基于 RSOM 树模型的机器学习原理与算法研究

夏胜平, 张乐锋, 虞 华, 张 静, 胡卫东, 郁文贤
(国防科学技术大学电子科学与工程学院 ATR 重点实验室, 湖南长沙 410073)

摘 要: 机器学习和识别可归结于一个高速、有效地搜索非常大的样本空间问题, 以实现对训练和识别样本的最佳拟合. 对于复杂背景的模式样本集, 同类型样本的独立同分布(i. i. d)特性通常难以保证, 统计理论无法有效应用. 本文将层次化思想和自组织映射(SOM)神经网络相结合, 采用递归实现技术实现了一种高效、高容量, 能够自适应增长的模式分类树(RSOM 树)生长方法, 用于模式识别和机器学习的基本建模. 通过对大量公用数据集的测试以及在实际的雷达目标识别系统中应用, 方法有效性得到了证明.

关键词: 模式识别; 分类树; 神经网络; SOM; RSOM; 机器学习
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2005) 05-0939-06

Theory and Algorithm of Machine Learning Based on RSOM Tree Model

XIA Sheng-ping, ZHANG Le-feng, YU Hua, ZHANG Jing, HU Wei-dong, YU Wen-xian
(*State Lab of Automatic Target Recognition, National University of Defense Technology, Changsha, Hunan 410073, China*)

Abstract: One useful perspective on machine learning and target recognizing is that they involve effectively and rapidly searching some very large sample space to determine one that best fits the observed data and trained samples. If the independent identical distribution (i. i. d) property of pattern samples of complexity background is not satisfied, statistic theory is of no effect. In this paper, a method on producing RSOM pattern recognition tree is proposed, in which the hierarchy model, SOM neural net algorithm and recursive technique are utilized. This is a basic model of pattern recognition in machine learning. The validity of this method is proven by testing lots of common data sets and practical application in radar target recognition system.

Key words: pattern recognition; classification tree; neural net; SOM; RSOM; machine learning

1 引言

独立同分布特性(i. i. d)是文献[1]理论框架的基本架设, 然而, 在复杂的模式识别和机器学习问题中, 很多情况下, 对同一类目标, 其模式样本来源于多种子模式, 如印刷体字符光学图像识别问题, 由于字体不同, 即使是同一个字, 也会有多种不同的表现, 如宋体、楷体、魏体、草书等等, 同一种字体的样本可以认为具有 i. i. d 特性, 但不同字体样本之间差异显著, 应分别看作几种不同的子模式, 样本之间的 i. i. d 特性不能保证, 这些子模式的样本对该类目标而言, 并非来源于同一个母体. 对这类问题, 核心在于能否找到相应的子模式, 从而可有效地应用统计理论.

另外, 机器学习和自动识别可以归结于一个搜索问题[2], 学习器或者说识别器通过搜索参与训练的样本构成的空间, 得到能够最佳拟合训练样本和识别样本的样本子集, 由此进行泛化, 因而, 算法能否高速、有效地搜索非常大的或无限大的样本空间, 以实现对学习和识别样本的最佳拟合是机器学习算法研究最核心的问题之一.

再者, 人脑学习是一个有师指导和无师指导相结合的、循序渐进的、自治的学习过程. 对一个学习系统而言, 如果该系统不能够自治地进行学习, 也就是不具备发现知识并学习之的能力, 这样的系统就不能说具有真正的"智能".

考虑到 SOM 网络作为一种无导师示范、具有自组织功能的神经网络[3,4], 通过对输入模式的反复学习, 可以使连接权矢量的空间分布密度与输入模式的概率密度趋于一致, 即连接权矢量的空间分布能反映输入模式的统计特征. 它在已知特征空间的样本分布或大致分布的情况下可以取得很好的自组织效果. 不过, 在许多实际的分类/识别问题中, 如前述的字符识别问题, 以及雷达目标识别问题, 样本空间的维数很高, 样本集概率分布情况非常复杂, 这时直接使用 SOM 模型往往无法取得很好的效果.

