基于阴影集数据选择的可拓神经网络性能改进

周 玉[1]，钱 旭[2]，王自强[2,3]
(1. 华北水利水电学院 电力学院，郑州 450011；2. 中国矿业大学(北京) 机电与信息工程学院，北京 100083；
3. 河南工业大学 信息科学与工程学院，郑州 450001)

摘 要： 为了改进可拓神经网络的性能，提出一种基于阴影集的数据选择方法. 通过该方法获取用于训练可拓神经网络的训练样本，进而改进可拓神经网络的性能. 针对可拓神经网络的特点，选择核数据和边界数据作为可拓神经网络的训练样本；利用基于阴影集的数据选择方法，可以自动获取核数据和边界数据. 实验结果表明，与传统可拓神经网络相比，改进的可拓神经网络不仅节约了训练时间，而且网络的泛化能力和分类识别准确度得到了有效提高.

关键词： 可拓神经网络；数据选择；阴影集；训练样本；泛化能力
中图分类号： TP 183 **文献标志码：** A **文章编号：** 0254－0037(2013)03－0430－08

Performance Improvement of Extension Neural Network Using Data Selection Method Based on Shadowed Sets

ZHOU Yu[1], QIAN Xu[2], WANG Zi-qiang[2,3]
(1. School of Electric Power, North China University of Water Conservancy and Hydroelectric Power, Zhengzhou 450011, China;
2. School of Mechanical Electronic & Information Engineering, China University of Mining & Technology (Beijing), Beijing 100083, China;
3. College of Information Science & Engineering, Henan University of Technology, Zhengzhou 450001, China)

Abstract: To improve the performance of extension neural network (ENN), a data selection method based on shadowed sets was proposed. This method was used to obtain training sample data for improving the performance of ENN. According to the characteristics of ENN, core data and boundary data were selected as training data for ENN; using shadowed-sets-based data selection method, core data and boundary data could be captured automatically. Experimental results indicate that the learning speed of the improved extension neural network (IENN) is faster than traditional ENN. Moreover, the generalization ability and the recognition accuracy are improved effectively.

Key words: extension neural network (ENN); data selection; shadowed sets; training samples; generalization ability

训练样本是影响神经网络性能的重要因素之一，好的训练样本能提高神经网络的性能[1-2]. 通过对训练样本进行选择，从而改进神经网络性能的研究工作，受到国内外学者的关注，而且有了可观的成

收稿日期：2011-06-21.
基金项目：国家自然科学基金资助项目(70701013)；教育部科技研究重点资助项目(107021)；华北水利水电学院高层次人才科研启动基金资助项目(201117).
作者简介：周 玉(1979—)，男，讲师，主要从事智能信息处理技术、智能控制与决策方面的研究，E-mail: zhouyu@ncwu.edu.cn.

果[3-6]. 文献[3]中，作者利用基于模糊聚类的训练样本数据选择方法提高了 BP 网络的监督学习性能；文献[4]中，作者采用均匀设计法构造样本中心，结合聚类方法对训练样本进行优选，有效提高了径向基函数神经网络的泛化能力；文献[5]中，作者利用基于主动学习的样本数据选择方法提高了学习向量量化神经网络；文献[6]中，作者针对多层感知器，提出基于统计的主动学习训练样本选择方法，提高了多层感知器的性能.

从以前的研究中可以发现，通过数据选择方法对训练样本进行预处理，可以更好地指导神经网络进行有效的训练，进而提高神经网络模型的性能. 基于此，本文提出一种基于阴影集[7-8]的训练样本数据选择方法用以提高和改进可拓神经网络(extension neural network，ENN)[9]的性能.

1　可拓神经网络的性能改进

可拓神经网络(ENN)是继模糊神经网络、遗传神经网络、进化神经网络等之后的又一类新的神经网络类型[10]，是可拓理论[11-12]与神经网络相结合的产物. ENN 对于基于区间的分类与识别等问题效果显著，并有了较好的应用[13-16].

针对 ENN 的特点，在阴影集的基础上提出核数据和边界数据 2 个概念，ENN 的训练样本将从这 2 类数据集中进行选择. 基于阴影集的数据选择方法可自动提取核数据和边界数据，有效剔除不必要的样本，保留典型样本，提高了训练样本质量，从而使 ENN 的性能得到提高.

1.1　阴影集概念

阴影集是由模糊集诱导而来的，目的是解决模糊集中使用具有精确数值的隶属度来描述模糊逻辑的缺陷问题，在于帮助观察和解释不确定现象. 图 1 为一个模糊集以及由它所诱导出的阴影集的一个示例. 通过提升部分隶属度到 1，降低部分隶属度到 0 以及维持整体不确定性平衡，可以将传统的隶属函数变换成具有三值逻辑的阴影集，分别是隶属度为 1、0 和不确定的 3 个区域. 由定义在论域 X 上的一个给定的模糊集 B 而诱导的阴影集 A 是 X 中的一个区间值集合，它将 X 中的元素映射到 0、1 和单位区间 $[0,1]$，即 $A: X\rightarrow\{0,1,[0,1]\}$. $A(x)=0$ 表示元素 x 完全排斥在 A 之外，表示为 exclusion(A)；$A(x)=1$ 表示元素 x 完全包含在 A 之内，也称满足 $A(x)=1$ 的区域为核，表示为 core(A)；而 $A(x)=[0,1]$ 表示关于 A 中的元素 x 的隶属度不明确，也就是说不能确定该区域中的元素 x 是否属于 A，表示为 shadow(A).

