基于熵值法的加权最小二乘支持向量机

刘 畅 范 彬
(中南大学机电工程学院 长沙 410012)

摘 要 支持向量机是一种以统计学习理论为基础的机器学习算法,着重解决小样本的建模问题,并且对非线性高维数据具有较好的处理能力。通常对于多维特征的数据,会对每一维数据做归一化处理以消除量纲的影响,但缺点在于忽视了各维特征的权重差异。提出了一种加权最小二乘支持向量机的建模方法,通过熵值法确定每一维特征的权重,根据特征权重对数据进行加权处理,最后由最小二乘支持向量机建立该系统模型。实验表明,对于多维特征的数据,所提方法具有更好的建模效果。

关键词 支持向量机,熵值法,多维特征,特征权重
中图法分类号 TP181 **文献标识码** A

Weighted Least Squares Support Vector Machine Based on Entropy Evaluation

LIU Chang FAN Bin
(College of Mechanical and Electrical Engineering, Central South University, Changsha 410012, China)

Abstract Support vector machine is a kind of machine learning algorithm based on statistical learning theory, which has a desirable modeling performance for nonlinear and high-dimensional data, even in the case of small samples. Typically, the data with multiple features would be normalized due to different dimensions. However, it ignores the dissimilarity of different features. A weighted least squares support vector machine was proposed. According to the entropy evaluation method, the feature weights may be determined so that the data could be normalized and weighted. Then the system model would be established through the least squares support vector machine. The experimental results demonstrate the effectiveness and superiority of the proposed method for the system with multiple features.

Keywords Support vector machine, Entropy evaluation method, Multiple features, Feature weight

1 引言

非线性、多维的系统建模问题是一个备受关注的话题,系统的强非线性通常导致了难以获得其准确的物理模型。因此近年来,许多以数据驱动为基础的优秀算法相继被提出。

众所周知,神经网络是一种十分经典的数据挖掘方法[1-2],然而传统的人工神经网络需要大量的训练样本和所调参数,当样本数量较少时,训练出的模型通常不具有推广性,并且容易陷入局部最优[3],支持向量机(SVM)的提出有效地解决了这一难题[4]。最初,SVM是作为一种处理二分类问题的方法被提出,通过最大化几何间隔确定出超平面,从而实现不同类别的划分,该方法被推广到回归问题的建模中,基于结构风险最小化的原理,引入的正则化项有效地避免了过拟合问题,保证了模型的泛化性。尤其是当面对小样本、非线性以及高维数据时,SVM表现出特有的优势。随后,最小二乘支持向量机(LS-SVM)的提出有效地解决了计算复杂度的问题,它将最优化问题的求解转化为线性方程组的形式[5-6],在模式识别[7]、文本分类[8]、图像识别[9]、时间序列预测[10]等领域都得到了广泛的应用。

当面对多输入的建模问题时,由于量纲的不同,容易造成输入的某些特征对模型的影响过大,进而影响建模的精度。因此,对输入数据进行归一化处理显得尤为重要[11]。通过归一化处理,可以消除数据的量纲,使物理系统数值的绝对值变成某种相对值关系。由此,输入数据的各特征便具有相同的权重,有效地避免了某些特征被“忽视”的问题。许多回归和分类的建模过程都将数据的归一化处理运用其中[12-14]。然而,尽管归一化处理可以消除量纲,实现数据各特征的“等价”,但是,在实际的多输入回归问题中,各个特征的权重并非一定相同。因此,如何确定不同特征的权重,建立更有效的模型仍是一个有待解决的问题。

在信息论中,熵是对不确定性的一种度量,若不确定性越大,则熵也越大[15]。同时,根据熵的特性,也可以用它来判断某个指标的离散程度,指标的离散程度越大,意味着该指标对该事件的影响越大。

