C_i$ (i=1，2，…，K)的平均值( $p$ 和 $\overline{m}_i$ 都是 m+n 维的)，$C_i$ (i=1，2，…，K)为聚类后的簇。

上述过程可用算法表示如下：

1.在样本数据集 D 中选择 K 个样本点，将 K 个样本点值分别赋给各聚类中心 $\overline{m}_i (i = 1, \ldots,$ k)；

2.对样本数据集 D 中样本点 $p_j$ $(j = 1, \cdots n)$ 依次计算到各簇中心 $\overline{m}_i$ 的距离 $d(i, j) = \sqrt{| p_j - \overline{m}_i |^2}$ ；

3.找出 pj(j=1,…,n)关于 $\overline{m}_i (i = 1, \cdots, k)$ 的最小距离 $d(i, j)$，则 $p_j \in C_i$ ；

4. 计算各簇中样本点的平均值：$\overline{m}_i = \frac{1}{n_i} \sum_{j_i=1}^{n_i} p_{j_j}$ ， $p_{j_i} \in C_i$ ， $i = 1, \cdots, k$ ；

5.按(1)式计算数据集 D 中所有点的平方误差的总和 E(t)，并与前一次误差 E(t-1)比较；

6.若 E(t)－E(t-1)<0，则转到 2；

7.结束。

关于簇的数目 K 值，可根据输出变量所需的模糊语言值个数来确定。一个模糊变量的语言值个数通常可取为 3、5、7、9，综合考虑模型精度及模型复杂度，语言值个数的缺省值可取为 7，即 k=7。

## 4 数据离散原理

当以 K －均值方法对样本数据聚类处理结束后，生成了 K 个簇。由于每一个样本点是由 m 个输入变量值 $\{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ 和一个输出变量值 $\{y_j\}$ 构成，j=1，…，n 。现在用第 i 个簇中的 $n_i$ 个样本点的 m+1 维的坐标值构成这样的数据组：

$$\begin{aligned} b_1^i &= \max\{x_{11}^i, x_{21}^i, \cdots, x_{n_i 1}^i\}; \\ b_2^i &= \max\{x_{12}^i, x_{22}^i, \cdots, x_{n_i 2}^i\}; \\ &\vdots \\ b_m^i &= \max\{x_{1m}^i, x_{2m}^i, \cdots, x_{n_i m}^i\}; \\ b_{m+1}^i &= \max\{y_1^i, y_2^i, \cdots, y_{n_i}^i\}; \end{aligned} \quad (2)$$

式中，$x_{21}^i$ 表示第 i 个簇中的第 2 个样本点关于第 1 个变量的(坐标)值，$y_2^i$ 是该样本点对应的输出值。式(2)对应了 m+1 个最大值和 m+1 个最小值，

不难看出，由这 m+1 个数据组中的各自的最大和

$$\begin{aligned} b_1^i &= \max\{x_{11}^i, x_{21}^i, \cdots, x_{n_i 1}^i\}; & s_1^i &= \min\{x_{11}^i, x_{21}^i, \cdots, x_{n_i 1}^i\}; \\ b_2^i &= \max\{x_{12}^i, x_{22}^i, \cdots, x_{n_i 2}^i\}; & s_2^i &= \min\{x_{12}^i, x_{22}^i, \cdots, x_{n_i 2}^i\}; \\ &\vdots & &\vdots \\ b_m^i &= \max\{x_{1m}^i, x_{2m}^i, \cdots, x_{n_i m}^i\}; & s_m^i &= \min\{x_{1m}^i, x_{2m}^i, \cdots, x_{n_i m}^i\}; \\ b_{m+1}^i &= \max\{y_1^i, y_2^i, \cdots, y_{n_i}^i\}; & s_{m+1}^i &= \min\{y_1^i, y_2^i, \cdots, y_{n_i}^i\}; \end{aligned}$$

最小值确定了相应变量的一个变化子区间$(a_l^i, b_l^i)$。因此，对K个簇进行依次处理，则在每一个变量的变化论域上会相应得到K(K缺省值为 7)个变化子区间$(a_l^1, b_l^1)$，…，$(a_l^7, b_l^7)$。

用上述方法产生的输入和输出空间划分，由于某一变量$x_l$ $(l=1,2,\cdots,m,m+1;\ x_{m+1}=y)$关于簇 i 的最大值$b_l^i$不可能与簇 i+1 的最小值$a_l^{i+1}$正好相等，这样两个相邻的区间会出现重叠或分离的情况如图 1 所示，故应进行相应的处理。