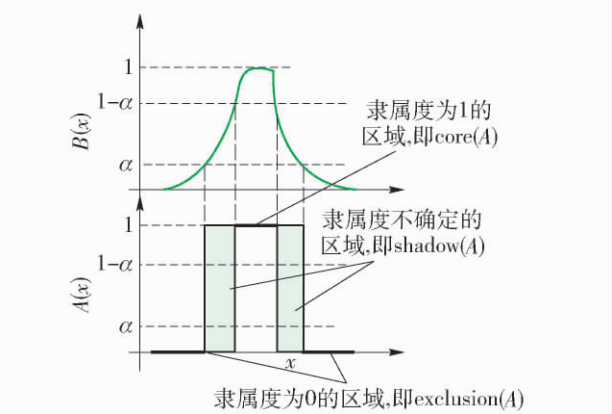


图 1　模糊集以及由它所诱导出的阴影集

Fig. 1　Fuzzy set and its induced shadowed set

阴影集构造过程中最关键的问题之一就是阈值 α 的计算. 文献[7-8]提出了基于不确定性平衡的优化方法. 如图 2 所示，为了尽量使隶属度改变量整体平衡(即不确定性平衡)，应该使得式(1)中的 V 达到最小，从而求得最优的阈值参数 α，如式(2)所示.

$$V(\alpha)=\left|\int_{-\infty}^{a}A(x)\,\mathrm{d}x+\int_{a_2}^{+\infty}(1-A(x))\,\mathrm{d}x-\int_{a_1}^{a_2}A(x)\,\mathrm{d}x\right| \tag{1}$$

$$\alpha_{\mathrm{opt}}=\arg\ \mathrm{Min}_{\alpha}V(\alpha) \tag{2}$$

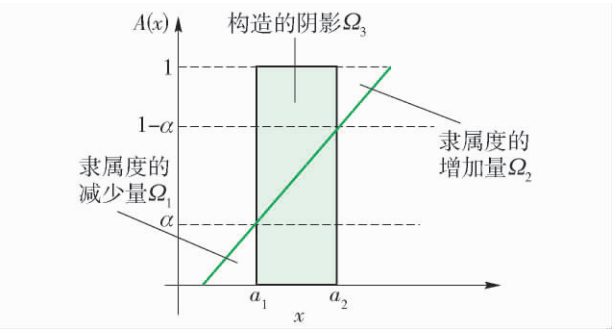


图 2　计算最优阈值

Fig. 2　Computing the optimal threshold

如果 $V=0$，表示达到完全平衡，即

$$\Omega_1+\Omega_2=\Omega_3 \tag{3}$$

对于论域是离散的情况，可用式(4)～式(5)求最优阈值参数：

$$V(\alpha)=\left|\sum_{i:A(x_i)<\alpha}A(x_i)+\sum_{i:A(x_i)>(1-\alpha)}[1-A(x_i)]-\mathrm{card}\{x_i\in X\mid\alpha<A(x)<(1-\alpha)\}\right| \tag{4}$$

$$\alpha_{\text{opt}} = \arg\ \text{Min}_{\alpha} V(\alpha) \tag{5}$$

阴影集与三值逻辑同构，所以阴影集中的逻辑运算与三值逻辑相同. 有关阴影集更多的运算规则、数学基础以及应用研究可以参见文献[7-8，17-20].

1.2 可拓神经网络

ENN 的结构如图 3 所示，由输入层和输出层组成. 连接输入层第 j 个节点和输出层第 k 个节点的 2 个权值分别用 w_{kj}^{L}（代表某一特征经典域的下限值）和 w_{kj}^{U}（代表相应特征经典域的上限值）表示，其中 L 表示下限，U 表示上限.

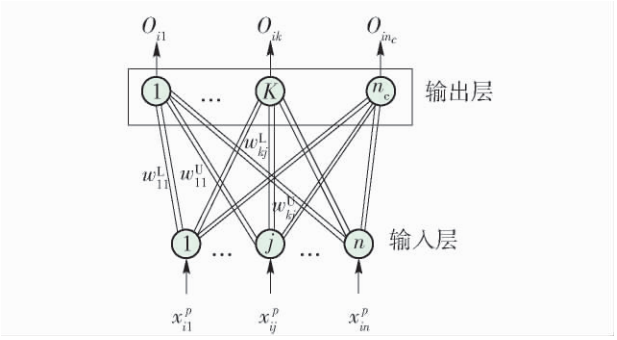


图 3 可拓神经网络结构

Fig. 3 Structure of ENN

ENN 是基于监督学习的，其算法如下.

假设训练样本集为 $X \equiv \{X_1, X_2, \cdots, X_{N_t}\}$，其中 N_t 是训练样本总个数. 第 i 个样本表示为 $X_i^p \equiv \{x_{i1}^p, x_{i2}^p, \cdots, x_{in}^p\}$，其中：$p$ 是样本标签；n 是样本特征的总个数. 为评价 ENN 分类性能，定义一个总的错误率 $E_t = N_m / N_t$，N_m 是错分类的总个数.

