文章编号:1671 - 6833(2017)03 - 0049 - 05

结合投影与近邻操作的支持向量快速筛选方法

李蒙蒙, 尚志刚, 李志辉

(郑州大学 电气工程学院,河南 郑州 450001)

摘 要: 为减少支持向量机(SVM)的计算负担并保证分类精度并提高运算效率,提出一种结合投影与近邻操作的支持向量快速筛选方法. 该方法利用 Fisher 投影轴的全局特性将其作为 SVM 最优分类面的近似法方向,在该方向快速筛除大量非支持向量,将分类边界附近的样本集作为备选支持向量集,同时为解决投影操作未考虑样本局部结构信息造成的误删支持向量的问题,结合近邻操作回选样本空间中备选支持向量的近邻样本更新扩充备选支持向量集,以该子集中的样本作为 SVM 的输入. 在多个 UCI 标准数据集上的实验结果表明,该方法在充分保证分类精度的前提下有效降低了 SVM 的计算负担,具有较好的推广性.

关键词: 支持向量机;支持向量;Fisher 投影;k-近邻;快速筛选

中图分类号: TP391.4 **文献标志码**: A doi:10.13705/j.issn.1671 - 6833.2016.06.003

0 引言

支持向量机(support vector machine,SVM)是建立在统计学习 VC 维理论和结构风险最小化基础上的机器学习方法,其学习过程实际是求解一个二次规划问题,需用到所有训练样本的 Hessian 矩阵,故遇到样本集较大的学习问题时,传统方法内存消耗过大且学习速度缓慢,从而影响了它的实用价值和推广. 针对这一问题,近年出现了许多改进算法提高 SVM 对大样本集的学习速度. 一种思路是改进优化方法,以 Keerthi 等提出的循环最近点算法[1]、Platt 提出的序贯最小优化算法最具代表性. 另一种思路是通过某种处理缩减样本集得到规模较小的替代集而又不影响分类精度,如文献[2 - 3]提出的局部支持向量机算法,利用 K 均值聚类生成的样本中心点集作为替代;文献[4]提出基于距离排序的快速支持向量机分类算法;文献[5]提出基于 k-近邻法的快速训练算法;文献[6]提出了基于 Fisher 鉴别分析的训练样本缩减策略. 上述算法筛除了冗余样本,减少了无谓运算,提高了运算速度. 但由于数据压缩往往会误删一部分支持向量,破坏原分类边界,造成一定程度的精度下降. 考虑到 SVM 确定的支持向量均靠近分类边界并最终决定最优分类面,而其他大量非支持向量则属于冗余样本,故如何在保持原样本集分类边界较完整的条件下,快速筛除与分类无关的非支持向量具有重要的研究价值.

为快速准确地筛除大量非支持向量,提高 SVM 确定支持向量的计算效率,笔者提出一种结合投影与近邻操作的方法,利用 Fisher 投影的全局特性,快速粗略地筛除大量远离分类边界的冗余样本以减少计算量,结合邻域选择可以保留数据局部结构信息的优势,进行样本回选以避免对支持向量的误删,保证分类边界信息的完整性.

1 支持向量机的原理与问题

SVM 基于结构风险最小化理论在样本空间中构造最优超平面. 对线性可分样本集,必然存在最优超平面保证在分类间隔最大的条件下正确划分所有训练样本,这同时保证了经验风险最小和结构风险最小,从而达到期望风险最小化[7—9].

设数据集中的两类样本分别可标记为 $\{\boldsymbol{x}_i, y_i\}, i=1,\cdots,l$,其中,$l$ 为样本的总数,y_i 为类别标签且 $y_i \in \{-1,1\}$. SVM 的求解问题是一个对于不等式约束的条件极值问题,引入非负的拉格朗日系数 α_i,可表述为如下的拉格朗日方程:

收稿日期:2016 - 06 - 08;**修订日期**:2016 - 08 - 18

基金项目:国家自然科学基金资助项目(U1304602,61473266);河南省高等学校重点科研资助项目(15A120016)

通信作者:尚志刚(1975—),男,甘肃兰州人,郑州大学副教授,博士,主要从事生物医学信息与模式识别研究,E-mail:zhigang_shang@ zzu.edu.cn.

$$L(\boldsymbol{w},b,\alpha_i) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l}\alpha_i(y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1). \quad (1)$$

将上式转化为较简单的"对偶"形式为:

$$W(\alpha_i) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i \cdot \boldsymbol{x}_j). \quad (2)$$

求解使 W 取最大值的 α,若 α^* 为最优解,则最优分类面的法向量可表示为 $\boldsymbol{w}^* = \sum_{i=1}^{l}\alpha^* y_i \boldsymbol{x}_i$ [10-11].