本文受教育部新世纪人才基金(NCET-13-0593)资助。
刘 畅(1992-),男,硕士生,主要研究领域为数据挖掘与支持向量机;范 彬(1979-),男,博士生,主要研究领域为过程建模与控制、鲁棒设计,E-mail:liuchang_csu@163.com。

基于此，提出了一种基于熵值法加权的支持向量机建模策略，重点解决了多维特征数据的建模问题。该方法能够有效地确定出不同指标的权重，区分不同特征的重要性，进而在使用 LS-SVM 建立该多输入模型时，能够侧重于依赖权重较大的特征输入，使其对模型产生更重要的影响。因此，该方法基于支持向量机的建模策略，结合特征加权的思想，在多输入的建模过程中，表现出更好的效果。

2　最小二乘支持向量机

在回归分析的问题中，LS-SVM 旨在求出一个最优的映射函数，假定样本的数据集为$\{(x_i,y_i)\mid i=1,\cdots,n\}$，$x_i,y_i$ 分别作为样本的输入值与输出值。该映射函数 f 能够反映输入与输出之间的关系，即可表示为 $y=f(x)$。在 LS-SVM 模型中，可具体表示如下：

$$y=\omega^{\mathrm{T}}\phi(x)+b \tag{1}$$

其中，ϕ 是一未知的非线性映射函数，将输入 x 从原始空间映射到特征空间，借此，将低维空间的非线性拟合问题转变为高维特征空间的线性问题；另外，ω 和 b 分别表示权重向量和偏置。

该模型可转化为带约束条件的最优化问题：

$$\begin{aligned}&\min_{\omega,b,e_i} J(\omega,b,e_i)=\frac{1}{2}\|\omega\|^2+\frac{C}{2}\sum_{i=1}^{n}e_i^2\\&\text{s.t. } y_i=\omega^{\mathrm{T}}\phi(x_i)+b+e_i,i=1,\cdots,n\end{aligned} \tag{2}$$

其中，e_i 代表建模误差，C 表示正则化参数，保证在模型复杂度与拟合误差之间取得折中。为求解上述问题，可引入拉格朗日因子，将带有约束条件的最优化问题转变为无约束的优化问题：

$$L(\omega,b,e_i;a_i)=J(\omega,b,e_i)-\sum_{i=1}^{n}a_i(\omega^{\mathrm{T}}\phi(x_i)+b+e_i-y_i) \tag{3}$$

其中，a_i 即为拉格朗日因子，根据 KKT 条件，则有：

$$\begin{cases}\dfrac{\partial L}{\partial\omega}=0\rightarrow\omega=\sum_{i=1}^{n}a_i\phi(x_i)\\[2ex]\dfrac{\partial L}{\partial b}=0\rightarrow\sum_{i=1}^{n}a_i=0\\[2ex]\dfrac{\partial L}{\partial e_i}=0\rightarrow a_i=\gamma e_i,i=1,\cdots,n\\[2ex]\dfrac{\partial L}{\partial a_i}=0\rightarrow\omega^{\mathrm{T}}\phi(x_i)+b+e_i-y_i=0,i=1,\cdots,n\end{cases} \tag{4}$$

联立以上方程组，可得：

$$\begin{bmatrix}0 & \mathbf{1}_N^{\mathrm{T}}\\ \mathbf{1}_N & \boldsymbol{Q}+C^{-1}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{a}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{Y}\end{bmatrix} \tag{5}$$

其中，$\boldsymbol{I}$ 为单位矩阵，$\mathbf{1}_N=[1,\cdots,1]\in R^n$，$\boldsymbol{a}=[a_1,\cdots,a_n]^{\mathrm{T}}$，$\boldsymbol{Y}=[y_1,\cdots,y_n]^{\mathrm{T}}$，$\boldsymbol{Q}_{ij}=\phi(x_i)^{\mathrm{T}}\phi(x_j)=K(x_i,x_j)$，$i,j=1,\cdots,n$；同时，引入满足 Mercer 定理的核函数 K 来代替特征空间的点积计算，而不需要知道映射 ϕ 的显示表达式。