首先利用可拓理论的物元模型[12] 确定 ENN 的初始权值，并计算每一类的初始中心点. 假如输入第 i 个训练样本，利用可拓距离计算样本 X_i^p 和第 k 类的距离：

$$D_{ik} = \sum_{j=1}^{n} \left[\frac{|x_{ij}^p - z_{kj}| - (w_{kj}^{\text{U}} - w_{kj}^{\text{L}})/2}{|(w_{kj}^{\text{U}} - w_{kj}^{\text{L}})/2|} + 1 \right]$$
$$k = 1, 2, \cdots, n_c \tag{6}$$

式中 $z_{kj} = (w_{kj}^{\text{U}} + w_{kj}^{\text{L}})/2$. 确定 k^*，使得 $D_{ik^*} = \text{Min}\{D_{ik}\}$. 若 $k^* = p$，再输入下一个训练样本，否则进行第 p 类和第 k^* 类所对应的类中心和连接权值的调整，如式(7) ~ (10) 所示. 一直循环输入训练样本，直到总 E_t 达到指定的值或指定的最大训练步数进行完毕.

1) 类中心的调整

$$z_{pj}^{\text{new}} = z_{pj}^{\text{old}} + \eta(x_{ij}^p - z_{pj}^{\text{old}}) \tag{7}$$

$$z_{k^*j}^{\text{new}} = z_{k^*j}^{\text{old}} - \eta(x_{ij}^p - z_{k^*j}^{\text{old}}) \tag{8}$$

2) 权值的调整

$$\begin{cases} w_{pj}^{\text{L(new)}} = w_{pj}^{\text{L(old)}} + \eta(x_{ij}^p - z_{pj}^{\text{old}}) \\ w_{pj}^{\text{U(new)}} = w_{pj}^{\text{U(old)}} + \eta(x_{ij}^p - z_{pj}^{\text{old}}) \end{cases} \tag{9}$$

$$\begin{cases} w_{k^*j}^{\text{L(new)}} = w_{k^*j}^{\text{L(old)}} - \eta(x_{ij}^p - z_{k^*j}^{\text{old}}) \\ w_{k^*j}^{\text{U(new)}} = w_{k^*j}^{\text{U(old)}} - \eta(x_{ij}^p - z_{k^*j}^{\text{old}}) \end{cases} \tag{10}$$

式中 η 表示学习速率，$0 < \eta < 1$.

1.3 基于阴影集的训练样本数据选择方法

在 ENN 实际分类应用中，通过网络的学习算法自动从训练样本中提取各个特征向量经典域范围，从而确定网络权值参数. ENN 通过可拓距离来判断一个模式与类中心的相似程度，进而决定该模式的类别. 决定可拓距离大小的是各个特征值的边界和中心数据，因此，中心数据和边界数据这 2 类数据对 ENN 而言是信息量最大的数据，它们对 ENN 的学习起着重要的作用.

一般来说，位于类中心附近的训练样本是代表该类别最典型的样本数据，利用这些数据可以使得训练出来的模式类区域更加紧凑，不同模式类区域间隔更大，但是如果这样的训练样本过多或者只用中心数据进行训练，那么容易导致训练出来的模式类区域过于狭小，从而增加了其他样本的错误识别，尤其是位于类边界的数据的误识别. 另外，从分类器学习的角度来说，学习的目的是从样本数据中学到或者建立一个好的映射关系，并根据这个映射关系对新的模式进行划分. 也就是说，通过学习找到一个分类面，将分类空间划分为不同的类区域，训练的作用在于分类超曲面的生成，从这个方面来说，边界样本就是位于理想分类超曲面附近的样本，所以边界数据对分类器的训练至关重要. 然而，如果只用边界数据作为训练集，训练过程中会出现过拟合现象，从而导致不能对新模式进行有效识别和分类，即分类器的泛化能力较差. 所以，根据 ENN 的特点和分类器的要求，在样本数据的选择上，应该把这 2 种数据结合起来作为训练数据，这样就会结合 2 类数据的优势.

为了有效提取样本数据中的核（中心）数据和边界数据，提出一种基于阴影集的数据选择方法. 在阴影集的基础上，结合 ENN 的分类识别问题，构造了 2 种关键的数据类型：核数据和边界数据.

定义 1 核数据：假设有 c 个类别，若待分类的模式属于一个核的数据集合，称之为核数据，即

$$\text{Core data} = \{x \mid \exists_i x \in \text{core}(A_i)\},\ i = 1, 2, \cdots, c \tag{11}$$

定义 2　边界数据：假设有 c 个类别，若待分类的模式至少属于 2 个阴影的数据集合，称之为边界数据，即

$$\begin{aligned}\text{Boundary data} = \{x \mid (\exists_{i,j} x \in \text{shadow}(A_i) \cap x \in \text{shadow}(A_j)) \cup (\exists_{i,j,k} x \in \text{shadow}(A_i) \cap x \in \text{shadow}(A_j) \cap x \in \text{shadow}(A_k)) \cup \cdots \cup (\exists_{i,j,k,\cdots} x \in \text{shadow}(A_i) \cap x \in \text{shadow}(A_j) \cap x \in \text{shadow}(A_k) \cap \cdots) \cup \cdots\} \end{aligned} \tag{12}$$

核数据和边界数据示意图如图 4 所示.

