文章编号: 1001-0920 (2012) 09-1421-04

基于 FCM 与神经网络的案例推理方法

韩 敏, 沈力华

(大连理工大学 电子信息与电气工程学部，辽宁 大连 116023)

摘 要: 目前有关案例推理 (CBR) 的研究主要集中在案例检索方面, 对案例库构造方法的研究则较为少见, 而好的案例库, 既可以提高案例检索效率, 又可以保证较好的检索准确率. 鉴于此, 针对 CBR 中的案例库进行研究, 引入模糊 C 均值方法去除原案例库中的冗余案例, 从而实现对神经网络-案例推理方法的改进. 最后通过对 UCI 数据进行的仿真实验表明了改进后的案例推理方法无论在案例检索精度还是在案例检索速度上均有所提高.

关键词: 案例库；模糊 C 均值；神经网络；案例推理

中图分类号: TP18 **文献标志码:** A

Case-based reasoning based on FCM and neural network

HAN Min, *SHEN Li-hua*

(Faculty of Electronic Information and Electrical Engineering，Dalian University of Technology，Dalian 116023, China．Correspondent：HAN Min，E-mail：minhan@dlut.edu.cn)

Abstract: Recently, most researches about case-based reasoning are focused on case retrieval, while little attention is paid for the method of construting a proper case base. Better case base can improve both the efficiency and accuracy of case retrieval. Therefore, the case base is constructed in this paper. By removing redundant cases by using FCM(fuzzy C-means), neural network-CBR(NN-CBR) is improved. The simulation results conducted with UCI data show the improvement of the proposed method in both accuracy and efficiency.

Key words: case base；fuzzy C-means algorithm；neural network；case-based reasoning

1 引 言

案例推理 (CBR) 方法是指将需要解决的案例与历史案例相对比, 找到与当前案例最相似的案例, 从而利用相似的历史案例的解决方案来解决当前案例[1]. 案例推理涉及到以下问题: 首先将先前的问题及其解决方案存入案例库; 然后通过在案例库中找到已有的知识来解决当前问题. 目前, 有关案例推理的研究主要集中在第 2 部分, 即案例检索过程的研究[2]. 在此过程中, 属性权重尤为重要, 许多研究都通过利用合理的方法确定属性权重来提高案例检索的精度. 文献 [3] 利用粗糙集确定属性权重, [4-5] 利用遗传算法确定属性权重, [6-7] 利用神经网络确定属性权重. 然而, 上述方法仅在属性权重方面进行了改进, 好的案例库不但可以提高案例检索速度, 而且可以保证较好的检索精度, 同时可以节省一定的案例存储空间.

本文针对上述问题进行研究, 利用模糊 C 均值 (FCM) 方法去除案例库中的冗余案例, 从而使神经网络在训练网络权值和案例检索时能够节省一定的时间. 案例库的存储也可以节省一定的空间, 同时能够保证较好的案例检索精度, 从而实现对已有神经网络-案例推理 (NN-CBR) 方法的改进.

2 神经网络确定属性权重的两种方法

相对于遗传算法与粗糙集方法确定属性权重, 神经网络确定属性权重不需要进行编码解码和属性离散化, 而且确定属性权重的方法也较多[8], 目前较为方便有效的方法是根据网络权值确定属性权值[7].

1) 活跃性. 节点的活跃性通过计算其在训练数据中的活跃水平方差得到, 第 j 个隐层节点的活跃性为

$$A_j = (w_j^{(2)})^2 \mathrm{var}\Big(g\Big(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\Big)\Big). \tag{1}$$

其中: w_j 为隐层到输出层的链接权值, var( ) 为方差函数, $g(\,)$ 为激活函数, w_{ij} 为输入层到隐层节点的连接权值, x_i 为案例 x 的第 i 个属性. 第 i 个输入节点的活

收稿日期: 2010-12-16；**修回日期:** 2011-03-14.

基金项目: 国家自然科学基金项目(61074096)；国家 863 计划项目(2007AA04Z158).

作者简介: 韩敏(1959−), 女, 教授, 博士生导师, 从事复杂工业系统建模与控制、智能技术及优化算法等研究；沈力华(1984−), 女, 硕士生, 从事案例推理相关算法的研究.

跃性为

$$A_i = \sum_{j=1}^{M} ((w_{ji}^{(1)})^2 A_j). \tag{2}$$

2) 显著性. 显著性通过估算误差相对于权重的二次导数进行计算, 权重的显著性值正比于权重的平方值, 因此采用它作为输入节点的重要性. 第 i 个输入节点的显著性为

$$Sa_i = \sum_{j=1}^{M} ((w_{ji}^{(1)})^2 (w_j^{(2)})^2). \tag{3}$$

其中: M 为隐层节点数, 其他变量含义与式 (1) 相同.

