文章编号:1001 - 9081(2006)12 - 2887 - 03

多超球面 OC-SVM 算法在隐秘图像检测中的应用

唐玉华,杨晓元,张敏情,韩 鹏
(武警工程学院 网络与信息安全重点实验室,陕西 西安 710086)
(tangyuhua8201@163.com)

摘 要:针对二类支持向量机分类器在图像密写分析应用中训练步骤复杂与推广性弱的缺点,把一类支持向量机(OC-SVM)引入算法,提出一种基于核的多超球面 OC-SVM 算法。算法利用核空间中样本特征差异突出的特性,首先对样本在核空间进行 K-均值聚类,然后使用 OC-SVMs 对各子类训练建立多超球面分类模型,实现分类判决。实验结果表明,算法有效地实现了对隐秘图像的盲检测,提高了检测精度。

关键词:盲检测;图像密写分析;核 K-均值聚类;多超球面;一类支持向量机
中图分类号: TP309 **文献标识码**:A

Detection of stego images using one-class support vector machines with multiple hyperspheres

TANG Yu-hua, YANG Xiao-yuan, ZHANG Min-qing, HAN Peng
(*Key Laboratory of Network and Information Security of the Engineering College of the APF, Xi'an Shaanxi* 710086, *China*)

Abstract: In order to reduce the complexity and weak generalization of classification method using two class support vector machines in images steganalysis, a new Kernel-based classification method using OC-SVMs with multiple hyperspheres was put forward. Considering that the data features were expected to be more separable in kernel space, we first performed the K-means clustering in kernel space, then trained the sub-class data separately using OC-SVMs and established a multiple hyperspheres classification model to decide the class label of new data. The experimental results show that this method has efficiently improved the classification precision.

Key words: blind detection; images steganalysis; Kernel-based K-means clustering; multiple hypersphere; One-Class Support Vector Machines (OC-SVM)

0 引言

支持向量机是建立在 VC 维理论和结构风险最小原理基础上的一种统计学习模型。它根据有限的样本信息在模型复杂性和学习能力之间寻求最佳折中,以获得最好的推广能力,在解决小样本,非线性及高维模式识别问题中表现出许多特有的优势[1]。

目前,在图像隐秘检测系统中,主要采用的是基于二类样本的支持向量机,这种方法需对每种隐秘算法分别建立正常图像和隐秘图像两类分类器,训练步骤复杂,系统效率低。当隐秘算法没有公开或难以获取时,这种方法将不能进行隐藏检测。本文提出的算法采用一类支持向量机[2]分类器,只需对正常真彩图像进行训练,提高了检测系统的效率,适用于多种隐秘算法。

把一类支持向量机应用于图像密写分析中,只需对一类样本训练建立超球面分类模型,有效地解决了隐藏图像样本缺乏的分类问题。但是,同一类样本的特征也存在着部分差异,如果只按照单超球面分类模型分类,则可能将一些非正常的样本错误地判别为正常样本。本文提出了一种基于核K-均值聚类的多超球面一类支持向量机分类算法,实验证明算法有效地降低了错分率,提高了分类精度。

1 一类超球面支持向量机

考虑 n 个 d 维训练样本点,记作$\{\vec{x}_1,\vec{x}_2,\cdots,\vec{x}_n\}$,一类支持向量机通过特征映射 $\phi:R^n \to F$ 将样本投影到一个高维特征空间中,并建立一个体积尽量压缩且包含尽可能多的训练样本点的球面,即超球面。超球面由球心 c 和半径 r 来描述。

原问题可描述为解如下二次优化问题(Primal Form):

$$\min_{c,r,\xi} r^2 + \frac{1}{vn}\sum_i \xi_i$$
$$\text{s.t.} \quad \|\phi(\vec{x}_i) - c\|^2 \leqslant r^2 + \xi_i, \xi_i \geqslant 0,\ i = 1,\cdots,n \tag{1}$$

