文章编号:1006 - 5911(2003) S0 - 0144 - 05

支持向量机增量学习的算法与应用

曾文华[1],马 健[2]
(1. 厦门大学计算机科学系,福建 厦门 361005;
2. 杭州电子工业学院计算机分院,浙江 杭州 310037)

摘 要:提出一种新的支持向量机的增量学习算法,分析了新样本加入训练集后支持向量集的变化情况。基于分析结论提出一种新的学习算法。研究了基于支持向量机的山羊绒和细支绵羊毛动物纤维图像识别问题,根据山羊绒和细支绵羊毛动物纤维图像的特点,分别采用自动阈值分割和 Top - Hat 变换,得到纤维边缘和鳞片边缘。仿真结果表明,基于支持向量机的动物纤维图像识别率高于传统的基于人工神经网络的识别率。

关键词:支持向量机;增量学习算法;动物纤维;图像识别
中图分类号:TP181 **文献标识码**:A

0 引言

由 V. Vapnik 等人提出的统计学习理论(Statistics Learning Theory,SLT),是一种小样本的机器学习理论,为机器学习问题提供了统一的框架[1~3]。通过对学习一致收敛问题的讨论,SLT 给出了学习机推广能力的界,进而给出了不同于传统学习中经验风险最小化准则(Empirical Risk Minimization,ERM)的结构风险最小化准则(Structure Risk Minimization,SRM),并在此基础上给出了一种新的学习算法——支持向量机(Support Vector Machines,SVM)。SVM 通过非线性核函数,将输入空间映射到高维线性特征空间,构造样本的最优分类面,最优分类面的构造过程即是实现 SRM 准则的过程。SVM 基于 SLT 坚实、严谨的理论基础,比传统学习方法具有较好的学习性能和泛化能力。SVM 学习采用优化方法得到的结果是全局最优解,不会产生过学习和局部最小等问题。

我国是纺织用动物纤维种类资源比较丰富的国家,山羊绒、细支绵羊毛等动物纤维产品在数量和质量上均居世界前列。由于各种纤维的品质和价格相差悬殊,这就使产品质检中纤维含量的检测十分重要。与其他纤维相比,山羊绒与细支绵羊毛在结构、外观形态、理化性能上都较为接近,因而正确、有效地鉴别这两种纤维就显得非常重要。

目前,鉴别山羊绒和细支绵羊毛的方法主要有四种:光学投影显微镜法、扫描电镜法、溶液解法和计算机图像分析法。前三种方法具有操作复杂、不能较好地反映特征以及鉴别操作带有主观性等不足之处。计算机图像分析法,通过对纤维数字图像的处理分析,从中提取多种定量特征,克服了其他方法的单一标准鉴别的缺点,客观、准确、快速,越来越受到关注,被认为具有很好的前景。

针对山羊绒与细支绵羊毛混纺产品质检中纤维含量的检测问题,本文提出了基于支持向量机的动物纤维图像识别方法。基于 SVM 寻优问题的 KKT 条件和样本之间的关系,分析了样本增加后支持向量集的变化情况,基于分析结论,提出了一种新的 SVM 增量学习算法。根据对采集的纤维图像特点的分析,分别利用自动阈值分割和灰值形态学方法,

基金项目:工业控制技术国家重点实验室开放实验室研究课题资助项目(K01001)。
作者简介:曾文华(1964 -),男,江苏兴化人,厦门大学计算机科学系博士,教授,主要从事智能控制、神经网络、机器学习等研究。
E - mail:whzeng @xmu. edu. cn。

设计了纤维边缘与鳞片边缘的自动预处理方法和特征提取算法。在此基础上,采用新的 SVM 增量学习算法,实现了这两种动物纤维的图像识别。

1　一种新的 SVM 增量学习算法

1.1　支持向量机

给定分类问题,其样本集为$\{X_i, y_i\}$, $i = 1, \cdots, l$, $X_i \in R^n$, $y_i \in \{\pm 1\}$。设超平面方程为:

$$wX + b = 0 \tag{1}$$

则决策函数为:

$$f(x) = \mathrm{sgn}(wX + b) \tag{2}$$

求最大分类间隔的超平面问题,可以转化为如下二次寻优问题[4]:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i \\ &\text{s.t.} \quad y_i(wX_i + b) \geq 1 - \xi_i \\ &X_i \in R^n, y_i \in \{\pm 1\}, i = 1, \cdots, l \end{aligned} \tag{3}$$

