高技术通讯 2014 年 第24 卷 第3 期:236 ~243

doi:10.3772/j.issn.1002-0470.2014.03.003

稀疏集 SVN 惩罚校正方法及其种质评价应用研究①

谭文学②* ** 赵春江③*** 吴华瑞***

(*北京工业大学计算机学院 北京 100022)
(**湖南文理学院计算机学院 常德 415000)
(***国家农业信息化技术研究中心 北京 100097)

摘 要 针对支持向量学习网络(SVN)学习稀疏样本数据集时,稀疏目标和非稀疏目标的分类器错误率严重失衡而实用性大大降低的问题,在拉格朗日乘数渐近分析基础上,引入惩罚校正因子、逆向训练样本和错误训练率等概念,提出了惩罚校正支持向量网络学习算法和校正方法,并将该方法应用于以 CT 图像特征数据集为基础的小麦籽种品质定级。等值分析说明该学习算法能有效地等级化籽种特征数据,准确率达 95%;和其他同源方法的对比试验显示:针对稀疏样本集,该算法在获得可观综合预测准确性的同时,能显著改善稀疏样本集各目标分类器的预测错误率的极性分布,并展现良好的学习性能。

关键词 惩罚校正,支持向量网络,错误训练,稀疏样本,逆向训练

0 引 言

机器自动分类和识别是一种当下非常流行的以机器学习为基础的数据挖掘和分析技术,其通过已知类标签的样本集训练分类器,然后对一组有着相同采样特征的样本集合标注类标签或者分组。特征值相近的样本归到一组,相异的分配到相应类标签,从而发现、认识新奇的模式特征或类集。基于"分隔样本且间隔最优"的支持向量网络(support vector network,SVN)已广泛应用于归类、评价和聚类等。分类(classification)、分隔(separation)、定级(gradation)3 个操作都蕴含了"根据一定规则把某个集合打散为几个子集"的过程,操作实质上没有区别,只是在不同应用场景中被冠以不同的名称。SVN 又称支持向量机(SVM),是年轻的计算型学习算法[1],因巧妙运用核函数理论解决诸多学习算法无法解决的低维线性不可分的问题而成为知识发现、数据挖掘领域的研究热点[2,3]。然而,对于现实世界中大量多目标、不均衡、有噪声和过失误差的数据集(有时限于实际,无法提供均匀的训练集),它无法达到满意的效果[4]:对于训练样本充足学习充分的目标类有很满意的低预测错误率,而对于另外一些稀疏样本却表现出难以接受的预测错误率。原因是它忽略了训练样本容量差异和均匀惩罚,致使学习机局部的分类器训练错误率高,本该学到的那部分知识被过量的惩罚、噪声"淹没",类边界、决策超平面偏移严重,抬高了该分类器的泛化错误率。文献[5]从样本个体和类标签个体引入权因子,提出了加权支持向量机算法。然而在样本集大容量的情形下,很难合理、科学地确定权重,类标签权重研判也缺乏指导算法,因而权赋值盲目,效果不明显。文献[6]提出了 ν - SVC 算法,引入参数 ν 渐近地控制错误率和支持向量百分比的上界,从而优化学习训练精度。实验数据显示,在样本维数低的情况下效果较好,而对高维样本集,错误率高达 20%。

农作物特性以及遗传信息都是通过作物籽种遗传至后代[7],因而,籽种品质评价及优选是农业增产的重要环节。籽种品质可通过其外观几何特征来反映,反之,通过外观几何特征参数来评价其品质也是可行的[8]。在正常情况下,品质空间的籽种样本容量是不平衡的,由此导致在学习算法指导机器学习过程中,从训练样本到分类器的知识、经验、信息

① 国家自然科学基金(61271257,61102126),国家科技支撑计划(2013BAJ04B04)和湖南省科技计划(2013GK3135)资助项目。
② 男,1973 年生,博士生,副教授;研究方向:农业信息智能分析与机器学习;E-mail:twxpaper@163.com
③ 通讯作者,E-mail:zhaocj@nercita.org.cn
(收稿日期:2013-05-06)

传递不平衡,致使错误训练和错误预测剧增。如何为数量比例失调的训练集构造合适的学习模型是机器籽种品质评价需解决的关键课题。

本文针对上述问题,研究了稀疏集 SVN 惩罚校正方法及其在种质评价上的应用。

1 问题描述

支持向量分类问题可归结为 n 维空间的二次优化问题。为清晰描述算法,以 2 组划分问题为例,在此基础上给出相关定义。

令 i 为整数且 $1 \leqslant i \leqslant l$; $\bar{x}_i$ 为样本向量; y_i 为样本 $\bar{x}_i$ 的类标签, $y_i \in \{1, -1\}$。如果 $y_i = 1$, 则 $\bar{x}_i$ 为来自"1"类样本,又名为"+"类样本;如果 $y_i = -1$, 则 $\bar{x}_i$ 为"-1"类样本,又名"-"类样本。今已知类标签的 2 维样本构成的集合,欲构造一超平面,又名分类器,使得其将该样本集分隔为 2 个子集,一个"+"类样本集,一个"-"类样本集,且满足 2 类边界超平面之间分隔间隔最大。