基于上述的思考, 本文作者于 2001 年初建立了由基本的 SOM 网络作为节点构成的 SOM 树模型, 由于算法实现采用了递归的 SOM 网络树生长方法, 由此得到 SOM 网络树, 记为 RSOM 树. 并在 Matlab 和 VC++ 环境下设计实现了两种版本的基于 RSOM 的智能机器学习与模式识别系统, 用[5]中的大量

收稿日期: 2004-02-09; 修回日期: 2004-12-20
基金项目: 国家 863 计划(No. 2003AA134030); 国防科技预研项目(No. Y018-6/Q200210)

数据集和常规的双螺旋线模式识别问题[6]进行了测试, 取得了良好的效果. 从 2003 年 5 月至今, RSOM 方法应用于多种型号、多部雷达的舰船目标自动识别系统中, 表现出良好的自学习能力、对不同场景识别问题的适应能力和优良的识别性能.

文献资料表明, 与之类似的研究有文献[7, 8]、文献[9]. 文献[7, 8]用熵减的方式来控制网络的逐层生长, 将每一类的样本当作同一母体的样本来对待, 利用输入样本来计算先验概率和条件概率, 并且由此计算相应的熵函数, 这一点对其所述的高分辨雷达一维距离像目标识别问题, 样本之间的 i. i. d 特性是不能满足的, 从而也无法对样本的先验分布进行有效的估计, 并计算其条件概率熵用于网络生长的控制. 事实上, 该文中已经遇到了由此带来的问题, 文中阐述到, 运动目标的一维距离像是很不稳定的, 它会随着目标视角的变化而变化, 使得相同的目标在不同视角下有不同的一维距离像, 而不同目标在不同视角下有可能有相似的甚至相同的距离像, 这就决定了识别雷达一维距离像在高维模式空间识别大样本集和具有复杂定义边界的问题, 从概率的角度来解释, 该问题的焦点之一就在于如何对待这些样本的非 i. i. d 性质; 文献[9]基于将模式样本与相应叶节点模式中心的均方误差之和控制到可以接受的误差范围来控制网络的生长, 极限情况是每个叶节点中只有一个样本, 从而使得, 所训练得到的网络的均方误差为 0, 这显然是一种过拟合的情况, 难以保证网络的范化能力. 在识别的过程中, 上述两种算法都是针对叶节点中的样本类型属性数目进行简单的表决, 因而一旦训练完毕, 相应叶节点的类型属性是固定的. 二者基于 SOM 的无导师聚类的思想, 基本停留在 SOM 树的简单生成方面, 对方法的基础性、普适性和巨大潜能尚无深刻认识, 其生长方法、识别方法均与本文方法有很大差异.

本文算法的基本思想是, 首先对所有带有类别属性标识的原始训练样本用一个 SOM 网络进行训练, 得到一组输出节点, 之后按最近邻原则将所有原始训练样本分配到相应的节点, 由此形成一个模式分类树的根节点. 考察根节点所属输出节点, 对分配到其中的样本进行可分性判决条件的检测, 若不可分, 则将该节点属性赋为叶节点, 停止该节点的分解. 此时, 叶节点中同类目标的样本数据可以近似认为具有 i. i. d 特性; 若可分, 则用与根节点 SOM 网络训练完全相同的算法对该节点进行训练, 得到相应的 SOM 网络, 并将该节点的样本分配到相应的输出节点, 由此, 通过采用递归的方法对所有节点进行类似的分析, 直到没有节点需要进一步生长为止. 这样就得到了一棵 RSOM 树. 在 RSOM 树的生长过程中, 有多种控制因子, 包括类内、类间可分性判拒、层数控制、样本数控制以及对训练样本或测试样本的识别正确率控制等因子, 以保证所训练得到的 RSOM 树具有优良的结构和对待识别样本的优良的范化能力. 之后, 对需要进行识别的样本采用动态加权投票的方式或其它方式进行识别处理.

本文在第二节对基本的 SOM 原理与算法进行了必要的阐述, 第三节给出了 RSOM 算法, 第四节给出了算法的应用实例.

2 基本的 SOM 原理与算法及其改进

SOM 网络是一种无导师示范、具有自组织功能的神经网络, 并具有保持拓扑的特点, 一般取为平面网格结构[9], 如图 1 所示, 由输入层和竞争层组成, 输入层神经元数为 n, 对应于输入模式矢量维数, 竞争层由 $M = m \times l$ 个神经元组成, 且构成一个二维平面阵列. 在竞争层中, 神经元的竞争是这样进行的: 对于获胜的神经元 g, 在其周围 N_g 的区域内, 神经元在不同程度上都得到兴奋, 而在 N_g 以外的神经元都被抑制, 形成墨西哥帽形状的兴奋刺激. 这个 N_g 区域可以是任何形状, 但一般是均匀对称的, 例如长方形或六角形. N_g 是时间的函数, 用 $N_g(t)$ 表示, 随着 t 增加, $N_g(t)$ 减少, 最后达到预定的范围. 以下称 $N_g(t)$ 为邻域, 表示邻域中所包含的神经元个数.

