面向传感网络多源数据融合的 SVM 方法*

陈　燕[1]，顾大刚[1]，陈亚林[2]
（1.贵阳学院 数学与信息科学学院，贵州 贵阳 550002；2.南京财经大学 管理科学与工程学院，江苏 南京 210046）

摘　要： 由于多源传感数据及其噪声构成复杂的非线性可分空间，数据融合是目前在资源受限的传感网络中安全、准确和高效地消除冗余数据的重要方法。结合 SVM 泛化能力强、凸优化的特点，侧重分析了非线性可分多源数据集转化为高维线性可分空间的可行性方法。仿真实验结果表明，宽度参数范围预估方法可以加速高斯核宽度参数的确定。针对多分类情形，仿真实验结果表明，通过控制误差积累，更能确保分类的有效性。

关键词： 数据融合；支持向量机(SVM)；高斯核函数；DAG-SVMs
中图分类号： TN925；TP391.4　　**文献标识码：** A　　DOI：10.16157/j.issn.0258-7998.201073

中文引用格式： 陈燕，顾大刚，陈亚林. 面向传感网络多源数据融合的 SVM 方法[J].电子技术应用，2021，47(11)：25-28.
英文引用格式： Chen Yan，Gu Dagang，Chen Yalin. SVM method for multi-source data fusion of sensor networks[J]. Application of Electronic Technique，2021，47(11)：25-28.

SVM method for multi-source data fusion of sensor networks

Chen Yan[1]，Gu Dagang[1]，Chen Yalin[2]
（1.School of Mathematics and Information Science，Guiyang University，Guiyang 550002，China；
2.School of Management Science，Nanjing University of Finance & Economics，Nanjing 210046，China）

Abstract： The complex nonlinear separable space is composed of multi-source sensing data and its noise. Data fusion is an important method for eliminating redundant data safely, accurately and efficiently in resource-constrained sensor networks. Because of SVM generalization ability and its convex optimization, this paper focuses on the feasibility of transforming nonlinearly separable multi-source data sets into high-dimensional linear separable spaces, based on the simulation experiment. The method based on the width parameter range estimation can accurately determine the width parameter of Gaussian kernel. For the multiple classification, the stimulation experiment show, by controlling the accumulation of errors,it is more effective to ensure the classification.

Key words： data fusion；support vector machine(SVM)；Gaussian kernel；DAG-SVMs

0 引言

为了提高传感网络观测数据的可靠性和准确度，需要复置传感器，使得多个节点有相互交叉的观测范围，这将导致采集的数据存在大量冗余，极大地增加了数据存储、处理和传输的资源消耗。如何在传感网络数据聚集过程中安全、准确和高效地消除冗余数据，是资源受限的传感网络应用研究中的核心问题之一，而数据融合是目前解决这一问题较为有效的一种方法。所谓数据融合是指将同一目标的多个观测结果整合成一个统一的结果[1]。数据融合技术主要是通过压缩数据、提取特征和数据关联等手段降低数据中的信息冗余，从而降低传感网络的资源消耗，增加置信度[2]。

数据融合的质量和效率主要体现在融合算法上。目前常用的数据融合算法有：基于时间序列的加权平均法，其方法简单，处理速度快，但融合质量较差[3-4]；利用概率分布求融合值的贝叶斯估计法，提高了融合结果的精度，但误差较难控制[5-6]；充分利用概率分布函数、似然函数和信任函数的 D-S 推理方法[7-8]，将多个信息融合的不确定性推理过程融合于模糊逻辑推理过程中，但信息描述依赖于主观因素，不利于特征提取[9-10]；神经网络算法，针对数据融合不确定性推理过程，通过训练，能拥有相应传感网络的规律，再利用规律进行数据融合[11-13]。

支持向量机 SVM 是一种基于统计学的二分类算法。其基本思想是通过学习得到一个边界函数，使得训练集中所有实例距离边界的最小距离最大化[14]。SVM 基于结构风险最小化原则，泛化能力强，且它是一个凸优化问题，只要局部最优解一定是全局最优解。

1 面向多源数据融合的 SVM 方法

控制空间的维数和设计高效的 SVM 多分类器是面向多源数据融合的 SVM 方法需要解决的核心问题。

* 基金项目：贵阳市科技局贵阳学院专项(GYU-KY-(2021))；教育部青年基金项目(18YJCZH016)

