文章编号:2095-0411(2017)03-0060-09

基于核 PCA 与 SVM 算法的木材缺陷识别

马 旭,刘应安,业 宁,闫 贺
(南京林业大学 信息科学技术学院,江苏 南京 210037)

摘要:木材缺陷是影响木材产业化推广的重要因素之一,通过合理的木材缺陷识别方法可以有效规避木材缺陷在实际应用中带来的资源浪费问题,同时大幅提高木材的实际利用率。针对木材节子非线性的特征,提出了一种新颖的木材缺陷识别方法。首先,通过核主成分分析方法(Kernel Principal Component Analysis,KPCA),采用多项式的核函数(Polynomial kernel function)对木材原始的非线性数据从低维映射到高维线性特征空间,然后再对映射空间中的线性样本进行降维处理,目的是为了提取到样本的特征参数。其次,结合 SVM 模型,选择多项式核函数,完成对木材缺陷的识别。最后,通过比较实验所得数据与实测数据,实验结果表明本文提出的方法有较高的识别精度和识别效率。

关键词:木材缺陷;核函数;主成分提取;支持向量机
中图分类号:TS 611 **文献标志码**:A **doi**:10.3969/j.issn.2095-0411.2017.03.009

Application of KPCA and SVM to Wood Defect Recognition

MA Xu, LIU Ying'an, YE Ning, YAN He
(College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China)

Abstract: Wood defect is an important factor affecting the wood industrialization promotion. A reasonable wood defect recognition method can effectively avoid the waste of resources caused by wood defects in the practical application. At the same time it can raise the actual utilization of wood. Considering the nonlinear characteristic of wood defects, a new wood defect recognition method is proposed. Firstly, mapping wood original nonlinear data from low dimensional to high dimensional linear feature space using the polynomial kernel function. And then the mapping space of linear dimension reduction processing samples. The purpose is to extract the feature parameters to the samples. Next by means of the SVM model, the polynomial kernel function is selected to complete the wood defect identification. The experimental results show that the proposed method has higher recognition accuracy and efficiency by comparing the data from experiment and the measured data.

Key words: wood defect; kernel function; PCA; SVM

中国森林资源逐渐减少,而木材的需求量日益增大,如何提高木材的合理利用率成了急需解决的一个重大问题。第七届中国木材保护工业大会暨人工林优化新技术发展高峰论坛上,中国工程院院士李坚、中国木材与木制品流通协会会长刘能文等木材工业领域的权威专家呼吁,中国应通过政策驱动和科技创新等手段提高木材综合利用、木材功能改良和废旧木材回收利用等能力,以缓解当前的木材供需矛盾。据统计,目前

收稿日期:2016-09-15。

作者简介:马旭(1993—),男,江苏仪征人,硕士生,主要从事数据挖掘、模式识别研究。通讯联系人:刘应安(1965—),E-mail: lyastat@163.com

我国的木材综合利用率仅为 40%～50%，而世界上比较先进的国家木材利用率达到 70%～80%[1]。如果能提高 1%，我国每年将节约木材 $7.5\times10^5\ m^3$。