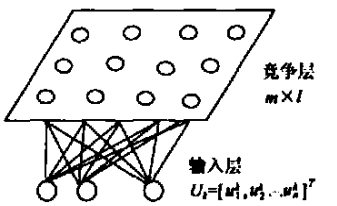

图 1 SOM 网络模型结构

设网络的输入模式为 n 维矢量 $\boldsymbol{U}_k = [u_1^k, u_2^2, \cdots, u_n^k]^T$, $k = 1, \cdots, p$, 共有 p 个模式矢量, 构成训练样本集 U. 竞争层神经元 j 的输出响应为 $v_j, j = 1, \cdots, M$. 竞争层神经元 j 与输入层神经元之间连接权矢量为 $\boldsymbol{W}_j = [\omega_{j1}, \omega_{j2}, \cdots, \omega_{jn}]^T, j = 1, 2, \cdots, M$.

在进行网络训练之前, 对所有输入矢量 $\boldsymbol{U}_k$, $k = 1, \cdots, p$ 按矢量维进行归一化. 本文采用了稳健统计的方法, 可有效抑制各维数据量级上的差异以及可能出现的奇异值对 SOM 网络训练造成的不良影响. 归一化方法为将输入矢量每维对应的全部数据的 ε 和 1- ε 分位点当作该维数据的极小、极大值(不妨设 ε 为 0.05), 对所有数据按维进行极差变换, 得到归一化的样本数据. 在此基础上, 采用两个矢量欧式距离最小的准则来确定获胜神经元, 即:

$$d_j = \| U_k - W_k \|_2 \tag{1}$$

$$d_g = \min(d_j), j = 1, 2, \cdots, M \tag{2}$$

网络在自组织学习过程中, 其连结权需要根据矢量 U_k 进行调整. 连结权的调整采用如下方程:

$$W_j(t+1) = \begin{cases} \boldsymbol{W}_j(t) + \eta(t)[\boldsymbol{U}_k - \boldsymbol{W}_j(t)], & j \in N_g(t) \\ \boldsymbol{W}_j(t), & j \notin N_g(t) \end{cases} \tag{3}$$

综上, 改进后的 SOM 网络训练步骤归纳如下:

(1)初始化. 将 w_{ji} 在[0, 1]区间取随机值; 确定学习速率 $\eta(t)$ 得初值 $\eta(0)$ ($0 < \eta(0) < 1$); 确定邻域 $N_g(t)$ 的初值 $N_g(0)$ 及总学习次数 T;

(2)对全部输入模式进行极差变换归一化:

$$u_s^k = \frac{u_s^k - Z_s}{Z_{1-\varepsilon} - Z_\varepsilon}, \quad s = 1, 2, \cdots, n \tag{4}$$

其中, $Z_{1-\varepsilon}^{\varepsilon}$、$Z_\varepsilon^s$ 分别为输入矢量第 s 维全部数据的 1- ε 和 ε 分位点;

(3)将 p 个训练样本随机乱序排列, 然后依次给网络提供输入模式 $U_k = [u_1^k, u_2^k, \cdots, u_n^k]^T$, 计算连结权矢量 $\overline{W}_j = [\overline{w_{j1}},$

$\bar{w}_{j2}, \cdots, \bar{w}_{jn}]^T$ 与输入矢量 U_k 之间的欧氏距离:

$$d_j = \| U_k - \bar{W}_j \|_2, \quad j = 1, 2, \cdots, M \tag{5}$$

(4) 找出最小距离 d_g, 确定获胜神经元 g:

$$d_g = \min(d_j), \quad j = 1, 2, \cdots, M \tag{6}$$

(5) 进行连结权的调整. 对竞争层邻域 $N_g(t)$ 内所有神经元与输入层神经元之间连接权进行修正

$$w_{ji}(t+1) = \begin{cases} \bar{w}_{ji}(t) + \eta(t)[\bar{u}_i^k - \bar{w}_{ji}(t)], & j \in N_g(t) \\ \bar{w}_{ji}(t), & j \notin N_g(t) \end{cases} \quad i = 1, 2, \cdots, n \tag{7}$$

(6) 将下一个输入模式提供给网络的输入层, 返回步骤 (3), 直到 p 个模式全部提供一遍. 称之为进行了一个批次的训练.