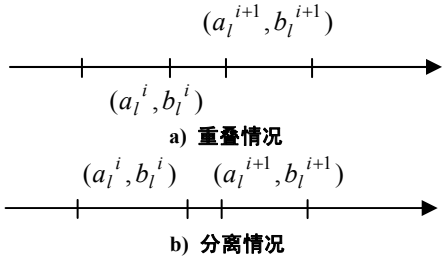


**Fig.1. The place distributing of conjoint area**

**图 1.相邻区间位置分布**

针对图 1 所示情况，我们给出如下的处理策略：

策略 1．若$b_l^i < a_l^{i+1}$，说明相邻两区间分离，则令$c_l^i = b_l^i + \frac{a_l^{i+1} - b_l^i}{2}$，此时两个新的相邻区间为$(a_l^i, c_l^i)$和$(c_l^i, b_l^{i+1})$；

策略 2．若$b_l^i = a_l^{i+1}$，说明相邻两区间恰好相连，则令$c_l^i = b_l^i$即可；

策略 3．若$b_l^i > a_l^{i+1}$，说明相邻两区间重叠，则首先比较两个区间长度的大小，并令 d 等于二分之一较小区间的长度值；若$d \geq (b_l^i - a_l^{i+1})$，则$c_l^i = b_l^i - \frac{b_l^{i1} - a_l^{i+1}}{2}$；否则两个相邻区间合并，产生一个新区间$(a_l^i, b_l^{i+1})$。

按照上述策略对每一个变量区间进行处理，便可完成输入输出空间的划分。

当输入输出空间划分结束后，即每一个变量区间被划分为$q_l$个子区间后，给每一个子区间赋一个名誉值，赋值方法可根据需要自行确定。在此，我们给出缺省的赋值方法为：按照子区间坐标值大小依次给子区间赋值为 1，2，…，$q_l$。这样，当变量的值落在某一区间上时，则该值就用该区间的名誉值表示。从而完成了数据的离散化，为模糊规则的提取作好准备。

## 5 模糊规则的提取

### 5.1 隶属函数的定义

以各个变量的离散点为基础，定义如下形式的隶属函数：

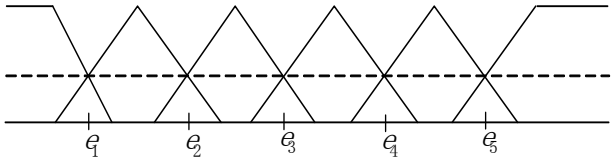


**Fig.2. The membership function of input error**

**图 2.输入误差的隶属函数**

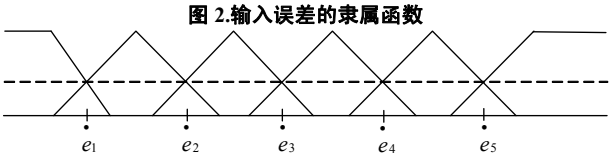


**Fig.3. The membership function of input error rate**

**图 3.输入误差变化率的隶属函数**

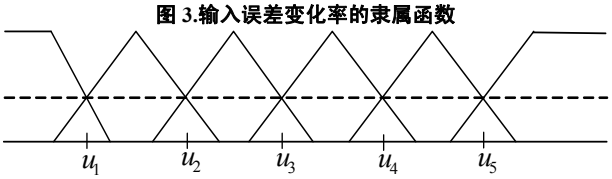


**Fig.4. The membership function of output**

**图 4.输出的隶属函数**

我们定义隶属函数的形状为三角形，模糊集合$FE_i, F\dot{E}_i, FU_i$是和清晰集合$E_i, \dot{E}_i, U_i$。其中，有如下的等式成立：

$$\mu FE_i(\frac{e_{i-1}+e_i}{2}) = 1$$
$$\mu FE_i(e_{i-1}) = \mu FE_i(e_i) = 0.4$$
$$\mu F\dot{E}_i(\frac{\dot{e}_{i-1}+\dot{e}_i}{2}) = 1$$
$$\mu F\dot{E}_i(\dot{e}_{i-1}) = \mu F\dot{E}_i(\dot{e}_i) = 0.4$$
$$\mu FU_i(\frac{u_{i-1}+u_i}{2}) = 1$$
$$\mu FU_i(u_{i-1}) = \mu FU_i(u_i) = 0.4$$

### 5.2 采样点的模糊化

在对系统的输入输出数据进行有限次采样(t 次，但是 t 的个数不定)，得到的输入输出采样值为如下式(3)序列：

$$\begin{matrix} e_1, \dot{e}_1, u_1 \\ e_2, \dot{e}_2, u_2; \\ \vdots \quad \vdots \quad \vdots \\ e_t, \dot{e}_t, u_t \end{matrix} \qquad (3)$$

经插值处理后，对插值后的输入输出采样值按如下原则进行模糊化处理：

若输入采样值$e_i$满足：

$\mu_{E_p}(e_i) = \max[\mu_{E_1}(e_i), \mu_{E2}(e_i), \ldots \mu_{E_m}(e_i)]$则模糊



978-1-935068-40-2 © 2010 SciRes.