1.1 非线性向量空间的转化

转化的实质是寻找一种映射关系将非线性向量空间映射为线性可分空间。由于SVM使用向量内积表示原始数据点，考虑到在映射后的新空间求内积，会使空间的维数急剧增加，甚至无法计算。因此，通常使用核函数的方法，在输入空间就完成向量内积的计算。使用核函数完成向量空间转化主要考虑两个问题：(1)选择合适的核函数；(2)确定核函数中的参数。

Mercer定理认为任何半正定的函数都可以作为核函数。可以根据泛化误差理论，计算模型的期望错误率 EP_e^{n-1} 为：

$$EP_e^{n-1}=\frac{1}{n}E(L(x_1,y_1,\cdots,x_n,y_n)) \tag{1}$$

式中，x_i、y_i 为数据集中的元素，L 为 x 到 y 的映射。针对SVM，使用非线性样本训练集对多项式核函数和高斯核函数计算 EP_e^{n-1}，得出高斯核的错误率较低，故此选择高斯核函数来完成向量空间的转化。

预处理后的数据集应用高斯核函数生成核矩阵过程如下：

当宽度参数 $\sigma>0$ 时，高斯核函数 K 可定义为：

$$K(x,z)=e^{-\frac{||x-z||^2}{2\sigma^2}} \tag{2}$$

由式(2)可知，σ 直接影响着核函数的拟合程度。因此，高斯核的使用重点应是确定其宽度参数 σ，通常采用梯度下降法或交叉验证法来确定 σ。但当初始 σ 确定不当，会极大地增加计算代价。陈洋洋等提出首先基于支持向量间的距离来限制 σ 的选择范围，再采用常规方法来确定 σ，从而加快 σ 参数的确定[15]。

首先，计算任意两类支持向量集 A、B 之间近似的宽度参数 σ_0：

$$\sigma_0=\sqrt{\underset{i\in A}{\text{mid}}\ (\min_{j\in B}||x_i-x_j||)^2} \tag{3}$$

其次，计算支持向量矩阵中任意两个向量集 C、D 之间的距离中值 d_{med}：

$$d_{med}=\sqrt{\underset{i\in C}{\text{mid}}\ (\min_{j\in D}||x_i-x_j||)^2} \tag{4}$$

然后，利用距离中值来确定参数 σ_i 的范围：

$$\sigma_i^2=k_i\cdot d_{med}^2,\ k_i=\frac{1}{8},\frac{1}{4},\frac{1}{2},1,2,4,8,\cdots \tag{5}$$

最后，使用梯度下降法或交叉验证法来确定最终的 σ。

1.2 线性多分类处理

根据SVM解决多分类的思路，对任意两个类构建一个二分类器，引入图的概念，从而能较好地控制积累误差[16]。

1.2.1 样本训练阶段构建多个二分类SVM

根据线性判别函数 $g(x)=w^Tx+b$，其中 w 为分类面的法向量，b 为分类面的偏移。设 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，$x_i\in R^n$，$y_i\in\{-1,+1\}$，则分类规则可紧凑表示为：$y_i(w^Tx+b)\geqslant 1$，$|g(x_i)|\geqslant 1$。

设 $g(x)=0$ 时有超平面 H，与 H 垂直的向量为 w，则任意样本向量 x 可表示为：

$$x=x_p+\frac{rw}{||w||} \tag{6}$$

其中，r 为几何间隔，则判别函数 $g(x)$ 变换为：

$$g(x)=w^T(x_p+\frac{rw}{||w||})+b=r||w|| \tag{7}$$

则：

$$r=\frac{g(x)}{||w||} \tag{8}$$

那么，两类间隔的距离则为 $\frac{2}{||w||}$，为了得到最大的类间距，则最优超平面应满足：$\min\frac{1}{2}w^Tw$，s.t. $y_i(w^Tx+b)\geqslant 1$。

当间隔最大化时，就可以找出支持向量：$\arg\max_{w,b}\{\min(y_i(w^Tx+b))\frac{1}{||w||}\}$。

那么问题就转化为寻找 w 和 b，引入拉格朗日乘子，用条件极值求解间隔最大时的 w 和 b：

$$\max\ L(w,b,a)=\frac{1}{2}(w^Tw)-\sum_{i=1}^{n}a_i[y_i(w^Tx_i+b)-1] \tag{9}$$

分别对 w、b 求偏导数，并令其为0：

$$w=\sum_{i=1}^{n}a_iy_ix_i \tag{10}$$

$$\sum_{i=1}^{n}a_iy_i=0 \tag{11}$$

代入拉格朗日函数，则有：

$$Q(a)=L(w,b,a)$$
$$=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jy_iy_jx_i^Tx_j \tag{12}$$