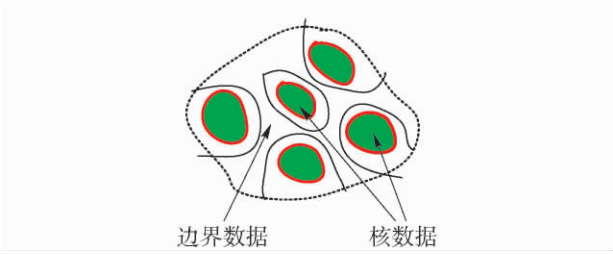


图 4　核数据和边界数据示意图

Fig. 4　Sketch map of core data and boundary data

基于阴影集的数据选择过程如下：首先对样本数据进行 FCM(fuzzy C-means) 聚类，获得最优模糊划分矩阵 $\boldsymbol{U} = [u_{ik}]$，$i = 1,2,\cdots,c$，$k = 1,2,\cdots,N$，这里 c 表示类别数，N 代表样本的个数. 由划分矩阵可以获得 c 个模糊集，由 c 个模糊集可以诱导对应的 c 个阴影集，再通过阴影集构建核数据和边界数据，最后通过在核数据和边界数据中进行数据提取以获取用于训练 ENN 的高质量的训练样本.

不同于其他的数据选择方法，基于阴影集的训练样本数据选择方法更可靠，这是因为对处于 exclusion(A) 和 core(A) 的数据具有信心，同时该方法效率更高，因为传统的数据选择方法是逐个数据进行选择，而该方法则是成批进行的. 最重要的是由于阴影集的最优阈值参数可以根据不确定性平衡的优化方法自动获得，所以核数据和边界数据的获取也可以实现自动化.

1.4　算法描述

基于阴影集数据选择的可拓神经网络性能改进的算法描述如下.

步骤 1　对原始样本数据进行 FCM 操作，得到最佳的模糊划分矩阵 $\boldsymbol{U} = [u_{ik}]$.

步骤 2　由划分矩阵 $\boldsymbol{U} = [u_{ik}]$ 得到 c 个模糊集.

步骤 3　根据不确定性平衡的优化方法，由式 (4) ~ (5) 分别求得 c 个最优的阈值 $\alpha_1,\alpha_2,\cdots\alpha_c$.

步骤 4　由 c 个模糊集和对应的 c 个最优的阈值诱导出 c 个阴影集.

步骤 5　根据诱导出的阴影集，将划分矩阵 $\boldsymbol{U} = [u_{ik}]$ 变换成阴影集对应的划分矩阵 $\boldsymbol{U}' = [u'_{ik}]$，即 $u'_{ik} \rightarrow \{0, 1, [0, 1]\}$.

步骤 6　由 $\boldsymbol{U}' = [u'_{ik}]$ 以及定义 1 和定义 2，构造核数据和边界数据.

步骤 7　数据提取. 从核数据集和边界数据集中选择数据，获得 N_t 个样本数据作为训练样本. 数据提取的原则：选择更接近类中心和类边界的数据. 基于此，引入一个 0 ~ 1 的控制参数 β_1 和一个 1 ~ 2 的控制参数 β_2 用来控制所选择数据的数量，继而更新三值矩阵 $\boldsymbol{U}' = [u'_{ik}]$，即

$$\begin{cases} u'_{ik} = 1, & u_{ik} > 1 - \beta_1\alpha_i \\ u'_{ik} = 0, & u_{ik} < \beta_2\alpha_i \\ u'_{ik} = [0, 1], & \beta_2\alpha_i < u_{ik} < 1 - \beta_1\alpha_i \end{cases} \tag{13}$$

根据定义 1 和定义 2 重新建立核数据和边界数据. 从建立的核数据和边界数据中选择数据作为 ENN 的训练样本.

由式 (13) 可知，β_1 越小，则 $u'_{ik} = 1$ 的数据会更少，即 $x \in \text{core}(A_i)$ 的机会越少，数据越靠近类中心；β_2 越大，则 $u'_{ik} = 0$ 的数据会更大，$x \in \text{shadow}(A_i)$ 的机会越少，数据越靠近各类数据的边界处；若 β_1、β_2 都等于 1，则此时的核数据和边界数据与步骤 6 中的核数据和边界数据一致. 控制参数的选择需要根据实际数据集的分布情况进行选择. 在以下的各项实验和应用中，参数 β_2 取值都为 1.

