具有合适拒识机制的高正确识别率分类器设计

杨国为[1,2],万鸣华[2,3],赖志辉[3],张凡龙[2],黄伟强[4]
(1. 青岛大学电子信息学院,山东青岛 266071;2. 南京审计大学信息工程学院,江苏南京 211815;3. 深圳大学计算机与软件学院,广东深圳518060;4. 香港理工大学,香港九龙 999077)

摘 要: 针对目前一些正确识别率高的SVM(Support Vector Machines)分类器、超球SVM分类器、深度学习分类器在一些典型样本集上应用时仍然有2%左右的错误识别率和增量学习功能不强的问题,本文提出了一种具有合适拒识机制的高正确识别率分类器设计方案和相应的增量学习算法,较好地解决了上述问题.主要工作包括:同类特征集合的紧密包裹集构造算法;基于同类特征集合和紧密包裹集的同类特征区域紧密包裹面的求解算法;设置所有紧密包裹面之外的公共区域为分类器的拒识区域的方法;当增加新类别、增减训练样本时,以上算法的增量学习算法.用uci数据集做对比实验表明,在拒识率小于1.3%的情况下,本文方法设计的分类器正确识别率大于99.13%.

关键词: 分类器; 模式识别; 支持向量机; 增量学习; 分类面; 包裹学习
中图分类号: TP3 **文献标识码:** A **文章编号:** 0372-2112(2021)08-1569-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20190518

High Correct Recognition Rate Classifier Design with Appropriate Rejection Mechanism

YANG Guo-wei[1,2], WAN Ming-hua[2,3], LAI Zhi-hui[3], ZHANG Fan-long[2], WONG Wai-keung[4]
(1. *School of Electronic Information, Qingdao University, Qingdao, Shandong* 266071, *China*;
2. *School of Information Engineering, Nanjing Audit University, Nanjing, Jiangsu* 211815, *China*;
3. *School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong* 518060, *China*;
4. *Hong Kong Polytechnic University, Kowloon, Hong Kong* 999077, *China*)

Abstract: At present, some SVM(Support Vector Machines) classifiers, hypersphere SVM classifier and deep learning classifier with high correct recognition rate still have about 2% false recognition rate and weak incremental learning function. In this paper, a high correct recognition rate classifier designed with appropriate rejection mechanism and incremental learning algorithm is proposed to solve the above problems. The main work include: the construction algorithm of compact packing set of homogeneous feature set; the algorithm for solving the compact packing surface of homogeneous feature region based on homogeneous feature set and compact packing set; the method of setting all the public areas outside the compact packing surface as the rejection area of the classifier; when adding new categories, increasing or decreasing training samples, the above algorithms are incremental learning algorithms. A comparison experiment with uci data sets shows that the correct recognition rate of the classifier is greater than 99.13%, when the rejection rate is less than 1.3%.

Key words: classifier; pattern recognition; support vector machine; incremental learning; classification surface; wrapping learning

1 引言

几十年来,分类器设计取得了很多很好的成果.然而,目前一些正确识别率高的SVM(Support Vector Machines)分类器、超球SVM分类器、深度学习分类器在一些典型样本集上应用时仍然有2%左右的错误识别率[1~19].这些分类器不能直接用于做重大疾病认证识别、人的身份认证识别、钞票认证识别、票据认证识别、恐怖分子认证识别等比较严肃的认证识别.在比较严

收稿日期:2019-05-13;修回日期:2020-11-19;责任编辑:孙瑶
基金项目:国家重点研发计划(No.2017YFC080-4000);国家自然科学基金面上项目(No.6177227,No.61876213);江苏省自然科学基金(No.BK20171494,No.BK20201397);江苏省高校自然科学重点项目(No.18KJA520005)

肃的认证识别场合,往往需要引入合适的拒识机制[1],以便分类器工作时,要么拒识,要么分类正确.也即要求:①拒识率(在公共测试样本库中,拒识的样本个数与测试样本库的总样本个数之比)很小;②正确识别率(在去掉拒识的样本之后的测试样本库中,正确分类样本个数与总样本数的比例)为100%或接近100%.若拒识率较大,分类器的实用范围和场合则受到限制.正确识别率不能接近100%,人们不敢直接仅用该识别器去认证一些特别重要的事物或事件.显然拒识率小与正确识别率高是矛盾事件,解决该矛盾十分困难.

要做到拒识率小、正确识别率高,实际上就是要设计识别器,使该识别器确定的同ω类样本的特征区域C_ω(把区域内的任一点视为ω类样本点,而把区域之外点视为别的类点或拒识点)几乎包含了ω类所有样本点形成的实际特征区域(几乎不损失自己领域),同时几乎不侵占别的已知类的特征区域和未知可能类的特征区域.

一个很好的识别系统应该有增量学习功能[2].增量学习功能使识别系统在更正训练样本、增减训练样本和增减识别类时,能继承系统已有的大部分知识,不断优化和升级.