木材资源供应不足与市场需求急剧增加是中国木材工业面临的主要矛盾[2]。随着中国的木材供需矛盾愈发突出，对木材缺陷识别已经成为一个热门话题。结合模式识别，不少新的模式识别的方法已被应用在木材缺陷识别上，并且取得了一定成就。在对木材识别上，许多学者作出了长足的贡献。2010 年牟洪波等[3]提出了基于 BP(Back Propagation)和 RBF 神经网络(Radial Basis Function Neural Network)的木材缺陷检测研究。首先对木材缺陷图像分别进行了灰度增强变换，改进的加权均值滤波处理、中值滤波处理，最大的保留了图像缺陷细节，易于后续的图像特征提取。运用常见的几个边缘检测算子对木材缺陷图像进行边缘检测，提取出清晰的木材缺陷边缘。然后对木材缺陷特征选择。2010 年王玉珏等[4]提出了一种基于颜色特征木材缺陷检测的研究方法，文中研究了两种木材缺陷的分割方法。2013 年徐姗姗等[5]提出了基于卷积神经网络的木材缺陷识别。2016 年白雪冰等[6]从木材的缺陷分割入手，提出了局部二值拟合模型在木材识别上的应用方法。综合上述的传统对木材缺陷识别的研究，可以得到木材缺陷识别的两个重点：一是特征抽取，二是分类识别。无论是基于颜色还是边缘识别算子的方法，这里总有一个问题：非线性的特征。而在卷积神经网络中，其实也正是巧妙避开了这个非线性的问题。这里针对非线性特征提出了一种新的识别方法，通过 KPCA 对木材数据降维，然后通过 SVM 训练分类。由于木材实验样本以及真实生产中木材样本的数据以及维数并非大量，使得 SVM 完全适合木材缺陷识别的分类工作。新方法充分利用了木材缺陷的非线性特点，无需过多的前期图像预处理，简化了传统的木材缺陷识别算法。考虑了木材小样本以及降维后维数不高的特性，利用 SVM(Support Vector Machine)分类，无论是在降维原始数据还是分类识别上都有了长足的进步，在对木材识别的精度与时间上都有所提高。

1　木材缺陷特性

1.1　木材缺陷

木材缺陷是指呈现在木材上能够降低其质量、影响其使用的各种缺点[7]。国家标准将木材缺陷分为 10 大类：节子、变色、腐朽、虫害、裂纹、树干形状缺陷、木材构造缺陷、伤疤、木材加工缺陷和变形。任何树种的木材都存在缺陷，这 10 大木材缺陷木材的形成通常分为 3 大种：①因树木生长特性或环境的影响而形成的缺陷，称为天然缺陷；②由于生物对树木的危害而产生的缺陷，称为生物危缺陷；③在木材的加工干燥等等人为干预过程中形成的缺陷，称为人为缺陷。天然缺陷、生物缺陷、人为缺陷中，后面 2 种缺陷都是人为可控的，所有在检验木材缺陷中，天然缺陷成了一个重点。在各种天然缺陷中，节子(如图 1 所示)是最普通最常见的一种木材缺陷。

图 1　木材节子缺陷

节子[8]在木材中是一种常见的缺陷，它不仅破坏木材纹理的均匀性和完整性，更会影响木材表面的视觉效果和加工难度。从图 1 可以发现，木材节子部分的像素值亮度与木材纹理背景有明显的反差。因此，从纹理上区分良好木材与缺陷木材是一种简单有效且可行的方法。纹理特征不是简单的从像素点出发，它需要在一个区域中对多个像素点进行统计计算。这种区域性不会因为局部的偏差导致模式匹配的不成功。纹理特征常具有旋转不变性，并且对于噪声有较强的抵抗力。

1.2　木材检验工作

木材检验的工作贯穿于木材的生产到木材的加工的整个过程，是一个必不可少的过程。因为木材的检验工作会关系到整个木材产品的质量与销售情况，直接影响企业的经济效益。木材能否被充分利用、体现其价值，前提就是做好木材的检验工作。

由于构造上的不规则不确定性质，木材会呈现出各种病态，这些病态宽泛来说都叫做木材缺陷。木材缺

陷是影响木材物理力学性能以及外表纹理的主要因素之一。木材的等级评定,除了根据木材的用途和树种之外,主要还是根据各种缺陷的允许存在程度来判定。

1.3 传统图像识别的特征提取特性

在传统的图像处理中,降维往往是图像处理的第一步。假如一幅图像用 32×32 的矩阵存储,那么一幅图片的维数就是 1024 维,这对计算形成了巨大的困难。所以特征提取成为图像识别的必经之路。