其中,$v \in (0,1)$ 控制落在超球面外训练样本的比例,对超球面的半径和它所包含的训练样本数目进行折中。$\|\cdot\|$ 为欧式距离;引入松弛因子 ξ_i,在保证超球面最压缩的情况下,使训练样本点尽可能多地被包含在超球面中。

原问题可转化为对偶问题,并描述如下:

$$\min_{\alpha} \sum_{i,j} \alpha_i\alpha_j\phi(\vec{x}_i)^T\phi(\vec{x}_j) - \sum_i \alpha_i\phi(\vec{x}_i)^T\phi(\vec{x}_i)$$
$$\text{s.t.} \quad 0 \leqslant \alpha_i \leqslant 1/vn,\ \sum_i \alpha_i = 1 \tag{2}$$

收稿日期:2006 - 06 - 26 **基金项目**:计算机网络与信息安全教育部重点实验室课题资助(200409);国家自然基金资助项目(60473029);武警部队军事科研项目(wjk200608)

作者简介:唐玉华(1982 -),男,湖南人,硕士研究生,主要研究方向:信息安全、密写分析、数字水印; 杨晓元(1959 -),男,湖南人,教授,主要研究方向:密码学、信息安全; 张敏情(1967 -),女,陕西人,教授,主要研究方向:密码学、信息安全; 韩鹏(1982 -),男,新疆人,硕士研究生,主要研究方向:信息安全、密写分析、数字水印.

α_i 为 Lagrange 乘数。

解上述二次优化问题,得到α_i,继而可得超球面球心和半径:

$$c = \sum_{i=1}^{n} \alpha_i \phi(\vec{x}_i) \tag{3}$$

$$r^2 = \| \phi(\vec{y}) - c \|^2 \tag{4}$$

将(3)式代入(4)可得:

$$r^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \phi(\vec{x}_i)^T \phi(\vec{x}_j) - 2 \sum_{i=1}^{n} \alpha_i \phi(\vec{x}_i)^T \phi(\vec{y}) + \phi(\vec{y})^T \phi(\vec{y}) \tag{5}$$

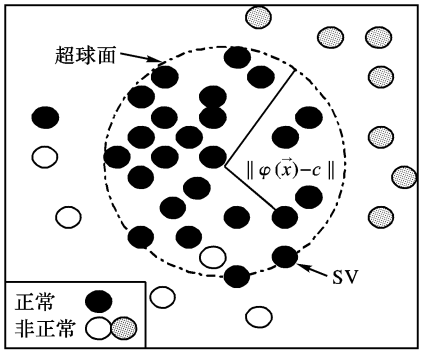

图1 单超球面一类支持向量机

图1中,黑点表示一类支持向量机训练样本,灰色和白色的点表示与训练样本不同类的样本,虚线圆表示超球面,其中球心为 c,半径为 r,任意一训练样本点 x 到球心的距离为 $\| \phi(\vec{x}) - c \|$,在虚线圆上的点称为支持向量点(SV)。

由超球面参数 c 和 r,建立决策方程 $f(\vec{x})$,判定样本点 $\vec{x}$ 是否在超球面内部,定义如下:

$$f(\vec{x}) = r^2 - \| \phi(\vec{x}) - c \|^2 \tag{6}$$

为了避免直接计算 $\phi(\vec{x})$,将(5)式代入,可得决策方程:

$$f(\vec{x}) = r^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \phi(\vec{x}_i)^T \phi(\vec{x}_j) + 2 \sum_{i=1}^{n} \alpha_i \phi(\vec{x}_i)^T \phi(\vec{x}) - \phi(\vec{x})^T \phi(\vec{x}) \tag{7}$$