上述计算需要用到所有训练样本的 Hessian 阵,而实际上非支持向量对应的 α^* 均为0,对于最优分类面的确定没有贡献,因此如何快速可靠地筛除非支持向量就成为提高 SVM 计算速度的关键.

2 结合 Fisher 投影与近邻法的支持向量快速筛选方法

SVM 中靠近分类边界的支持向量对确定最优分类面较重要,而远离边界的样本可视作冗余样本进行筛除. Fisher 最佳投影轴可近似视为 SVM 分类面的法方向,故可将其引入 SVM 中快速筛除冗余样本以缩减样本集,从而更快筛选出支持向量.但由于 Fisher 投影侧重样本集的全局结构信息,而未考虑局部结构信息可能误删支持向量,故笔者拟采用结合投影和近邻操作的方法解决这一问题,以快速准确筛选出潜在支持向量.

2.1 Fisher 投影筛除非支持向量

Fisher 投影基于 Fisher 准则[12]寻求最佳方向,使所有特征点在该方向得到最好的分类.以二分类为例,类均值向量为 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$,原空间类内散度矩阵为 $\boldsymbol{S}_w$,类间散度矩阵为 $\boldsymbol{S}_b$,所有样本投影到一维空间后类内散度和类间散度变为具体值,以 $\tilde{S}_w$ 和 $\tilde{S}_b$ 表示,此时 Fisher 准则函数为:

$$\max J(\boldsymbol{w}) = \frac{\tilde{S}_b}{\tilde{S}_w} = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}. \quad (3)$$

为求满足上式的投影方向 $\boldsymbol{w}$,引入拉格朗日乘子求解无约束极值问题,得到最佳投影方向:

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2). \quad (4)$$

所有样本向 $\boldsymbol{w}$ 投影得对应投影值,但这种向一维空间的特征压缩不像 SVM 考虑了分类面附近的局部结构信息,故推理知在此基础上筛除冗余样本会不可避免地误删一部分潜在支持向量.

笔者采用数值实验验证上述推理,Fisher 投影筛除方法如下:对 $\boldsymbol{X}$ 中任一样本 $\boldsymbol{x}$,其投影表示为 $z = \boldsymbol{w}^T\boldsymbol{x}$.选取两类样本最靠近边界者的均值为基准点,记作 z_0,逻辑推理知远离 z_0 的是非支持向量冗余样本,反之则更有可能是 SVM 确定的支持向量.记 z_0 两侧投影点对应的原始样本集为 $\boldsymbol{X}^+$ 和 $\boldsymbol{X}^-$,分别计算它们与 z_0 的距离,记:

$$\boldsymbol{D}^+ = \{|\boldsymbol{w}^T\boldsymbol{x} - z_0| \mid \boldsymbol{x} \in \boldsymbol{X}^+\} = \{D_1^+, D_2^+, \cdots, D_{n_1}^+\}. \quad (5)$$

$$\boldsymbol{D}^- = \{|\boldsymbol{w}^T\boldsymbol{x} - z_0| \mid \boldsymbol{x} \in \boldsymbol{X}^-\} = \{D_1^-, D_2^-, \cdots, D_{n_2}^-\}. \quad (6)$$

其中,1,2,…,n_1;1,2,…,n_2 是排序前的序号.

对 $\boldsymbol{D}^+$ 和 $\boldsymbol{D}^-$ 排序,以(1),(2),…,(k_1),…,(n_1);(1),(2),…,(k_2),…,(n_2)作为排序后的新序号,记:

$$D_{(1)}^+ \leqslant D_{(2)}^+ \leqslant \cdots \leqslant D_{(k_1)}^+ \leqslant \cdots \leqslant D_{(n_1)}^+. \quad (7)$$

$$D_{(1)}^- \leqslant D_{(2)}^- \leqslant \cdots \leqslant D_{(k_2)}^- \leqslant \cdots \leqslant D_{(n_2)}^-. \quad (8)$$