根据线性方程组(5)，可以求出拉格朗日因子 a 和偏置项 b，进而可得到最终的预测模型：

$$\hat{y}=\sum_{i=1}^{n}a_iK(x,x_i)+b \tag{6}$$

由此可以看出，LS-SVM 可以归结为一个线性方程组的求解问题，相比于标准的 SVM，计算复杂度更小，速度更快。

3　数据加权

3.1　归一化处理

在上述的建模问题中，当输入变量 x_i 是多维特征时，即 $\boldsymbol{x}_i=[x_{i1},\cdots,x_{id}]$，这里假定是 d 维，那么通常首先对输入数据做预处理，以消除量纲，避免各个特征之间的量级差异。通过归一化方法，使得每一维的变量都落入[0,1]的区间上，归一化后的数据为：

$$t_{ip}=\frac{x_{ip}-\min\{x_{1p},\cdots,x_{np}\}}{\max\{x_{1p},\cdots,x_{np}\}-\min\{x_{1p},\cdots,x_{np}\}} \tag{7}$$

其中，$i=1,\cdots,n;p=1,\cdots,d$。

经过归一化处理之后，如若不考虑不同特征之间的权重关系，便可将处理后的输入、输出数据运用至 LS-SVM 建模中。

3.2　熵值法确定特征权重

当考虑不同特征之间的权重关系时，可以根据每个特征的离散程度，通过熵值法，判断该特征对整体的影响，并确定其权重。当某个指标的离散程度较大时，说明该指标提供的信息量较大，其熵值较小；反之，当某个指标的离散程度越小，甚至完全相同时，说明该指标几乎不提供任何有效信息，可以剔除该指标，不考虑其影响，此时，熵值接近 1。

此处，将 n 个 d 维特征的输入数据表示为：

$$\boldsymbol{X}=\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}=\begin{bmatrix}x_{11} & \cdots & x_{1d}\\ x_{21} & \cdots & x_{2d}\\ \vdots & \cdots & \vdots\\ x_{n1} & \cdots & x_{nd}\end{bmatrix} \tag{8}$$

确定这 d 维特征权重的具体步骤如下。

1)计算数据所占比重

首先，计算各个特征下每个输入数据占该指标的比重：

$$f_{ip}=\frac{x_{ip}}{\sum_{i=1}^{n}x_{ip}},p=1,\cdots,d \tag{9}$$

由此，得到 $\boldsymbol{F}$ 矩阵：

$$\boldsymbol{F}=\begin{bmatrix}f_1\\ f_2\\ \vdots\\ f_n\end{bmatrix}=\begin{bmatrix}f_{11} & \cdots & f_{1d}\\ f_{21} & \cdots & f_{2d}\\ \vdots & \cdots & \vdots\\ f_{n1} & \cdots & f_{nd}\end{bmatrix} \tag{10}$$

2)计算各特征熵值

用 E_p 表示所有输入对第 p 项指标的贡献总量，则第 p 维特征的熵值为：

$$E_p=-K\sum_{i=1}^{n}f_{ip}\ln(f_{ip}) \tag{11}$$

其中，常数 $K=\dfrac{1}{\ln(n)}$，这样可以保证 $0\leqslant E_p\leqslant 1$。由此可以看出，当 $x_{1p}=x_{2p}=\cdots=x_{np}$ 时，$f_{1p}=f_{2p}=\cdots=f_{np}=\dfrac{1}{n}$，此时，第 p 维特征熵值 $E_p=1$。这意味着某项指标的数据相等时，该项指标无参考价值，可以不考虑其影响。