(7) 更新学习速率 $\eta(t)$ 及邻域 $N_g(t)$

$$\eta(t) = \eta_0(1 - \frac{t}{T}) \tag{8}$$

$$N_g(t) = \mathrm{INT}[N_g(0)\exp(-\frac{t}{T})] \tag{9}$$

式中, η_0 为学习速率初值; t 为学习次数; T 为总学习批次; $\mathrm{INT}[x]$ 表示对 x 取整.

(8) 令 $t = t + 1$, 返回步骤(3), 直到 $t = T$ 为止, 得到一个对训练集 U 进行了 T 批次学习的 SOM 网络, 记作 SOM_Net, 并将归一化参数随同网络一起保存;

(9) 对 U 中的样本用该 SOM 网络用最近邻法则进行分配, 所有样本被分配到 SOM 网络的 M 个输出节点中, 记作 $U^j, j = 1, 2, \cdots, M$, 有:

$$U = \bigcup_{j=1}^{M} U^j \tag{10}$$

其中:

$$U^j = \{ U_k \in U, k \in \{1, 2, \cdots, p\}\}$$

$$U^j \cap U^i = \Phi, i \neq j, i, j \in \{1, 2, \cdots, M\}$$

设 U^j 中分配了 N_j 个样本, 对其中的样本重新标号为:

$$U^j = \left\{ \begin{array}{l} U^j_{k'} = U_k \in U, k \in \{1, 2, \cdots, p\} \\ k' = 1, 2, \cdots, N_j \end{array} \right\}$$

接下来进一步给出 *RSOM* 树训练与识别算法.

3 RSOM 算法

3.1 基本定义

对于常见的模式识别问题, 其训练集除包含输入模式 U 之外, 每个模式样本还包括其相应的类别属性标识, 与 U_k 相对应, 该模式样本的类别属性标识记作 T_k, 得到类别属性标识集 $\boldsymbol{T} = \{T_k, k = 1, \cdots, p\}$, 由此, 一个完整的训练集记作 $\Psi = \{U, \boldsymbol{T}\}$.

设某一样本集 U 有 p 个样本 $\boldsymbol{U}_k = [u_1^k, u_2^k, \cdots, u_n^k]^t$, 分属 c 个模式类. $\omega_i = \{U_k^i, k = 1, 2, \cdots, N_i\}, i = 1, 2, \cdots, c$ 将样本集 U 分成 c 个子集, 则有:

$$U = \bigcup_{i=1}^{c} \omega_i \tag{11}$$

记每个子集 ω_i 表示同一模式类的样本组成的集合; $\bar{m}_i$ 表示 ω_i 模式类的样本均值; $\bar{m}$ 为所有模式样本的均值; P_i 为 ω_i 类在所有样本中所占的频度; S_{wi} 为 ω_i 类的类内离差阵; S_{wi} 为 ω_i 类的类内距离; $\bar{d}^2(\omega_i, \omega_j)$ 为 ω_i 类与 ω_j 类的类间距离; 总的类内离差阵为 S_w; 以上定义可参见文献[4, 9], 则类别可分性判据定义为:

$$J = \frac{Tr[S_B]}{Tr[S_w]} \tag{12}$$

可知, J 越大, 样本可分性越好; J 越小, 样本可分性越差. 不妨设定某个阈值 ϑ, 当 J 小于阈值 ϑ 时, 样本集不可分. RSOM 算法用该可分性判据来控制 RSOM 树的生长, ϑ 一般取值为 0.1 左右.

3.2 RSOM 算法步骤

考虑到一棵模式分类树不可能无限制地生长, 并且在进行 SOM 网络训练时应有一定的样本数目的要求; 另外, 如果一个节点中某一个模式类的样本频度 P_i 占绝对优势, 如接近或等于 1.0, 有理由不再对该节点进行训练, 称这样的节点为确定性节点. 因此, 训练时除前述的可分性判据外, 另外附加最大层数限制、可分节点最小样本数限制以及确定性节点判定三个条件对 RSOM 树生长进行控制(见算法 S4 中 a), b), c)), 基本的 RSOM 训练算法归纳如下:

