DOI：10.13989/j.cnki.0517-6611.2010.29.012
安徽农业科学，Journal of Anhui Agri Sci 2010，38(29)：16105-16106　　责任编辑　常俊香　责任校对　况玲玲

改进的近似支持向量机在葡萄酒质量鉴定中的应用

徐海涛　(聊城大学，山东聊城 252059)

摘要　在介绍标准支持向量机(SVM)的基础上，引出近似支持向量机(PSVM)方法的基本原理，并提出一个改进其分类精度的新算法(PSVM-2)。针对葡萄酒质量鉴定这一实际问题，比较了SVM与PSVM及PSVM-2的表现能力，分析了3种算法的复杂性及其分类性能。
关键词　支持向量机；PSVM；葡萄酒；质量鉴定
中图分类号　TS201　　**文献标识码**　A　　**文章编号**　0517-6611(2010)29-16105-02

PSVM-2 Method and Its Application on Identification of Wine Quality

XU Hai-tao　(Liaocheng University，Liaocheng，Shandong 252059)

Abstract　Based on introducing standard support vector machine(SVM)，the basic rationale of proximal support vector machine(PSVM) method was led out and a new algorithm(PSVM-2) was proposed to improve its classification precision. For the practical problem of quality assessment of grape wine，the performance abilities of SVM，PSVM and PSVM-2 were compared and their complexities and classification performances were analyzed.
Key words　Support vector machine；PSVM；Grape wine；Quality assessment

支持向量机(SVM)由Vapnik[1-2]等创立，是以统计学习理论为基础的机器学习方法，近年来已在许多领域得到成功应用。SVM分类器的分类效果很好，是最好的分类器之一。标准的支持向量机在使用时具有较高的精度，然而对样本个数较多的数据进行分类时，由于其时间复杂度高的显著缺陷，分类性能大大下降。近似支持向量机(psvm)是由Fung[3]和Mangasarian[4]提出的一种支持向量机，通过对标准的支持向量机进行参数和约束调整，将原问题转化为求一个无约束严格凸二次规划，由最优性条件知其最优解在目标函数梯度为0处取得，且最优解唯一。近似支持向量机算法简洁明了，并且由于使用了对分类算法中数据样本点个数与维数进行置换的技巧，使其在处理样本点个数远远大于其特征维数的分类问题时具有非常明显的优势。

程伟等[5]与韩勇鹏[6]分别就支持向量机在粮食产量预测与乳制品分类问题中的应用进行了详尽的研究。由于在葡萄酒质量鉴定过程中产生的合格样本和不合格样本的数目存在很大差距，这就造成了不平衡分类问题，针对此问题，笔者考虑使用改进的PSVM对葡萄酒的质量进行分类鉴定，并同时比较使用标准SVM和PSVM算法所用的时间及分类情况。

1　近似支持向量机(PSVM)

首先用PSVM对样本数据进行初步训练，得到其分类超平面的法向量，然后将样本数据投影到法向量上，设定一个新的惩罚参数矩阵G，用含有不同惩罚参数的改进PSVM对样本数据进行2次分类，得出比第1次结果更好的分类超平面。由于篇幅限制，该研究仅讨论线性核分类，且只考虑2类问题。

首先介绍标准SVM[7-8]，设已知训练集 $T=\{(x_1,y_1),\dots,(x_l,y_l)\}\in(X\times Y)^l$，式中，$x_i\in X=R^n$，$y_i\in Y=\{-1,1\}$，$i=1,\dots,l$选择惩罚参数 $C>0$，构造并求解对变量 w，b 和 $\xi=(\xi_1,\dots,\xi_l)^T$ 的最优化问题：

$$\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t}\quad y_i(<w,x_i>+b)+\xi_i\geqslant 1,\quad \xi_i\geqslant 0,\ i=1,\dots,l \tag{1}$$

求得最优解 w^*，b^* 和 ξ^*；构造分类超平面 $<w^*,x>+b^*=0$，由此求得决策函数。

针对上述问题，构造 l行 n列矩阵 $A=(x_1,\dots,x_l)^T$ 表示输入集，m行 m列对角矩阵 D，其中 $D_{ii}=y_i$，将上述问题(1)写成向量形式，并进行调整，即得近似支持向量机(PSVM)：

$$\min_{w,b,\xi}\ \frac{C}{2}\|\xi\|^2+\frac{1}{2}\|w,b\|^2$$
$$\text{s.t}\quad D(Aw+eb)+\xi=e \tag{2}$$