定义 1 某训练样本集,其为所涉及的类标签提供的样本数量不平衡,样本或携带错误及噪声,且次大位容量相对于最大容量的 2 类标签的样本容量之比低于某个值 τ, 如 0.5,则称该样本集为稀疏训练集,又名不平衡训练集。

定义 2 $\bar{w}$ 为超平面法向量, b 为决策平面截距, ξ_i 为截距漂移,则称 $\{y_i[(\bar{w} \cdot \bar{x}_i) + b] - 1 + \xi_i\} = 0$ 为类标签 y_i 的边界超平面(Boundary Hyperplane)。

定义 3 当 $\bar{x}_i$ 处在 $\xi_i = 0$ 边界超平面上,即两类边界超平面截距差为 2,则称 $\bar{x}_i$ 属于在边界上的支持向量,简称为边界支持向量(On-Boundary-SV)。

定义 4 当 $\bar{x}_i$ 处在 $1 > \xi_i > 0$ 的超平面上,换言之, $\bar{x}_i$ 被包围在由 $\{y_i[(\bar{w} \cdot \bar{x}_i) + b] - 1\} = 0$ 定义的边界超平面和由 $(\bar{w} \cdot \bar{x}_i) + b = 0$ 定义的决策超平面中间,则称 $\bar{x}_i$ 为边界内支持向量(Within-Boundary-SV)。

定义 5 当 $\bar{x}_i$ 处在 $\xi_i = 1$ 的超平面上,即 $\bar{x}_i$ 落在决策超平面 $[(\bar{w} \cdot \bar{x}_i) + b] = 0$ 上,则称样本向量 $\bar{x}_i$ 为决策支持向量(Decision-SV)。

定义 6 当样本 $\bar{x}_i$ 处在 $\xi_i > 1$ 的超平面,对于"+"样本,它已经下击穿越过决策超平面,进入类标签"-1"的区域;对于"-"样本,已经上击穿决策超平面,进入类标签"1"的区域,则称 $\bar{x}_i$ 为穿透决策超平面的支持向量(Through-Decision-SV)。

图 1 描述了不同类型的支持向量相对于决策超平面的分布。

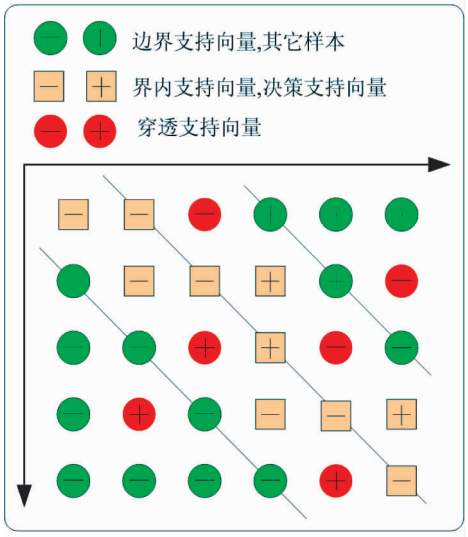

图 1 不同情形的支持向量

定义 7 若某学习机在训练过程中有部分样本,为使被训练的分类器能对他们正确分类,它们中的每个都对分类器施加微量的调整,然而这些微量调整对受训分类器的预测准确性有负面影响,势必使最优分类间隔变窄,对分类器的训练,产生误导。此时称该微量的调整为错误训练(Mis-training),称这些样本为逆向训练样本(adverse training samples)。一般而言,边界内支持向量、决策支持向量、穿透支持向量都是逆向训练样本。正例训练样本总体中,逆向训练样本数量在总量中所占的比例定义为"+"分类器的错误训练率(Mis-training Rate),记为 $error_{\text{train}}(1)$。

定义 8 "+"类预测样本总体中,在预测过程中,被某学习机错误分类的样本数量所占的比例定义为该学习机对于"+"类的错误预测率,记为 $error_{\text{predict}}(1)$。

2 支持向量网络模型

Cortes 提出的支持向量网络(SVN)算法[1]的数

学形式化模型为

$$\begin{cases}\min\limits_{\bar{w},\bar{\xi}} \Phi(\bar{w},\bar{\xi}) = \dfrac{1}{2}(\bar{w}\cdot\bar{w}) + C\sum\limits_{i=1}^{l}\xi_i \\ s.t.\ y_i[(\bar{w}\cdot\bar{x}_i)+b] \geqslant 1-\xi_i \\ \xi_i \geqslant 0, i = 1,2,\cdots,l \\ C > 0\end{cases} \tag{1}$$

转换拉格朗日极值[5]问题如下：

$$\begin{aligned} L(\bar{\alpha},\bar{\xi},\bar{w},b,\bar{\beta}) = {} & \frac{1}{2}(\bar{w}\cdot\bar{w}) + C\sum_{i=1}^{l}\xi_i \\ & - \sum_{i=1}^{l}\alpha_i\{y_i[(\bar{w}\cdot\bar{x}_i)+b]-1 \\ & + \xi_i\} - \sum_{i=1}^{l}\beta_i\xi_i \end{aligned} \tag{2}$$