求 $Q(a)$ 的最大值，得到最优 a^*，则最优向量 w^*：

$$w^*=\sum_{i=1}^{n}a_i^*y_ix_i \tag{13}$$

选用任一支持向量，得到最优偏置量 b^*：

$$b^*=y_i-w^{*T}x_i \tag{14}$$

综上，通过求解最优拉格朗日乘子，就可以得到最优超平面，完成一个二分类SVM的构建。在样本训练阶段，共需构建 $\frac{n(n-1)}{2}$ 个二分类SVM。

1.2.2 决策阶段构建有向无环判定树

将前面得到的 $\frac{n(n-1)}{2}$ 个分类器作为节点构造有向无环决策树。图1为5个二分类器(SVM_{1-5})所构造的二叉决策树。

由上述分类决策过程可以看出DAG-SVM算法的关键是控制误差积累。因此，从根节点起，确定每个节点都为其所处层级最易分类的二分类器。

首先，计算 $\frac{n(n-1)}{2}$ 个分类器之间的距离。设有 N 个

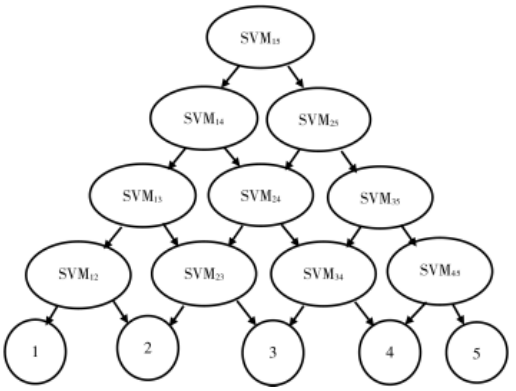

图 1　DAG-SVM 算法示例

样本的训练样本集 X 的均值为 $\boldsymbol{m}$，根据马氏距离则得：

$$\boldsymbol{m}=\frac{1}{N}\sum_{k=1}^{N}(x_1(k),x_2(k),\cdots,x_d(k))^{\mathrm{T}} \tag{15}$$

则各二分类器的训练样本集的均值中心 $C_i(i=1,2,\cdots,n)$ 为：

$$C_i=\frac{1}{N_i}\sum_{x\in N_i}\boldsymbol{x} \tag{16}$$

那么任意两个分类器样本中心的距离 d_{ij} 为：$d_{ij}=||C_i-C_j||$。考虑类分布情况，计算每个类的标准差 σ_i：

$$\sigma_i=\frac{1}{N_i-1}\sum_{x\in N_i}||\boldsymbol{x}-C_i|| \tag{17}$$

计算任意两个类的分离程度 S_{ij}：

$$S_{ij}=\frac{d_{ij}}{\sigma_i+\sigma_j} \tag{18}$$

其次，确定有向无环决策树的根节点，将 S_{ij} 最大的两类作为根节点。

然后，对不同分类节点，根据分类函数的值判断下一级所属的分类集合，直至叶节点，最终完成样本的多分类。

2 算法验证实验分析

2.1 高斯核函数宽度参数实验分析

实验使用 PASCAL VOC 数据集中 bird 训练集，图像大小为 500×375，正样本数量为 330，事先对训练样本归一化处理。实验在处罚因子 $C=1$ 和 $C=300$ 的情况下，分别采用交叉验证法和刘翔提出的宽度参数范围预估法进行确定宽度参数的对比实验，测试其在不同宽度参数取值的分类错误率。

采用交叉验证法，在高斯核函数宽度参数 $\sigma^2=\{0.125,0.25,0.5,1,2,4,8\}$，$C$ 取 1 和 300 时，其分类错误率如图 2 所示。

由图 2 可知，随着支持向量机个数的增加，高斯核函数中宽度参数选择对分类错误率的影响趋于平缓，通过训练集和测试集交叉验证，可以确定较为合适的宽度参数，但需要的训练时间和宽度参数的选取标签较难控制。

