

一种新的分级混合聚类法

马宝萍

(南京师范大学控制科学与工程系, 210042, 南京)

[摘要] 为了克服模糊聚类算法的不足, 提出了一种新的分级混合聚类法, 利用自组织神经网络对数据初步进行特征提取, 再利用基于熵的聚类算法进行聚类, 从而既提高了聚类过程的效率, 又保证了聚类结果的有效性.

[关键词] 聚类, 自组织神经网络, 熵

[中图分类号] TP18, [文献标识码] A, [文章编号] 1672- 1292- (2003) 01- 0022- 04

聚类分析是模式识别的基本内容之一, 常用的聚类方法有系统聚类法、模糊聚类法、自组织特征映射等等, 系统聚类法是以多元统计分析为基础的数学分类方法, 其产生与应用已有很长的历史, 而模糊聚类法和自组织特征映射则是目前模式识别中的研究热点.

模糊聚类算法(如FCM、PCM 算法等等)的一个共同特点就是要求事先确定分类数目, 而这个要求在很多实际情况中是难以满足的, 因此使其应用受到一定的限制, 同时上述算法还具有对初始条件敏感、容易陷入局部极小值等不足. 文献[1]采用Kohonen自组织网络与FCM结合的聚类方法, 提高了FCM算法的速度, 但仍不可避免地会陷入局部极小值. 文献[2]提出了一种基于熵的聚类方法, 无需预先指定聚类数, 并且待定参数较少, 但文中没有对聚类有效性进行直接检验, 难以保证聚类结果为最优.

大量的文献报导都旨在改进聚类算法的性能, 却极少研究如何提高聚类过程的效率. 事实上在进行聚类有效性检验时, 需要对多种可能的聚类数进行重复地计算和比较, 计算量非常大, 耗时很长.

为了提高聚类的速度及效率, 同时保证聚类有效性, 本文提出一种新的分级混合聚类法(Hierarchical Hybrid Clustering Method, HHCM), 通过实例说明该方法的聚类结果较好, 并且克服了陷入局部极小值的问题.

1 分级混合聚类法

本文提出的分级混合聚类法的流程图如图1所示.

其中HHCM算法分两步进行, 第一步利用Kohonen自组织网络对数据进行特征提取, 第二步采用基于熵的聚类法. 该方法的特点是无需事先指定分类数目, 并且计算量较小. HHCM的结构如图2所示.

其中 $X = \{x_1, x_2, \dots, x_n\} \subset R^p$ 为待分类的数据样本, $T = \{t_1, t_2, \dots, t_m\} \subset R^p$ 为第一级聚类的结果, 通常 $\sqrt{n} \leq m \ll n$, 也就是说这一步起到了数据压缩的效果, 为下一级聚类作好准备, $V = \{v_1, v_2, \dots, v_c\} \subset R^p$ 为第二级聚类得到的类的中心. 需要说明的是在聚类之前, 要对数据进行标准化处理, 例如对 X 中的第 i 个变量:

计算样本均值
$$x = \frac{1}{n} \sum_{j=1}^n x_{ij} \tag{1}$$

样本极差
$$R_i = \max_{j \in \{1, \dots, n\}}(x_{ij}) - \min_{j \in \{1, \dots, n\}}(x_{ij}); \quad i = 1, 2, \dots, p \tag{2}$$

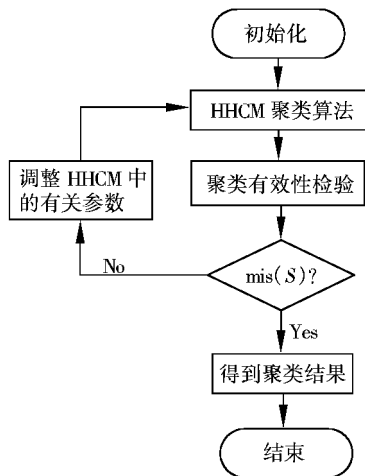


图1 分级混合聚类算法流程图

收稿日期: 2003- 02- 18.