国内外学者对特征提取进行了大量的研究。PCA(Principal Component Analysis)主成分分析由 Pearson K 于 1901 年发明[9],用于分析数据及建立数理模型。其方法主要是通过对协方差矩阵进行特征分解,以得出数据的主成分(即特征向量)与它们的权值(即特征值)。PCA 是最简单的以特征量分析多元统计分布的方法。Fisher 线性判别[10]将 d 维空间的数据点投影到 m 维的低维空间去,使不同类的样本点在低维空间的投影尽量分离,同类的样本点在低维空间中尽量紧凑。除了上面最常见的 PCA,FLD(Fisher Linear Discriminant),还有其他一些特征提取的方法。但是如图 1 所示,这些样本的不均匀分布并非线性,就是说无法映射到一个特征空间使其与正常纹理区分开来。所以用线性的特征提取来做此方面的工作,效果不会很好。基于此,实验中采用了一种非线性的核 PCA 方法做了特征提取,实验表明,非线性的核方法对木材缺陷的特征提取效果显著。

2 核 PCA 与支持向量机背景

2.1 PCA 主成分分析

PCA(Principal Component Analysis,主成分分析)[11]是应用最为广泛的特征提取之一。其主要思想是提取样本的主要特征,减少冗余低效的信息,使得高维数据能够在低维空间处理,从而解决维数过高导致的计算性能瓶颈问题,PCA 本质就是 K-L 变换,保留最大方差。

假设给定的样本

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

若能用较少的几个综合变量来代替原来较多的变量(即降维处理),处理问题将会变得简单。在降维处理中,所选取的投影空间投影后的数据应该能尽可能多的反映原数据的信息,同时,综合变量之间需要相互独立。所以寻找的特征空间就是原多个变量的线性组合,使满足上面 2 个条件的系数组合。

首先给出样本的中心点

$$\bar{x}=\frac{1}{N}\sum_{n=1}^{N}x_n \tag{2}$$

假设 $\boldsymbol{u}$ 为投影的向量,那么投影之后的方差

$$S=\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{u}^{\mathrm{T}}x_n-\boldsymbol{u}^{\mathrm{T}}\bar{x})^2=\boldsymbol{u}^{\mathrm{T}}S\boldsymbol{u} \tag{3}$$

满足约束

$$\boldsymbol{u}^{\mathrm{T}}\cdot\boldsymbol{u}=1 \tag{4}$$

通过拉格朗日乘子法,构造拉格朗日函数

$$L_{(u)}=\boldsymbol{u}^{\mathrm{T}}S\boldsymbol{u}+\lambda(1-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}) \tag{5}$$

求解此函数可得

$$S\boldsymbol{u}=\lambda\boldsymbol{u} \tag{6}$$

到这就是一个标准的特征值表达式了,λ 对应的特征值,$\boldsymbol{u}$ 对应的特征向量。式(6)的左边取得最大值的条件就是 λ 最大,也就是取得最大的特征值的时候。假设要将一个 D 维的数据空间投影到 M 维的数据

空间中($M \leqslant D$)，那取前 M 个特征向量构成的投影矩阵就是能够使得方差最大的矩阵了。这就得到了投影空间。

主成分分析作为一个特征提取办法，需要在降维的维数与原信息保留之间找到一个平衡。在计算原信息保留的方法中，传统 PCA 给出了一种计算方法称之为主成分贡献率。主成分 z_i 的贡献率

$$a_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{7}$$

式中：a_i 为 z_i 主成分的贡献率，λ_i 为对应特征值。其本质就是对因特征值所占整体比重。那么推广下就可以等到累计的贡献率

$$a = \sum_{i=1}^{m} a_i = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{8}$$

式(8)表示的累计贡献率即所选择的特征值之和占总特征值的比重。在一般的工程应用中，累计贡献率一般要求在 85%～95%。

2.2　核 PCA 主成分分析

2.2.1　核方法

核方法(kernel methods)[12]是解决非线性模式分析问题的一种有效途径，核心思想在于通过某种线性的映射将原始数据映射到合适的高维空间中，再通过线性的学习器在新的高维空间中分析处理模式。核函数的形式

$$k(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle \tag{9}$$

式中：k 函数即核函数，ψ 代表了数据从原始低维向高维的映射。$\langle , \rangle$ 为内积。可以看出，核函数将高维空间的内积运算转化为了低维空间的核函数计算，巧妙的解决了在高维空间的“维数灾难”等问题。