步骤 8　把所选择的 N_t 个数据当作 ENN 的训练样本，利用可拓理论中物元模型确定权值，则

$$\boldsymbol{R}_k = \begin{bmatrix} N_k & c_1, & V_{k1} \\ & c_2, & V_{k2} \\ & \vdots & \vdots \\ & c_n, & V_{kn} \end{bmatrix}, \quad k = 1,2,\cdots,n_c \tag{14}$$

在可拓理论中，$\boldsymbol{R}_k$ 表示物元，N_k 表示研究对象，c_j 表示 N_k 的第 j 个特征，$V_{kj} = \langle w_{kj}^{\mathrm{L}}, w_{kj}^{\mathrm{U}} \rangle$ 表示第 k 类关于特征 c_j 的经典域. 其中

$$w_{kj}^{\mathrm{L}} = \underset{i \in N_t}{\mathrm{Min}}\{x_{ij}^k\} \tag{15}$$

$$w_{kj}^{\mathrm{U}} = \underset{i \in N_t}{\mathrm{Max}}\{x_{ij}^k\} \tag{16}$$

步骤 9　计算每个类别的初始中心点

$$\boldsymbol{Z}_k = \{z_{k1}, z_{k2}, \cdots, z_{kn}\} \tag{17}$$

$$\boldsymbol{Z}_{kj} = (w_{kj}^{\mathrm{L}} + w_{kj}^{\mathrm{U}})/2, k = 1,2,\cdots,n_c; j = 1,2,\cdots,n \tag{18}$$

步骤 10　输入训练样本. 假设输入第 i 个样本

及对应的类标签 p: $X_i^p = \{x_{i1}^p, x_{i2}^p, \cdots, x_{in}^p\}$, $p \in n_c$.

步骤 11 利用式(6)得 X_i^p 和第 k 类的可拓距离.

步骤 12 确定 k^*,使得 $D_{ik^*} = \text{Min}\{D_{ik}\}$. 如果 $k^* = p$,则转到步骤 14,否则执行步骤 13.

步骤 13 调整权值,即按照式(7)~(10)进行第 p 类对应权重的调整和第 k^* 类对应类中心的调整.

步骤 14 重复步骤 10~步骤 13,如所有的样本都训练完,那么一次学习完成.

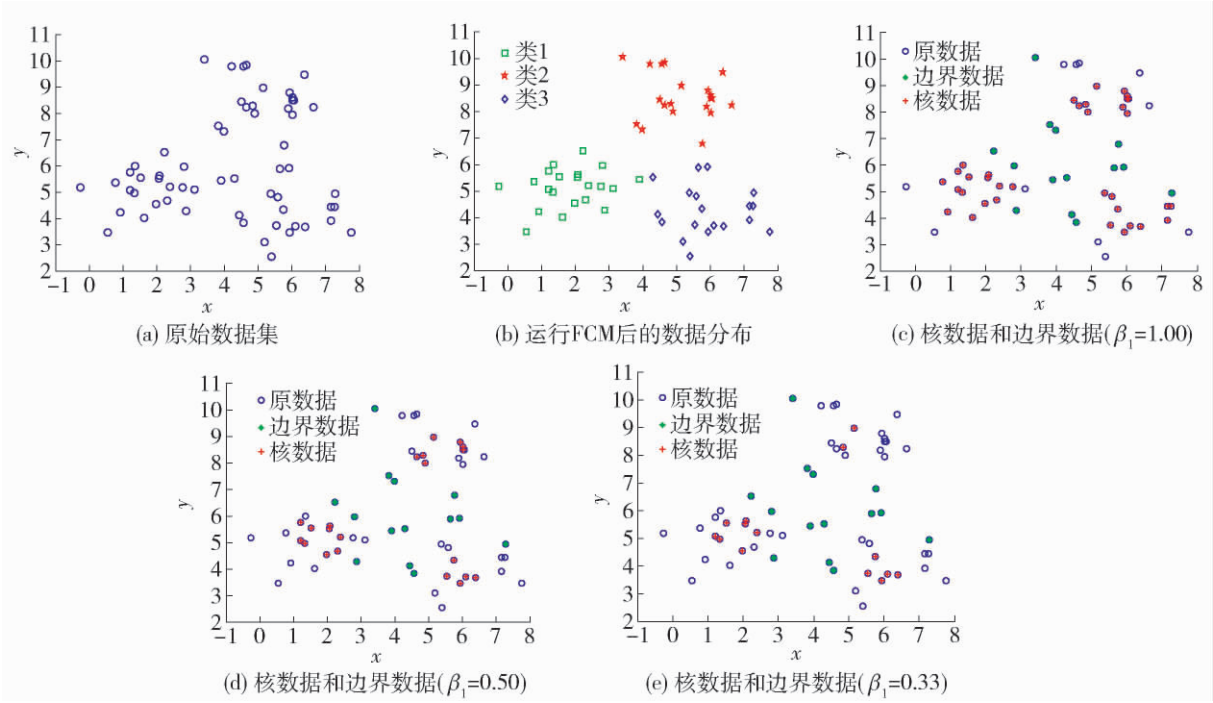
步骤 15 如果总误差 E_t 达到指定的值,或设定的训练次数进行完毕,则结束;否则转到步骤 10.

2 实验与分析讨论

通过实验研究来验证 ENN 的性能提高. 实验中分别采用不同的数据集,包括人工数据集、UCI 数据集和实际工程中的数据集. 其中 UCI 数据集和实际工程中的数据集与文献[9]中的数据集一样,即 IRIS 数据集和用于汽轮发电机组振动故障诊断问题数据集. 为了更好地与传统的 ENN 进行对比,除了进行与文献[9]相同的实验外,还进行了交叉验证实验.