则潜在支持向量与 $D_{(1)}^+$ 和 $D_{(1)}^-$ 对应的样本相邻较近,故可设定一定的样本筛除参数 $r(0 < r < 1)$筛除非支持向量,即满足以下条件的点对应的样本,其集合记为 $\boldsymbol{F}$:

$$\begin{cases} P\{D_{(k_1+1)}^+, D_{(k_1+1)}^+, \cdots, D_{(n_1)}^+\} = r; \\ P\{D_{(k_2+1)}^-, D_{(k_2+1)}^-, \cdots, D_{(n_2)}^-\} = r. \end{cases} \quad (9)$$

剩余的点组成的并集可表示为 $\boldsymbol{D} = \{D_{(1)}^+, D_{(2)}^+, \cdots, D_{(k_1)}^+\} \cup \{D_{(1)}^-, D_{(2)}^-, \cdots, D_{(k_2)}^-\}$,它们对应的样本构成备选支持向量集 $\boldsymbol{G}$.

以 Iris 数据集为例做上述处理,结果见图1.其中 l_1 为样本集 $\boldsymbol{X}$ 在 Fisher 最佳投影轴上的分布,l_2 为 SVM 确定的支持向量集 $\boldsymbol{S}$ 的分布,l_3 为用上述方法处理得到的备选支持向量集 $\boldsymbol{G}$ 的分布.

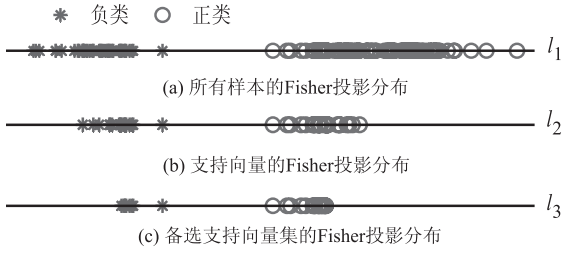


图1 仅采用 Fisher 投影筛除非支持向量的效果

Fig.1 The effect of filtering out non-support-vector using Fisher projection only

从图1可看出,仅采用 Fisher 投影筛除冗余样本会误删部分支持向量,故需考虑 $\boldsymbol{G}$ 的扩充更新,为此引入 k-近邻法从样本空间回选备选支持向量的近邻补充合并为新的备选支持向量集.

2.2 k-近邻法回选扩充备选支持向量集

k-近邻法是基于距离的度量方法[13].采用欧

氏距离作为确定近邻的测度,在原始样本空间中回选 k 个与备选支持向量距离最近的样本更新扩充备选集. 遍历备选支持向量集 $\boldsymbol{G}$,对其中任一元素 $\boldsymbol{g}$,计算其与 $\boldsymbol{F}$ 中样本之间的距离,记:

$$d_j = \mathrm{norm}(\boldsymbol{g} - \boldsymbol{F}_j) = \{d_1, d_2, \cdots, d_j\}. \quad (10)$$

对 d_j 排序,记:$d_{(1)} \leqslant d_{(2)} \leqslant \cdots \leqslant d_{(j)}$,设定近邻样本回选参数 $k(1 \leqslant k)$,从中选取前 k 个较小值对应的样本,即距离 $\boldsymbol{g}$ 最近的 k 个样本组成回选更新集$\boldsymbol{g}^*$,合并所有备选支持向量的更新集作唯一化处理剔除重复样本,得到的合集记为$\boldsymbol{G}^*$,则 $\boldsymbol{G} \cup \boldsymbol{G}^*$ 即为补偿更新后的备选支持向量集.

结合近邻操作回选更新备选支持向量集后的非支持向量筛除效果如图 2 所示,l_4 为采用上述方法更新备选支持向量集后的投影分布.

对比图 1 与图 2 表明对 $\boldsymbol{G}$ 进行上述回选更新后可更准确地提取潜在支持向量,进而更好地保留数据集的边界信息,为确定最优分类面提供可靠支持. 笔者将这种结合 Fisher 投影和 k-近邻法快速筛选支持向量的方法称为 Fisher Projection-k Nearest Neighbor_Support Vector Filter,记作 FP-kNN_SVF.

