文章编号： 1005 - 3662(2000)06 - 0016 - 03

神经网络分类器的特征提取和优选

马少华，高 峰，李 敏，吴成东
（沈阳建筑工程学院 自控系，辽宁 沈阳 110015）

摘 要： 结合木质胶合体缺疵质检过程，论述了用类内方差、类间方差，特征相关度三个目标函数综合测度图像特征差异的技术方法，从而实现了对神经网络输入层神经元的优选，降低了网络规模，提高了分类性能。

关 键 词： 特征；特征提取；特征选择；神经网络
中图分类号： TP183 文献标识号： B

1 引 言

人工神经网络是理论化的人脑神经网络的数学模型。它实际上是由大量简单元件相互连接而成的复杂网络，具有高度的非线性，能够进行复杂的逻辑操作和非线性关系实现的系统。而特征是对模式类不同性质的基本描述。模式识别中的特征选择和提取目的是多重的。一是为了降低模式表达维数；二是源于工程的考虑；三是辨识出与特定分类任务不相干的或多余的测量值。除工程约束外，降维对提高分类性能也有益处，有助于误识率的减低。本文将介绍一种判别方法，以‘计量’每一特征（输入）对模式判定的贡献，从而发现并去除多余或不良的特征，减少神经网络输入层神经元数，从而使网络规模得以降低，提高了计算效率和分类器诊断性能。

2 特征提取与优选

一个模式分类系统要正常工作，必须对这些模式类有适当描述，以确认潜在的有用特征。下列原则可以用来选取特征：①可区分性 对于同属一类的模式，特征应有相近值；②唯一性 分属不同类的模式，其特征应具有显著差异；③不相关性 不应有取自同一特定类的两个特征反映该类的同一属性。④数量少 一个模式分类器的复杂程度随着使用特征数的增加而急剧增大。另一方面，若所用特征数过少，分类器性能也降低。

特征选取可以看作是一个剪除不期望的噪声输入（特征）的过程。通过从可用特征中去除噪声输入，压缩特征集，从而降低神经网络规模。实现该过程的方法主要有强迫法和统计法。

1）强迫法 该方法基于检验特征的所有组合以构成特征子集，不需对特征进行详尽研究。由基于特征子集的分类器性能确定最佳特征。强迫法虽然有效，但其最大的缺点是特征的所有组合数十分巨大。例如 5 个特征的组合数为 31，且随着特征的增加，组合数增长很快。实际应用时，会增加试验次数，延长试验周期，耗费大量时间、精力。有时还难以达到预期的效果，甚至无法完成全部试验。

2）统计法 研究特征的统计规律以确定其适切性。用于测度特征适切性的三个目标函数分别称为类内方差 类间方差 特征相关度。这三个函数与前面提到的特征选取原则相对应。

①类内方差 类内方差是某类一个特定特征的方差。如图 1 所示，类 1 和类 2 的特征测量值具有不同的类内方差。类 1 的方差大于类 2 的方差，所以类 2 较易被确认。假设类 j 共有 P_j 个训练模式，x_{ij} 为类 j 的特征 x 中第 i 个模式，类 j 的特征 x 的均值为 μ_{xj}。值得注意的是，这些均值基于训练样本，而不是实际均值。即：