在案例推理过程中, 首先通过神经网络训练得到输入层到隐层以及隐层到输出层的网络连接权值, 再利用这两种方法得到描述案例的各个属性的权值. 精简的案例库不但可以节省上述方法训练网络权值的时间和准确率, 同时可以节省案例存储空间, 因此利用 FCM 方法对案例库进行相应处理, 可以实现对上述两种方法的改进.

3 由 FCM 改进的神经网络-案例推理方法

3.1 FCM 聚类算法

FCM 聚类算法是基于目标函数的聚类算法, 目标函数[9]为

$$J_{\mathrm{m}} = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (d_{ik})^2. \tag{4}$$

其中: c 为所有类别数目, m 为平滑参数, n 为案例库中案例的数目, u_{ik} 为第 k 个案例相对于第 i 个类别的隶属度, d_{ik} 为第 k 个案例和第 i 个类别的典型案例间的距离. 最终目标是使得目标函数达到最小值, 并将此时的聚类结果作为最终结果.

FCM 算法具体步骤如下.

Step 1: 直接确定出聚类类别数目为案例库中案例的类别数目, 计算或更新隶属度矩阵 $U^{(b)}$, 从而得到新的划分矩阵

$$u_{ik}^{(b+1)} = 1 \Big/ \sum_{j=1}^{c} \left(d_{ik}^{(b)} / d_{jk}^{(b)} \right)^{\frac{2}{m-1}}. \tag{5}$$

利用式 (5) 可得到新的划分矩阵中的各个元素. 假设存在 i 和 k 使得 $d_{ik}^{(b)} = 0$, 则 $u_{ik}^{(b+1)} = 1$. 能够证明每个案例对应于每一类别的隶属度总和是 1.

Step 2: 新的聚类原型模式矩阵为

$$p_i^{(b+1)} = \sum_{k=1}^{n} (u_{ik}^{(b+1)})^m x_k \Big/ \sum_{k=1}^{n} (u_{ik}^{(b+1)})^m, \tag{6}$$

其中 x_k 为第 k 个案例的案例表述. 利用式 (6) 可得到矩阵中各分类的新聚类中心.

Step 3: 若满足停止条件则停止算法, 同时输出划分矩阵和聚类原型 P; 否则, 设 $b = b+1$, 转至 Step 2 继续进行.

3.2 FCM-神经网络模型

FCM 与神经网络相结合不但具有神经网络的学习能力, 而且具有 FCM 提取有效数据的能力. 通过 FCM 将冗余案例去除, 得到精简的案例库, 有利于案例存储和案例检索过程中速度的提高.

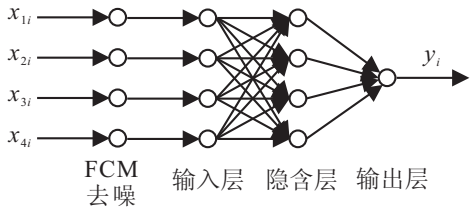

图 1 FCM-神经网络模型

神经网络为 4 层结构, 去噪层和输入节点个数与案例特征维数一致, 隐层节点个数根据经验设置为输入节点个数的 2 倍. 首先, 将各案例利用 FCM 方法进行检验, 验证其是否可以作为确定各属性权值的训练数据, 即:

1) 利用 FCM 方法对未经处理的案例库进行聚类处理, 聚类的最佳类别数直接设置为原案例库包含案例的类别数;

2) 得到存储各类别案例索引的元胞数组, 通过查看各单元中大部分案例所属类来确定该单元所属类;

3) 确定各单元所属类别后与原案例库对比, 找到各单元中聚类错误的案例索引;

4) 将错误案例从原有案例库中去除, 得到新的案例库.

然后, 将新的案例库作为神经网络的训练样本对网络进行训练, 激活函数采用 sigmoid 函数的形式, 即

$$g(a) = 1/(1 + \mathrm{e}^{-a}). \tag{7}$$

设输入层为 d 个神经元, 即案例有 d 个属性表示, 隐含层为 M 个神经元, 则输出值为

$$y = g\Big(\sum_{j=0}^{M} w_j^{(2)} g\Big(\sum_{i=0}^{d} w_{ji}^{(1)} x_i \Big) \Big). \tag{8}$$

最后, 按照神经网络确定权值的方法得到表述案例的各属性权值.

3.3 基于 FCM-神经网络的案例推理模型

利用上述 FCM-神经网络模型得到表示案例的各个属性的权值, 再利用距离公式得到最相似案例为

$$d_{ij} = (x_i - x_j)^{\mathrm{T}} W (x_i - x_j). \tag{9}$$

其中: d_{ij} 为第 i 个案例与第 j 个案例的距离; x_i 和 x_j 两个列向量分别为第 i 个案例和第 j 个案例; W 为对角型矩阵, 对角线上各元素为上述得到的各属性权值.