必须满足(KKT)条件[6]：

$$\begin{cases}\dfrac{\partial L}{\partial \bar{w}} = \bar{w} - \sum\limits_{i=1}^{l}\alpha_i y_i \bar{x}_i \\ \dfrac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \\ \dfrac{\partial L}{\partial b} = \alpha_i y_i = 0 \\ \alpha_i\{y_i[(\bar{w}\cdot\bar{x}_i)+b]-1+\xi_i\} = 0, \forall i \\ \beta_i\cdot\xi_i = 0, \forall i \\ \alpha_i,\beta_i,\xi_i \geqslant 0, \forall i\end{cases} \tag{3}$$

对于 $\{y_i[(\bar{w}\cdot\bar{x}_i)+b]-1+\xi_i\} \neq 0\}$ 的样本，$\alpha_i = 0$；反之，满足 $\{y_i[(\bar{w}\cdot\bar{x}_i)+b]-1+\xi_i = 0\}$ 的 $\bar{x}_i$ 称为支持向量；它们线性组合可得法向量 $\bar{w}$，$\alpha_i y_i = 0$，$\beta_i\cdot\xi_i = 0$，$\alpha_i+\beta_i = C$，代入式(2)得

$$\begin{aligned} L(\bar{\alpha},\bar{w}) = {} & \frac{1}{2}(\bar{w}\cdot\bar{w}) + C\sum_{i=1}^{l}\xi_i - (\bar{w}\cdot\bar{w}) \\ & - \sum_{i=1}^{l}\alpha_i\{by_i - 1 + \xi_i\} - \sum_{i=1}^{l}\beta_i\xi_i \\ = {} & -\frac{1}{2}(\bar{w}\cdot\bar{w}) + \sum_{i=1}^{l}\alpha_i \end{aligned} \tag{4}$$

$$\begin{cases}\max \tilde{L}(\bar{\alpha}) = \sum\limits_{i=1}^{l}\alpha_i - \dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}\alpha_i\alpha_j y_i y_j \bar{x}_i\cdot\bar{x}_j \\ s.t. \quad \sum\limits_{i=1}^{l}\alpha_i y_i = 0, C \geqslant \alpha_i \geqslant 0; \forall i\end{cases} \tag{5}$$

可以看出，式(4)是一个仅仅关于 α_i 的函数，且 $0 \leqslant \alpha_i,\beta_i \leqslant C$，$\alpha_i y_i = 0$；这是对 α_i 的约束。

得到式(1)的对偶问题如式(5)，根据 Wolfe 对偶定理，两者有相同的可行解。

3 SVN 的错误训练率分析

模型中得参数 C 决定了学习算法在追求分类器的最大化分类间隔时，对样本训练错误的最小化“容忍和纵容”的程度，其尺度和分类器无关，对所有的参与训练的样本，无论类别一视同仁，这是问题根源之所在[9,10]。当训练集是稀疏训练集时，例如“1”类训练样本在数量上远远小于“-1”类时，就会导致两者的错误训练率失调。

令 NSV^+、NSV^- 分别表示“1”类和“-1”类支持向量数量，N^+、N^- 分别表示两类的样本数，α_i^{out+}、α_i^{in+}、α_i^{on+} 分别表示“1”向量中穿透支持向量、界内支持向量、边界支持向量 $\bar{x}_i$ 的拉格朗日系数，记 $\sum\limits_{y_i=1}\alpha_i = D$，$\sum\limits_{y_i=-1}\alpha_i = E$。

定理 1：对于式(1)定义的 SVN 学习机，有下式成立：

$$error_{\text{train}}(1) \approx \frac{D}{N^+\cdot C} \tag{6}$$

$$error_{\text{train}}(-1) \approx \frac{E}{N^-\cdot C} \tag{7}$$

$$\frac{error_{\text{train}}(1)}{error_{\text{train}}(-1)} \approx \frac{N^-}{N^+} \tag{8}$$

证明：分析 KKT 条件 $\alpha_i\{y_i[(\bar{w}\cdot\bar{x}_i)+b]-1+\xi_i\} = 0, \beta_i\xi_i = 0, \forall i$。

情形 1：$\alpha_i = 0$，则 $\beta_i = C \neq 0$，必有 $\xi_i = 0$，则有 $y_i[(\bar{w}\cdot\bar{x}_i)+b]-1 \neq 0$。样本 $\bar{x}_i$ 被正确分类。

情形 2：如果 $0 < \alpha_i < C$，则 $\beta_i = C-\alpha_i \neq 0$，必有 $\xi_i = 0$，$y_i[(\bar{w}\cdot\bar{x}_i)+b]-1 = 0$。$\bar{x}_i$ 处在截距无漂移边界超平面上，为边界支持向量，分类正确。

情形 3：$\alpha_i = C$，则 $\beta_i = 0$，又 $\xi_i > 0$，则有 $\{y_i[(\bar{w}\cdot\bar{x}_i)+b]-1+\xi_i\} = 0$。$\bar{x}_i$ 在相对于边界超平面有漂移的超平面上，此时又分为 3 种情况：

如果 $1 > \xi_i > 0$，则 $\bar{x}_i$ 夹在边界超平面和决策超平面之间，为边界内支持向量，分类结果属正确，但精度不理想，就训练分类器而言，是有误差训练。