$$\mu_{xj} = \frac{1}{P} \sum_{i=1}^{p_j} x_{ij}$$

收稿日期： 2000 - 03 - 03
作者简介： 马少华（1954 - ），男，辽宁沈阳人，副教授，研究生，主要从事智能控制等方面的教学与科研工作。

则类内方差可由下式确定：

$$\sigma_{xj}^2 = \frac{1}{P}\sum_{i=1}^{p_j}(x_{ij} - \mu_{xj})^2$$

理想情况下，对于同一类中模式的特征应有相近值，因此类内方差应较小。较小类内方差可做为目标函数之一用于类的最佳特征的选取。

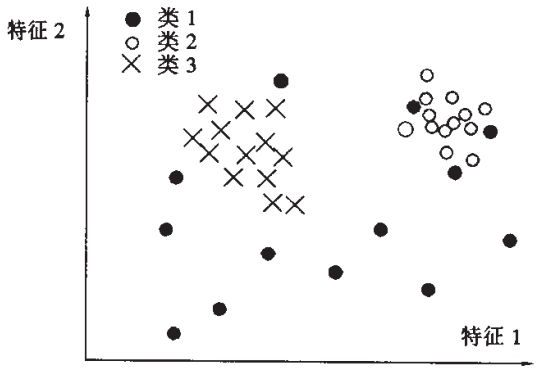

图 1　类内方差和类间方差

② 类间方差　类间方差测度特征的划分类的能力。类距均值经标准化等于类间方差。对于类 j 和类 l 的特征 x，类距可如下定义：

$$D_{xij} = \frac{|\mu_{xj} - \mu_{xl}|}{[\sigma_{xj}^2 + \sigma_{xl}^2]^{\frac{1}{2}}}$$

同样划分类的最佳特征所具有的均值相差的越大越好。若类间方差较小，那么不同类的特征趋于交迭，类内方差较大，也是一样。这是因为类间方差要给出较好的类别划分需依赖类内方差。显然，类内方差、类间方差都应该被作为目标函数给予考虑。

③ 特征相关度　特征相关度能确定类中含有冗余信息的特征，高度相关特征只反映同一特性，因此只需要一个特征。类 j 特征 x 和 y 的相关系数如下计算：

$$D_{xjl} = \frac{\left\{P_j\sum_{i=1}^{p_j}(x_{ij} - \mu_{xj}) \cdot (y_{ij} - \mu_{yj})\right\}^{-1}}{\sigma_{xj} + \sigma_{yj}}$$

其值在 -1 与 1 之间。0 表示两特征完全不相关。1 表示较高相关度，-1 表示较高负相关度。如果相关系数绝对值比 0 大许多，将有一个特征被抛弃。由此可见，通过计算特征的相关度，可以降低特征集的规模。

3　检验木材胶合板样本特征集

木质胶合板缺疵质检过程实质上就是利用神经网络进行模式识别。即把由胶合板的图像提取的特征提供给神经网络以识别缺疵。胶合板的数字图像由 512×512 个像素构成，每个像素伴以 0(黑) 到 255(白) 灰度值。用分段方法能够判别缺疵区，且与非缺疵相区分。一旦缺疵区被找到，一个 X 轴方向的 60 个像素，Y 轴方向 85 个像素的窗口被置于缺疵之上。窗口的原点位于缺疵的中间，这个窗口的尺寸与 3 cm^2 区域相对应，足以覆盖除树皮以外的任何缺疵。灰度值和它们的频率由特征提取窗口来记录。同一类缺疵样本的灰度直方图有相似的形状。这种用窗口提取特征的方法被多位研究者使用。典型灰度直方图如图 2 所示。第一、第二类特征可以由窗口提取。第一类特征是基本特征，且可由窗口的灰度直方图直接计算；第二类特征是构造特征，能由阈值和边缘检测处理图像而获得。由训练和检验神经网络的样本共提取了代表胶合板缺疵的 17 个特征。这些特征见表 1。

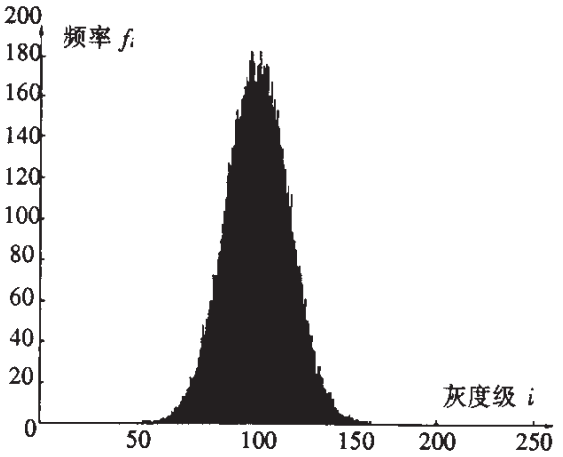

表 1　胶合板典型特征

	特　征		
1	灰度均值	10	最高灰度值
2	灰度中值	11	直方图黑区尾部长度
3	灰度最频值	12	直方图白区尾部长度
4	灰度标准差	13	均值为阈值的边缘像素数
5	畸变	14	以 $\mu - \delta$ 为阈值后像素数
6	峰度	15	特征 14 的边缘像素数
7	黑像素数	16	以 $\mu + \delta$ 为阈值后像素数
8	明亮像素数	17	特征 16 的边缘像素数
9	最低灰度值		