基于 FCM-神经网络的案例推理方法流程如图 2 所示, 具体步骤如下:

Step 1: 搜集已有的历史案例及其解决方案, 找到可以表述案例的各个属性, 建立原始案例库.

Step 2: 利用FCM方法对原始案例库进行去噪处理, 得到新的精简的案例库, 同时获得输入到神经网络中的训练样本.

Step 3: 利用距离公式 (9) 检索得到最相似案例及其相应的解决方案, 在分类问题中, 解决方案即为案例所属类别.

Step 4: 将得到的检索结果与专家经验相结合得到最终解决方案, 并形成新的案例.

Step 5: 专家对新案例进行评价, 如果案例解决成功, 同时相似度符合要求, 则将新案例存储到案例库中; 否则, 放弃该案例.

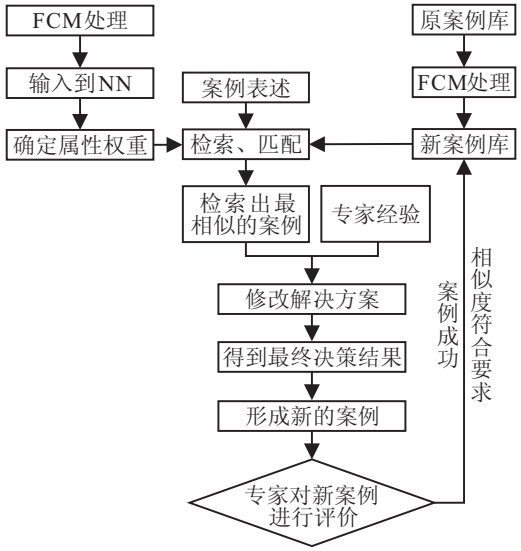

图 2　基于FCM-神经网络的案例推理流程

4　仿真实例

利用UCI中breasts数据和Iris数据进行仿真实验, 以验证本文方法的有效性. 实验过程如下:

1) 将输入、输出数据变换到 [0,1] 之间, 即

$$\overline{x_i} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \tag{10}$$

其中: x_i 为需要预处理的数据, $x_{\min}$ 为数据变换范围的最小值, $x_{\max}$ 为数据变换范围的最大值.

2) 按照 1:1 分配案例库中的案例和测试案例, 利用FCM方法对案例库进行预处理. FCM提取前后案例库中的案例数如表 1 所示.

表 1　FCM提取前后案例库中案例数

案例库	Breasts 数据	Iris 数据
提取前	639	88
提取后	518	73

从表 1 可以看出, FCM方法将聚类错误的案例从案例库中去除, 减小了案例库规模, 从而使得之后的案例检索效率得到提高.

3) 利用神经网络确定属性权值. 设输入节点为 n, 则输出节点设为 $2n-1$, 最大迭代次数设为 2 000, 训练误差设为 0.001. 根据式 (2) 和 (3) 得到属性权值 W. 以breasts为例, 经FCM处理前后, 案例库中案例的训练误差随迭代次数的收敛曲线如图 3 和图 4 所示.

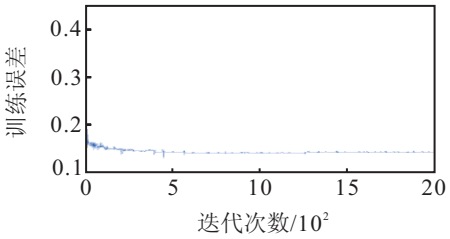

图 3　未经FCM提取的训练误差变化曲线

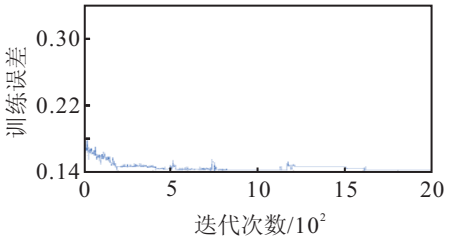

图 4　FCM提取后的训练误差变化曲线

从图 3 和图 4 可以看出, 经FCM提取后的案例库中的样本具有更小的训练误差, 同时由于案例库中的数据量减少, 训练速度会更快.

4) 在案例调整过程中采用下式得到待测案例 x_q 的最终所属类别[9]:

$$C(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{K} w_i \delta(v, f(x_i)). \tag{11}$$

其中: $w_i \equiv 1/d(x_q, x_i)^2$, V 为类别的有限集合, $f(x_i)$ 为案例 x_i 的所属类别. 当 $x_i = v$ 时, $\delta = 1$; 否则, $\delta = 0$.