可以看出，由于(2)式的约束条件改成了等式约束，2个超平面 $<w,x_i>+b=\pm1$ 之内和之外的点均有可能产生误差，所以(2)式的损失函数改成了 $\frac{C}{2}\|\xi\|^2$，即PSVM所求的问题为使得正负类点分别在 $<w,x_i>+b=\pm1$ 附近聚集，且 $<w,x_i>+b=\pm1$ 的间隔最大。将(2)式约束条件中的 ξ 代入目标函数，得到与(2)式等价的问题：

$$(\text{PSVM})\ \min_{w,b}\ \frac{C}{2}\|e-D(Aw+eb)\|^2+\frac{1}{2}\|w,b\|^2 \tag{3}$$

令矩阵 $H=(A\quad e)$，$\beta=\begin{pmatrix}w\\b\end{pmatrix}$，则上式可简化为关于 β 的无约束严格凸二次规划：

$$\min_{\beta}\ f(\beta)=\frac{C}{2}\|e-DH\beta\|^2+\frac{1}{2}\|\beta\|^2 \tag{4}$$

则由解得最优性条件知，(4)式的最优解必在其梯度 $\nabla f(\beta)=0$ 处得到。由此可知：$\nabla f(\beta)=C(-DH)^T(e-DH\beta)+\beta=0$，解之得：

$$\beta=\begin{pmatrix}w\\b\end{pmatrix}=\left(\frac{I}{C}+H^TH\right)^{-1}H^TDe \tag{5}$$

2　改进PSVM的2次分类法(PSVM-2)

对不平衡数据样本进行分类时，分类超平面偏移性存在的原因有2种：一是2类样本数目有差异；二是与其分散程度有关。分类超平面会向分散程度较大的类偏移，因此，可用投影标准差表示正负类样本点在 w 上的分散程度。基于由(4)式求出的法向量 w，得到样本点在 w 上的投影 $<w,x_i$

作者简介　徐海涛(1979-)，男，山东平度人，硕士，馆员，从事数据挖掘与机器学习和模式识别研究。
收稿日期　2010-06-30

>, i=1, …, l令 l_+为正类点个数,分别计算 2类样本点的投影均值 μ_+、μ_-和投影标准差 σ_+、σ_-。其中, $\mu_+=\frac{1}{l_+}\sum_{i=1}^{l_+} w^T x_i$, $\mu_-=\frac{1}{l-l_+}\sum_{i=l_++1}^{l} w^T x_i$, $\sigma_+=\sqrt{\frac{1}{l_+-1}\sum_{i=1}^{l_+}(w^T x_i-\mu_+)^2}$, $\sigma_-=\sqrt{\frac{1}{l-l_+-1}\sum_{i=l_++1}^{l}(w^T x_i-\mu_-)^2}$。另外,由于分类超平面会向样本数目少的类偏移,为纠正其偏移性,该研究针对正负类样本分别采取不同的惩罚因子 C_+, C_-,并希望正、负类点产生的误差之和相等,即$\sum_{i=1}^{l_+}(C_+\xi_i)^2=\sum_{i=l_++1}^{l}(C_-\xi_i)^2$,同时希望相应于正负类点的损失 ξ的期望值相等,得$\frac{C_+}{C_-}\propto\frac{\sqrt{l-l_+}}{\sqrt{l_+}}$,同时又有$\frac{C_+}{C_-}\propto\frac{\sigma_+}{\sigma_-}$,所以可令: $C_+=C\frac{\sigma_+}{\sqrt{l_+}}$, $C_-=C\frac{\sigma_-}{\sqrt{l-l_+}}$。设新的惩罚参数矩阵 G为 l×l阶对角阵,且

$$\begin{cases} G_{ii}=C_+ & \text{if } D_{ii}=1 \\ G_{ii}=C_- & \text{if } D_{ii}=-1 \end{cases} \quad (6)$$

根据(6)式, PSVM可改进为:

(改进的 PSVM)

$$\min_{\beta} f(\beta)=\frac{1}{2}\| G(e-DH\beta)\|^2+\frac{1}{2}\|\beta\|^2 \quad (7)$$

同 PSVM解法,有$\nabla f(\beta)=C(-GDH)^T G(e-DH\beta)+\beta=0$,解之得:

$$\beta=\begin{pmatrix} w^* \\ b \end{pmatrix}=(I+H^T DGGDH)^{-1}H^T DGGe \quad (8)$$

得到改进后的决策函数 sgn(<w^*, x>+b^*)。

综上所述,可给出改进的 PSVM 2次分类算法: (PSVM-2)