SVM思想是支持向量机将所有特征向量映射到一个很高维的空间里.在这个空间里建有一个最大间隔超平面,该超平面对应的原始空间曲面就是分类决策面.在分开两类特征向量(数据)的超平面的两边建有两个互相平行的超平面.分隔超平面使两个平行超平面的距离最大化.显然,SVM确定的同一类特征区域往往是无界区域,而实际的同类特征区域是有界的.SVM确定的同一类特征区域侵占了其他类的实际特征区域或未知类的特征区域,而且侵占较严重,有较大错分样本的风险.因此SVM不适合直接用于做重大疾病认证识别、通过生物特征对人身份认证识别、钞票认证识别、票据认证识别、恐怖分子认证识别等分类器设计工作.又由于当增加训练样本或增加新类别时,求解SVM分类决策面的工作需要重新进行,因此SVM没有增量学习功能.在多分类的SVM分类器设计中[3,4],改变一个训练样本或增加一个类别,相应分类器学习训练过程需要重头开始,分类器无法继承以前训练学习的任何结果,因此SVM多分类器也没有增量学习功能.在许多改进的SVM方法[5~8](考虑到有时不同类特征区域不平衡等特性)中,也没有引入合适拒识机制:实际上是不方便确定拒识区域;因为当前方法下,确定了拒识区域并不一定带来正确识别率的提高.SVM确定的同一类特征区域可能会侵占未知类的特征区域,有把未知类样本错判为某已知类的风险.一般情况下,SVM分类器存在2%左右的错误识别率,正确识别率不能逼近100%:SVM确定的已知类特征区域都尽可能最大化占据未知特征空间.SVM分类器没有增量学习功能:当增减类别时,学习训练工作要彻底重来.当增加或减少训练样本时,学习训练工作也要彻底重来.

超球SVM分类算法思想[9~13]是将所有特征向量映射到一个很高维的空间里,在这个空间里建立一个满足某种约束的半径最小超球面,超球面包裹几乎所有同类样本点.该超球面或同心超球面对应的原始空间曲面就是分类决策面.第二节实验表明,超球SVM分类决策面包裹区域可能侵占了未知类别的特征区域,即分类决策面没有紧密包裹同类样本实际特征区域.因此在超球SVM分类器中,也存在类似问题:

(1)没有引入合适拒识机制.因为不方便确定恰当的拒识区域,或者勉强确定了拒识区域,但并不一定带来正确识别率的提高.

(2)分类决策面包裹区域可能侵占未知类的特征区域,有把未知类错判为某已知类的风险.在一般情况下,分类器还有2%左右的错误识别率,正确识别率不能接近100%.

(3)分类器没有增量学习功能.当增减类别时,学习训练工作要彻底重来.当增加或减少训练样本时,学习训练工作也基本上要彻底重来.

在大数据时代,通过不断改进各种深度学习分类器来提高正确识别率,有些对象正确识别率甚至达到99%.以前街景文字识别几乎是不可能的事情,现在街景文字正确识别率也很高[14~20].很遗憾,这些分类器在一些典型样本集上的应用仍然存在2%左右的错误识别率,不适合直接做重大疾病认证识别、通过生物特征对人身份认证识别、钞票认证识别、票据认证识别、恐怖分子认证识别等比较严肃的认证识别工作,也很难通过引入拒识机制等方法改造它们.实际上,在以上深度学习分类器中也存在问题:

(1)很难确定恰当的拒识区域.分类决策面很难表达,恰当拒识区域也无法描述.

(2)分类决策面包裹区域可能会侵占未知类的特征区域,有把未知类错判为某已知类的风险.通常情况下,分类器仍有2%左右错误识别率,正确识别率不能接近100%.

(3)增量学习功能有限.当增减类别时,学习训练工作基本上要重来.当增加或减少一些训练样本时,学习训练的很多工作也要重来.

本文针对以上各种分类器设计中存在的难题,提出同类特征集合的紧密包裹集构造算法;给出基于同类特征集合、紧密包裹集的同类特征区域紧密包裹面模型和求解算法;设置了所有紧密包裹面之外的公共

区域为分类器的拒识区域,从而得到一种具有合适拒识机制的高正确识别率分类器设计算法. 该算法设计的分类器几乎没有错误识别率,正确识别率接近100%.

2 超球SVM分类决策面不紧密包裹同类特征的例子

例1 如图1所示,二维空间有正三角形围成的某事物类实际特征域. 用均匀采样方法采集样本300个,用高斯核函数,并用带惩罚系数$C'=0.5$的超球SVM训练得出该事物决策面. 该曲面近似为一个"圆". 决策面不紧密包裹同类特征. 圆内是该事物的特征域,显然该特征域侵占了实际的别的已知类或未知类特征域(即正三角形外的部分领域),同时损失了自己应有的部分领域(即正三角形两个角的小领域).