核函数的价值在于它虽然也是将样本从低维到高维进行转换，但是它事先是在低维上进行计算，而将实际上的映射效果表现在高维中，避免了直接在高维空间中的复杂计算。

2.2.2　核 PCA

核方法的特性使之成为处理非线性问题的热门方法。核方法已经应用在了各个数据挖掘的领域中，如支持向量机，KLDA，KPCA 中。核主成分分析(KPCA)是 PCA 算法的一种非线性处理改进。KPCA 在进行特征提取之前，先将原始数据通过核函数向高维映射，这样就能将非线性的原始数据在高维中线性表示。即通过非线性函数 Φ 映射到高维特征空间 F，$F = \{\Phi(x), x \in R^N\}$，在这样的高维空间中，在对映射后的数据进行相同的 PCA 处理。实验表明，KPCA[13-14]比传统 PCA 识别性能更加，尤其在处理非线性数据上效果显著。

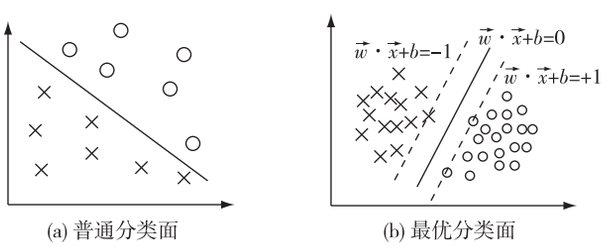
假设低维向高维的映射是 Φ，同传统 PCA 一样，需要首先对数据进行中心化，在中心化过后得到协方差矩阵

$$\bar{\boldsymbol{C}} = \frac{1}{N}\Phi(x_i)\Phi(x_i)^{\mathrm{T}} = \frac{1}{N}\Phi(X)\Phi(X)^{\mathrm{T}} \tag{10}$$

PCA 投影可以用内积运算表示，因此当我们把 $K_{ij} = x_i^{\mathrm{T}} x_j$ 推广到映射的投影空间后就变成 $K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi(x_i)^{\mathrm{T}} \Phi(x_j)$ 时，通过 PCA 的分析结果并不会改变。

2.3　支持向量机

SVM 由 20 世纪 90 年代 Vapnik[15]提出，支持向量机寻找的不只是一个能分类的超平面，而是一个最优的超平面。普通分类面和最优分类面如图 2 所示。



(a) 普通分类面　　(b) 最优分类面

图 2　普通分类面和最优分类面

对比 2 个超平面可以发现,图 2(b)的分类面使得 2 类数据间隔最大,所以这是一个最优分类面。构造目标函数

$$\min \frac{1}{2} ||\omega||^2, \text{s.t. } y_i(\omega^{\mathrm{T}} x_i + b) \geqslant 1, i = 1, \cdots, n \tag{11}$$

式中:ω, b 是需要确定的分类面,x_i 为第 i 个样本,y_i 为对应的类别标号。定义拉格朗日函数(α 为拉格朗日乘子)

$$L_{(\omega,b,\alpha)} = \frac{1}{2} ||\omega||^2 - \sum_{i=1}^{n} \alpha_i (y_i(\omega^{\mathrm{T}} x_i + b) - 1) \tag{12}$$

分别对 ω, b 求导,得到:

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{13}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{14}$$

这样 ω 求出之后通过 2 个支撑面可求出

$$b = -\frac{\max_{i;y=-1}^{(i)} w^{*\mathrm{T}} x^{(i)} + \min_{i;y=1}^{(i)} w^{*\mathrm{T}} x^{(i)}}{2} \tag{15}$$