如果 $\xi_i > 1$，则 $\bar{x}_i$ 为穿透边界，被错误分类，属于越界支持向量，施加了错误的训练。

如果 $\xi_i = 1$，则 $\bar{x}_i$ 属于决策支持向量，分类完全不可信，训练有益性不可靠。

令 NSV^{out+}、NSV^{in+}、NSV^{on+} 分别表示“1”类的穿

透支持向量、界内及决策支持向量、边界支持向量的数量,则所有拉格朗日系数之和为

$$D = \sum_{y_i=1}^{NSV_{out}^{+}} \alpha_i^{out+} + \sum_{i=1}^{NSV_{in}^{+}} \alpha_i^{in+} + \sum_{i=1}^{NSV_{on}^{+}} \alpha_i^{on+} \tag{9}$$

由于 $\alpha_i^{out+} = C$, $\alpha_i^{in+} = C$, $0 < \alpha_i^{on+} < C$, 则有 $NSV^{+} \cdot C \geqslant D \geqslant NSV^{out+} \cdot C$。两边同除以 $N^{+} \cdot C$ 得式

$$\frac{NSV^{out+}}{N^{+}} \leqslant \frac{D}{N^{+} \cdot C} \leqslant \frac{NSV^{+}}{N^{+}} \tag{10}$$

分析中间项的 D, 如果所有“1”类样本都是支持向量且非边界支持向量,则分母表示所有“1”类样本的拉格朗日系数之和,这意味着所有的“1”类样本不是被错误分类,就是分类效果不理想、不可靠、有误差,两种情形均为错误训练,或者机器实施了逆向学习,且 $D = N^{+} \cdot C$, 比值为 1。

其下边界为“1”类穿透支持向量在“1”类样本数量中所占的比例,为最小错误训练率;其上边界为全部“1”类支持向量在“1”类样本数量中所占的比例,为最大错误训练率。不难理解,真实的错误训练率应该是夹于其间的某个值,中间项越靠近 1,说明当前情形越靠近这个临界状态。它是训练错误率的渐近,记为 $error_{\text{train}}(1) \approx \frac{D}{N^{+} \cdot C}$, 即式(6)。

同理,可以得到:

$$\frac{NSV^{out-}}{N^{-}} \leqslant \frac{E}{N^{-} \cdot C} \leqslant \frac{NSV^{-}}{N^{-}} \tag{11}$$

$error_{train}(-1) \approx \frac{E}{N^{-} \cdot C}$, 即式(7)。由于 $\alpha_i y_i = 0$, 故 $\sum \alpha_i y_i = 0$, 则

$$\sum_{y_i=1} \alpha_i y_i + \sum_{y_i=-1} \alpha_i y_i = D - E = 0 \tag{12}$$

故 $D = E$, 两者相比得到 $\frac{error_{\text{train}}(1)}{error_{\text{train}}(-1)} \approx \frac{N^{-}}{N^{+}}$, 即式(8)。

证明毕。

定理 1 得到了针对 2 组分类问题训练的不同类标签的目标之间的错误训练率 $error_{\text{train}}$ 的比值的渐近。训练和预测本质上都是归类,不过是分类器在不同的阶段完成的同一操作。通常情况下,学习机的 2 个错误率和随机抽样无关。故我们提出假设: $error_{\text{test}}(*) \approx error_{\text{train}}(*)$。换言之,某个类标签分类器的错误训练率 $error_{\text{train}}(*)$ 越高,则其预测错误率 $error_{\text{test}}(*)$ 越大;反之,控制前者能有效抑制后者。

从应用和泛化的角度,用户不希望得到偏向性严重的学习机,即期望实践中,各个目标类标签的判别错误率相近,而不是出现对于某个类的样本则极高,而对于其它类样本则非常低的极端分布态势。从定理 1 的结论可知:稀疏训练样本集中数量处于弱势的样本对应的类标签分类器的错误训练率高,错误预测率大,训练输出“极性”分类器,可见,不适合用该方法训练。然而,稀疏训练集却相当常见,比如在常规体检过程中,正常个体相对于病态个体往往有压倒性优势。

4 SVN 的惩罚校正

通过前节分析,SVN 由稀疏样本集训练得到错误率失衡的分类器模型,根本原因在于,学习算法为包容线性不可分的样本而施加惩罚时忽略了不同类标签所拥有的学习样本数量上的差异,实行了均匀惩罚。今令 $\vartheta^{*} \geqslant 1$ 为惩罚校正因子(对于 2 组分类问题则为 ϑ^{+} 对应“1”类样本, ϑ^{-} 对应“-1”类样本),则校正后的 SVN 相对应的优化问题的数学模型如式

$$\begin{aligned} L(\bar{\alpha}, \bar{\xi}, \bar{w}, b, \bar{\beta}) = {} & \frac{1}{2}(\bar{w} \cdot \bar{w}) + C\vartheta^{*} \sum_{i=1}^{l} \xi_i \\ & - \sum_{i=1}^{l} \alpha_i \{ y_i [(\bar{w} \cdot \bar{x}_i) + b] - 1 \\ & + \xi_i \} - \sum_{i=1}^{l} \beta_i \xi_i \end{aligned} \tag{13}$$