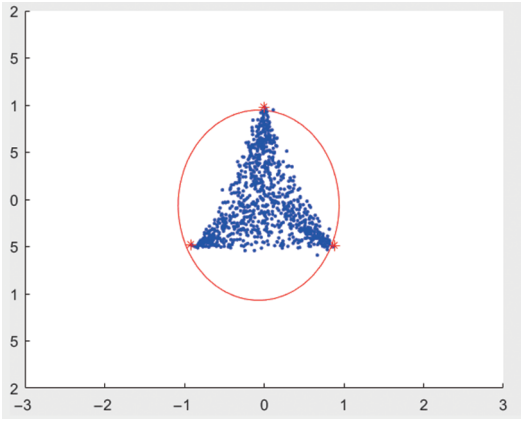

图1　超球SVM分类决策面不紧密包裹同类特征的例子

3 同类特征集合的紧密包裹集的存在性证明

设C是N维特征空间$\mathbf{R}^N$的一点集,ε是大于零的小常数,r是大于1的常数. 若对于C中任一两点X,Y,总存在C中不同点$X_1,\cdots,X_h$,使距离$\delta(X,X_1),\delta(X_1,X_2),\cdots,\delta(X_h,Y),\delta(X,Y)$都不大于$\varepsilon/\sqrt{rN}$,则称$C$是$\varepsilon/\sqrt{rN}$致密连通. 设$C_{\varepsilon/\sqrt{rN}}$是$C$加上所有$\mathbf{R}^N$中到$C$的距离不大于$\varepsilon/\sqrt{rN}$的点的区域,当$C_{\varepsilon/\sqrt{rN}}$是数学意义上的单连通区域时,则称$C$是$\varepsilon/\sqrt{rN}$致密单连通. 形象地说,单连通是没有"洞"的区域. 若以C中任意两点为端点的连线都在$C_{\varepsilon/\sqrt{rN}}$中,且这些不同连线(段)上任意两点连线也在$C_{\varepsilon/\sqrt{rN}}$中,则称$C$是$\varepsilon/\sqrt{rN}$致密凸集. 这时$C$中到$C_{\varepsilon/\sqrt{rN}}$的边界的距离等于$\varepsilon/\sqrt{rN}$的点叫做$C$的$\varepsilon/\sqrt{rN}$致密边界点. 令$C_\varepsilon$是$C$加上所有$\mathbf{R}^N$中到$C$的距离不大于$\varepsilon$的点的区域.

定理1(紧密包裹集存在定理) 设C是N维特征空间$\mathbf{R}^N$的一个$\varepsilon/\sqrt{rN}$致密单连通有界点集,且C是$\varepsilon/\sqrt{rN}$致密凸集. 令C的$\varepsilon/\sqrt{2N}$致密边界点个数为Γ. 对于任意点$X=(x_1,x_2,\cdots,x_i,\cdots,x_N)\in C$,定义$2N$个点$(x_1\pm\varepsilon,x_2,\cdots,x_i,\cdots,x_N)$,$\cdots$,$(x_1,x_2,\cdots,x_{i-1},x_i\pm\varepsilon,x_{i+1},\cdots,x_N)$,$\cdots$,$(x_1,x_2,\cdots,x_i,\cdots,x_{N-1},x_N\pm\varepsilon)$. 设$\varepsilon/\sqrt{rN}$—$\varepsilon$紧密包裹集为

$$I(C)=(C_\varepsilon-C_{\varepsilon/\sqrt{rN}})\cap \{(x_1,x_2,\cdots,x_i\pm\varepsilon,\cdots,x_N)\,|\,(x_1,x_2,\cdots,x_i,\cdots,x_N)\in C\}$$

则当$r>\dfrac{(\sqrt{N}+\sqrt{2})^2}{N}$时,$I(C)$至少有$\Gamma$个点,且对应于每一个$C$的边界点,上述$2N$个点中至少有一个点在$I(C)$中,而且$X=(x_1,x_2,\cdots,x_i,\cdots,x_N)\in I(C)$到$C_{\varepsilon/\sqrt{rN}}$内的立体表面距离大于$\varepsilon/\sqrt{rN}$.

证明略.

注:$\dfrac{(\sqrt{N}+\sqrt{2})^2}{N}$是$N$的递减数,不大于最大值5.8289.

4 同类特征集合的紧密包裹集构造算法

由定理1可知,本文所谓$\varepsilon/\sqrt{rN}$—ε紧密包裹集为

$$I(C)=(C_\varepsilon-C_{\varepsilon/\sqrt{rN}})\cap \{(x_1,x_2,\cdots,x_i\pm\varepsilon,\cdots,x_N)\,|\,(x_1,x_2,\cdots,x_i,\cdots,x_N)\in C\}$$

是关于集合C的一个集合. 该集合内的点被限制在一个厚度不超过$\varepsilon-\varepsilon/\sqrt{rN}$的皮腔(区域)上,皮腔包裹在集合$C$的外侧.

4.1 致密性参数的优化算法

设有从同类特征区域$T\subset\mathbf{R}^N$采集的同类特征点集合为C,该集合是$\varepsilon/\sqrt{5.9N}$致密有界η凸集. 不妨设C中的点足够多,$r\geq5.9$,使$C\subset T\subset C_\varepsilon$. 因此$C_{\varepsilon/\sqrt{rN}}$也称为同类特征区域$T$的近似覆盖区域. 显然$C_{\varepsilon/\sqrt{rN}}$的体积与$T$的体积之差小于$2N(2\eta+2\varepsilon/\sqrt{rN})^{N-1}(\varepsilon/\sqrt{rN})$. 即当$\varepsilon$充分小时,该差值也充分小. 在包含$C$的同类特征区域$T$固定的情况下,$\varepsilon$越小,$C$中点越多,$C$中点分布越紧密(即从$T$中采集的样本越多). 已知$T$是凸集,给定$\varepsilon$,构造满足$\varepsilon/\sqrt{5.9N}$致密凸集性质的$C$并不困难. 反过来,已知点集$C$且假定$C$有$\varepsilon/\sqrt{5.9N}$致密凸集性,要求出最小的$\varepsilon$十分困难. 本文根据$\mathbf{R}^N$中不在同一超平面上$N+1$个点决定的单纯形是一个包含这$N+1$点的最小体积凸多面体的理论,给出一种估计$\varepsilon$的算法. 设$C$有$M$个点,用$X_1,X_2,\cdots,X_M$表示. 估计$\varepsilon$算法如下所示.

第一步:计算X_j的第一近邻X_{j_1}:$j_1=\arg\min\limits_{i\neq j}\|X_j-X_i\|$.