2.1 随机产生的人工数据集

该数据集包含 60 个数据,类别数是 3,数据分布如图 5(a)所示. 在该实验中,利用全部 60 个数据作为训练样本训练 ENN,同时用相同的数据集当作测试数据集测试网络性能;而对于 IENN,首先对训练样本进行数据选择,利用选择后的样本数据训练 ENN,用同样的测试集检查网络性能. 图 5(b)所示为运行 FCM 后的数据聚类结果. 运行 FCM 后将数据集分为 3 类,并且获得一个最佳的模糊划分矩阵 $\boldsymbol{U} = [u_{ik}]$,这里 $i = 1,2,\cdots,3$, $k = 1,2,\cdots,60$. 这 3 个模糊集可以诱导出相应的阴影集,并计算最优阈值,可得 $\alpha_1 = 0.1258$, $\alpha_2 = 0.1051$ 和 $\alpha_3 = 0.1471$. 利用控制参数可以控制核数据和边界数据量的大小. 当 $\beta_1 = 1.00$ 时,可以获得 35 个核数据和 14 个边界数据,如图 5(c)所示;当 $\beta_1 = 0.50$ 时,可以获得 21 个核数据和 14 个边界数据,如图 5(d)所示;当 $\beta_1 = 0.33$ 时,可以获得 14 个核数据和 14 个边界数据,如图 5(e)所示. 表 1 显示了 ENN 和 IENN 的性能对比结果.



(a) 原始数据集 (b) 运行FCM后的数据分布 (c) 核数据和边界数据(β_1=1.00)

(d) 核数据和边界数据(β_1=0.50) (e) 核数据和边界数据(β_1=0.33)

图 5 二维人造数据集以及对应的核数据和边界数据

Fig. 5 Two-dimensional synthetic dada set and its core data and bandary data

从表 1 可看出,利用少于一半的训练样本数据至少可以获得相同的识别精度. 证明了基于阴影集数据选择方法的有效性,即可以保留信息量大的和典型的数据. 同时,说明利用基于阴影集的数据选择方法要比最常用的样本数据选择方法——随机选择方法效果好.

2.2 IRIS 数据集分类问题

这里分 3 种情况进行实验,其中前 2 种情况与

文献[9]中的保持一致，而第 3 种情况是为了更好地验证 IENN 的分类性能和泛化能力而进行的十折交叉检验实验.

表 1　实验结果

Table 1　Experimental results

实验方法	训练数据个数	学习时间/步	识别精度/%
无数据选择	60	2	100.00
核数据和边界数据($\beta_1=0.33$)	28	3	100.00
随机选择	28	4	98.33

情况 1　随机选择一半数据作为训练样本数据，剩余的一半数据作为测试数据集. 对于 IENN，对相同的训练样本进行数据选择，然后利用所选择的数据训练 ENN，最后利用相应的测试数据集来验证 IENN 的性能. 进行数据选择，控制参数 $\beta_1=\beta_2=1$，分别含有 49 个核数据和 18 个边界数据. 表 2 是在文献[9]的基础上总结的各种基于监督学习的神经网络分类器性能对比. 可以看到，ENN 在网络结构、训练性能以及泛化能力方面相比于其他几类神经网络分类器均有了提高. 而通过选择样本后所得到的 IENN 在性能上又有所提高，主要表现在泛化能力(测试误差为 0.027). 图 6 为 IENN 学习曲线. 另外，ENN 和 IENN 虽然都是 9 步完成学习，但是 IENN 每一步遍历的数据少于 ENN 遍历的数据，实际所用的时间更少.

表 2　各种神经网络分类器性能对比(情况 1)

Table 2　Comparison of the classification performance of various neural networks(Case 1)

网络类型	网络结构	训练数据个数	连接权个数	学习时间/步	训练误差	测试误差
感知器(perceptron)	4–3	75	12	200	0.173	0.213
多层感知器(MLP)	4–4–3–3	75	37	50	0.027	0.040
概率神经网络(PNN)	4–75–3	75	525	1	0.000	0.053
学习向量量化神经网络(LVQ)	4–15–3	75	105	20	0.080	0.053
对向传播神经网络(CPN)	4–20–3	75	140	60	0.107	0.160
可拓神经网络(ENN)	4–3	75	24	9	0.000	0.040
改进可拓神经网络(IENN)	4–3	67	24	9	0.000	0.027

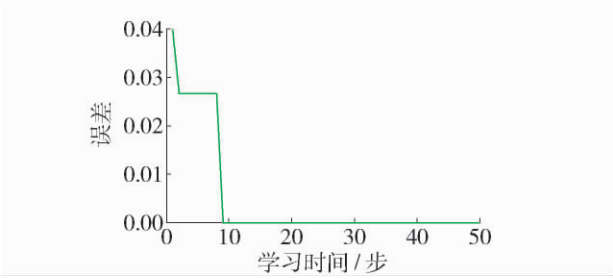


图 6　IENN 的学习曲线(情况 1)

Fig. 6　Learning curves of IENN (Case 1)

情况 2　这里用全部数据训练 ENN 和从全部数据中进行选择，再对 ENN 进行训练，然后用全部数据测试分类器性能. 表 3 是 ENN 和 IENN 的性能对比. 从表 3 可以看出，IENN 可以用少的样本数据获得较好的分类准确率，说明该方法能保障分类器的泛化能力，同时减少了训练时间. 图 7 为 IENN 学习曲线图.