必须满足约束条件:

$$\begin{cases} \dfrac{\partial L}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = C\vartheta^{*} - \alpha_i - \beta_i = 0 \\ \dfrac{\partial L}{\partial b} = \alpha_i y_i = 0 \\ \alpha_i \{ y_i [(\bar{w} \cdot \bar{x}_i) + b] - 1 + \xi_i \} = 0, \forall i \\ \beta_i \cdot \xi_i = 0, \forall i \\ \alpha_i, \beta_i, \xi_i \geqslant 0, \forall i \end{cases} \tag{14}$$

定理 2:对于式(13)定义的 SVC 学习机,有:

$$\frac{error_{\text{train}}(+)}{error_{\text{train}}(-)} \approx \frac{N^{-} \cdot \vartheta^{-}}{N^{+} \cdot \vartheta^{+}} \tag{15}$$

证明: $\bar{w} - \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i = 0$, 代入式(13)得到:

$$L(\bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + C\vartheta^{*} \sum_{i=1}^{l} \xi_i - (\bar{w} \cdot \bar{w})$$

$$-\sum_{i=1}^{l}\alpha_i\{by_i-1+\xi_i\}-\sum_{i=1}^{l}\beta_i\xi_i \quad (16)$$

变换得到

$$L(\bar{\alpha})=-\frac{1}{2}(\bar{w}\cdot\bar{w})+C\vartheta^*\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i by_i+\sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}(\alpha_i+\beta_i)\xi_i \quad (17)$$

将 $C\vartheta_*-\alpha_i-\beta_i=0$ 代入式(17)得到

$$L(\bar{\alpha})=-\frac{1}{2}(\bar{w}\cdot\bar{w})+C\vartheta^*\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i by_i+\sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}C\vartheta^*\xi_i \quad (18)$$

消去同类项,代入 $\alpha_i y_i=0$ 得

$$L(\bar{\alpha})=-\frac{1}{2}(\bar{w}\cdot\bar{w})+\sum_{i=1}^{l}\alpha_i \quad (19)$$

该函数是一个仅仅关于 α_i 的函数,且 $\alpha_i y_i=0$, $C\vartheta^*=\alpha_i+\beta_i$, 为约束。得到对偶问题如下式:

$$\max\tilde{L}(\bar{\alpha})=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\bar{w}\cdot\bar{w} \quad (20)$$

分析其约束条件。

情形 1: $\alpha_i=0$, 则 $C\vartheta^*=\beta_i$ 不为 0,则 $\xi_i=0$, $y_i[(\bar{w}\cdot\bar{x}_i)+b]-1>0$, $\bar{x}_i$ 远离边界,被正确分隔。

情形 2: $0<\alpha_i<C\vartheta_*$, 则 $C\vartheta^*-\alpha_i=\beta_i$ 不为 0,则 $\xi_i=0$. $y_i[(\bar{w}\cdot\bar{x}_i)+b]-1=0$, $\bar{x}_i$ 处在无漂移的边界上,两边界截距之差为 2,分类正确。

情形 3: $\alpha_i=C\vartheta_*$, 则 $\beta_i=0$, 则 $y_i[(\bar{w}\cdot\bar{x}_i)+b]-1+\xi_i=0$。此情形下,分 2 种情况进行讨论:

如果 $0<\xi_i<1$, 则 $\bar{x}_i$ 夹在边界平面和决策平面中间,被正确归类,精度不理想,属边界内支持向量。

如果 $\xi_i>1$, $\bar{x}_i$ 被错误分类,属穿透支持向量。

分析错误训练率,同理可以得到

$$\frac{NSV^{out+}}{N^+}\leqslant\frac{D}{N^+\cdot C\cdot\vartheta^+}\leqslant\frac{NSV^+}{N^+} \quad (21)$$

$$error_{\text{train}}(1)\approx\frac{D}{N^+\cdot C\cdot\vartheta^+} \quad (22)$$

$$error_{\text{train}}(-1)\approx\frac{E}{N^-\cdot C\cdot\vartheta^-} \quad (23)$$

由于 $D=E$, 可以得到 $\frac{error_{train}(1)}{error_{train}(-1)}\approx\frac{N^-\cdot\vartheta^-}{N^+\cdot\vartheta^+}$, 即式(15)。证明毕。

定理 2 指出:对于稀疏训练集,可以通过调节校正因子 ϑ^* 来平衡训练错误率,即调节 ϑ^* 以维持

$$N^-\cdot\vartheta^-\approx N^+\cdot\vartheta^+ \quad (24)$$

根据错误率近似假设,类标签的 $error_{\text{test}}$ 就能趋近,从而表现出大体一致的预测准确率。

5 籽种影像特征提取

5.1 籽种样本标识

随机选取栽种小麦的实验田地块,收割后提取部分颗粒样本,组织育种专家组对籽种性状及品质等级进行评判[9]。品质分为 3 个等级,上品为甲等,良好为乙等,不满意为丙等。样本等级的最终确定取专家组的表决结果,编号并记录等级标志。