第二步：计算 X_j 的第二近邻 X_{j_2}：$j_2=\arg\min\limits_{i\neq j,j_1}\left\|X_j-X_i\right\|$.

……

第 $N+1$ 步：计算 X_j 的第 $N+1$ 近邻 $X_{j_{N+1}}$：$j_{N+1}=\arg\min\limits_{i\neq j,j_1,\cdots,j_N}\left\|X_j-X_i\right\|$.

第 $N+2$ 步：计算 X_j 与近邻的最大距离：$g(X_j)=\max\limits_{i=j_1,\cdots,j_{N+1}}\left\|X_j-X_i\right\|$.

第 $N+3$ 步：计算 ε 的次优估计：$\varepsilon\geq\sqrt{5.9N}\max\limits_{1\leq j\leq M}g(X_j)=\sqrt{5.9N}\max\limits_{1\leq j\leq M}\max\limits_{i=j_1,\cdots,j_{N+1}}\left\|X_j-X_i\right\|$.

按照以上算法计算出来的 ε，可以证明 C 是 $\varepsilon/\sqrt{5.9N}$ 致密凸集.

容易看到，当 M 有限时，以上致密性参数的优化算法收敛，且算法复杂度为 $O(M^2)$.

4.2 紧密包裹集构造算法

下面构造设 $\varepsilon/\sqrt{5.9N}$—ε 紧密包裹集

$$I(C)=(C_\varepsilon-C_{\varepsilon/\sqrt{5.9N}})\cap\{(x_{j1},x_{j2},\cdots,x_{ji}\pm\varepsilon,\cdots x_{jN})\,|\,(x_{j1},x_{j2},\cdots,x_{ji},\cdots,x_{jN})\in C\}$$

第一步：构造 M 个点的 $\varepsilon/\sqrt{5.9N}$ 超球邻域判别函数 $f_j(X)=\varepsilon/\sqrt{5.9N}-\left\|X-X_j\right\|,1\leq j\leq M.$

这里当 $f_j(X)\geq 0$ 时，可以判断 X 在以 X_j 为中心，$\varepsilon/\sqrt{5.9N}$ 为半径的超球邻域 $\Pi\left(X_j\right)$ 内.

第二步：由每一个 $X_j=(x_{j1},x_{j2},\cdots,x_{ji},\cdots,x_{jN}),1\leq j\leq M,$ 派生出 $2N$ 个点 $(x_{j1},x_{j2},\cdots,x_{ji}\pm\varepsilon,\cdots,x_{jN}),1\leq i\leq N.$

第三步：检测所有派生点 $(x_{j1},x_{j2},\cdots,x_{ji}\pm\varepsilon,\cdots,x_{jN})$，$1\leq i\leq N,1\leq j\leq M$ 是否在 $\Pi\left(X_j\right),1\leq j\leq M$ 中，把不在任何一个超球邻域 $\Pi\left(X_j\right)$ 内的派生点集合起来就得到 $\varepsilon/\sqrt{5.9N}$—ε 紧密包裹集 $I(C)$.

由定理1可知，$I(C)$ 中点的个数不少于 C 的边界点个数. 而且当 ε 很小时，$C_{\varepsilon/\sqrt{rN}}$ 的体积与 T 的体积之差也很小，小于 $2N(2\eta+2\varepsilon/\sqrt{rN})^{N-1}(\varepsilon/\sqrt{rN})$. 形象地说，紧密包裹集紧密包裹 $C_{\varepsilon/\sqrt{rN}}\supset T\supset C$ 周围.

5 基于紧密包裹点集的同类特征区域紧密包裹曲面的求解算法

作变换 $\phi:\mathbf{R}^N\to H$，把特征空间映射为更高维空间. $k:\mathbf{R}^N\times\mathbf{R}^N\to\mathbf{R}$ 是对应核函数. 如图2所示，小超球半径为 r，同心大超球半径为 $\sqrt{r^2+\rho^2}$，小超球内的“*”点是 C 中点变换到高维空间的对应点，大超球外的“+”点是 $I(C)$ 中点变换到高维空间的对应点. 我们的目标是找到合适变换 $\phi:\mathbf{R}^N\to H$ 使小超球几乎包含所有“*”点且 r 最小，$\sqrt{r^2+\rho^2}$ 最大，即 ρ^2 最大. 这样对应于高维空间小超球面的原始空间曲面就是 C 的紧密包裹曲面.

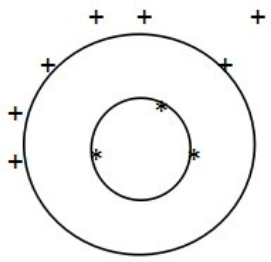

图2 紧密包裹曲面示意图

为了方便，设 C 中有 m_1 个点，$I(C)$ 中有 $m_2=n-m_1$ 个点，c 是高维空间的球心. 我们建立以下优化模型，通过求解优化解来构造同类特征区域紧密包裹曲面.

$$\begin{aligned}&\min_{r,c,\rho,\xi}\ r^2-\upsilon\rho^2+\frac{1}{\upsilon_1 m_1}\sum_{i=1}^{m_1}\xi_i+\frac{1}{\upsilon_2 m_2}\sum_{j=m_1+1}^{n}\xi_j\\&\text{s.t.}\quad\left\|\phi(X_i)-c\right\|^2\leq r^2+\xi_i,\quad 1\leq i\leq m_1\\&\qquad\left\|\phi(X_j)-c\right\|^2\geq r^2+\rho^2-\xi_j,\quad m_1\leq j\leq n\\&\qquad 0\leq\xi_k,\quad 1\leq k\leq n\end{aligned}\tag{1}$$