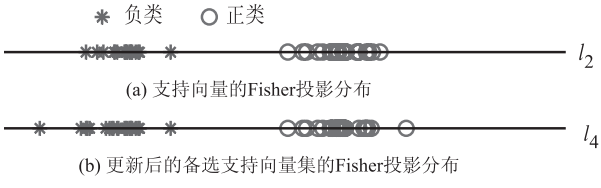


图 2　结合近邻操作回选更新后的非支持向量筛除效果

Fig. 2　The effect of filtering out non-support-vector combining nearest neighbor selection with Fisher projection

2.3　FP-kNN_SVF 方法的步骤

笔者提出的 FP-kNN_SVF 方法流程见图 3. 具体步骤表述如下:

(1) 输入数据集,设定筛除率 r 和近邻参数 k;

(2) Fisher 投影得到投影值并计算基准点;

(3) 计算各投影值与基准点的距离并排序,按 r 由式(9)筛除非支持向量得到备选支持向量集;

(4) 计算每个备选支持向量与筛除样本间的距离并排序,按 k 回选备选支持向量的近邻更新备选支持向量集,得到新的备选支持向量集.

3　数值实验与结果分析

为考察 FP-kNN_SVF 方法在不同数据集上的应用性能,包括其计算速度、分类精度与鲁棒性,以及参数 r 和 k 对实验结果的影响,笔者开展了相关数值实验. 文中所有计算均采用 MATLAB 软件编程实现,实验条件为 Intel(R) Pentium(R) D CPU 3.20 GHz/4.00 GB/ Windows XP/MATLAB 8.0.

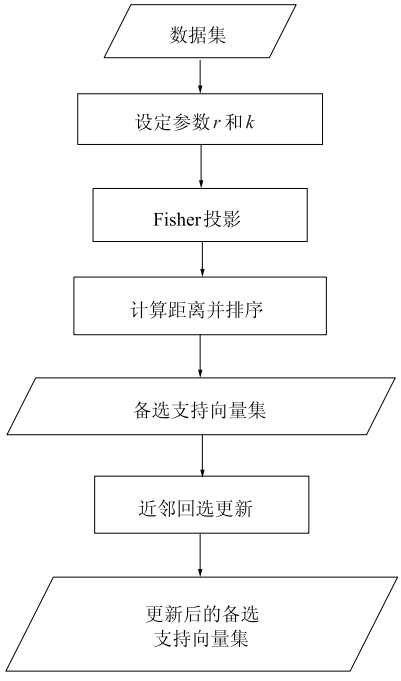


图 3　FP-kNN_SVF 方法流程图

Fig. 3　The flow chart of the FP-kNN_SVF method

3.1　FP-kNN_SVF 方法的有效性评估

考虑在样本数量和特征数量上多样化的条件下对比结果差异,选取 UCI 数据库中 6 个数据集进行实验. 采用交叉互验证,随机选取每个数据集 70% 的样本作训练集,30% 作测试集,随机抽样 10 次分别实验,以 10 次的平均结果作为 3 种方法的最终结果加以比较,分类效果如表 1 所示. (其中 Iris 和 Wine 数据集均原有 3 类,分别将两者的第一类作为正类,其余两类合并作为负类.)

实验结果表明:①F_SVM 和笔者提出的 F-k_SVM 都可以快速筛除样本,并较大幅度降低运行时间,F_SVM 运行时间只占 SVM 的 28% 以下,F-k_SVM 的运行时间也都只占 SVM 的 64% 以下,F_SVM 在的表现更突出;②3 种方法中,F_SVM 的分类精度最低,而 F-k_SVM 的分类精度高于 F_SVM,与原始 SVM 相当甚至高出 SVM,且其鲁棒性也高于其他两种方法.

综合以上两点说明:F-k_SVM 在不同数据集上均有良好的表现,能够充分保证分类的准确性和鲁棒性,并提高计算效率.

3.2 筛除率 r 对结果的影响

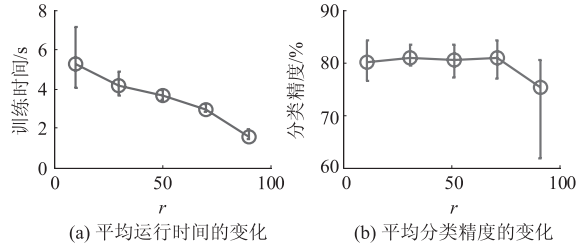
对于样本筛除率 r 的选择,笔者以 Australian 数据集为例,分别令 $r=10,30,50,70,90$,各进行10次重复随机抽样实验,记录10次的平均运行时间与分类精度,取 $k=5$,结果如图4所示.