当 $f(\vec{x}) \geqslant 0$ 时,判定样本点在超球面内,否则在超球面外。

2 基于核的 K-均值聚类算法

本文利用超球面一类支持向量机提出了基于核的 K-均值聚类改进算法,求得近似最优的聚类中心。

2.1 K-均值聚类算法

设有 N 个未知标记的样本 $\{x_1, x_2, \cdots, x_N\}$,K-均值聚类算法根据样本的特征向量将样本分为 K 类。设第 k 类的样本数目为 N_k,则样本总数 $N = \sum_{k=1}^{K} N_k$,每类的均值为 $(m_1, m_2, \cdots, m_K)$,则 $m_k = \frac{1}{N_K} \sum_{i=1}^{N_K} x_i, k = 1, \cdots, K$。K-均值聚类基于误差平方和准则,即 K-均值聚类最小化的目标函数为 $J_k = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \| x_i - m_k \|^2$。

2.2 核 K-均值聚类改进算法

核 K-均值算法利用一个非线性映射 $\phi: R^n \to F, \vec{x} \phi(\vec{x})$,将样本空间 R^n 中的样本 $\vec{x}$ 映射到一个高维的核空间 F 中,在核空间中进行 K-均值聚类。核空间中待分类样本为:$(\phi(\vec{x}_1), \cdots, \phi(\vec{x}_n))$,进行核 K-均值聚类就是最小化核空间中样本点 $\phi(\vec{x})$ 到聚类中心 c 的欧式距离:

$$D = \| \phi(\vec{x}) - c \|^2 = \| \phi(\vec{x}) - \sum_{i=1}^{n} \alpha_i \phi(\vec{x}_i) \|^2 \tag{8}$$

Kernel 函数定义为:

$$k(\vec{x}, \vec{y}) = \phi(\vec{x})^T \phi(\vec{y}) \tag{9}$$

典型的核函数类型有:Liner kernel、Polynomial kernel、RBF kernel、Sigmoid kernel。

将 Kernel 函数式(9)代入式(8),化简得:

$$D = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\vec{x}_i, \vec{x}_j) - 2 \sum_{i=1}^{n} \alpha_j k(\vec{x}_i, \vec{x}) + k(\vec{x}, \vec{x}) \tag{10}$$

改进的 K-均值聚类算法为:

1)初始化聚类中心:在核空间中任意选取 k 个样本,作为样本的初始聚类中心;

2)计算核空间中样本 $(\phi(\vec{x}_1), \cdots, \phi(\vec{x}_n))$ 到各聚类中心的欧式距离,根据最近邻原则分配到各类别中;

3)利用一类支持向量机对 k 个子类进行训练,求出各子类的中心 $c_m = \sum_{i=1}^{n_m} \alpha_i \phi(\vec{x}_i), m = 1, \cdots, k$;

4)重复步骤2和3,直至连续 n 次迭代后 c_i 不再改变,或者是核空间中样本 $(\phi(\vec{x}_1), \cdots, \phi(\vec{x}_n))$ 到各聚类中心的欧式距离不再减小(或变化很小)。

实验结果表明初始聚类中心的随机选择会对聚类结果产生较大影响。为此,采用文献[3]提出的核 K-均值聚类算法,来初始化聚类中心。步骤如下:

1)在核空间中任意选取 k 个样本,作为样本的初始聚类中心;

2)计算核空间中样本 $(\phi(\vec{x}_1), \cdots, \phi(\vec{x}_n))$ 到各聚类中心的欧式距离,根据最近邻原则分配到各类别中;

3)在各子类中,分别以每个样本为类中心,计算类内其他样本点到类中心的距离之和,记录距离之和最小值(*sum*)与所对应的样本点;

4)重复循环步骤1)~3)m 次(通常 $m \ll n$),这样每个类得到 m 个 *sum* 值,取 m 个 *sum* 值中最小的一个,取该点作为该类的中心。

通过此算法可以得到一组近似最优的初始聚类中心,这样就为一类支持向量机计算中心点 c 降低了计算复杂度。

3 基于 K-均值聚类的多超球面算法

在解决一类支持向量机多超球面问题时必须对初始样本进行聚类,采用普通的聚类方法不能体现样本在核空间的相似性,因此本文提出利用基于核的 K-均值聚类算法来建立多超球面。