3)计算各项指标的权值

得到各维特征的熵值后，首先计算各特征下所有输入的

一致性程度，此处用 D_p 表示第 p 维特征的一致性程度：

$$D_p=1-E_p \tag{12}$$

可以看出，一致性越高，D_p 越小，信息量越小，权重越小；反之，权重越大。因此，可求出各维特征的权重：

$$w_p=\frac{D_p}{\sum_{p=1}^{d}D_p},p=1,\cdots,d \tag{13}$$

3.3 归一化数据的加权处理

将归一化之后的数据与求出的特征权重相乘，即可得到所需的输入数据：

$$u_{ip}=t_{ip}w_p,i=1,\cdots,n;p=1,\cdots,d \tag{14}$$

由此可以看出，经过处理的输入数据 u_{ip} 不仅消除了量纲的影响，而且能够体现不同特征之间的权重关系，从而对建模产生或大或小的作用。将处理后的输入、输出数据$\{(u_i,y_i)|i=1,\cdots,n\}$运用至 LS-SVM 建模当中，即可建立多输入系统的回归模型：

$$\hat{y}=\sum_{i=1}^{n}a_iK(u,u_i)+b \tag{15}$$

其中，拉格朗日因子 a_i 和偏置项 b 的求解过程在第 2 节已作说明。

4 实验验证

为检验所提方法对多输入回归问题建模的有效性，将该方法与标准 LS-SVM、归一化 LS-SVM 方法以及同样确定特征权重的 TM-SVM[16]方法作比较，以均方根误差(RMSE)作为评估标准，其定义如下：

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{16}$$

其中，n 为样本个数，y_i 和$\hat{y}_i$ 分别为实际和预测输出。

实验数据来自 UCI 机器学习数据库的“联合循环发电厂”(Combined Cycle Power Plant)实验。2006 年－2011 年，当该发电厂满载工作时，收集了 9568 组样本数据，数据集的 4 个输入与 1 个输出分别为：Temperature(T)，Ambient Pressure(AP)，Relative Humidity(RH)，Exhaust Vacuum(EV)，Electrical Energy Output(EP)，具体属性信息如表 1 所列。

表 1 数据集属性信息

	属性	范围
输入 1	T	1.81～37.11/℃
输入 2	AP	992.89～1033.30/milibar
输入 3	RH	25.56%～100.16%
输入 4	EV	25.36～81.56/(cm/Hg)
输出	EP	420.26～495.76/MW

可以看到，该数据集可视为一个多输入的回归系统，4 个输入变量具有不同的量纲，且离散程度也各不相同。从中随机选取 5000 组作为训练数据，剩余 4568 组作为测试数据，并用 3 种方法分别建模并进行比较。

根据 5000 组训练数据，通过熵值法分别计算出 4 个输入的权重：0.6025(T)，0.0001(AP)，0.1879(RH)，0.2094(EV)。进而将输入数据归一化，并乘以相应权重，最后将处理后的输入数据用于 LS-SVM 建模中。

为了更加清楚地展示建模效果，分别从训练和测试样本中随机抽出 100 组数据样本进行观测，如图 1 和图 2 所示。可以看出，使用所提方法建立的模型不仅在训练样本上具有较好的逼近能力，而且在测试样本上同样能够较好地拟合实际输出。

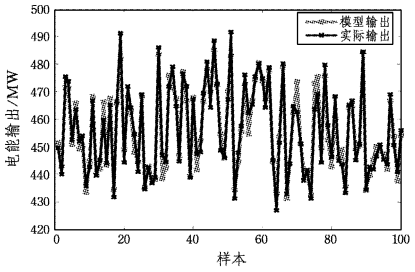

图 1 训练样本模型

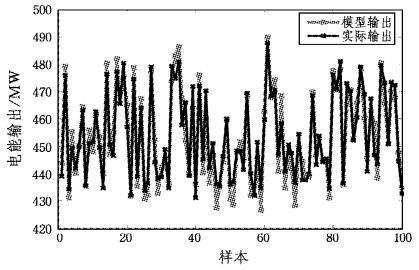

图 2 测试样本模型

此外，为展示所提方法的有效性与优越性，将其与另外 3 种常用的 LS-SVM 建模方法进行比较，分别为标准 LS-SVM、归一化 LS-SVM 以及同样确定特征权重的 TM-SVM。此处用相对误差反映模型的精确程度，其定义如下：