通过 SMO 算法求解对偶问题中的拉格朗日乘子 α ,即可得到 ω 和 b 的真实解。

由式(13)得到的 $\omega = \sum_{i=1}^{n} \alpha_i y_i x_i$,将 ω 带入分类函数中得到

$$\sum_{i=1}^{n} \alpha_i y_i < x_i, x > + b \tag{16}$$

在式(16)中,产生了一个重要的性质:内积。这一点至关重要,这就是使用 kernel 进行非线性推广的基本前提了。这样通过核函数,分类面就变为

$$\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b \tag{17}$$

通过核函数,可以将数据映射到高维以便区分,但却仍直接在原来的低维空间中进行计算,而不需要显示的写出映射的结果,避开了直接在高维空间中进行计算。但是由于噪声的存在,数据往往并非线性可分,也不应该是投影到高维空间中。对于偏离正常位置很远的这些数据点,称之为 outlier。outlier 会导致分类面的偏移影响整体分类效果。定义 $\xi \geqslant 0$ 为松弛变量,对应数据点允许偏离支撑平面的量。那么点与支撑平面的最小距离就是 $1-\xi$ 。要求间隔与所有松弛变量和最小,那么目标函数就变成

$$\min \frac{1}{2} \|w\|^2 + C\sum_{i}^{n} \xi_i, \text{s.t. } y_i(w^{\mathrm{T}} x_i + b) \geqslant 1 - \xi_i, \xi_i \geqslant 0, i = 1, \cdots, n \tag{18}$$

C 为自定义的一个正数,用以平衡两者之间的权重。同样构造拉格朗日函数:

$$L(\omega, b, \xi, \alpha, \gamma) = \frac{1}{2} ||\omega||^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (y_i(\omega^{\mathrm{T}} x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \gamma_i \xi_i \tag{19}$$

分别对 ω, b, ξ 求导可得:

$$\begin{aligned} &\frac{\partial L}{\partial \omega} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ &\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad\qquad i = 1, \cdots, n \\ &\frac{\partial L}{\partial \xi} = 0 \Rightarrow C - \alpha_i - \gamma_i = 0 \end{aligned} \tag{20}$$

其整个对偶问题可以写作:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle$$
$$\text{s.t.},\sum_{i=1}^{n}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C,i=1,\cdots,n \tag{21}$$

可以发现与原来对偶问题的差别在于 α_i 多了个上限 C，如果再将核函数带入，求出 ω,b，完整的支持向量机模型完成。

3　核 PCA 与 SVM 的木材缺陷识别算法研究

3.1　KPCA 对木材样本时间复杂度研究

从算法时间复杂度分析上解释[16]，对于由 $n\times m$ 的原始数据（n 为样本个数，m 为样本维数），PCA 中计算协方差矩阵需要 $\boldsymbol{O}(nm)$，而对于 $m\times m$ 的协方差矩阵进行特征分析需要 $\boldsymbol{O}(m^3)$，所以 PCA 总共需要的时间复杂度为 $\boldsymbol{O}(nm+m^3)$，KPCA 对 $n\times n$ 的矩阵进行特征分析需要 $\boldsymbol{O}(n^3)$，计算核需要 $\boldsymbol{O}(mn^2)$，所以 KPCA 的时间复杂度时 $\boldsymbol{O}(mn^2+n^3)$。对比 2 个时间复杂度，在实际的木材缺陷识别中，样本图片拉成一维后维数极高，比样本数高出太对，所以 PCA 的速度相对 KPCA 数度慢了许多。对于 LDA 算法，计算协方差矩阵需要 $\boldsymbol{O}(nm)$，对 $m\times m$ 的协方差阵再次进行特征是分解时间复杂度为 $\boldsymbol{O}(m^3)$，所以 LDA 的总计时间复杂度为 $\boldsymbol{O}(nm+m^3)$。相比较于 KPCA 的 $\boldsymbol{O}(mn^2+n^3)$，大大超出。对于木材样本，往往是维数超级大，一般都在10 000维以上，而百十个的样本数量相对超大的维数计算上完全可以忽略。

验证对于各算法的时间复杂度，抽取了 30 个测试样本用以比较算法之间时间关系。对比发现对于木材样本而言，KPCA 相比 PCA 速度明显快了不少，而 LDA 因为维数过大导致内存超出。