5) 根据CBR检索原理计算案例检索的准确率为

$$\text{accu} = \frac{\sum_{i=1}^{n} m_i}{n} \times 100\%. \tag{12}$$

其中: n 为待分类的案例总数; 当第 i 个待分类案例检索正确时, m_i 取值为 1, 否则为 0.

根据式 (12) 得到方法案例检索准确率, 将其与传统CBR方法 (记为C-CBR) 和两种神经网络案例推理方法 (NN-CBR$_1$, NN-CBR$_2$) 进行比较, 比较结果如表 2 和图 5 所示.

表 2　Breasts数据各方法准确率比较结果

方 法	K				
	1	3	5	7	9
C-CBR	0.921 3	0.914 7	0.932 6	0.938 7	0.897 5
NN-CBR$_1$	0.963 1	0.954 8	0.984 2	0.945 2	0.961 7
NN-CBR$_2$	0.971 7	0.941 3	0.951 3	0.974 2	0.971 8
本文方法	0.986 2	0.972 5	0.955 2	0.968 4	0.973 9

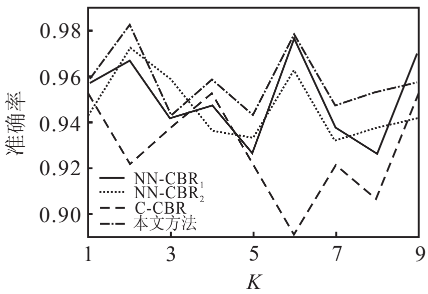

图 5 Iris 数据各方法比较结果

由表 2 和图 5 可以看出, 由于引进了 FCM 方法对案例库进行处理, 去除了冗余案例, 得到了相对较好的案例库, 从而使得案例检索精度得到了一定的提高.

表 3 给出了 4 种算法处理时间的对比. C-CBR 方法由于是人工指定权值, 处理时间最短; NN-CBR_1 和 NN-CBR_2 的时间消耗主要集中在神经网络计算权值的过程中; 本文所提出的方法虽然有较多的计算环节, 但由于对案例库进行了去噪, 降低了权值计算负担, 速度较 NN-CBR 方法有明显提升.

表 3 Breasts 数据各方法速度比较结果

方 法	C-CBR	NN-CBR_1	NN-CBR_2	本文方法
时间消耗/s	0.28	63.15	62.82	50.78

5 结 论

本文对案例推理中的案例库进行研究, 利用模糊 C 均值算法对案例库进行去噪处理, 使得案例存储节省了一定空间, 同时得到了更有效的神经网络训练样本, 使得网络的训练更加有效和充分. 案例库中的案例总数减少, 但有效案例数不变, 因此, 节省了案例检索的时间, 而且保证了一定的检索精度. 实验结果表明了所提出方法的真实有效性.

参考文献(References)

[1] Castro J L, Navarro M, Sánchez J M, et al. Loss and gain functions for CBR retrieval[J]. Information Sciences, 2009, 179(11): 1738-1750.

[2] Wang Hyuk Im, Sang Chan Park. Case-based reasoning and neural network based expert system for personalization[J]. Expert Systems with Applications, 2007, 32(1): 77-85.

[3] 韩敏, 张俊杰, 彭飞, 等. 一种基于多决策类的贝叶斯粗糙集模型[J]. 控制与决策, 2009, 24(11): 1615-1619.
(Han M, Zhang J J, Peng F, et al. Bayesian rough set model based on multiple decision classes[J]. Control and Decision, 2009, 24(11): 1615-1619.)

[4] Hyunchul Ahn, Kim Kyoung-jae, Ingoo Han. Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system[J]. Expert Systems, 2006, 23(5): 290-301.

[5] Hyunchul Ahna, Kim Kyoung-jae. Global optimization of case-based reasoning for breast cytology diagnosis[J]. Expert Systems with Applications, 2009, 36(1): 724-734.

[6] Shin Chung-kwan, Ui Tak Yun, Huy Kang Kim, et al. A hybrid approach of neural network and memory-based learning to data mining[J]. IEEE Trans on Neural Networks, 2000, 11(3): 637-646.

[7] Li Hui, Sun Jie, Sun Bo-liang. Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors[J]. Expert Systems with Applications, 2009, 36(1): 643-659.

[8] Pal S K, De R K, Basak J, Unsupervised feature evaluation: A neuro-fuzzy approach[J]. IEEE Trans on Neural Networks, 2000,11(2): 366-376.

[9] 高新波. 模糊聚类分析及其应用[M]. 西安: 西安电子科技大学出版社, 2004.
(Gao X B. Fuzzy cluster analysis and its applications[M]. Xian: Xidian University Press, 2004.)

下 期 要 目