其对偶问题为:

$$\begin{aligned} &\max \sum_{i=1}^{l} \alpha_i - \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j X_i X_j \\ &\text{s.t.} \quad 0 \leq \alpha_i \leq C \\ &\sum_{i=1}^{l} \alpha_i y_i = 0, i = 1, \cdots, l \end{aligned} \tag{4}$$

α为 Lagrange 乘子。

由上述问题得到最优解α,则 SVM 的分类函数为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i (X \cdot X_i) + b\right) \tag{5}$$

如果分类问题是非线性的,则采用一种称为核函数(kernel function)的方法,使输入空间映射到高维的核函数特征空间,将非线性问题转化为该空间中的线性分类问题。此时对偶问题为:

$$\max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(X_i, X_j) \tag{6}$$

决策函数为:

$$f(X) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(X_i, X) + b\right) \tag{7}$$

式中:$K(\ )$为核函数。

1.2　KKT 条件

对偶问题的最优解$\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_l]$,使得每个样本 X 满足优化问题的 KKT 条件[5]为:

$$\alpha_i = 0 \Rightarrow y_i f(X_i) \geq 1 \tag{8}$$

$$0 < \alpha_i < C \Rightarrow y_i f(X_i) = 1 \tag{9}$$

$$\alpha_i = C \Rightarrow y_i f(X_i) \leq 1 \tag{10}$$

式中:非零的α_i为支持向量。

考虑函数系 $f(X) = h$,可知 $f(X) = 0$ 为分类面,$f(X) = \pm 1$ 为分类间隔的边界,其上的样本即为支持向量。

定理 1　对样本集进行训练得到 SVM 分类器,α为 Lagrange 乘子。$\alpha = 0$ 对应的样本分布在分类器分类间隔之外,$0 \leq \alpha \leq C$ 的样本位于分类间隔之上,$\alpha = C$ 位于分类间隔之内[5]。即是:

$$\alpha_i = 0 \Rightarrow | f(X_i) | \geq 1 \tag{11}$$

$$0 < \alpha_i < C \Rightarrow | f(X_i) | = 1 \tag{12}$$

$$\alpha_i = C \Rightarrow | f(X_i) | \leq 1 \tag{13}$$

1.3　增量学习后支持向量集变化分析

当核函数类型及其参数确定后,支持向量集可完全描述整个样本集的分类特征,支持向量集和训练样本集之间的等价关系可以得到证明。通常,支持向量集只是样本集的一小部分,因此,通过对支持向量集的研究可知,增量学习是可行且有效的。如果新增样本带有原样本集不包含的分类信息,则学习后支持向量集必然发生变化,以体现新信息的加入。如果考虑新增样本对支持向量集的影响,则应考虑以下问题:①什么样的新增样本使支持向量集发生变化?②这种变化是怎样的?③新增样本集是如何组成的?

定理 2　$f(X)$为 SVM 分类决策函数,$\{X_i, y_i\}$为新增样本。满足 KKT 条件的新增样本将不会改变支持向量集;而违背 KKT 条件的新增样本将使支持向量集发生变化。违背 KKT 条件的样本可以分为三类(见图 1):

(1)位于分类间隔中,与本类在分类边界同侧,可被原分类器正确分类的,满足 $0 \leq y_i f(X_i) < 1$;

(2)位于分类间隔中,与本类在分类边界异侧,被原分类器错误分类的,满足 $-1 \leq y_i f(X_i) \leq 0$;

(3)位于分类间隔外,与本类在分类间隔异侧,被原分类器错误分类的,满足 $y_i f(X_i) < -1$。

图 1 中,方形标记代表'$y = +1$';十字标记代表'$y = -1$';$A1$, $A2$ 和 $A3$ 是新增样本。

定理 2 表明,对新增样本再学习得到新的 SVM 分类器时,KKT 条件比分类函数的分类判断更合理,分类错误是样本违反 KKT 条件的特定情况。定理 2 告诉我们,只有违背 KKT 条件的样本,才会影响增量学习后的支持向量集。因此,新增样本可