表 1　算法时间对比

KPCA/s	PCA/s	LDA
0.030 0	0.230 0	内存溢出

3.2　KPCA 对木材样本降维性能度量

设训练样本 X 为 $n\times m$ 维数据，包含 n 个 m 维的原始数据样本。规定单位维数与样本个数下，信息保留量（A_C）与训练时间（t），降维后的维数 N 之比为衡量木材降维效果的度量衡。即

$$Q_{(X,N)}=\frac{A_C}{t} \tag{22}$$

式(22)中，Q 越高，则比较的算法性能则越好。在木材样本的测试中，固定样本为 35×112 000 的木材样本。算法性能如下图所示：

图 3 中，数据显示这种衡量标准具有稳定性，且清楚表明在木材缺陷识别的样本中，KPCA 的效率值远远大于 PCA 的效率值。

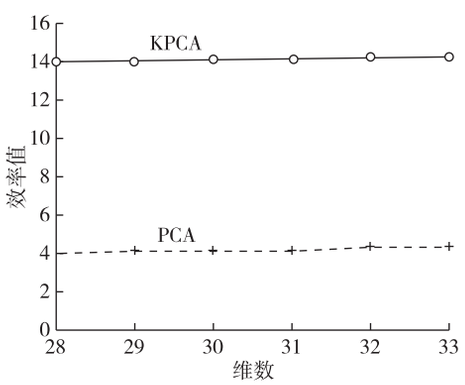


图 3　效率值对比图

3.3　SVM 对木材缺陷数据的识别优化

SVM 主要由学习能力和泛化能力两大块组成，对于木材缺陷数据，则需要再学习能力与泛化能力上寻求一个最优的平衡。常见的平衡调节是通过惩罚因子 C 来调节，通过 C 的大小来选择哪种能力占主要地位。

$$\min\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)+C\sum_{i=1}^{n}\xi_i \tag{23}$$

但是通过惩罚因子 C 的调节只能是调节两者比重大小，而不影响两者本身。

因此，本文目的在于在核函数中寻求学习能力与泛化能力的一种新的平衡。RBF 核是局部核函数，学习能力好，泛化能力差。多项式核函数是全局函数，学习能力差，泛化能力好。在构造核函数时，通过预测样

本类别与实测样本类别的比较,选择预测精度高的核函数作为模型的核函数选择。实验表明,在与实测样本对比的情况下,多项式的核函数相比较 RBF 核函数有较高的预测精度。

通过对不同核函数的选择与实验,能够为木材缺陷样本在分类识别上提供了更加精确的预测。使得学习能力与泛化能力之间的平衡调节不再单一。

4 实 验

4.1 KPCA 对数据样本降维处理优化

实验中选取了 300 个训练样本,每个样本为 280×400 的数据。部分样本图片如图 4 所示。为了在降维时使数据保留最大的信息,同时保证维数足够小,实验时分别统计了 PCA 和核 PCA 不同的维数时所保留的信息量。

图 4 部分训练样本图片

在图 5 各个方法维数与贡献率构成的二维图中,观察可知 KPCA 使用多项式核时,调整参数 $d=3$ 时,效果最佳,相同的维数比较其他方法、参数及核,能够保留更多的信息。

从图 6 中可以发现,通过核 PCA 降维后,在 177 维前保留的累计贡献率基本稳定增长,在 177 维之后增幅缓慢。且在 177 维时累计贡献率可达到 0.971 2,足够后续实验的进行。