3.1 一类支持向量机多超球面算法思想

我们所得到的正常样本特征的相似度不强,所以建立的单超球面不能包含更多的正常样本[4]。因此,本文研究建立多个子类超球面,替代一类支持向量机的单超球面,以更有效地识别样本。

利用上文的核 K-均值聚类改进算法,对样本在核空间聚类,得到近似最优聚类。然后,通过一类支持向量机对每个子类进行训练,建立子类超球面分类模型。检测时,用子类分类模型实现对待识样本的判别。

设定训练样本的标记都为 +1,测试用正常样本的初始化标记也为 +1,异常样本的初始化标记为 -1。判别规则如下:

当所有子类超球面分类模型对待识样本的判别标记都为 -1 时,待识样本为异常样本;否则,为正常样本(即 k 个子类超球面中任意一个或几个对待识样本判别为 +1 时)。

图 2 中虚线圆表示一类支持向量机的单超球面的平面示意图。从图中可以看出,此超球面错误地判别了一些非正常样本,而改进的一类支持向量机子类超球面(实线圆)则可以有效地识别出原来落于单超球面(虚线圆)中的错分样本。

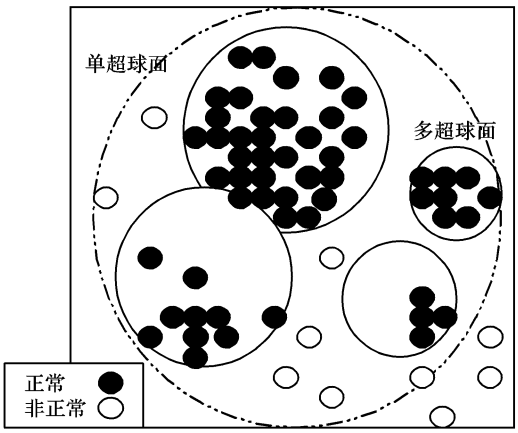

图 2　多超球面一类支持向量机

3.2　算法实现

本文在 Matlab 6.5 环境下,通过改进文献[5]的软件包,实现了多超球面一类支持向量机算法。

Libsvm 结合 SMO[6] 与 SVM^{light}[7] 算法实现了快速解决二次优化问题。它利用 SVM^{light} 工作集的思想,在每次循环中,将训练样本分为两个集合 B 和 N,B 为工作集 $B \subset \{1,\cdots,n\}$,$N \equiv \{1,\cdots,n\} \backslash B$。Libsvm 采用工作集的极端情况,每次只优化两个 Lagrange 乘数 α_i,α_j (SMO 算法),从而降低了解二次优化问题的难度。其主要优势在于可以从中得到一个解析解,而不需要优化软件求解。

把 Libsvm 的思想应用到一类超球面支持向量机上,可快速建立一类超球面支持向量机分类器。

4　实验结果及分析

表 1　多超球面一类支持向量机检测精度

软件	训练	测试	$k=1$	$k=2$	$k=3$	$k=4$	$k=6$	$k=8$
F5r11	Data0	Data0 1656	85.45	100.00	100.00	100.00	100.00	93.01
	1000	Data1 2650	20.29	63.21	44.27	39.53	32.98	17.62
H4PGP	Data0	Data1 1656	85.44	100.00	100.00	100.00	100.00	100.00
	1000	Data0 2650	13.87	63.32	64.75	44.27	39.53	11.67
Jsteg41	Data0	Data0 1656	80.13	100.00	100.00	100.00	100.00	100.00
	1000	Data1 2656	24.32	50.64	52.00	60.74	45.23	42.21