表 3　ENN 和 IENN 性能对比(情况 2)

Table 3　Performance comparison between ENN and IENN(Case 2)

神经网络模型	训练数据个数	训练时间/步	训练误差	分类准确率/%
ENN	150	4	0.026 7	97.33
IENN	132	4	0.026 7	97.33

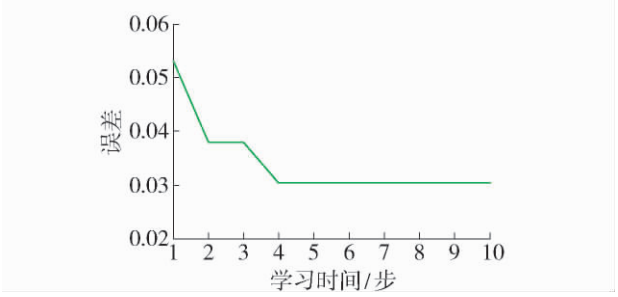


图 7　IENN 的学习曲线(情况 2)

Fig. 7　Learning curves of IENN (Case 2)

情况 3 为了客观评价 ENN 和 IENN 的性能，这里采用十折交叉检验实验. 通过十折交叉检验所得到的结果具有客观性. 实验结果如表 4 所示. 从表 4 可以看到，ENN 需要 13.1 步就能收敛，训练误差达 0.022 2，测试精度是 96.67%；而 IENN 只需要较短的学习时间(7.7 步)就能收敛，并且训练误差达到 0.018 5，测试精度达到 97.33%. 该实验进一步说明：IENN 不仅可以缩短训练时间、提高分类精度，而且神经网络的泛化能力也有所提高. 同时从测试标准差可以看到，IENN 在分类识别问题上具有稳定性.

表 4 ENN 和 IENN 十折交叉检验结果(情况 3)

Table 4 Experimental results of ENN and IENN using 10-fold cross validation (Case 3)

神经网络模型	学习速率	平均训练数据个数	平均学习时间/步	平均训练误差	平均测试精度/%	测试精度标准偏差
ENN	0.05	135.0	13.1	0.022 2	96.67	0.044 7
IENN	0.05	120.2	7.7	0.018 5	97.33	0.044 3

2.3 汽轮发电机组振动故障诊断问题

汽轮发电机组振动信号的频域特征包含了大量的机组故障信息. 与文献[9,16]一致，选择频谱中小于 0.4f，0.4f～0.5f，f，2f，3f 和大于 3f 等 6 个不同频段上的幅值分量能量作为特征向量；同时选择了机组常见的 3 种故障：油膜振荡(F1)、轴系不平衡(F2)、轴系不对中(F3). 该数据集包含了 15 个样本数据. 下面分 2 种情况进行实验，情况 1 与文献[10]一致，情况 2 是采用留一法交叉验证.

情况 1 用全部的 15 个数据当作训练数据对不同种类的神经网络进行训练，同时用这 15 个数据对训练好的各种网络模型进行测试；而对于 IENN，先进行数据选择，再同样用这 15 个数据对 IENN 进行测试. 表 5 显示了利用不同类型的神经网络模型学习结果. 从表中可以看到，传统的神经网络模型虽然可以很好地对故障类型进行识别，但它们的学习时间过长，而且网络结构复杂；而 ENN 和 IENN 不仅在结构上更加简单，在训练时间上要远远低于传统的神经网络模型，分别为 2 步和 1 步. IENN 的学习时间有所减少，但更重要的是 IENN 的泛化能力得到了提高，就是 IENN 使用 12 个数据即可对全部的 15 个数据进行正确的识别诊断，即 IENN 利用较少的训练数据即可实现系统的泛化能力. 这同时也说明了基于阴影集的样本数据选择方法可以保留信息量更大的数据，剔除冗余和信息量不大的数据.

表 5 利用不同类型神经网络模型实验结果(情况 1)

Table 5 Experimental results using different NN (Case 1)

神经网络模型	网络结构	训练数据个数	学习时间/步	识别精度/%
MLP	6-13-3	15	2 561	100
AWN	6-13-3	15	900	100
ENN	6-3	15	2	100
IENN	6-3	12	1	100

情况 2 为了充分说明 IENN 的整体性能，尤其是泛化能力和诊断识别精度得到提高，进行交叉验证法来比较 ENN 和 IENN 的性能. 由于汽轮发电机组振动故障诊断样本数据是小样本数据，故而采用留一法交叉验证. 实验结果如表 6 所示. 从表中可以明显看出，IENN 可以使用更少的样本数据却能保持较高的诊断识别精度，而且平均训练时间也得到了减少. 这说明利用样本数据选择方法可以很好地提高样本质量，从而可以提高 ENN 泛化能力、识别精度以及更快的学习速度.

表 6 ENN 和 IENN 留一法交叉验证结果(情况 2)

Table 6 Experimental results of ENN and IENN using leave-one-out cross-validation (Case 2)

神经网络模型	平均训练数据个数	平均学习时间/步	平均识别率/%
ENN	14.0	3.2	93.33
IENN	11.3	1.0	100.00

2.4 分析讨论

通过以上相关实验研究可以发现：

1) 提取核数据和边界数据作为 ENN 的实际训练样本数据是可取和有效的.