其中，ξ_i,ξ_j 为松弛变量，$\frac{1}{\upsilon_1 m_1},\frac{1}{\upsilon_2 m_2}$ 为惩罚系数. 为求解该模型，我们引入 Lagrange 函数

$$\begin{aligned}L(r,c,\xi,\alpha,\beta)=&\ r^2-\rho^2+\frac{1}{\upsilon_1 m_1}\sum_{i=1}^{m_1}\xi_i+\frac{1}{\upsilon_2 m_2}\sum_{j=m_1+1}^{n}\xi_j\\&+\sum_{i=1}^{m_1}\alpha_i\left(\left\|\phi(X_i)-c\right\|^2-r^2-\xi_i\right)\\&-\sum_{j=m_1+1}^{n}\alpha_j\left(\left\|\phi(X_i)-c\right\|^2-r^2-\rho^2+\xi_j\right)\\&-\sum_{k=1}^{n}\beta_k\xi_k\end{aligned}\tag{2}$$

Lagrange 函数的极值点应满足

$$\frac{\partial L}{\partial r}=2r\left(1-\sum_{i=1}^{m_1}\alpha_i y_i\right)=0\tag{3}$$

$$\frac{\partial L}{\partial\rho}=2\rho\left(-\upsilon+\sum_{j=m_1+1}^{n}\alpha_i y_i\right)=0\tag{4}$$

$$\frac{\partial L}{\partial\xi_i}=\frac{1}{\upsilon_1 m_1}-\alpha_i-\beta_i=0,\quad 1\leq i\leq m_1\tag{5}$$

$$\frac{\partial L}{\partial\xi_j}=\frac{1}{\upsilon_2 m_2}-\alpha_j-\beta_j=0,\quad m_1+1\leq j\leq n\tag{6}$$

$$\frac{\partial L}{\partial c}=2c\sum_{i=1}^{m_1}\alpha_i y_i-2\sum_{i=1}^{m_1}\alpha_i y_i\phi(X_i)=0\tag{7}$$

因此

$$c=\frac{\sum_{i=1}^{n}\alpha_i y_i\phi(X_i)}{\sum_{i=1}^{n}\alpha_i y_i}=\sum_{i=1}^{n}\alpha_i y_i\phi(X_i) \tag{8}$$

得对偶问题模型

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i y_i\phi(X_i)\cdot\phi(X_i)-\sum_{i=1}^{n}\sum_{j=m_1+1}^{n}\alpha_i\alpha_j y_i y_j\phi(X_i)\cdot\phi(X_j)$$

$$\text{s.t.}\quad 0\le\alpha_i\le\frac{1}{\upsilon_1 m_1},\quad 1\le i\le m_1 \tag{9}$$

$$0\le\alpha_j\le\frac{1}{\upsilon_2 m_2},\quad m_1+1\le j\le n$$

$$\sum_{i=1}^{n}\alpha_i y_i=1 \tag{10}$$

$$\sum_{i=1}^{n}\alpha_i=3\upsilon+1 \tag{11}$$

用$K(X_i,X_j)$替代$\phi(X_i)\cdot\phi(X_j)$,由核函数性质可知

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i y_i k(X_i,X_i)-\sum_{i=1}^{n}\sum_{j=m_1+1}^{n}\alpha_i\alpha_j y_i y_j k(X_i,X_j)$$

$$\text{s.t.}\quad 0\le\alpha_i\le\frac{1}{\upsilon_1 m_1},\quad 1\le i\le m_1 \tag{12}$$

$$0\le\alpha_j\le\frac{1}{\upsilon_2 m_2},\quad m_1+1\le j\le n$$

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

$$\sum_{i=1}^{n}\alpha_i=2\upsilon+1 \tag{14}$$

对偶问题是一个二次规划问题,可用二次规划问题求解方法,如序贯最小优化(Sequential Minimal Optimization, SMO)算法来求解.

在求解以上问题之后,为求出r、ρ^2和$r^2+\rho^2$,考虑两集合

$$S_1=\{x_i\mid 0<\alpha_i<\frac{1}{\upsilon_1 m_1},\quad 1\le i\le m_1\} \tag{15}$$

$$S_2=\{x_j\mid 0<\alpha_j<\frac{1}{\upsilon_2 m_2},\quad m_1+1\le j\le n\} \tag{16}$$

令$n_1=\left|S_1\right|,n_2=\left|S_2\right|$. 由KKT条件知

$$r^2=\frac{1}{n_1}P_1,\quad \rho^2=\frac{1}{n_2}P_2-\frac{1}{n_1}P_1 \tag{17}$$

其中

$$P_1=\sum_{x_i\in S_1}\left\|\phi(X_i)-c\right\|^2=\sum_{x_i\in S_1}\left(k(X_i,X_i)-2\sum_{k=1}^{n}\alpha_k y_k k(X_i,X_k)+<c,c>\right) \tag{18}$$

$$P_2=\sum_{x_j\in S_2}\left\|\phi(X_i)-c\right\|^2=\sum_{x_j\in S_2}\left(k(X_j,X_j)-2\sum_{k=1}^{n}\alpha_k y_k k(X_j,X_k)+<c,c>\right) \tag{19}$$

$$<c,c>=\left\langle\sum_{i=1}^{n}\alpha_i y_i\phi(X_i),\sum_{j=m_1+1}^{n}\alpha_j y_j\phi(X_j)\right\rangle=\sum_{i=1}^{n}\sum_{j=m_1+1}^{n}\alpha_i\alpha_j y_i y_j k(X_i,X_j) \tag{20}$$

分类决策函数为

$$f(x)=\operatorname{sgn}(r^2-\left\|\phi(X)-c\right\|^2)=\operatorname{sgn}\left(r^2+2\sum_{k=1}^{n}\alpha_k y_k k(X,X_k)-k(X,X)-<c,c>\right) \tag{21}$$

分类决策曲面W为

$$r^2-\left\|\phi(X)-c\right\|^2=r^2+2\sum_{k=1}^{n}\alpha_k y_k k(X,X_k)-k(X,X)-<c,c>=0 \tag{22}$$

根据紧密包裹集和分类决策曲面W的构造过程,可得以下紧密包裹学习定理.