实验对F5r11、H4PGP、Jsteg41进行测试,分别对2560幅真彩图像进行随机消息嵌入,嵌入的消息大小从最大消息嵌入量的10% ~90% 之间随机选取。对两类图像进行特征提取,提取方法采用彩色小波包预测模型[4],每幅图像提取出一个108 维特征向量。对正常图像进行基于核的多超球面 OC-SVM 训练,训练时的核函数取 Linear 线性核,训练样本使用数据的前 1 000 个样本,建立基于单样本的分类器,聚类个数 $k \in \{1,2,3,4,6,8\}$。使用分类器对正常图像和隐秘图像进行分类,结果如表 1 所示。

表 1 中,$k=1$ 表示单超球面一类支持向量机算法的检测精度,$k=n, n>1$ 表示多超球面一类支持向量机算法的检测精度。结果显示,当 k 取适当值时($k>1$),多超球面一类支持向量机算法要比单超球面算法具有更好的检测精度。在此需要说明,k 的合理选取有赖于数据集本身的特性,可以通过反复实验记录一个经验值。

5　结语

本文提出一种基于核空间 K-均值聚类的多超球面一类支持向量机分类算法。算法将多超球面思想应用到隐秘图像盲检测分类问题当中,解决了一类支持向量机建立超球面模型中样本特征差异问题带来的不便。利用 Kernel 核空间聚类,突出样本特征差异,更好地实现了多超球面模型的建立。实验数据表明,该算法比单超球面一类支持向量机具有更高的精度。

参考文献:

[1] VAPNIK V. The Nature of Statistical Learning Theory[M]. Springer Verlag, 1995.

[2] SCHÖLKOPF B, SMOLA A, WILLIAMSON R, *et al*. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207 - 1245.

[3] 孔锐, 张国宣, 施泽生, 等. 基于核的 K-均值聚类[J]. 计算机工程, 2004, 30(11): 12 - 13.

[4] LYU S, FARID H. Steganalysis Using Color Wavelet Statistics and One-class Support Vector Machines[A]. SPIE Symposium on Electronic Imaging[C]. San Jose, CA, 2004.

[5] CHIH-CHUNG CHANG, CHIH-JEN LIN. LIBSVM: a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf, 2001 -06 -08.

[6] PLATT J. Using Sparseness and Analytic QP to Speed Training of Support Vector Machines[A]. Advances in Neural Information Processing Systems 11[C]. Cambridge, MA: MIT Press, 1999.

[7] JOACHIMS T. Making large - Scale SVM Learning Practical[A]. Advances in Kernel Methods - Support Vector Learning[C]. MIT-Press, 1999.

(上接第 2886 页)

种变换方式是变换图形的透视点,保持图形相对坐标轴不动,从而得到不同视角的透视图。在绘制完成后,还可以对图形进行明暗效果处理,并进行色彩填充,以获得真实的效果。该绘制算法可以进一步推广到其他三维图形的绘制,如凸多面体、凹多面体等。其中的消隐算法对于采用面集表示法的其他三维物体(如立方体等)也是适用的。

参考文献:

[1] 孙正兴, 周良, 郑宏源. 计算机图形学基础教程[M]. 北京: 清华大学出版社, 2004. 173 - 175.

[2] JANSSEN LT. A simple efficient hidden line algorithm[J]. Computer and Structures, 1983, 17(4): 563 - 571.

[3] SOUKKERS R, LAW HKW. On the hidden line removal problem [J]. Computers and Structures, 1987, 26(4): 709 - 717.

[4] 严蔚敏, 吴伟民. 数据结构[M]. 北京: 清华大学出版社, 1997. 273 - 277.

[5] PRESS WH, TEUKOLSKY SA, VETTERLING WT, *et al*. C++ 数值算法[M]. 第 2 版. 胡健伟, 赵志勇, 等译. 北京: 电子工业出版社, 2005. 248 - 251.

[6] SCHNEIDER PJ, EBERLY DH. 计算机图形学几何工具算法详解[M]. 周长发, 译. 北京: 电子工业出版社, 2005. 171 - 174.

[7] 孙家广. 计算机图形学[M]. 第 3 版. 北京: 清华大学出版社, 1998. 486 - 488.