2) 基于阴影集的数据选择方法具有较高的效率，而且可以从样本数据的整体上对中心数据和边界数据进行自动划分.

3) 训练神经网络的首要和根本任务是确保训练好的网络模型对非训练样本具有好的泛化能力，即有效逼近样本蕴含的内在规律. 从实验中可以发现，IENN 在泛化能力上得到提高.

4) 由于本方法剔除了冗余和信息量少的数据，对典型数据进行了保留，使得训练样本大大减小，因此训练时间减少，并相应减小了网络的学习负担.

3　结论

神经网络性能与训练样本质量有着直接的关系，故而对样本数据进行有效选择是提高分类器性能的关键因素之一. 针对 ENN 的特点，利用基于阴影集的数据选择方法来获取高质量的训练样本，从而提高和改进了 ENN 的性能. 实验结果表明：选择核数据和边界数据作为训练 ENN 的实际训练数据是科学和可靠的，改进的可拓神经网络不仅在分类准确度、泛化能力方面有了提高，而且减少了 ENN 的训练时间.

参考文献：

[1] 魏海坤，徐嗣鑫，宋文忠. 神经网络的泛化理论和泛化方法 [J]. 自动化学报，2001，27(6)：806-815.
WEI Hai-kun，XU Si-xin，SONG Wen-zhong. Generalization theory and generalization methods for neural networks [J]. Acta Automatica Sinica，2001，27(6)：806-815. (in Chinese)

[2] ZHOU Yu，WU Ya-li. Analyses on influence of training data set to neural network supervised learning performance [J]. Advances in Computer Science，Intelligent System and Environment，2011，106：19-25.

[3] GUAN D，YUAN W，LEE Y K，et al. Improving supervised learning performance by using fuzzy clustering method to select training data [J]. Journal of Intelligent & Fuzzy Systems ，2008，19：321-334.

[4] 马翔，陈新楚，王邵伯. 均匀设计法在 RBF 神经网络样本优选中的应用 [J]. 模式识别与人工智能，2005，18(2)：252-255.
MA Xiang，CHEN Xin-chu，WANG Shao-bo. Application of the uniform design to the optimal selection of samples for RBF neural network [J]. Pattern Recognition and Artificial Intelligence，2005，18(2)：252-255. (in Chinese)

[5] PEREIRA C E. Learning vector quantization with training data selection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2006，28(1)：157-162.

[6] FUKUMIZU F. Statistical active learning in multilayer perceptrons [J]. IEEE Transactions on Neural Networks，2000，11(1)：17-26.

[7] PEDRYCZ W. Interpretation of cluster in the framework of shadowed sets [J]. Pattern Recognition Letters，2005，26(15)：2439-2449.

[8] PEDRYCZ W. From fuzzy sets to shadowed sets：interpretation and computing [J]. International Journal of Intelligence System，2009，24(1)：48-61.

[9] WANG M H，HUNG C P. Extension neural network and its applications [J]. Neural Networks，2003，16(5)：779-784.

[10] 周玉，钱旭，张俊彩. 可拓神经网络研究综述 [J]. 计算机应用研究，2010，27(1)：1-5.
ZHOU Yu，QIAN Xu，ZHANG Jun-cai. Survey and research of extension neural network [J]. Application Research of Computers，2010，27(1)：1-5. (in Chinese)

[11] CAI Wen. Extension theory and its applications [J]. Chinese Science Bulletin，1999，44(17)：1538-1548.

[12] 杨春燕，蔡文. 可拓工程 [M]. 北京：科学出版社，2007：18-97.

[13] ZHOU Yu，PEDRYCZ W，QIAN Xu. Application of extension neural network to safety status pattern recognition of coal mines [J]. Journal of Central South University of Technology，2011，18(3)：633-641.

[14] WANG W H. Partial discharge pattern recognition of current transformers using an ENN [J]. IEEE Trans on Power Delivery，2005，20(3)：1984-1990.

[15] LAI Y H，CHE H C. Modeling patent legal value by extension neural network [J]. Expert Systems With Applications，2009，36(7)：10520-10528.

[16] WANG M H. Extension neural network-type 2 and its applications [J]. IEEE Trans on Neural Networks，2005，16(6)：1352-1361.

[17] CATTENEO G，CIUCCI D. An algebraic approach to shadowed sets [J]. Electronic Notes in Theoretical Computer Science，2003，82(4)：64-75.

[18] MITRA S，PEDRYCZ W，BARMAN B. Shadowed *C*-means：integrating fuzzy and rough clustering [J]. Pattern Recognition，2010，43(4)：1282-1291.

[19] BARMAN B，MITRA S，PEDRYCZ W. Shadowed clustering for speech data and medical image segmentation [J]. Lecture Notes in Computer Science，2008，5306：475-484.

[20] GÜRKAN E，ERKMEN İ，ERKMEN A M. Two-way fuzzy adaptive identification and control of a flexible-joint robot arm [J]. Information Sciences，2002，145(1/2)：13-43.

(责任编辑　梁　洁)