检验木材胶合板的 17 个图像特征共使用

232个训练样本。计算各类各个特征类内方差，但还不能用于比较。因为这些值带有不同的计量单位。因此，需要标准化(除以特征均值)。

类 j 与其他 k 个类关于特征 x 的类间方差计算式如下：

$$\sum_{i=1}^{k}\frac{|\mu_{xj}-\mu_{xl}|}{[\sigma_{xj}^2+\sigma_{xl}^2]^{\frac{1}{2}}}$$

从类间方差定义可以看出，对于特征集 $X^t=(x_1\,x_2\cdots x_n)$ 类 j 和类 t 的可分离度由下式计算：

$$\sum_{i=1}^{n}\frac{|\mu_{xmj}-\mu_{xml}|}{[\sigma_{xmj}^2+\sigma_{xml}^2]^{\frac{1}{2}}}$$

使用该公式计算每对类的可分离度。

现在可以依照类内方差、类间方差选取最佳特征。理论上，类内方差要尽可能小，且类间方差尽可能大。类内方差的上限和类间方差的下限可作为选取最佳特征的标准。然而，这里没有新特征可用，于是引进去除最差原则。在可用的特征中剪除不良特征。

选取的标准是如果特征的类间方差相对较小，则舍弃具有最大类内方差的特征。应用标准值和类间方差进行计算，结果满足舍弃标准的特征被注以“0”，以表示应从特征集中删除；而其他特征注“1”，以示可包含在特征集内。通过计算特征相关度对选取的特征进一步检查其冗余性。具有较高相关度的特征对中要去除一个。本文中，特征至少具有0.9的相关系数才认为相关。使用这一阈值，特征的冗余性判定结果见表2。“1”表示非冗余，“0”表示冗余。表2结果总结如下：

① 特征1与特征2、特征9相关；② 特征2与特征1相关；③ 特征3与其他特征都不相关。

表2中信息可以合并以判断类1特征的冗余性。本文使用合并图来处理。如图3a所示。节点代表特征。当有较高相关度时，用线连接，两相连接节点中具有较小类间方差的被剪除。例如，特征14、15中特征15将被去除。当一个节点与几个节点相连时，那么末端节点被保留，以减少删除的节点数，如节点1、2、9，节点2、9保留，而节点1被剪除。此过程始于节点1，顺时针进行，直至所有特征都被处理过为止。值得注意的是，

表2 类特征间相关度

	特征																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
6	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
9	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
16	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1

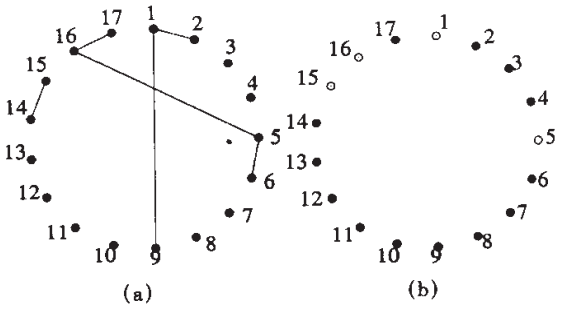

图3 特征冗余性合并图

当一个节点被删除(3b中的空心圆)，该节点所连接的线也被去除。结果节点2、3、4、6、7、8、9、10、11、12、13、14、17得以保留。同理可获得其他各类的非冗余特征。

4 结 语

本文着重介绍了特征选取的统计方法。引入类内方差、类间方差、特征相关度3个目标函数测度木材胶合板图像的特征差异，进而阐述了将三种方差综合来测度特征差异的方法，实现了对特征的优选。

参考文献：

〔2〕 戚飞虎，等译. 模式识别与图像处理 M. 上海交通大学出版社，1989.

〔3〕 赵荣椿. 数字图像处理导论 M. 西北工业大学出版社，1995.6.

Feature Extraction and Selection of Neural Network

MA Shao-hua, GAO Feng, LI Ming, WU Cheng-dong

(Autamation Dept. Shenyang Architechtural and Civil Engineering Institut, 110015, China)

Abstract: This paper presents the method of feature extraction and feature selection face wood veneer inspection application of neural network classifiers.

Key words: feature; feature extraction; feature selection; neural network