定理2(紧密包裹学习定理) 设有从同类特征区域$T\subset\mathbf{R}^N$采集的同类特征点集合为C,该集合是$\varepsilon/\sqrt{rN}$致密有界η凸集. 设$\delta>0$,$r\ge 5.9$,$C\subset T\subset C_{\varepsilon/\sqrt{rN}}\subset C_{\varepsilon}$和$\varepsilon/\sqrt{rN}$—$\varepsilon$紧密包裹集$I(C)$包含足够多个点. 若$\varepsilon<\frac{\delta}{2^N N(\eta+1)^{N-1}}$,那么曲面$W$包裹区域体积与同类特征区域$T$的体积差小于$\delta$.

证明略.

例2 图3是例1中正三角形围成的某事物类实际特征域(识别对象的同类特征域)的紧密包裹点集和紧密包裹面. 十分明显,本文方法获得的包裹面比用超球支持向量机(也称为支持向量域描述, SVDD)获得的包裹面包裹正三角要紧密得多(参见例1).

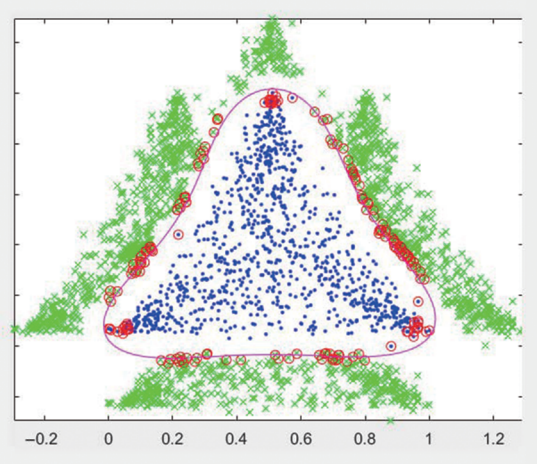

图3　特征域的紧密包裹点集和紧密包裹面

6　多类分类器的合适拒识机制设置

对于η类$\omega_1,\omega_2,\cdots,\omega_\eta$分类问题,用上述方法找到尽

可能小的参数ε. 用以上方法求出每类样本对应的判别函数

$$
\begin{aligned}
{}_jf(X)&=\operatorname{sgn}({}_jr^2-\left\|\phi(X)-{}_j\boldsymbol{c}\right\|^2)\\
&=\operatorname{sgn}({}_jr^2+2\sum_{k=1}^{n}{}_j\alpha_k y_k k(X,X_k)-k(X,X)-<{}_j\boldsymbol{c},{}_j\boldsymbol{c}>)\\
&1\le j\le\eta
\end{aligned}
\tag{23}
$$

和分类决策紧密包裹曲面

$$
\begin{aligned}
&{}_jr^2-\left\|\phi(X)-{}_j\boldsymbol{c}\right\|^2\\
&={}_jr^2+2\sum_{k=1}^{n}{}_j\alpha_k y_k k(X,X_k)-k(X,X)-<{}_j\boldsymbol{c},{}_j\boldsymbol{c}>\\
&=0\\
&1\le j\le\eta
\end{aligned}
\tag{24}
$$

当${}_jf(X)\ge 0$时,判定X是ω_j类. 为了防止同一点属于不同类(即不同紧密包裹区域有重叠)现象出现,规定${}_jf(X)$有优先级,标号小优先级高. 当$j=\arg\min_i{}_if(X)\ge 0$时,判定$X$是$\omega_j$类.(注:当对同一点$X$存在多个$j$使${}_jf(X)\ge 0$时,也可用近邻方法类分类)

设置所有紧密包裹曲面之外的公共区域为分类器的拒识区域,即当$\max_j{}_jf(X)<0$时,分类器拒识X.

7 增量学习算法

当增加新类别时,以前的所有工作结果都可保留. 只需要对新类别样本集求包裹点集和基于新类别样本集和包裹集求新类别的分类决策函数和分类决策紧密包裹曲面,同时对拒识区域做相应调整.

当增加训练样本且原来分类器能正确分类时,分类器不用做任何调整. 否则仅对错分和拒识样本(可以判断是新边界点)所对应类别集合重新求包裹点集(以前大部分计算结果可以用上)和基于该新类别样本集及包裹集求新类别的分类决策函数与分类决策紧密包裹曲面,同时对拒识区域做相应调整.

当减去错分样本时,仅对错分样本(可以判断边界点)所对应类别集合重新求包裹点集(以前大部分计算结果可以用上)和基于该新类别样本集和包裹集求新类别的分类决策函数和分类决策紧密包裹曲面,同时对拒识区域做相应调整.