表1 SVM、F_SVM 和 F-k_SVM 在6种数据集上的分类效果比较

Tab.1 Comparison of classification effect of SVM、F_SVM and F-k_SVM on 6 data sets

数据集	样本数量	特征维数	筛除率 r/%	运行时间/s			分类精度/%		
				SVM	F_SVM	F-k_SVM	SVM	F_SVM	F-k_SVM
Iris	150	4	70	0.38	0.01	0.08	98.89±1.15	96.00±3.70	99.33±1.07
Wine	178	13	70	0.24	0.02	0.08	92.07±3.69	73.77±4.38	92.26±3.56
Heart	270	13	30	0.34	0.05	0.16	70.00±5.63	62.47±6.92	71.23±4.90
Breast	683	10	25	27.62	7.70	13.31	96.00±1.33	95.46±1.26	96.15±1.20
Australian	690	14	25	17.73	3.44	4.60	79.61±3.23	75.80±3.00	80.05±1.43
Pima	768	8	25	35.54	8.04	15.91	68.57±2.40	66.96±3.49	68.91±2.16

注:不筛除样本的方法记作 SVM,文献[8]中仅采用 Fisher 投影筛除样本后结合 SVM 的方法记作 F_SVM,笔者提出的 FP-kNN_SVF 方法结合 SVM 记作 F-k_SVM.



(a) 平均运行时间的变化 (b) 平均分类精度的变化

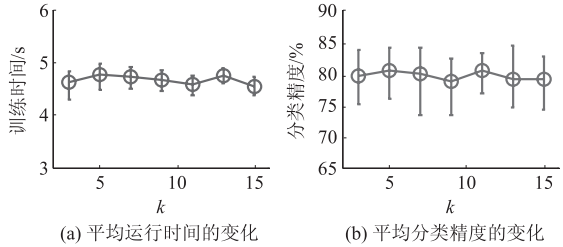
图4 不同 r 值下的实验结果($k=5$)

Fig.4 The test results with different r ($k=5$)

图4表明,平均运行时间随 r 的增大而降低,这是由于 r 越大,保留的需处理样本越少,运算时间也对应降低;r 为10~70时平均分类精度变化不大,而 r 为90时精度明显下降,这是由于筛除了过量样本造成支持向量损失严重引起的.故针对不同数据集,应根据其实际的数据结构设置不同的 r 以获得时间与精度之间最好的折衷.

3.3 参数 k 对结果的影响

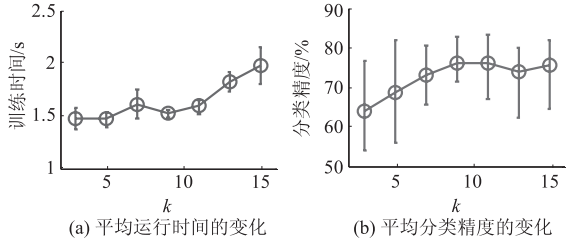
对近邻参数 k 的选择,笔者仍以 Australian 数据集为例,分别令 $k=3,5,7,9,11,13,15$,各进行10次随机抽样实验,记录平均运行时间与分类精度,r 为25和90时的结果如图5、图6所示.



(a) 平均运行时间的变化 (b) 平均分类精度的变化

图5 不同 k 值下的实验结果($r=25$)

Fig.5 The test results with different k ($r=25$)



(a) 平均运行时间的变化 (b) 平均分类精度的变化

图6 不同 k 值下的实验结果($r=90$)

Fig.6 The test results with different k ($r=90$)

图5、图6的结果表明:r 取25时,平均运行时间和分类精度随 k 的变化仅发生较小波动,即 k 对结果无较大影响;r 取90时,平均运行时间和分类精度随 k 的变化均呈上升趋势,即此时 k 对结果产生了一定的影响.这是因为,r 取25时样本筛除结果相对安全,较多支持向量被保留,而近邻回选得到的多数结果相互重合,造成不同 k 值下更新后的备选支持向量集相差不大,故实验结果波动不大;而 r 取90时由于筛除过程过于贪婪,误删了一部分支持向量造成分类精度的下降,而通过逐渐增大 k,可以扩充备选支持向量集恢复已被破坏的数据结构,这一方面增加了分类的准确性,但也因运算量的增加造成运行时间的上升.故实际操作中应根据不同的 r 下的结果确定 k 的取值以改善算法性能,即 r 取值合适时应尽量选取较小的 k 减少运算;r 取值过大时应尽量选取较大的 k 值以保证精度.