5.2 籽种内核 CT 影像获取

采用无损伤的 X 射线断层扫描技术取得以白色为背景的若干整齐有序籽种颗粒排列的 CT 图像,冲洗得到 13cm×18cm 的黑白胶片之后,用汉王 7600 扫描仪扫描胶片得到灰度图像。扫描参数为 72DPI,256 级灰度。

采纳专家意见,通过图像处理技术处理获得了 280 个籽种颗粒内核的几何参数数据。特征参数设计如下:

(1)籽种投影到水平面得到近似椭圆,该椭圆的面积 A;

(2)籽种椭圆的周长 G;

(3)致密性系数 $C=4\pi A/G^2$;

(4)长轴长度 L;

(5)短轴长度 W;

(6)不对称系数 $Coef$, 即长轴和短轴之比;

(7)籽种牙槽长度。牙槽(颗粒表面长而窄的沟槽)长度 Lg。

图 2 展示了 198 号样本的影像及其特征参数值及类标识。

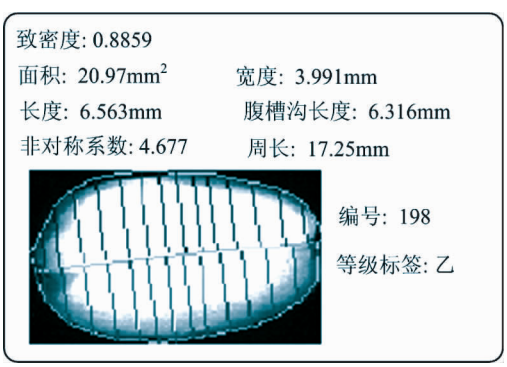

图 2 198 号籽种样本的特征参数

6 实验及结果分析

6.1 实验过程

6.1.1 数据预处理

算法代码测试是以在某个实数区间[0,1]的值为输入完成的。为了避免异常,拟将样本各个属性数据缩放到该测试区间。样本数据均非负值,则最小值映射为0,最高值映射为1进行缩放,整个数据集合共享某个缩放系数,这不会影响算法和样本对分类器的训练[11,12]。部分样本预处理后,在2个最大的主分量属性坐标系上投影得到散点图,如图3所示。

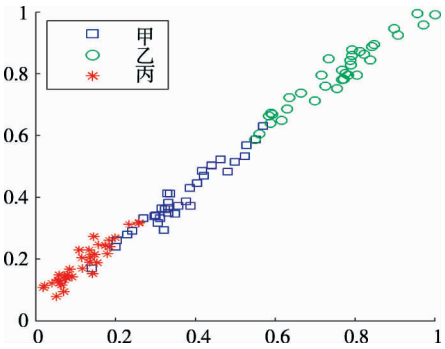

图3 部分籽种样本的2极大主分量散点图

6.1.2 稀疏比例设置

总体上,3类籽种的样本分布是不平衡的,为了多方面地开展对比实验,本次采样依然遵循"等量"原则。但是,为偏于观察实验效果,通过设置稀疏比例模拟得到稀疏训练样本集。甲、乙、丙3级别样本数量之比设定为10∶2∶1,借此来对比观察呈极性数量分布的样本集通过算法训练得到的甲、丙类的分类器校正效果。

6.1.3 核函数选定及参数优化

核函数采用径向基函数[13] $K(\bar{x}_1,\bar{x}_2)=e^{-\gamma|\bar{x}_1-\bar{x}_2|^2}$。这里需要考虑采用什么样的参数组合$(C,\gamma)$方能达到最好的定级效果。在事先没有经验的前提下,我们采用交叉验证的网格化搜索的办法优化组合(C,γ)。该方法的基本原理是"梯度下降法",即朝着准确率增加的方向寻找最优参数组合,给出准确率的等值线,为优化参数选择提供参考。

此外,本问题为多目标(multi-group)归类问题。多目标分类方法采用1-to-1方法[14],目标数量为3,训练过程将得到3个决策面,再采用表决多数胜出的方式确定类属。

参数优化得到准确率等值线如图4所示。可以看出:$C=128$,$\gamma=0.0078$,总体上可达到97.5%的准确率,较低准确率也达到95%,算法有效。

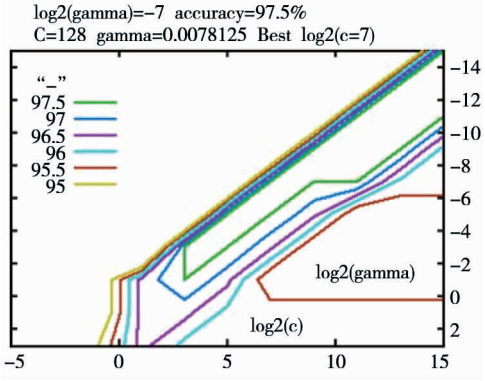

图4 交叉验证准确率 accuracy(C,γ)等值线图

6.2 结果分析、对比及讨论

6.2.1 错误率对比

参照上述参数,对模拟的稀疏训练集开展了文献[6]提出的方法和本方法(用This表示)的对比实验。结果显示:用文献[1]的方法训练得到的模型对等量的3类样本分别进行测试,丙类样本70有32个定级正确,38个错误,$error_{test}$(丙)=54.29%;而根据本文提出的方法,在惩罚校正系数$\vartheta^{丙}=10$时,$error_{test}$(丙)=17.14%,测试样本的错误率均值8.09%,其他数据见表1。和文献[1]算法、文献[6]算法相比,本文方案由稀疏样本集训练得到的分类器的预测错误率有明显改善。