(S1) 初始化. 选择网络结本结构 m, l; 确定 RSOM 树生长最大层数 L; 确定可进行 SOM 训练的节点中样本的最小个数 $p_{\min}$; 确定可分性判决阈值 ϑ; 给定确定性节点样本频度阈值 P_T; 选定训练集 $\Psi = \{U, \boldsymbol{T}\}$;

(S2) 用 2.1 节中的 SOM 训练算法进行第一次训练, 得到 RSOM 树的根节点, 将其置入 RSOM 树中;

(S3) 将根节点中训练得到的 M 个节点组成一个待训练节点集合 $\{U^j\}$;

(S4) 若待训练节点集 $\{U^j\}$ 为空, 转(S8); 按照树的左序遍历算法选取 $\{U^j\}$ 中的待训练节点进行检测, 考察是否需要进行 *SOM* 网络训练, 包括四方面的检测, 如下:

(a) 层数限制: 考察当前节点所在层数 Layer, 若 Layer = L, 转(S7);

(b) 最小样本数限值: 考察 U^j 中样本数量是否大于 $p_{\min}$, 若小于, 转(S7);

(c) 确定性节点限值: 计算当前节点中各类样本的 P_i, 若存在 $P_i \geqslant P_T$, 则转(S7);

(d) 可分性判据: 计算类别可分性判据指标 J, 若 $J \leqslant \vartheta$, 则转(S7);

(S5) 对当前节点采用用 2.1 节中的 SOM 训练算法进行 SOM 网络训练, 将得到的 SOM_Net 添加到 RSOM 树中;

(S6) 将该节点训练得到的节点添加到待训练节点集 $\{U^j\}$ 中, 转(S4);

(S7) 将当前节点属性赋为叶节点, 将其添加到 RSOM 树中, 且将其从节点集 $\{U^j\}$ 中删除;

(S8) 终止训练, 生成训练完毕的 RSOM 树, 如图 2 所示.

至此, 训练得到了一棵层次化的 RSOM 树, 然而训练的根本目的还是在于能够对模式样本进行分类识别. 设有待识别模式样本 U_t, 输入 RSOM 树, 要求通过 RSOM 树确定该样本的模式类别属性, 并以类别属性隶属度向量 $\mathscr{R} = [r_1, r_2, \cdots, r_c]^t$ 的形式给出.

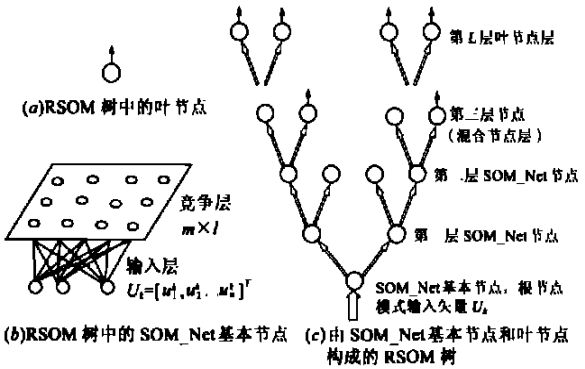

图 2 RSOM 树结构

下面, 给出基本的识别算法:

(S1)初始化. 给定参与泛化识别的样本数 K.

(S2)将根节点 SOM 网络当作获胜节点网络 SOM_Net;

(S3)用获胜节点中 SOM_Net 保存的归一化参数对 U_t 归一化为 $\overline{U}_t$;

(S4)将 $\overline{U}_t$ 输入当前 SOM_Net, 采用 SOM 网络获胜节点求解方法求得相应的获胜节点;

(S5)若获胜节点为叶节点, 转 S7;

(S6)转(S3);

(S7)设当前叶节点中包含 p_{lf} 个样本, 对当前样本进行如下处理:

(S7.1) 若叶节点为确定性节点, 则当前样本被识别为该叶节点中所含样本的类属性 ω_i, 则 $\mathscr{R}=[r_1, r_2, \cdots, r_c]^T=[0.0, 0.0, \cdots, 1.0, \cdots, 0]^T$, $\mathscr{R}$中除 ω_i 对应的 r_i 为 1.0 外, 其余均为 0.0. 转(S7.4);

(S7.2) 求当前样本与叶节点中样本归一化后的欧氏距离:

$$d_k=\|\overline{U}_t-\overline{U}_k\|,\quad k=1, 2, \cdots, p_{lf} \tag{25}$$

(S7.3) 对 p_{lf} 个距离按升序排列, 选定 K 个最小的距离, 及其样本属性 ω_i;

(S7.4) 对类属性值以距离倒数(为防止 d_k 为 0, 不妨以 $1.0/(d_k+\lambda)$加权投票(其中 $\lambda\in(0, 1)$), 并对权值向量进行归一化处理, 得到向量 $\mathscr{R}$;

(S7.5) 结束叶节点数据处理, 输出 $\mathscr{R}$, 结束叶节点处理;

(S8) 输出 $\mathscr{R}$, 并根据 $\mathscr{R}$中元素按 winner take all 原则确定当前样本所属目标类别, 作出判决.