4 结论

笔者利用 Fisher 投影的全局性质与 k-近邻法保留局部结构信息的作用,提出了一种结合投影

与近邻操作的支持向量快速筛选新方法,通过在实际数据集上的应用得到如下结论:

(1)笔者提出的 F-k_SVM 可以快速准确筛选支持向量以提高运行速度,并能获得更高的分类精度和鲁棒性,有效解决了 F_SVM 精度下降的问题,具有较好的推广性.

(2)对不同数据集设置合适的筛除率 r 可同时保证较少的运行时间与较高的分类精度.

(3)设定不同的筛除率时,应根据实际情况进行参数 k 的选择以获得较好的实验效果.

笔者提出的 F-k_SVM 基于线性层面解决大样本集问题,扩展到核空间可能会更好解决复杂非线性问题;另外,文中筛除率 r 是事先设定的,如何根据数据集的结构自适应地确定 r 也很有意义,故上述两点将成为笔者下阶段的主攻方向.

参考文献:

[1] KEERTHI S S, SHEVADE S K, BHATTACHARYYA C, et al. A fast iterative nearest point algorithm for support vector machine classifier design[J]. Neural networks IEEE transactions on, 2000, 11(1): 124 – 136.

[2] 田新梅,吴秀清,刘莉. 大样本情况下的一种新的 SVM 迭代算法[J]. 计算机工程, 2007, 33(8): 205 – 207.

[3] 浩庆波,牟少敏,尹传环,等. 一种基于聚类的快速局部支持向量机算法[J]. 山东大学学报(工学版), 2015, 45(1): 13 – 18.

[4] 胡志军,王鸿斌,张惠斌. 基于距离排序的快速支持向量机分类算法[J]. 计算机应用与软件, 2013, 30(4): 85 – 87 + 100.

[5] 孙发圣,肖怀铁. 基于 K 最近邻的支持向量机快速训练算法[J]. 电光与控制, 2008, 15(6): 44 – 47.

[6] 饶刚,刘琼荪. 基于 Fisher 鉴别分析的支持向量机训练样本缩减策略[J]. 计算机工程与应用, 2012, 48(3): 156 – 157.

[7] 顾亚祥,丁世飞. 支持向量机研究进展[J]. 计算机科学, 2011, 38(2): 14 – 17.

[8] NANDHINI K, SANTHI B. Retrospection of SVM classifier[J]. Journal of theoretical and applied information technology, 2012, 38(1): 83 – 88.

[9] 张震,张英杰. 基于支持向量机与 Hamming 距离的虹膜识别方法[J]. 郑州大学学报(工学版), 2015, 36(3): 25 – 29.

[10] VAIDYA J, YU H, JIANG X. Privacy-preserving SVM classification[J]. Knowledge & information systems, 2008, 14(2): 161 – 178.

[11] 张炎亮,刘阳,王金凤. 基于改进 SVM 的煤矿水灾害救援组织系统可靠性预测[J]. 郑州大学学报(工学版), 2015, 36(3): 115 – 119.

[12] 陈立江,毛峡, ISHIZUKA M. 基于 Fisher 准则与 SVM 的分层语音情感识别[J]. 模式识别与人工智能, 2012, 25(4): 604 – 609.

[13] 曹根,葛孝堃,杨丽琴. 基于 K-近邻法的局部加权朴素贝叶斯分类算法[J]. 计算机应用与软件, 2011, 28(9): 267 – 268.

Fast Method to Filter Support Vectors Combined with Operation of Projection and Nearest Neighbors' Selection

LI Mengmeng, SHANG Zhigang, LI Zhihui

(School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: To reduce computational burden and improve operation efficiency of support vector machine (SVM) while ensuring classification accuracy, a fast method to filter support vectors combined with operation of projection and nearest neighbors' selection was proposed. Considering the global characteristics of Fisher projection, it can be viewed as the approximate normal directions of SVM optimal hyperplane and filter out a large number of non-support-vectors in this direction. The samples near the classification obtained boundary were regarded as alternative support vectors set. Neighborhood operation was combined to solve the problem that some support vectors might be filtered out mistakenly regardless of the local structure information. A number of nearest neighbors of the alternative support vectors were selected backward from the samples space to update and expand the alternative support vectors set. The sets was treated as the SVM input. The experimental results on several UCI standard data sets showed that the fast method had good generalization performance and reduced the computational burden effectively under the premise of fully guaranteed classification accuracy.

Key words: SVM; support vector; Fisher projection; k-nearest neighbor; rapid filter