(1)设已知训练集 $T=\{(x_1, y_1), \ldots, (x_l, y_l)\}\in (X\times Y)^l$,其中 $x_i\in X=R^n$为输入集, $y_i\in Y=\{-1, 1\}$为输出指标集, i=1, …, l构造 l行 n列矩阵 $A=(x_1, \ldots, x_l)^T$表示输入集, m行 m列对角矩阵 D,其中 $D_{ii}=y_i$;

(2)用原始的 PSVM(3)对输入集进行分类,得到初始的分类超平面的法向量 w;

(3)作出输入集样本点在 w上的投影 <w, x_i>, i=1, …, l令 l_+为正类点个数,分别计算 2类样本点的投影均值 μ_+、μ_-和投影标准差 σ_+、σ_-;

(4)设定新的惩罚参数矩阵 G,用改进的 PSVM(7)进行分类,得到改进后的分类超平面的法向量 w^*和决策函数 sgn(<w^*, x_i>+b^*)。

3 试验

为了验证改进的 PSVM方法的有效性,该研究从 UCI数据库中选取 1组关于葡萄酒质量鉴定的数据(Wine Quality Data Set数据集下载地址: http://archive.ics.uci.edu/ml/datasets/Wine+Quality)[9]进行试验,其中针对数据均选取全部数据同时作为训练集和测试集,并将多类问题通过合并转化为 2类问题。对于该数据集,其中样本总数为 1 599,维数为 12维,前 11维表示输入集,第 12维表示输出集。各维数分量分别表示葡萄酒的固定酸度、挥发酸度、柠檬酸含量、残余糖分、氯化物含量、自由二氧化硫含量、总二氧化硫含量、浓度、pH值、硫酸盐含量、酒精含量,最后 1维的输出表示葡萄酒的质量(分别用从数字 1～10表示葡萄酒的质量,数字越大表示质量越好)。为简单起见,该研究将葡萄酒质量参数不小于 5的均鉴定为合格品(输出设定为 1),其他的鉴定为非合格品(输出设定为 -1)。分别用 ERR+、ERR-、ERR表示支持向量机产生的正类错分率、负类错分率及总错分率,其中总错分率为正负类错分率的几何平均值。试验运行环境为 PC机 AMD双核 CPU4400+, 1 G内存, MATLAB6.5,其中惩罚参数 C=100。分别用 SVM、PSVM和 PSVM-2对数据进行分类,结果见表 1。

表 1 不同 SVM对葡萄酒质量鉴定结果

Table 1 Identification of wine quality with different SVM method

分类方法 Classification method	算法所用时间/s Time	错分率/% Error Rate ERR+	ERR-	ERR
SVM	610	15	18	17.4
PSVM	0.109	100	0	100
PSVM-2	0.329	34.82	11.11	23.89

由表 1可知,原始 SVM算法的分类能力最强,出现的错分点最少,但此算法所用时间为 610 s是后面 2种算法耗时的 10^3倍。PSVM算法虽然极大的提高了程序计算的效率,但分类结果非常差,竟出现了将正类点全部分错的情况。改进的算法 PSVM-2是目前耗时相对较少(0.329 s)且分类精度较高的一种算法,基本上达到了原始 SVM算法的分类精度。在现有的计算复杂度下,如何进一步提高 PSVM-2算法的分类精度将成为下一步的研究方向。

参考文献

[1] VAPNIK VLADIMIR N. Statistic learning theory[M]. New York: John Wiley & Sons, 1998.

[2] VAPNIK VLADIMIR N. The nature of statistical learning theory[M]. 2nd. New York: Springer, 2000.

[3] GLENN FUNG, MANGASARIAN O L. Proximal support vector machine classifiers[M]. New York: Association for Computing Machinery, 2001: 77-86.

[4] MANGASARIAN O L, MUSICANT DAVID R. Lagrangian support vector machines[J]. Journal of Machine Learning Research, 2001(1): 161-177.

[5] 程伟,张燕平,赵姝.支持向量机在粮食产量预测中的应用[J].安徽农业科学, 2009, 37(8): 3347 - 3348.

[6] 韩勇鹏. SVM方法及其在乳制品分类问题上的应用[J].安徽农业科学, 2009, 37(8): 3345 - 3346.

[7] 邓乃扬,田英杰.数据挖掘中的新方法—支持向量机[M].北京:科学出版社, 2004.

[8] INGO STEINWART, ANDREAS CHRISTMANN. Support vector machines[M]. Springer, 2008.

[9] ASUNCION A, NEWMAN D J. UCI Machine Learning Repository[M]. Irvine, CA: University of California, School of Information and Computer Science, 2007.