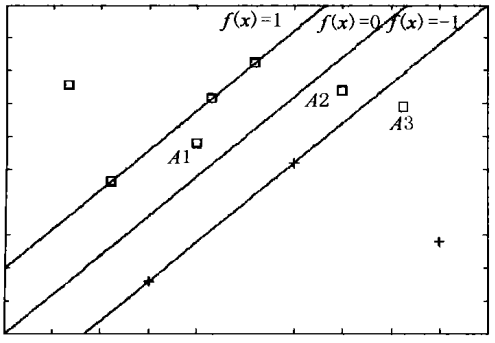


图1 新增样本与KKT条件的关系

以分为两部分:违背 KKT 条件的样本和满足 KKT 条件的样本,而后者由于包含的信息已经被原分类器所反映,因此可以不被学习。

定理 3 新增样本违背 KKT 条件,则原样本集中非支持向量可能转化为支持向量[5]。

定理 3 的证明可以由图 2 所示的特例来说明。原来的支持向量集由 $S1$～$S5$ 组成,$A1$～$A3$ 为新增样本。对新样本训练后,由图 2 可以直观地判断:$S1$,$A3$ 和原来的样本 $N1$ 组成新的支持向量集。因此增量学习中,只考虑新增样本和原来的支持向量集的做法虽然符合定理 2,但定理 3 表明,这样可能丢失原来样本集中的信息,而且不能有效地舍弃无用样本。

图 2 中 $S1$～$S5$ 为原支持向量集,$A1$～$A3$ 为新增样本集。训练后的支持向量集由 $S1$, $A3$ 和 $N1$ 组成,其中 $N1$ 为原样本集中的样本。

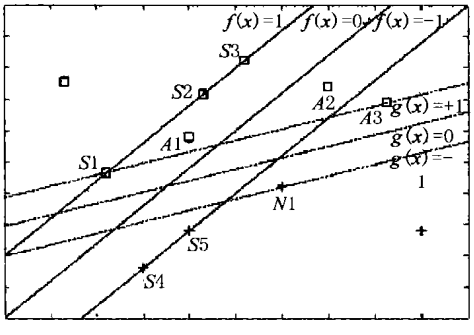


图2 对新增样本训练后支持向量集可能发生的变化

实际上,不是把新增样本加入原来的样本集,而是将它们结合在一起,两者的地位是等同的。因此,如果由新增样本集训练得到一个新的 SVM,则原来的样本集可以看作新 SVM 的新增样本,其中可能存在违背新 SVM 的 KKT 条件的样本,而这些样本可能成为增量学习后的支持向量。

1.4 增量学习算法

基于以上分析,我们提出一种新的增量学习方案。首先检验是否存在新增样本违背原来的 KKT 条件,如果存在,则由新增样本得到一个新的 SVM,接着检验是否存在原来的样本违背新的 KKT 条件,如果仍然存在,则由两个 SVM 的支持向量集加上两次违背条件的样本,构成新的训练集。最后由新的训练集得到最终的 SVM。

本方案考虑的不仅是分类错误,而且检验样本是否违背 SVM 的 KKT 条件,原来样本集中新的样本集的信息均被获取,而不违背 KKT 条件的样本被舍弃,保留可能有价值的样本,用来进行新的训练。

在方案中,可能需要对新增样本单独作一次训练,来得到新的 SVM,而整个过程对全部样本只做一次违背 KKT 条件的验证。如果 SVM 的决策函数为 $f(X_i)$, $y_i \in \{\pm 1\}$ 为类别号,违背 KKT 条件等价于 $y_i f(X_i) < 1$。接下来我们将给出一种增量学习算法,由于多类问题可以转化为多个二类问题,因此,这里只考虑基本的二类问题。具体算法如下:

6前提:

原样本集 X_0 和由其训练得到的 SVM 分类器 Ω^0, X_0^{SV} 表示 Ω^0 的支持向量集, X_I 表示新增样本。

6运算过程:

(1) 检验 X_I 中的样本是否违背 Ω^0 的 KKT 条件,如果样本没有违背,则算法停止,Ω^0 为增量学习结果。否则,根据检验结果, X_I 被分为 X_I^V 和 X_I^S。X_I^V 表示违背 Ω^0 的 KKT 条件的样本集合, X_I^S 表示满足 Ω^0 的 KKT 条件的样本集合。

(2) 由 X_I 得到新的 SVM 分类器 Ω^I, X_I^{SV} 表示 Ω^I 的支持向量集。

(3) 检验 X_0 中的样本是否违背 Ω^I 的 KKT 条件,如果没有样本违背,则算法停止,Ω^I 为增量学习结果。否则, X_0 被分为 X_0^V 和 X_0^S 两部分, X_0^V 表示违背 Ω^I 的 KKT 条件的样本集, X_0^S 表示满足 Ω^I 的 KKT 条件的样本集。

(4) X_U 表示 X_0^{SV} Y X_I^{SV} Y X_0^V Y X_I^V,由 X_U 得到新的 SVM 分类器 Ω 为增量学习结果, $X_H = X_0^S$ Y X_I^S 作为历史信息将被舍弃。