变量 $e_i$ 的语言值取为 $E_p$；

第二，若输入采样值 $\dot{e}_i$ 满足：

$$\mu_{E_q}(\dot{e}_i)=\max[\mu_{E_1}(\dot{e}_i),\mu_{E_2}(\dot{e}_i),\ldots,\mu_{E_n}(\dot{e}_i)]$$

则模糊变量 $\dot{e}_i$ 的语言值取为 $\dot{E}_q$；

第三， 若输出采样值 $u_i$ 满足：

$$\mu_{U_j}(u_i)=\max[\mu_{U_1}(u_i),\mu_{U2}(u_i),\ldots\mu_{U_l}(u_i)]$$

则模糊变量 $u_i$ 的语言值取为 $U_j$。

经模糊化处理后，式(3)变为如下形式：

$$E_{1p},\dot{E}_{1q},U_{1j};$$
$$E_{2p},\dot{E}_{2q},U_{2j};$$
$$\vdots\quad\vdots\quad\vdots$$
$$E_{tp},\dot{E}_{tq},U_{tq}。$$

### 5.3 模糊规则的一致性处理

一致性是对模糊系统模型的一项基本要求。也是粗集理论中值约简算法在模糊规则提取中的一种体现。在不一致的模型中，由于存在着相互矛盾的规则，有可能导致控制系统的响应发散。对于这一点无论是在静态的系统建模中，还是在动态地对规则所进行的实时修改中，都是必须加以判定的。

例如，有这样的一条规则：

*if* $e$ *is NB and* $\dot{e}$ *is* ZE *then u is PB*

它意味着当误差是NB，误差变化率为ZE时，输出为PB；

还有这样的一条规则：

*if e is NB and* $\dot{e}$ *is ZE then u is PS*

它意味着当误差是 *NB*，误差变化率为 *ZE* 时，输出为 *PS*。

那么这是两条矛盾的规则，因此，为了考虑规则的一致性，即去除矛盾规则，我们定义决策（结果）属性 $u_k$ 与前件（条件）属性 $e_i\times\dot{e}_j$ 的占有度如下：

$$O_{e_i\times\dot{e}_j}(u_k)=\frac{\left|u_k\cap e_i\times\dot{e}_j\right|}{\left|e_i\times\dot{e}_j\right|}\qquad(4)$$

其中，$u_k$ 是一个由决策(结果)属性 $u$ 划分的等价类，也就是说 $u_k\in U/I(u)$。$|e_i\times\dot{e}_j|$ 表示包含在子空间 $e_i\times\dot{e}_j$ 中的数据数。

我们从式(4)中可以看出，占有度表示前件子空间是 $e_i\times\dot{e}_j$ 而决策(结果)属性是 $u_k$ 的对象数与所有前件子空间是 $e_i\times\dot{e}_j$ 的对象数之比。

如果输入和输出空间已经被模糊分割，则生成模糊规则与确定如下的一个映射具有相同的意义 $f$: $F_e\times F_{\dot{e}}\to F_u$，最优模糊规则表可以通过确定等式 $F_e\times F_{\dot{e}}\to F_u$ 的最优映射来获得。我们构造映射等式如下：

$$f(e_i,\dot{e}_j)=u_k$$

其中 $u_k=\max_{u_d\in U/I(D)}O_{e_i\times e_j}(u_d)$，D是决策（结果）变量。

这样，我们就去除了矛盾规则，生成了一致规则表。

## 6 结论

本文利用样条插值思想对所采集的模糊系统样本数据进行空缺值的填充，消除了缺损数据对后续处理的影响。为了自动确定模糊规则，运用基于K—均值的聚类方法对插值后的输入—输出样本数据进行聚类；然后根据样本空间聚类在各维上的投影，对数据进行离散化处理，定义了“占有率”的概念，为模糊规则的化简和生成一致性模糊模型提供理论依据。

## References

[1] L.A.Zadeh.Fuzzy Sets.Information and Control， 1965，8:338-353.
[2] Sugeno M. An Introductory survey of fuzzy control Information Sciences，1985，36:59-832.
[3] Jiawei Han，Micheline Kamber. Data Mining Concepts and Techniques. China Machine Press，2001.8.