表1 不同 SVC 算法聚类结果

方法	文献[1]	文献[6]	本文	ϑ^*
甲	1.43%(69/70)	1.43%(69/70)	4.29%(67/70)	1
乙	8.57%(64/70)	4.28%(67/70)	2.86%(68/70)	5
丙	54.29%(32/70)	27.14%(51/70)	17.14%(58/70)	10
均值	21.42%	10.95%	8.09%	—

6.2.2 SV 分布对比

选取第二组人工数据500个样本,稀疏比例同前,在此场景下,观察分类器的决策超平面获得的支持向量的改变。3目标(等级甲、乙、丙分别用1,2,3表示)的3个决策模型分别标记为12,13和23;BSV_{13}:13模型的夹在2个最佳边界平面之间的SV(即:界内SV、决策SV、穿透SV)数量;SV_{13}:13模型

的支持向量总数；ρ_{13} = BSV_{13}:SV_{13}，其含义即逆向训练样本在支持向量中所占的比例；其他符号意义类推。实验数据表明文献[1]算法的 ρ_{13}，ρ_{23} 最高，文献[2]居中，本文方案最低。可见：校正惩罚之后，错误、不精确的 SV 在 SV 全体中所占比例有明显改善。本质上就是减少了错误训练的次数，因而推动模型的预测错误率下行。表 2 给出了 SV 分布。

表 2 不同 SVC 算法的 SV 分布

算法	SV_{12}	BSV_{12}	SV_{13}	BSV_{13}	ρ_{13}	SV_{23}	BSV_{23}	ρ_{23}
文献[1]	15	13	13	11	0.84	6	4	0.66
文献[6]	8	6	8	7	0.62	2	1	0.5
本文	14	10	15	9	0.60	3	0	0

6.2.3 校正系数 ϑ^* 压力实验和过度校正

校正公式(式(24))描述的是一种近似关系，最优校正量仍需要在一定范围内搜索。如果不断增大稀疏样本对应类的校正系数 ϑ^*，需考虑其错误率会不会持续改善，两者会呈现什么样的波动。故此，我们设计了校正系数压力实验，结果曲线如图 5。

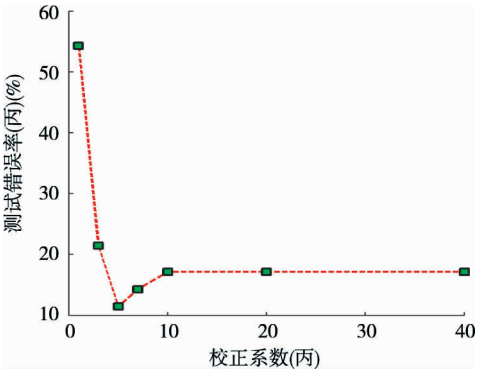

图 5 $error_{test}$(丙) – $\vartheta^{丙}$的压力试验折线图

可以看出：开始阶段，$error_{test}$(丙) 随着 $\vartheta^{丙}$的增加迅速减少；当降到一个局部极小值后，小幅度增加，而后收敛于一个稳定值。这称为过度校正，同过度拟合(Over-fitting)有类似的含义。它说明：预测错误率不会随着惩罚 ϑ^* 增加而无限改善，过度校正无益于改善分类器的泛化能力。校正只是一种补救性技术，实践中，应从学习效果角度，结合可接受的错误率合理选择校正系数，避免过度校正而对非稀疏类分类器训练造成不利的影响。

6.2.4 稀疏样本容量 N 压力实验

类似地，稀疏训练样本数量也影响着分类器的错误率。图 6 描述了错误率对于稀疏样本容量的压力实验数据。初期，随着 $N^{丙}$增加，$error_{test}$(丙) 迅速改善，而后平缓地改善，最后，趋于某个收敛值。SVN 是基于监督学习的网络，监督样本越多，机器能从通过样本训练学习到更多的“目标概念”的知识，从而更加准确地理解“目标概念”，降低错误率[15,16]，这吻合归纳学习的一般规律。即使在样本稀疏失衡的背景下，从学习和知识传播的角度，应当鼓励用户尽可能增加稀疏样本容量，以期待更好的训练效果[16]。