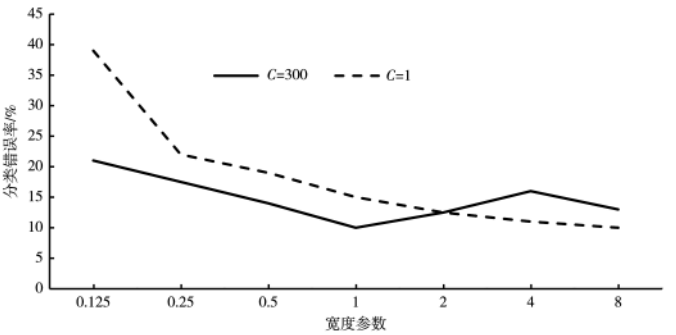

图 2　交叉验证法选择宽度参数的错误率

采用宽度参数范围预估方法，先计算两个支持向量之间的距离中值来明确宽度参数的范围，再采用交叉验证法或梯度下降法来确定最终的宽度参数取值。当 $C=300$ 时，分别对两种切片的样本图像计算其距离中值 $d_{\mathrm{med}}^2=\{5.12,9.11\}$，$\sigma^2=\{0.125,0.25,0.5,1,2,4,8\}$，其分类错误率如图 3 所示。

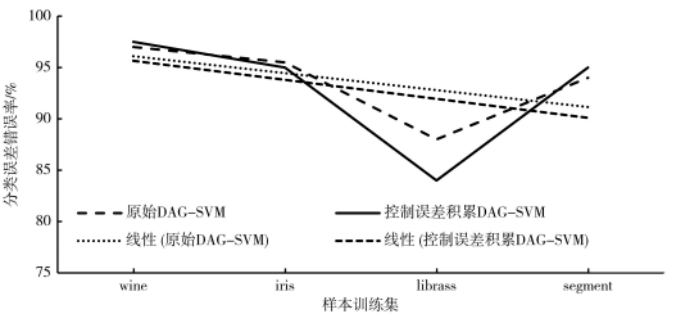

图 3　宽度参数范围预估方法选择宽度参数的错误率

由图 3 可知，图像样本向量的距离中值越大，则高斯核支持向量的宽度参数越小，分类错误率越大。通过计算两个支持向量集的距离中值，可以较快地确定高斯核的宽度参数。

2.2 DAG-SVM 线性多分类算法实验分析

实验使用 UCI 数据集中的 wine、iris、librass-movement 和 segment 4 个训练样本。分别采用原始 DAG-SVM 方法和控制误差积累的 DAG-SVM 方法进行 5 分类实验，其训练仿真实验对比结果如表 1 所示。

采用原始 DAG-SVM 方法和控制误差积累的 DAG-SVM 方法进行 5 分类的训练时间趋势如图 4 所示。

表 1　训练仿真实验对比

数据	宽度参数 σ	训练时间/h		分类精度/%	
		原始 DAG-SVM	误差积累 DAG-SVM	原始 DAG-SVM	误差积累 DAG-SVM
wine	3	1.287	1.515	97.83	98.12
iris	1	0.921	0.895	97.08	97.09
librass-movement	5	38.074	22.013	89.33	84.67
segment	0.3	85.714	55.312	96.9	97.67

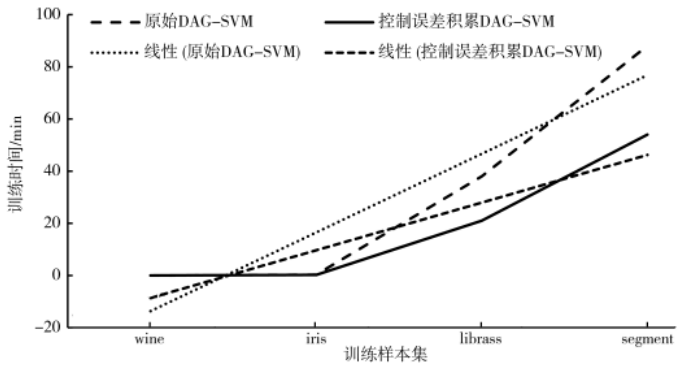

图 4　两种方法进行 5 分类的训练时间变化趋势

根据图 4 中移动平均趋势线可知，训练样本集规模越大，所需的训练时间快速增加。而采用控制积累误差的 DAG-SVM 方法的训练时间增势控制明显优于原始 DAG-SVM 方法，说明其在大规模样本集上能加速分类。