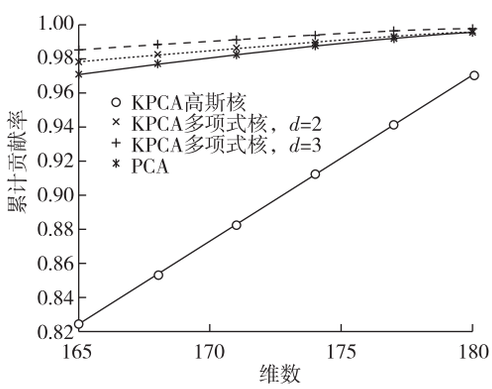


图 5 各算法维数-贡献率关系图

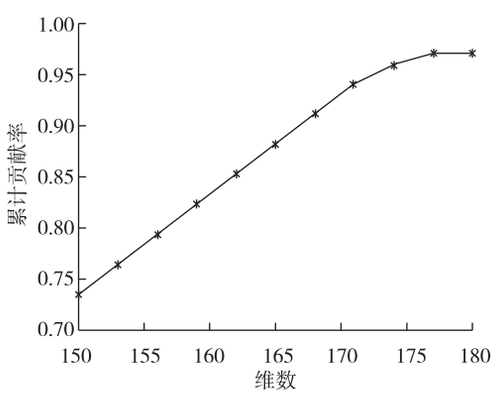


图 6 KPCA 多项式核维数-贡献率关系图

4.2 SVM 对木材分类应用

在对数据进行降维处理之后,实验采用了 841 个样本作为训练样本,其中包括 439 个缺陷木材的样本,402 个正常木材样本,然后又采集了 438 个缺陷木材样本与 121 个正常样本作为测试数据,通过 SVM 来选择一个最优模型[17]。实验木材样本数据见表 2。

通过 841 个真实训练数据样本与 559 个真实测试样本,保证了数据的公平,降低偶然性。为了确定最好的模型,下面分别对 3 种核函数参数进行试验选择。通过交叉验证的正确率预测精度,以及支持向量点数选择最优模型。

多项式核函数、RBF 核函数、Sigmoid 核函数见表 3~表 5。

表 2 实验木材样本数据表

训练数据	缺陷样本	439	841
	正常样本	402	
测试数据	缺陷样本	438	559
	正常样本	121	

表 3　多项式核函数$(g \cdot u'v + r)^d$

参数 g	参数 r	参数 d	预测精度/%	支持向量点	交叉验证(10 折)正确率/%	最优参数
0.01	1	1	98.235 4	121	92.271 1	
0.01	1	2	98.235 4	101	90.725 3	
0.01	1	3	98.465 1	90	90.725 3	
0.01	5	3	98.235 4	75	91.795 5	
0.01	10	3	98.568 8	73	92.508 9	
0.001	10	3	99.463 3	54	93.460 2	√
0.000 1	10	3	99.477 8	59	93.698 0	

表 4　RBF 核函数参数调优结果 $\exp(-g|u-v|^2)$

参数 g	预测精度/%	支持向量点	交叉验证(10 折)正确率/%
0.000 1	92.508 9	756	68.371 0
0.000 5	90.725 3	756	52.318 7
0.001 0	90.725 3	756	52.318 7

表 5　Sigmoid 核函数参数调优结果 $\tanh(g \cdot u'v + r)$

参数 g	参数 r	预测精度/%	支持向量点	交叉验证(10 折)正确率/%
0.001	0	91.795 5	720	52.318 7
0.001	10	91.795 5	720	52.318 7
0.002	0	91.795 5	720	52.318 7
0.010	0	91.795 5	720	52.318 7

表(3)～表(5)是部分具有代表性的实验结果统计。在实验中，先固定所有参数，然后对每一个参数依次调节，当找到最优参数时，再对下一个参数进行调节，依次找到每个参数的最优值。通过对 3 个核函数参数选择与调优，对比表(3)～表(5)，得出结论，在对木材样本的分类中，多项式核函数显然更具优势。在多项式核参数调优过程中，实验表明预测精度基本都已经到了一个很高的水平，在 559 个样本测试过程中，对参数的调整只会影响极个别的样本点的分类错误。所以多项式核是很符合木材分类的一个核函数。在对支持向量点和平均交叉验证正确率的参考下，最终选择参数使 $(0.001x_i^{\mathrm{T}}x_j + 10)^3$ 作为核函数，构建出对于木材缺陷识别的最优模型。

4.3　实验方法对比

在 KPCA 与 SVM 结合的木材缺陷识别分类中，实验选择最优结果的多项式核函数作为 KPCA 的核函数。实验选取了 BP-RBF 混合神经网络、卷积神经网络等近年来对木材识别应用的创新方法作为对比。各木材缺陷识别方法精度对比结果见表 6。