8 实验及对比分析

本文采用uci数据集来测试评价支持向量机SVM、支持向量数据描述(Support Vector Data Description, SVDD)、小球大间隔(Small Sphere Large Margin, SSLM)、深层感知网络、本文包裹学习算法性能. 如表1所示,其中pos,neg分别表示正负类样本的数量,m_1、m_2表示实验随机选取的样本个数,d表示样本的特征维度.

实验中所有的参数通过网格搜索以及交叉验证法获得. 对于所有的算法均采用高斯核函数:

$$K(u,v)=\exp(-\frac{1}{\delta}\|u-v\|^2) \tag{25}$$

其中,高斯核函数中的参数δ在网格$\left\{\frac{\sigma_0^2}{16},\frac{\sigma_0^2}{8},\frac{\sigma_0^2}{4},\frac{\sigma_0^2}{2},\sigma_0^2,2\sigma_0^2,4\sigma_0^2,8\sigma_0^2,16\sigma_0^2\right\}$中搜索选取,$\sigma_0^2$为样本的平均范数.

对于SVM模型,惩罚系数C在网格{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500}中搜索选取.

对于SVDD,C_1同样从{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500}中搜索,C_2从$\left\{\frac{1}{4}\times\frac{m_1}{m_2},\frac{1}{2}\times\frac{m_1}{m_2},2\times\frac{m_1}{m_2},4\times\frac{m_1}{m_2}\right\}$中搜索.

对于SSLM和包裹学习算法,参数v在网格{10,30,50,70,90}中搜索选取,v_1,v_2在集合{0.001, 0.01,0.1}中选取,如表2所示.

表1 实验数据集表

Dataset	pos	neg	m_1	m_2	d
Iris	50	100	45	90	4
Wine	59	119	53	107	13
Blood	570	178	487	15	4
Cancer	444	239	398	12	9

表2 学习算法的几个参量

Dataset	v	v_1	v_2
Iris	10	0.1	0.1
Wine	10	0.1	0.1
Blood	50	0.1	0.1
Cancer	30	0.1	0.01

按照文献[13]设定从数据集中取一定正类样本和负类样本,分别运行SVM,SVDD,SSLM,深层感知网络,以及包裹学习算法进行训练,余下的样本用于测试. 对于用包裹学习算法进行两分类的问题,实验做法是:先对训练样本比较多的类别(正类)执行包裹学习算法,并且执行包裹学习算法期间把另一类样本加入包裹样本集进行训练;其次用初始判别函数进行测试,优化分类判别方法. 具体要缩放调节判别函数中r,或把出错样本加入训练集重新训练,使测试结果满意为止. 若测试结果正确识别率为100%或相当满意,那么不用再对另一类(负类)样本执行包裹学习算法,当然也不设置拒识区域. 若对测试结果不够满意,那么对另一类(负类)样本执行包裹学习算法. 重复上述过程,并且设置该类(负类)判别区域为该类包裹面包裹的区域去掉另一类(正类)判别区域,即正类判别优先. 或者设置该正类判别区域为正类包裹面包裹的区域去掉另一类(负类)判别区域,即负类判别优先. 取正类判别优先还是负类判别

优先,要根据测试结果来定. 然后设置拒识区域为正类判别区域和负类判别区域之外的区域. 最后计算拒识率. 一般拒识区域比较大,甚至很大,但拒识率都比较小. 是否优化致密参数,对识别结果影响不大. 一般来说,优化致密参数后,正确识别率会有一点提高. 表3是实验结果. 实验结果表明,包裹学习算法有效,而且可以用来设计有合适拒识区域的高正确识别率的分类器.

表3　实验结果

Dataset	SVM	SVDD	SSLM	深层感知网络	包裹学习(不设置拒识区域)	包裹学习(设置拒识区域)	
						正确识别率	拒识率
Iris	96.3%	96.5%	98.2%	98.8%	98.5%	优化致密参数后100%	优化致密参数后1.3%
						优化致密参数前99.3%	优化致密参数前1.2%
Wine	93.75%	95.34%	97.1%	98.3%	97.7%	优化致密参数后100%	优化致密参数后2.1%
						优化致密参数前99.1%	优化致密参数前1.9%
Blood	68.67%	71.12%	71.33%	77.25%	72.67%	优化致密参数后99.32%	优化致密参数后2.3%
						优化致密参数前99.1%	优化致密参数前1.9%
Blood	68.67%	71.12%	71.33%	77.25%	72.67%	优化致密参数后99.32%	优化致密参数后2.3%
						优化致密参数前99.15%	优化致密参数前2.1%
Cancer	94.75%	93.78%	95.25%	98.54%	97.45%	优化致密参数后99.15%	优化致密参数后1.8%
						优化致密参数前99.1%	优化致密参数前1.7%

9　结束语

论文讨论了面向重大疾病认证识别、通过生物特征对人身份认证识别、钞票认证识别、票据认证识别等高正确识别率分类器设计要求. 指出了现有一些高正确识别率SVM分类器、超球SVM分类器、深度学习分类器在一些典型样本集上应用时仍然存在没有合适拒识机制和正确识别率不能接近100%等问题. 举出了超球SVM分类决策面不紧密包裹同类特征的实例. 证明了同类特征集合的紧密包裹集的存在定理. 提出了致密性参数的优化算法、同类特征集合的紧密包裹集构造算法、基于同类特征集合和紧密包裹集的同类特征区域紧密包裹面的求解算法,以及紧密包裹学习定理. 设置了较合适的多类分类器的拒识区域. 同时给出了当增加新类别,增减训练样本时,有关算法的增量学习算法. 以uci数据库做对比实验,实验表明本文提出的分类器设计算法是一种有效的高正确识别率分类器设计方法.