实验中采用原始 DAG-SVM 方法和控制误差积累的 DAG-SVM 方法进行 5 分类的分类精度变化趋势如图 5 所示。

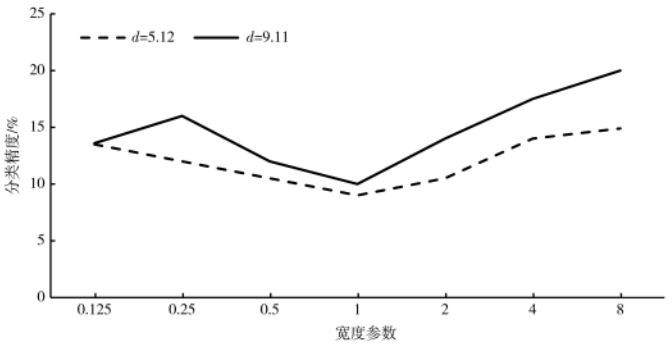

图 5　两种方法进行 5 分类的训练精度变化趋势

根据图 5 中精度趋势不难看出，当训练样本集的规模较小时，两种方法的训练精度区别不大；但当训练样本规模增大时，训练精度则会产生较大的抖动，而采用控制积累误差的 DAG-SVM 方法，其训练精度的稳定趋势要强于原始 DAG-SVM 方法，说明其在大规模样本集上分类精度更易控制。

3 结论

根据传感网络多源数据融合要求，结合 SVM 泛化能力强，凸优化的特点，本文侧重分析了非线性可分多源数据集转化为高维线性可分空间的可行性方法和针对多分类情形如何更好地控制误差积累，确保分类的有效性。

通过在训练样本集上的仿真实验，可以得出非线性可分空间转换为高维线性可分空间时，通过向量间距离最优策略能加速确定高斯核宽度参数。但通过两个支持向量间宽度参数来确定向量间的距离中值，有可能因为初始参数有误而造成宽度参数范围确定错误，这将极大地增加计算代价，因此后续将重点研究如何控制初始宽度参数的确定误差。

参考文献

[1] 刘同明，夏祖勋，解洪成.数据融合技术及其应用[M].北京：国防工业出版社，2000.
[2] 滕召胜，罗隆福，童调生.智能检测系统与数据融合[M].北京：机械工业出版社，2000.
[3] 多传感器数据融合技术研究与展望[J].物联网技术，2015，5(5)：23-25.
[4] 刘永星，赵涓涓.基于数据融合的无线传感器网络林火监控算法[J].计算机科学，2015(11)：158-163.
[5] 曾瑛，李星南.电力通信大数据并行化聚类算法研究[J].电子技术应用，2018，44(5)：1-4，24.
[6] GOLESTAN K，JUNDI A，NASSAR L.Capabilities，challenges in information gathering and data fusion[J].Autonomous and Intelligent Systems，LecUire Notes in Computer Science，2012，7326：34-41.
[7] 孙振东.面向多源数据融合的贝叶斯估计方法[J].齐鲁工业大学学报，2018，32(1)：73-76.
[8] 朱丹，李连登，童艳.外测实时数据融合算法应用研究[J].测控技术，2014，33(3)：56-58.
[9] 丁晗，侯瑞春.基于粗糙集和改进 D-S 证据理论的故障诊断方法[J].计算机与数字工程，2019，47(3)：543-549.
[10] 解从伟，张野.基于多测元融合的实时数据处理算法研究[J].仪表技术，2018(2)：24-27.
[11] 解从伟，项树林.光测实时数据融合算法研究[J].计算机技术与发展，2017(6)：304-306.
[12] DENG L，YU D，PLATT J.Scalable stacking and learning for building deep arcitectures[C]//2002 IEEE International Conference on Acoustic，Speech and Signal Processing.IEEE，Kyoto，Japan，2012：2133-2136.
[13] HU M，CHEN Y，KWOK J T Y.Building sparse multiple-kernel SVM classifiers[J].IEEE Transactions on Neural Networks，2009，20(5)：827-839.
[14] Ye Yuyun，Tian Miao，Liu Qiyu.Pulmonary nodule detection using V-net and high-level descriptor based SVM classifier[J].IEEE Access，2020，8：176033-176041.
[15] 陈洋洋.基于多尺度核加权融合的支持向量机核函数优化方法的研究[D].杭州：杭州电子科技大学，2017.
[16] 陈思羽，宁芊.DAG-SVM 结构优化研究及其在故障诊断中的应用[J].四川大学学报(自然科学版)，2015，52(2)：299-305.

(收稿日期：2020-11-04)

作者简介：

陈燕(1974-)，女，硕士，副教授，主要研究方向：数据分析、计算机仿真。

顾大刚(1968-)，男，硕士，教授，主要研究方向：大数据处理。

陈亚林(1978-)，女，博士，副教授，主要研究方向：数据分析。

扫码下载电子文档

版权声明