$$Relative\ Error(x)=\frac{|\hat{y}(x)-y(x)|}{y(x)}\times 100\% \tag{17}$$

同样，为了清楚地观测对比效果，从测试样本中随机抽取 30 组数据，通过式(17)分别求得 30 组样本的相对误差，建模效果的对比如图 3 所示。可以看出，加权 LS-SVM 具有更小的相对误差。

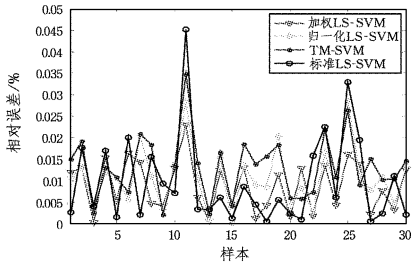

图 3 建模效果对比

此外，通过式(16)分别计算出 3 种建模方法的训练样本与测试样本的 RMSE，其结果如表 2 和表 3 所列。可以看到，在训练和测试样本上，所提方法都具有更小的建模误差，因此相比于标准 LS-SVM、归一化 LS-SVM 以及 TM-SVM，所提加权 LS-SVM 具有更高的建模精度。

表 2 训练样本均方根误差

方法	标准 LS-SVM	归一化 LS-SVM	TM-SVM	加权 LS-SVM
RMSE	4.3368	4.2473	4.1826	4.0495

表 3　测试样本均方根误差

方法	标准 LS-SVM	归一化 LS-SVM	TM-SVM	加权 LS-SVM
RMSE	4.9880	4.7770	4.7035	4.6715

结束语　本文提供了一种熵值法加权的最小二乘支持向量机的建模策略，主要针对多变量输入的回归分析问题。利用熵能够反映信息量的特性，根据每一维特征的离散程度，确定相应特征的权重大小；同时，对数据进行归一化处理，以消除量纲，避免各个特征之间的量级差异，将归一化后的数据与对应的特征权重相结合，得到能够充分反映特征信息的输入数据；最后，运用最小二乘支持向量机建立多输入的回归模型。仿真结果表明，所提方法取得了比常规方法更高的建模精度。

参 考 文 献

[1] HAYKIN S. Neural Networks: A Comprehensive Foundation [M]. Macmillan, 1998: 71-80.

[2] 焦李成. 神经网络系统理论[M]. 西安电子科技大学出版社, 1990.

[3] 张木想，马缚龙，肖国镇. 神经网络优化计算的新方法[J]. 电子学报，1993(7)：1-7.

[4] UKIL A. Support Vector Machine[J]. Computer Science, 2002, 1(4): 1-28.

[5] SUYKENS J A K, GESTEL T V, BRABANTER J D, et al. Least Square Support Vector Machine[J]. Euphytica, 2002, 2(2): 1599-1604.

[6] 阎威武，邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究[J]. 控制与决策，2003，18(3)：358-360.

[7] 李盼池，许少华. 支持向量机在模式识别中的核函数特性分析[J]. 计算机工程与设计，2005，26(2)：302-304.

[8] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：164-167.

[9] 何江平，文俊浩，邓恬洁，等. 基于支持向量机的图像识别[J]. 重庆大学学报(自然科学版)，2006，29(1)：57-60.

[10] 叶美盈，汪晓东，张浩然. 基于在线最小二乘支持向量机回归的混沌时间序列预测[J]. 物理学报，2005，54(6)：2568-2573.

[11] 柳小桐. BP 神经网络输入层数据归一化研究[J]. 机械工程与自动化，2010(3)：122-123.