(S9)结束对一个样本的识别.

综上, 得到了 RSOM 方法的基本算法. 下面从 RSOM 容量、训练和识别效率三方面对本文的方法进行定性分析. 由训练算法可知, 根节点的 SOM 网络的训练是整个 RSOM 树训练的瓶颈, 只要根节点 SOM 网络训练能够顺利完成, 其它节点训练所需消耗的计算机资源将远小于根节点所消耗的资源, 特别随着生长层数的增加, SOM 网络训练的负担将按指数级衰减. 而在一定训练样本数条件下, SOM 网络的个数将有限增长, 不妨设 SOM_Net 中网络基本结构为 3×3, 若生成一个 5 层的网络, 所需训练的 SOM 网络最大个数为 $1+9+9^2+9^3+9^4=7381$, 此时叶节点数最大可以达到 66429 个, 设平均每个节点中包含 100 个样本, 则整棵 RSOM 树样本容量为 6642900, 达 6 百多万, 增大网络基本结构, 则网络容量将进一步快速增加, 如网络结构变为 2×5, 其它条件相同的情况下, 网络容量可增加到 10^7, 达到千万量级.

通过对 47 维输入矢量 10 万个训练样本集的情况做过测试, 用 P4 1.8GHz, 内存 1G Byte 的计算机进行训练可在几秒钟内完成一棵 RSOM 树的训练, 而对一个样本进行识别时, 以最大 5 层的 RSOM 树为例, 只需进行 5 次求网络获胜节点的处理, 以及含 100 个样本的叶节点处理, 因此识别实时性容易保证, 表明 RSOM 方法具有很强的可实现性.

另外, 通过对 RSOM 的规模进行限定并对超过样本容量的叶节点进行另外一棵 RSOM 树的级联生长, 对巨型样本的情况, 可方便地生成多个 RSOM 树级联的网络群. 这种 RSOM 树或 RSOM 群落模型是开放式的系统, 其输出端对应的目标类型数目可以充分扩展, 如印刷体汉字识别, 每个字算一个模式类, 而常用的汉字以 5000 以上计, 每个汉字还有多种字体, 如常见的宋体、楷体、魏体、隶书等, 子模式种类达 20,000 以上, 对此, RSOM 方法具有很强的适应能力.

4 试验分析

4.1 基于 IRIS 数据进行的试验

IRIS 数据是 Fisher, R. A. 于 1936 年给出的一组较早用于识别领域算法测试的一种标准试验数据[11]. 该数据包括 3 种类型共 150 个样本, 每种类型分别含有 50 个样本数据, 每个样本由 4 维特征组成. 这四维数据的混叠程度不一, 不同组合下的聚类性能也不同, 图 3 给出了 *IRIS* 数据在部分特征组合下的示例. 其中"+"、"×"、"•"分别代表第一、二、三类目标; 图 3*a*、3*b* 中的横坐标 nIndex 代表数据序号, 纵坐标表示该维特征的取值; 图 3*c*、3*d* 中的横纵坐标则分别表示相应维特征的取值.

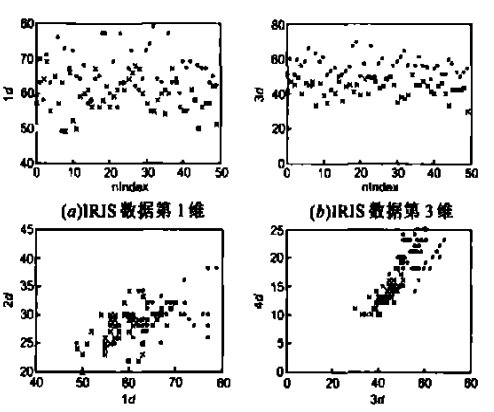

图 3 部分 IRIS 数据及其组合示例

针对 IRIS 数据的基于 4 种单维特征、6 种二维特征组合、4 种三维特征组合以及全部四维特征的 15 个训练样本集合, 运用 RSOM 方法训练产生了 15 棵 RSOM 树. 表 1 给出了上述 15 棵分类树对所有训练样本的识别结果.