表 6　各木材缺陷识别方法精度对比

方法	精度(10 次平均值)/%
RBF 神经网络	96.2
BP-RBF 混合神经网络	98.0
卷积神经网络	96.9
KPCA-SVM	98.6

同样采用上面的 841 个的真实训练数据样本与 559 个的真实测试数据进行训练测试。为了减小误差等带来的不稳定因素，实验对每一种方法进行了 10 次试验测试，通过取到各方法精度的平均值来进行对比。

在 RBF 神经网络方法中，实验构建的神经网络模型为 3 层。在 BP-RBF 混合神经网络中利用 BP 神经网络良好的数据压缩能力与 RBF 神经网络对数据较好的逼近效果，使得分类精度有不错提升。在卷积神经网络中，由于迭代次数与参数选择可能会导致方法精度有较大波动。在本文提出的方法中，10 次试验的精度都很稳定，偏差极小。

从以上方法对比中，试验表明本文提出的 KPCA 与 SVM 在木材缺陷识别的方法应用具有较高的精度，并且多次试验偏差极小。

5 结 论

核方法是对维数处理、解决非线性问题的一个高效便捷的方法。在木材识别上,核 PCA 降维方法能够极大限度地保留样本的原始信息。然后再通过基于多项式核函数的支持向量机,构造出了一个分辨率达到99%以上的模型。由于现实生活中符合线性条件的数据模型少之又少,而 KPCA 和 SVM 都通过非线性的方法来实现降维和分类,效果相对普通常用线性方法有显著提高。在实验中预测正确率虽然有了很大的提高,但是在保证预测精度的同时优化时间复杂度仍是今后研究的重点。

参考文献:

[1]郭益轩. 关于提高木材利用率的几点建议[J]. 林业经济,1988(2):61-62.
[2]陈志林,傅峰,叶克林. 我国木材资源利用现状和木材回收利用技术措施[J]. 中国人造板,2007(5):1-3.
[3]牟洪波. 基于 BP 和 RBF 神经网络的木材缺陷检测研究[D]. 哈尔滨:东北林业大学,2010.
[4]王玉珏. 基于颜色特征木材缺陷检测的研究[D]. 哈尔滨:东北林业大学,2010.
[5]徐姗姗,刘应安,徐昇. 基于卷积神经网络的木材缺陷识别[J]. 山东大学学报(工学版),2013(2):23-28.
[6]白雪冰,许景涛,郭景秋,等. 基于局部二值拟合模型的板材表面节子与虫眼的图像分割[J]. 浙江农林大学学报,2016(2):306-314.
[7]方超. 木材缺陷的图像检测技术[D]. 哈尔滨:哈尔滨工程大学,2010.
[8]王光林,范玉玲,邵新建. 浅析木材检验中节子缺陷对材质的影响[J]. 科技创新与应用,2013(8):35-36.
[9]PEARSON K. Onlines and planes of closest fit to systems of points in space[J]. Philosophical Magazine,1901, 2 (6):559-572.
[10]FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics,1936, 7 (2): 179-188.
[11]于成龙. 基于 PCA 的特征选择算法[J]. 计算机技术与发展,2011(4):123-125.
[12]牟少敏. 核方法的研究及其应用[D]. 北京:北京交通大学,2008.
[13]高绪伟. 核 PCA 特征提取方法及其应用研究[D]. 南京:南京航空航天大学,2009
[14]李庆震,祝小平. 基于核 PCA 的智能图像分析算法[J]. 弹箭与制导学报,2007(5):189-192.
[15]CORTES C, VAPNIK V. Support vector networks[J]. Machine Learning, 1995, 20:273-295
[16]吴晓婷,闫德勤. 数据降维方法分析与研究[J]. 计算机应用研究,2009(8):2832-2835.
[17]张召,业宁,业巧林. 基于纹理提取和 SVM 技术的自动木材缺陷识别[J]. 计算机工程与应用,2009,23:219-223.

(责任编辑:李艳)