作者简介: 马宝萍, 女, 1973- , 工学博士, 南京师范大学控制科学与工程系讲师, 主要研究方向为模糊建模与控制、神经网络.

$$\text{则标准化后的} \quad x'_{ij} = \frac{x_{ij} - x_i}{R_i}, \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, n \quad (3)$$

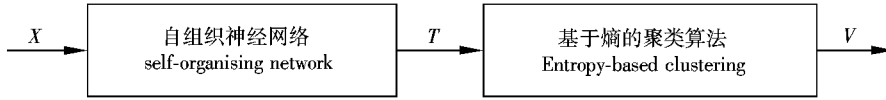


图2 分级混合聚类法示意图

1.1 自组织神经网络

图2中的自组织神经网络由输入层和竞争层组成,如图3所示:

将待分类模式 X 依次输入网络,经过竞争学习,在输出侧得到其特征向量 t_1, t_2, \dots, t_m . 在此采用改进的自组织算法^[1]对网络进行训练,具体步骤如下:

第1步:产生竞争层的第一个神经元,将模式 $x_1 = (x_{11}, x_{21}, \dots, x_{p1})$ 输入到输入层,迭代次数 $l = 1$,令第一个权值向量 $W_l^1 = x_1$,即 $w_{il}^1 = x_{i1}, i = 1, 2, \dots, p$,确定竞争层神经元的有效半径 δ ,神经元个数 $N_l = 1$,第 l 个单元的激活次数 $N_{sl} = 1$.

第2步:对于第 l 个输入模式,找出与当前模式距离最小的单元 J ,即

$$D(W_J^l, x_l) = \|W_J^l - x_l\| = \min_{i=1,2,\dots,N_l} \|W_i^l - x_l\| \quad (4)$$

$$\text{其中:} \quad \|W_J^l - x_l\| = (W_J^l - x_l)(W_J^l - x_l)^T \quad (5)$$

第3步:确定获胜单元

$$\text{If} \begin{cases} D(W_J^l - x_l) \leq \delta, & J \text{ is the winner} \\ D(W_J^l - x_l) > \delta, & \text{create a new unit} \end{cases}$$

如果单元 J 获胜,则修正该单元的权值

$$W_J^l = W_J^{l-1} + \alpha \|x_l - W_J^l\|, \quad \alpha = \alpha_0 / (N_{sl} + 1), \alpha_0 \in [0, 1] \quad (6)$$

$$N_{sl} = N_{sl} + 1; l = l + 1$$

如果生成新的单元,则

$$N_l = N_l + 1, \quad W_{N_l}^l = x_l, \quad l = l + 1$$

若 $l < n$,转第2步;否则 $m = N_l$,转下一步.

第4步:剔除散点,先找出竞争层各个单元的最大激活次数

$$N = \max\{N_{s1}, N_{s2}, \dots, N_{sm}\}$$

再将激活次数较小,如 $N_{sj} < rN$, $r \in (0, 1)$, $j \in \{1, 2, \dots, m\}$ 的单元除去.重置 $m = m - N_r$, N_r 表示被除去的单元个数.最后竞争层单元的输出为:

$$t_i = W_i, \quad i = 1, 2, \dots, m; \quad W_i = (w_{1i}, w_{2i}, \dots, w_{pi}) \quad (7)$$

即每个节点代表输入——输出空间的一个类.可见,如果某一个类中点的密度过低时,说明该类的分布具有一定的随机性,可能是由噪声引起的,除去这样的类之后,也就意味着受噪声污染的点不会进入下级分类.

1.2 基于熵的聚类算法

图2中的第二阶段的聚类采用基于熵的方法.根据熵的理论可知,基于熵的聚类方法尤其适用于类与类之间的界限不十分明显的情况.文献[3]中指出聚类过程实质上也可以看作是由高熵状态向低熵状

态的转变. 在多维数据空间中, 两点之间的熵的取值范围为 $0.0 \sim 1.0$, 聚类中心点的熵比其它点的要小.