可知, 用 1、2、3 维特征构成输入模式向量得到的训练集

可以得到最优的识别性能, 平均正确识别率可达到 96%.

表 1　IRIS 数据生成不同特征组合 RSOM 树的识别结果

训练样本 \ 训练结果	最大层数	叶节点数	平均正确率(%)			
			Ⅰ类	Ⅱ类	Ⅲ类	综合
1d	2	10	72	4	30	35.3
2d	2	9	44	0	0	29.3
3d	3	11	96	44	82	74
4d	3	11	100	90	92	94
1- 2d	2	10	98	0	0	32.7
1- 3d	3	15	98	84	86	89.3
1- 4d	4	17	100	88	86	91.3
2- 3d	3	18	100	84	84	89.3
2- 4d	4	21	98	86	88	90.7
3- 4d	3	13	100	92	90	94
1- 2- 3d	3	17	100	96	92	96
1- 2- 4d	4	17	94	86	90	90
1- 3- 4d	3	15	100	90	90	93.3
2- 3- 4d	3	13	100	86	86	90.7
1- 2- 3- 4d	3	18	100	88	94	94

4.2　算法在某型雷达目标识别系统中的应用

以下实验使用的数据是甲、乙两个雷达站的目标识别试验中对目标视频回波高精度采样后提取回波展宽、肩宽、跳动、扭动、凹口、倾斜度以及傅式变换等时频域特征[12~15]构成的 47 维的模式输入矢量集, 通过采用爬山法或说遗传算法进行进化式特征选择, 总计训练 600 余个 RSOM 树, 得到由 17 维特征训练而成的识别性能最优的 RSOM 树, 识别结果如表 2、表 3 所示.

在甲站, 得到 6 类舰船目标 14711 个样本, 其中 10700 个作为训练集, 4011 个作为测试集, 为保证测试有效性, 训练集在采集时间顺序上先于测试集采集的时间, 识别结果如表 2 所示.

在乙站, 共得到 5 类舰船目标的 11445 个样本, 其中的 8600 个样本作为训练集, 2845 个样本作为测试集, 识别结果如表 3 所示.

表 2　甲站目标识别样本与测试结果

目标类型	训练样本	测试样本	测试样本正确识别率
C1	800	279	0.897
C2	3000	1386	0.853
C3	1500	515	0.873
C4	3000	815	0.858
C5	1500	588	0.911
C6	900	428	0.905
合计	10700	4011	

表 3　乙站目标识别样本与测试结果

目标类型	训练样本	测试样本	测试样本正确识别率
C1	1800	648	0.901
C2	1800	524	0.898
C3	2500	739	0.905
C4	1500	402	0.929
C5	1000	532	0.916
合计	8600	2845	

对于 RSOM 树, 在训练过程中, 如果将层数生长最大值限制为 1, 就退化为基本的 SOM 网络. 对小样本的情况, RSOM 树的结果和单个 SOM 的情况基本是一致的. 对于雷达目标识别大样本的情况, SOM 网络表现出三方面的不足: (1) 由于目标样本分布属性的复杂性, 网络的结构难以确定, 所得到的 SOM 网络其连接权矢量的空间分布不能反映输入模式的统计特征; (2) 随着样本数 N 的增加, 其识别时间相对于 RSOM 树识别时间之比为 $O(N^2)/(100\sim 200)$; (3) 从测试的情况来看, 其平均正确识别率相对 RSOM 方法要低近 10 个百分点, 而且每次训练的得到的 SOM 网络其识别性能还很不稳定一致, 具有一定的随机性.

5　结束语

本文所研究 RSOM 树方法的有效性、可行性在实验室测试和实际运用中都得到了充分的验证. 所建模型在网络自增长、网络的巨容量以及生长的简单可重复性等方面都具有优良的性能.

该方法所建 RSOM 树以及级联的 RSOM 树群落模型与生物体神经网络具有很强的相似性, 对样本集 i.i.d 特性不满足、机器学习和自动识别之高速/有效地搜索非常大或无限大样本空间的问题、有师指导和无师指导相结合的、循序渐进并自主发现知识问题以及充分发挥 SOM 优势避免其弱势等多个方面的问题来说, 是一种有效的基本建模.