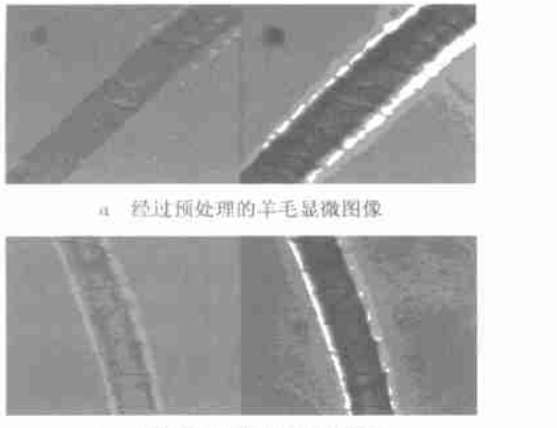
2 动物纤维图像的预处理和特征提取

2.1 图像的采集

实验所用图像由普通配有数码摄像头的显微镜采集。采集图像的格式为 BMP 图像,经过处理后,得到一组 256 ×256 象素 256 色的位图。

2.2　图像的预处理

由于采集环境和设备条件,图像中纤维边缘和鳞片边缘存在光影,同时图像受到了噪声的影响。因此,必须对采集的图像进行预处理。图 3 为经过预处理的图像效果。



a　经过预处理的羊毛显微图像

b　经过预处理的羊绒显微图像

图3　图像预处理效果

对以上得到的图像进行灰值形态学的变换,数字图像中的鳞片边缘表现为脉冲信号。当对图像中的边缘信息采用传统的边缘提取时,由于边缘卷积算子的窗口固定,而图像中的边缘宽度不定,极易产生一条边缘提取出两条边界线,以及受到微小噪声干扰的现象。通过分析不难看出,图像中边缘数字信号同构于数字形态学中的峰谷结构,因此,可以采用 Top - Hat 变换来提取。

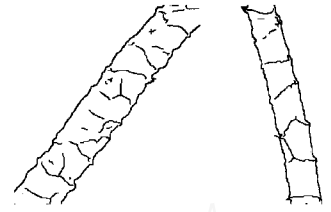
在 Top - Hat 变换中,采用常用的扁平结构元素,为了保证开运算中腐蚀作用于整个图像,假设图像外的值为无穷大。Top - Hat 变换后,图像的灰度级较低,图像较暗,通过加大图像的亮度,再采用阈值变换,将纤维的图像提取后,细化得到最终的结果,如图 4 所示。

a　羊绒鳞片毛　　b　羊绒鳞片

图4　鳞片提取效果

在采用 Top - Hat 变换中,动物纤维的外边缘没有提取出来。从图像的灰度直方图中可以发现,该图中存在典型的双峰结构。因而可以采用自动阈值方法,将纤维图像从背景中分离出来,再得到图像的外边缘,即为纤维的边缘图像。

最后,将鳞片边缘图像和纤维边缘图像叠加,得到自动处理的最终结果,如图 5 所示。



a　羊毛　　b　羊绒

图5　羊毛和羊绒的鳞片和边缘叠加图片

2.3　图像特征的提取

预处理结果的两个图像,含有纤维边缘线条和鳞片边缘线条。将图像中的线条按照八连通方法进行跟踪,得到全部线条的跟踪链表,链表中含有线条的长度、起始点和终止点的坐标。这将和原图像作为图像特征计算的基础。

(1)纤维边缘数据提取　从边缘预处理结果中不难发现,纤维的边缘是图像中最长的两条线条。因此,首先找到所有线条中最长的两条即为纤维的边缘,得到起始坐标。结合链表的起始坐标和原图像,采用同样的八连通跟踪,可以找到边缘上的所有点集合。纤维图像的边缘数据包括纤维宽度、边缘线与水平倾角和边缘线长度。

(2)鳞片特征数据提取　鳞片边缘图中的全部线条都被记录到,根据经验,太短的线条不是鳞片边缘,根据经验阈值,将它们从链表删除。根据始末点判断鳞片的方向,与前面得到的纤维方向比较,小于纤维宽度一定倍数的,认为不是鳞片边缘,将其删除。最后得到的链表就是鳞片数据,链表节点数目就是鳞片边缘数目。由于处理结果为鳞片边缘数目基本等于鳞片数目,因此可以将链表节点数目作为鳞片的数目。这将作为进一步特征提取计算的基础。纤维图像的鳞片特征数据包括鳞片平均间隔、鳞片的弯曲度、鳞片间平均接触长度和鳞片的边缘形态。