[12] 刘冲，赵海滨，李春胜，等. 基于频带能量归一化和 SVM-RFE 的 ECoG 分类[J]. 仪器仪表学报，2011，32(3)：534-539.

[13] 常军，李祯，朱业玉，等. 基于支持向量机(SVM)方法的冬季温度预测[J]. 气象科技，2005(s1)：102-106.

[14] 黄安民，焦淑菲，任海青，等. 支持向量机结合近红外光谱法测定杉木木质素的含量[J]. 林产化学与工业，2009，29(5)：1-5.

[15] 郭显光. 熵值法及其在综合评价中的应用[J]. 财贸研究，1994(6)：56-60.

[16] PENG X, XU D. Twin Mahalanobis distance-based support vector machines for pattern recognition[J]. Information Sciences, 2012, 200(1): 22-37.

(上接第 399 页)

CPU 的处理能力、内存处理能力、磁盘读写处理能力、带宽等 5 个参数来反映实际工作负载，并根据变异系数法求出相关参数权重。实验结果表明，该副本放置改进策略比默认策略更好地实现了各节点的负载均衡，同时也提高了集群的工作效率。

参 考 文 献

[1] 周江，等. 面向大数据分析的分布式文件系统关键技术[J]. 计算机研究与发展，2014，51(2)：382-394.

[2] 罗鹏，等. HDFS 数据存放策略的研究与改进[J]. 计算机工程与设计，2014，35(4)：1127-1131.

[3] 孙知信，等. 基于云计算的数据存储技术研究[J]. 南京邮电大学学报，2014，34(4)：13-19.

[4] 王海荣，等. 基于 Hadoop 的海量数据存储系统设计[J]. 科技通报，2014，30(9)：127-130.

[5] 张柄虹，等. 空间高效的分布式数据存储方案[J]. 计算机应用研究，2015，32(5)：1508-1511.

[6] 马晓亭. 数字图书馆大数据分布式存储架构模式与策略研究[J]. 新世纪图书馆，2015(5)：43-46.

[7] 康承昆，等. 一种基于多衡量指标的 HDFS 负载均衡算法[J]. 四川大学学报，2014，51(6)：1163-1169.

[8] 熊安萍，等. 一种基于混合索引的 HDFS 小文件存储策略[J]. 重庆邮电大学学报，2015，27(1)：97-102.

[9] 英昌甜，等. 一种面向低延迟的内存 HDFS 数据存储策略[J]. 微电子学与计算机，2014，31(11)：160-166.

[10] 尹颖，等. HDFS 中高效存储小文件的方法[J]. 计算机工程与设计，2015，36(2)：406-409.

[11] 卢美莲，等. 基于 CMM 模型的 HDFS 负载均衡策略[J]. 北京邮电大学学报，2014，37(5)：20-25.

[12] 孟祥萍，等. 基于 hadoop 云平台的智能电网 HDFS 资源存储技术研究[J]. 电测与仪表，2014，51(19)：24-30.

[13] 张华伟，等. 基于多目标优化的云存储副本分布策略的研究[J]. 计算机科学，2015，42(4)：44-50.

[14] 肖达，苏丽裕，王俊龙，等. CSFS：云存储服务支撑的文件系统设计与实现[J]. 北京邮电大学学报，2015，38(6)：77-82.

[15] 罗芳，等. 基于多属性的海量 Web 数据关联存储及检索系统[J]. 计算机工程与科学，2014，36(3)：404-410.

[16] 宋宝燕，等. 基于范德蒙码的 HDFS 优化存储策略研究[J]. 计算机学报，2015，38(9)：1825-1837.

[17] 肖玉泽，等. HDFS 下海量小文件高效存储于索引方法[J]. 小型微型计算机系统，2015，36(10)：2218-2223.

[18] MISHNE G, DALTON J, LI Z H, et al. Fast data in the era of big data: Twitter's real-time related query suggestion architecture[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. 2013: 1147-1157.