基于该模型可以建立巨型容量的、复杂的分布式智能系统. 由于 RSOM 树的结构、功能都是同构的, 可进一步开发 RSOM 树神经网络芯片, 在同一个或多个集成电路板上包括上万个甚至更多的、有限规模的标准 RSOM 树的基本硬件单元, 通过有效的 RSOM 群落管理, 进行 RSOM 树单元的重构, 增强系统的可扩展性, 实现复杂的分布式的、可并行处理的智能系统.

经过近 4 年的研究, 本文作者已经建立了基于此 RSOM 树模型的有效特征选择、增量式学习/更新学习、长时记忆与短时记忆、RSOM 树的裁剪/合并与再生长、RSOM 树群落生长模型以及“积极”学习和“消极”学习相结合的模型, 并建立了具有自主发现新的模式类别, 即发现“新知识”能力的模型, 实现了基于多 CPU 系统的 RSOM 并行训练算法等方面的内容, 建立了具有人脑神经网络结构和进化特征的, 能进行循序渐进式学习、有师指导和无师聚类相结合的学习以及能自治地学习的认知模拟学习模型, 并开发了基本的智能学习系统.

参考文献:

[1] Vladimir N Vapnik. 统计学习的理论与本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000.

[2] Tom M Mitchell. Machine Learning[M]. McGraw-Hill Companies, Inc. 1997.

[3] Simon Haykin. Neural Networks: A Comprehensive Foundation (2nd Edition)[M]. Prentice Hall, Inc, 2001.

[4] Richard O Duda, et al. Pattern Calssification[M]. Second Edition(ISBN: 0-471-05669-3), Jhon Wiely & Sons, Inc, 2001.

[5] Http://www.ics.uci.edu/pub/machine-learning-databas[DB/OL].

[6] 孙功星, 朱科军, 戴长江, 戴贵亮. 层次式多级子网级联神经网络[J]. 电子学报, 1998, 27(8): 49-51.
[7] 涂志江, 刘国岁. 基于熵的自组织神经网络树[J]. 计算机学报, 2000, 23(11): 1126-1129.
[8] 孙光明, 沈兰荪, 刘国岁, 等. 树型级联SOM网络用于雷达目标一维距离像识别[J]. 北京工业大学学报. 1998. 24(4): 17-24.
[9] Rauber A, Merkl D. The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data[J]. IEEE Trans Neural Networks, 2002, 13(6): 1331-1341.
[10] 孙即祥, 等. 现代模式识别[M]. 长沙: 国防科技大学出版社, 2002.
[11] Fisher R A. The use of multiple measurements in taxonomic problems [J]. Annual Eugenics, 1936(7), Part II, 179-188.
[12] 郁文贤. 智能化识别方法及其在舰船雷达目标识别系统中的应用[D]. 长沙: 国防科技大学, 1992.
[13] 郭桂蓉, 庄钊文, 陈曾平. 电磁特征抽取与目标识别[M]. 长沙: 国防科技大学出版社, 1996.
[14] 郁文贤. 海上雷达目标识别技术研制技术报告[R]. 长沙: 国防科技大学科研部, 2003. 11.
[15] 宋锐. 雷达舰船目标识别系统实现技术研究[D]. 长沙: 国防科技大学, 2003.

作者简介:

夏胜平 男, 1969 年生于湖南益阳, 控制理论与应用专业工学博士, 电子学与通信一级学科博士后, 副教授. 研究方向为智能信号处理、雷达自动目标识别、机器学习等. 已发表学术论文 40 余篇. E-mail: xiasp1227@sohu.com.

张乐锋 男, 1973 生, 博士生, 研究方向: 雷达信号处理、雷达自动目标识别.

虞 华 男, 1978 生于江苏常州, 硕士, 研究方向为雷达目标识别等.

张 静 女, 1977 年生于重庆市, 博士生, 研究方向为机器学习、雷达目标识别等.

胡卫东 男, 1967 年生于辽宁葫芦岛, 博士, 副教授, 研究方向为雷达目标识别、数据融合等.

郁文贤 男, 1964 生于上海, 博士生导师, 总装备部军事电子信息系统综合技术专业组特邀专家, 863 计划信息获取与处理技术主题专家组组长, 研究方向为智能信号处理、目标识别、数据融合等.