以上纤维边缘和鳞片间隔加起来,总共得到的向量为七维向量。

3　基于 SVM 的动物纤维图像识别方法

我们将实际系统分为基本功能层、内部处理层和人机界面三层,其中,①基本功能层,包含基本图像读写、处理、特征提取模块函数库和智能函数库两部分;②内部处理层,包含图像读写模块、图像预处理、图像特征提取三部分;③人机界面,负责人机交

互。系统的每个模块功能介绍如下:

(1) 基本图像读写处理特征提取模块　对不同格式图像文件的读写,显示采集到的图像;具有基本的图像处理功能,包括直方图操作、图像阈值变换、二值、灰度形态学、边缘检测等功能;具有基本的特征提取函数库。

(2) 图像预处理模块　提供针对特种动物纤维识别的处理方法。

(3) 特征提取模块　由图像处理模块的结果,根据上面介绍的方法得到图像的特征数据,并对数据进行管理。

(4) 智能识别模块　对特征提取模块的数据,根据训练好的学习模型进行识别,以及对数据模型再训练。其由智能函数库支持,包括 SVM,MLP 的学习训练算法,以及其他的相关算法。可以扩充其他的学习算法。

(5) 人机界面　提供以上全部模块的人机接口,可以对以上各部分进行干预,为人自由操作系统提供基础。

只需给定符合要求的图像,系统即可按照内部处理层中的图像读取、预处理、特征提取和智能识别的顺序,依次完成处理和识别,最后得到识别结果。整个流程由底层的图像处理函数库、特征提取函数库、智能技术函数库支持。另外,人机交互模块使系统的功能更加灵活。设计采用的整体分层和功能分块具有很好的扩展性,便于实现、管理和功能扩充。

4　结束语

我们讨论了样本与 KKT 条件之间的关系,基于讨论结果,分析了新增样本加入后,可能引起的支持向量集的变化,进而提出了一种新的 SVM 增量学习算法。实验证明,本算法可以在增加新增样本后有效压缩样本集的大小,舍弃无用样本,同时分类能力和泛化性能不受影响。

对于某种问题,经过几次增量学习后,新增样本含有的有用信息将越来越少,这时采用增量学习的开销收效甚微。如何提取少量信息,将是需要进一步研究的问题。在本算法中,新样本按先后顺次训练,而当训练集为一个大型的样本库时,可以将其分为几个合适的小部分,对各部分再采用并行学习方法。因此,本算法在并行学习上的应用,也是今后进一步研究的方向。

参考文献:

[1] RATSABY J. Incremental learning with sample queries[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 883 - 888.

[2] WANG E H C, KUH A. A smart algorithm for incremental learning[J]. International Joint Conference on Neural Networks, 1992, 3: 121 - 126.

[3] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer - Verlag, 1995.

[4] CHRISTOPHER J, BURGES C. A tutorial on support vector machines for pattern recognition[M]. Boston: Kluwer Academic Publishers, 1998.

[5] ZHOU Weida, et al. An analysis of SVMs generalization performance[J]. Acta Electronica Sinca, 2001, 29(5): 590 - 594 (in Chinese). [周伟达, 等. 支撑矢量机推广能力分析[J]. 电子学报, 2001, 29(5): 590 - 594.]

[6] CHANG Chihchung, LIN Chihjen. LIBSVM: a library for support vector machines [DB/OL]. http://citeseer.nj.nec.com/chang01libsvm.html, 2001 - 09 - 07.

An Incremental Learning Algorithm for Support Vector Machine and its Application

ZENG Wen - hua[1], *MA Jian*[2]

(1. Dep. of Computer Sci., Xiamen Univ., Xiamen 361005, China;
2. Sch. of Computer Sci., Hangzhou Inst. of Electronics Engineering, Hangzhou 310037, China)

Abstract: In this paper we present a learning algorithm for incremental support vector machine (SVM). We analysis the possible changes of support vector set after new samples are added to training set. Based on the analysis result, an algorithm is presented. This algorithm is applied for the image recognition of two special animal fibers. With the algorithm, the useless sample is discarded and knowledge is accumulated. The experiment result shows that this algorithm is more effective than the traditional SVM while the classification precision is also guaranteed.

Key words: support vector machine; incremental learning algorithm; animal fiber; image recognition

Foundation item: Project supported by the National Key Lab. of Industrial Control Tech., China (Grant No. K01001).