本文提出了带“紧密包裹集”约束超球支持向量机方法,其中紧密包裹集是本文作者提出并算法实现的独特概念. 带“紧密包裹集”约束也是本文的创新点之一. 正因为如此,才有紧密包裹学习方法,才有比现有一些覆盖“更紧密覆盖”,才有比现有一些分类器“更高正确识别率”的效果.

参考文献

[1] 王守觉. 仿生模式识别(拓扑模式识别): 一种模式识别新模型的理论与应用[J]. 电子学报, 2002, 30(10): 1417 – 1420.
Wang S J. Bionic (topological) pattern recognition—A new model of pattern recognition theory and its applications[J]. Acta Electronica Sinica, 2002, 30(10): 1417 – 1420. (in Chinese)

[2] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
Zhou Z H. Machine Learning[M]. Beijing, China: Tsinghua University Press, 2016. (in Chinese)

[3] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York, NY, USA: Springer New York, 1995.

[4] Tang F Z, Adam L, Si B L. Group feature selection with multiclass support vector machine[J]. Neurocomputing, 2018, 317(23): 42 – 49.

[5] Tian Y J, Qi Z Q, Ju X C, et al. Nonparallel support vector machines for pattern classification[J]. IEEE Transactions on Cybernetics, 2014, 44(7): 1067 – 1079.

[6] Ding L Z, Liao S Z. An approximate approach to automatic kernel selection[J]. IEEE Transactions on Cybernetics, 2017, 47(3): 554 – 565.

[7] Cano A, Zafra A, Ventura S. Weighted data gravitation classification for standard and imbalanced data[J]. IEEE Transactions on Cybernetics, 2013, 43(6): 1672 – 1687.

[8] Jayadeva, Khemchandani R, Chandra S. Twin support vector machines for pattern classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905 – 910.

[9] Xu Y T, Yang Z J, Pan X L. A novel twin support-vector

machine with pinball loss[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(2): 359 – 370.

[10] Wu M R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088 – 2092.

[11] Xu Y T, Liu C M. A rough margin-based one class support vector machine[J]. Neural Computing and Applications, 2013, 22(6): 1077 – 1084.

[12] Chorowski J, Wang J, Zurada J M. Review and performance comparison of SVM- and ELM-based classifiers [J]. Neurocomputing, 2014, 128: 507 – 516.

[13] Xu Y T, Yang Z J, Zhang Y Q, et al. A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification[J]. Knowledge-Based Systems, 2016, 95: 75 – 85.

[14] Xu Y T. Maximum margin of twin spheres support vector machine for imbalanced data classification[J]. IEEE Transactions on Cybernetics, 2017, 47(6): 1540 – 1550.

[15] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436 – 444.

[16] Shi C M, Panoutsos G, Luo B, et al. Using multiple-feature-spaces-based deep learning for tool condition monitoring in ultraprecision manufacturing[J]. IEEE Transactions on Industrial Electronics, 2019, 66(5): 3794 – 3803.

[17] Tsygvintsev A. On the overfly algorithm in deep learning of neural networks[J]. Applied Mathematics and Computation, 2019, 349: 348 – 358.

[18] Alharbi A, Doncker E D. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information[J]. Cognitive Systems Research, 2019, 54: 50 – 61.

[19] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55 – 75.

[20] 舒坚, 张学佩, 刘琳岚, 等. 基于深度卷积神经网络的多节点间链路预测方法[J]. 电子学报, 2018, 46(12): 2970 – 2977.
Shu J, Zhang X P, Liu L L, et al. Multi-nodes link prediction method based on deep convolution neural networks [J]. Acta Electronica Sinica, 2018, 46(12): 2970 – 2977. (in Chinese)

作者简介

杨国为 男,1964 年 2 月生,江西樟树人. 教授、博士生导师、中国人工智能学会理事、中国电子学会高级会员. 1985 年、1988 年和 2004 年分别在江西师范大学、北京科技大学获理学学士、硕士和工学博士学位. 现为青岛大学信号与信息处理研究所所长,主要从事模式识别、智能信息处理、智能控制等方面的研究工作.

E-mail:ygw_ustb@163.com

万鸣华 男,1978 年 3 月生,江西南昌人. 校聘教授、硕士生导师、中国计算机学会高级会员. 2003 年、2007 年和 2011 年分别在南昌航空大学、南京理工大学获工学学士、工学硕士和工学博士学位. 现为南京审计大学软件工程系教师,主要从事模式识别、机器学习和特征提取等方面的研究工作.

E-mail:wmh36@nau.edu.cn

赖志辉 男,1979 年 6 月生. 博士、教授、博士生导师. 2002 年在华南师范大学获学士学位,2007 年在暨南大学获硕士学位,2011 年在南京理工大学获博士学位. 2015 年入选深圳市海外高层次人才"孔雀计划",现任国际 SCI 期刊 International Journal of Machine Learning and Cybernetics 的编委. 主要从事图像处理与模式识别、深度学习与大数据哈希学习等方面的研究工作.

张凡龙 男,1985 年 9 月生,山东泰安人. 2007 年、2010 年在聊城大学获理学学士、硕士学位,2015 年在南京理工大学获工学博士学位. 现为南京审计大学信息工程学院副教授、硕士生导师,主要从事图像处理、压缩感知和特征提取等方面的研究工作.