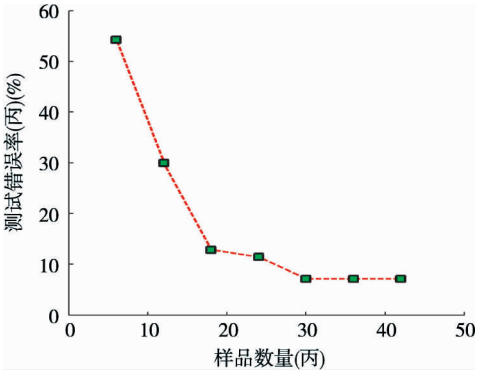

图 6 $error_{test}$(丙) – $N^{丙}$的压力试验折线图

7 结 论

本文探讨了用支持向量网络(SVN)聚类方法训练处理稀疏样本集时，分类器预测错误率呈现极性分布的问题，引入了界内支持向量、穿透支持向量、逆向训练样本和错误训练率等概念，基于拉格朗日系数分析方法，提出了惩罚校正的支持向量算法和校正方法，并以基于小麦籽种的 CT 图像特征的籽种品质定级为实验平台对该方法进行了测试。结果显示：通过该改进的机器学习方法，稀疏样本目标的准确率显著提高，整体定级性能也得到明显改善，这表明了该方法的有效性和适用性。具体而言，其具有以下特性：(1)对于稀疏样本，惩罚因子校正方法能在稳定整体性能的前提下大大改善学习效果；等值分析说明该方法能有效地处理籽种图像特征数据，准确率可达 97%。(2)压力测试数据说明其具有较好的收敛性，错误率和学习机的泛化能力会随着校正系数和样本容量较快收敛。本方法能显著改善 SVN 对于稀疏样本集合的学习性能，并具有较好的普适性，对于开发籽种品质智能评价系统和研究智能评价方法具有积极的现实意义。

参考文献

[1] Cortes C, Vapnik V. Support-vector network. *Machine*

Learning,1995. 20:273-297

[2] 周武杰,蒋刚毅,郁梅. 基于块内容和支持向量回归的图像质量客观评价模型. 高技术通讯,2012,22(11):1117-1123

[3] 杨志民,田英杰,刘广利. 城市空气质量评价中的模糊支持向量机方法. 中国农业大学学报. 2006,11(5):92-97

[4] Charytanowicz M,Niewczas J. A complete gradient clustering algorithm for features analysis of X-ray images,*advances in intelligent and soft computing*,69:15-24

[5] 杨志民,梁静,刘广利. 强模糊支持向量机在稻瘟病气象预警中的应用. 中国农业大学学报,2010,15(3):122-128

[6] Scholkopf B,Smola A,Williamson R C,et al. New support vector algorithms. *Neural Computation*,2000,12:1207-1245

[7] Wang J C,Hu J,Zhang C F. Assessment on evaluating parameters of rice core collections constructed by genotypic values and molecular marker information. *Rice Science*,2007,14(2):101-110

[8] Chang T T,Liu H W,Zhou S S. Large scale classification with local diversity AdaBoost SVM algorithm. *Journal of Systems engineering and electronics*. 2009,20(6):1344-1350

[9] 朱旻,李雪玲,李效来等. 基于元学习和叠加法的双层支持向量机算法. 人工智能和模式识别,2012,25(6):943-949

[10] Keerthi S S,Lin C J. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*,2003,15(7):1667-1689

[11] Segata N,Blanzieri E. Fast and scalable local kernel machines. *Journal of Machine Learning Research*,2010,11:1883-1926

[12] Dorff K C,Chambwe N,Srdanovic M,et al. Reproducible large-scale predictive model development and validation in high-throughput datasets. *Bioinformatics*, 2010, 26(19):2472-2473

[13] 王晓明,王士同. 最小类方差支持向量机与零空间分类器的集成. 模式识别与人工智能. 2010,23 (4):441-449

[14] Kowalski P, Lukasik S, Charytanowicz M. Data-driven fuzzy modeling and control with kernel density based clustering technique. *Polish Journal of Environmental Studies*,2008. 17:83-87

[15] 丁晓剑,赵银亮,李远成. 基于 SVM 的二次下降有效集算法. 电子学报,2011,39(8):1766-1770

[16] 朱婷婷,王丽娜,胡东辉等 基于不确定性推理的 JPEG 图像通用隐藏信息检测技术. 电子学报,2013,41(2):233-238

Research on a penalty regularization method for SVN based on sparse training set and its application to seed quality evaluating

Tan Wenxue[*,**], Zhao Chunjiang[***], Wu Huarui[***]

([*] College of Computer Science, Beijing University of Technology. Beijing, 100022, China)

([**] School of Computer Science, Hunan University of Arts and Science. Changde, 415000, China)

([***] National Engineering Research Center for Information Technology in Agriculture. Beijing, 100097, China)

Abstract

Aiming at the problem when an unbalanced, sparse training set is trained by using support-vector networks (SVN), the output classifiers are badly imbalanced in their mis-prediction rate, so they are badly unavailable, the novel concepts of penalty regularization coefficient, through-bound-SV, adverse training sample, mis-training rate and so on were introduced into this study based on the analysis of Lagrange multiplier, and the Regularized Penalty SVN learning algorithm was proposed; a method for regularizing penalty coefficient was designed. The algorithm was applied to grading wheat seeds quality based on a feature data set of CT image of the seeds. The contour analysis demonstrates it can effectively grade image features data with an accuracy rate of 95%. The results of the comparison experiment against some prior and congeneric algorithms suggests where a sparse training set is concerned. This method can reform the polar distribution of the mis-prediction rate of concerned classifiers, achieve an amusing accuracy of prediction, and present a very good global performance of machine learning.

Key words: penalty regularization, support-vector network, mis-training, sparse sample, adverse training