定义 1 x_i 与 x_j 之间的相似性测度 S 为:

$$S(x_i, x_j) = e^{-D_{ij}} \quad (8)$$

其中 D_{ij} 为 x_i 与 x_j 之间的欧氏距离, S 的取值范围为 $0.0 \sim 1.0$, S 趋于 1 表示两点距离很近, 可能落入同一类中, S 趋于 0 表示两点距离较远, 应属于不同的类. $\alpha = -\ln 0.5/D$, D 为所有点之间距离的平均值.

定义 2 点 x_i 的熵 E_i 可以表示为香农函数的形式

$$E_i = - \sum_{j \neq i} (S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})) \quad (9)$$

基于熵的聚类法具体步骤如下:

第 1 步: 对于 $T = \{t_1, t_2, \dots, t_m\}$, 计算每个点 t_i 的熵 E_i , $i = 1, 2, \dots, m$;

第 2 步: 找出 $\min(E_i)$ 对应的点 t_s , 作为一个类的中心;

第 3 步: 将点 t_s 以及满足

$$S(t_i, t_s) > \beta, \quad t_i \in T \quad (10)$$

的点从 T 中除去, 如果 T 仍不为空集, 则返回第 2 步.

最后, 将聚类结果表示为 $V = \{v_1, v_2, \dots, v_c\}$. 至此将 m 类进一步分成 c 类.

需要说明的是从 T 中选出熵最小的一点作为一个类的中心后, 与该点的相似性测度大于 β 的点在后面的聚类过程中将不再被考虑作为类的中心, 也就是说相似度高的点属于同一类的可能性更大. 此外为了避免将边缘点作为聚类中心, 在选取某一点为类的中心之前, 先计算满足 (10) 式的点数, 若小于 γ (例如取 $\gamma = m \cdot 5\%$, m 为总的数据点数) 的话, 表示该点可能为边缘点, 不宜作为类的中心.

(10) 式中 β 决定着聚类的数目和类的半径, β 越大, 类的数目越多, 半径越小, 因此适当调整 β 可以进行聚类的有效性检验, 从而确定模糊模型中的规则数. 在此选择 Sugeno-Yasukawa's 准则^[4], 令其中的模糊指数 $m = 0$, 则成为:

$$S(V, c) = \sum_{k=1}^n \sum_{i=1}^c (\|x_k - v_i\|^2 - \|v_i - x\|^2) \quad (11)$$

其中 x 为样本均值, 从上式中两项的含义不难看出, S 越小, 意味着每一类的密度越大, 类与类之间的距离越大, 分类效果也就越理想, 因此在有效性检验过程中改变 β 使 S 逼近最小值即可得到较优的分类. 由上述算法的步骤可以看出, 改变 β 后无需再计算熵和相似性测度, 因此与其它方法相比, 该方法的计算量要小得多, 而且需要人为确定的参数很少.

2 实例

下面通过实例来比较本文提出的分级混合聚类法与模糊 C 均值法 (FCM)、自组织特征映射法 (SOFM) 的聚类效果.

例: 选取二维空间的一组数据样本, 共 40 个点, 分布情况如图 4 所示:

采用 FCM 方法聚类, 假定聚类数 $c = 4$, 模糊度指数 $m = 2$, 分类结果如图 4(b) 所示, 图中数据所在位置为类的中心, 不同的符号表示不同的类. SOFM 方法数据分成 3 类, 如图 4(c) 所示. 分级混合聚类法中的参数 $\beta = 0.7$, 分类数为 4, 如图 4(d) 所示. 直观地看, 分成 4 类比 3 类更合理, 因此无需再对 β 进行调整, 而且从图中可以看出, 在对第 1 类与第 2 类的划分上, 分级混合聚类法的结果比前两种方法要好.

3 种聚类方法的 Sugeno-Yasukawa 指标分别为:

$$S_{\text{FCM}} = 13.2976; \quad S_{\text{SOFM}} = 18.6661; \quad S_{\text{HHCM}} = 12.3657$$

由此可见, HHCM 方法的聚类结果更为理想.

3 结论

本文提出的分级混合聚类法无需事先指定分类数目, 并且尽可能地减少了聚类有效性检验时的重

复计算, 能够在很大程度上克服数据中的噪声对聚类的不良影响, 保证聚类结果的有效性.

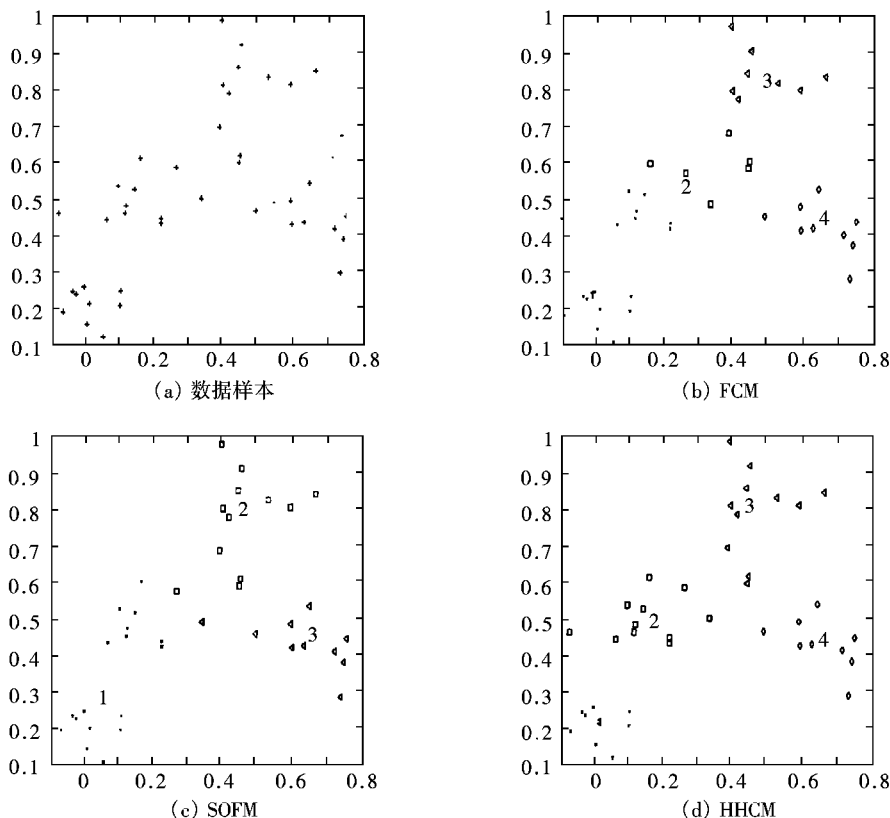


图 4 聚类方法的比较

[参考文献]

- [1] Linkens D A, M Y Chen. Input selection and partition validation for fuzzy modeling using neural network[J]. Fuzzy Sets and Systems, 1999, 107(3): 299~ 308.
- [2] Yao J, Dash M, Tan S T, *et al.* Entropy-based fuzzy clustering and fuzzy modeling[J]. Fuzzy Sets and Systems, 2000, 113(4): 381~ 388.
- [3] Shimoji S, Lee S. Data clustering with entropical scheduling[C]. Proceeding of IEEE Conference on Fuzzy Systems, 1994, 10: 2423 ~ 2428.
- [4] Sugeno M, Yasukawa T. A fuzzy logic based approach to qualitative modeling[J]. IEEE Trans Fuzzy Systems, 1993, 1(1): 7~ 31.
- [5] 马宝萍. 模糊建模与神经网络控制的研究及其在循环流化床锅炉中的应用[D]. 东南大学, 2001.

A New Hierarchical Hybrid Clustering Method

Ma Baoping

(Department of Control Science and Engineering, Nanjing Normal University, 210042, Nanjing, PRC)

Abstract: A new hierarchical hybrid clustering method is proposed to overcome the disadvantage of fuzzy clustering algorithm. By adopting the method, with the self-organizing network used to have the preliminary acquisition of the character of data, the data was then clustered by using an algorithm based on entropy. As a result, the efficiency and validity of clustering can be enhanced.

Key words: clustering, self-organizing neural network, entropy

[责任编辑: 刘健]