预测结果存在显著性差异。秩计算结果如表2所示,其中BP神经网络模型*RMSLE*值和宽度&深度模型*RMSLE*值之差的负秩数量为12,占总数的24%;正秩数量为38,占总数的76%,也说明了宽度&深度模型要优于BP神经网络模型。

表2 秩计算结果

Table 2 Result of the rank calculation

指标	数量 N	秩均值	秩和
负秩	12	8.83	106.00
正秩	38	30.76	1 169.00
总数	50		

由此可见,宽度&深度模型相较于BP神经网络模型,增加了宽度部分,因此提升了模型的稳定性,得到更好的预测效果。宽度&深度模型结合了神经网络的泛化能力和线性模型的记忆能力,通过在宽度部分输入小区的特征,对深度模型的预测结果进行修正,因而得到比BP神经网络更准确的预测结果。

4.3.2 宽度&深度模型与SARIMA模型对比

ARIMA系列模型常被用来预测时间序列,因此将本文方法与SARIMA模型进行对比。由于SARIMA模型采用网格搜索的方式定阶,因此不存在随机性。将训练得到的50个宽度&深度模型的预测结果的平均均方根对数误差与SARIMA模型的预测结果进行对比,如表3所示。由表3可见,本文方法的预测效果明显优于SARIMA模型。

表3 宽度&深度模型和SARIMA模型的预测结果比较

Table 3 Comparison of prediction results of Wide & Deep model and SARIMA model

模型	*RMSLE*
宽度&深度模型	0.985
SARIMA模型	2.095

4.3.3 宽度&深度模型与LSTM模型对比

按第2节方法对数据进行预处理,为每个小区建立一个LSTM模型。模型的输入为连续63 d的流量序列,输出为第64天的流量。通过31次单步预测来预测未来31 d的流量。模型由两层LSTM层和一层全连接层组成,两层LSTM层的单元数分别设置为64和32,激活函数采用双曲正切函数,全连接层的神经元个数为1,作为输出层。模型的损失函数设置为*RMSLE*,采用Adam优化器。实验结果如表4所示。由表4可见,本文方法的预测效果明显优于LSTM模型,且本文方法为所有小区建立统一模型,更易于应用。

表4 宽度&深度模型和LSTM模型的预测结果比较

Table 4 Comparison of prediction results of Wide & Deep model and LSTM model

模型	*RMSLE*
宽度&深度模型	0.985
LSTM模型	3.281

5 结论

本文提出一种基于宽度&深度模型的基站网络流量预测方法。首先,利用S-H-ESD算法和窗口平滑方法处理非平稳的流量时间序列数据。然后,将流量数据作为模型的深度部分(神经网络)输入,将RRC连接数和PRB利用率作为模型的宽度部分(线性模型)输入,将两部分进行联合训练获得流量预测模型,用于预测网络流量。该方法为所有基站小区流量建立单一模型,具有简单和易于实施的特点。实验结果表明,该方法优于当前广泛采用的SARIMA、BP神经网络和LSTM模型。

下一步的研究工作包括:进一步优化模型的宽度部分的特征,提高预测准确率;与更多的预测模型进行对比分析。

参考文献:

[1] 赵一权.无线网络运营数据分析与预测研究[D].北京:北京邮电大学,2018.

[2] 蒋品.基于机器学习的蜂窝网络基站流量分析与预测研究[D].北京:北京邮电大学,2019.

[3] 危彦.蜂窝网络中基于流量预测的节能关键技术研究[D].杭州:浙江大学,2012.

[4] 张佳鑫,张兴,李永竞,等.蜂窝网络中基站关系与业务关系网络与应用[J].中国科学:信息科学,2017,47(5):648-663.

[5] MADAN R,MANGIPUDI P S.Predicting computer network traffic:a time series forecasting approach using DWT,ARIMA and RNN[C]//2018 Eleventh International Conference on Contemporary Computing (IC3). Piscataway:IEEE,2018:1-5.

[6] LU H P, YANG F. Research on network traffic prediction based on long short-term memory neural network[C]//2018 IEEE 4th International Conference on Computer and Communications (ICCC). Piscataway: IEEE,2018:1109-1113.

[7] HUANG C W,CHIANG C T,LI Q H.A study of deep learning networks on mobile traffic forecasting[C]// 2017 IEEE 28th Annual International Symposium on

Personal, Indoor, and Mobile Radio Communications (PIMRC). Piscataway: IEEE, 2017: 1-6.

[8] GUI Y H, WANG D S, GUAN L Y, et al. Optical network traffic prediction based on graph convolutional neural networks[C]//2020 Opto-Electronics and Communications Conference (OECC). Piscataway: IEEE, 2020: 1-3.

[9] HOCHENBAUM J, VALLIS O S, KEJARIWAL A. Automatic anomaly detection in the cloud via statistical learning[EB/OL]. (2017-04-24)[2021-05-01]. https://arxiv.org/abs/1704.07706.

[10] YANG Z Y, ZHANG D Z, TANG J. Predicting PON networking traffic flow based on LSTM neural network with periodic characteristic data[C]//2020 IEEE 5th Optoelectronics Global Conference (OGC). Piscataway: IEEE, 2020: 39-42.

[11] LIU H Z, YANG L T, CHEN J J, et al. Multivariate multi-order Markov multi-modal prediction with its applications in network traffic management[J]. IEEE transactions on network and service management, 2019, 16(3): 828-841.

[12] HAN Y, JING Y W, LI K, et al. Network traffic prediction using variational mode decomposition and multi-reservoirs echo state network[J]. IEEE access, 2019, 7: 138364-138377.

[13] 骆凯,罗军勇,尹美娟,等.一种基于动态阈值的突发流量异常检测方法[J].信息工程大学学报, 2016, 17(4): 509-512.

[14] GHOSH D, VOGT A. Outliers: an evaluation of methodologies[C]//Proceedings of the Survey Research Methods Section-JSM 2012. Washington DC: ASA, 2012: 3455-3460.

[15] MIRZARGAR M, WHITAKER R T, KIRBY R M. Curve boxplot: generalization of boxplot for ensembles of curves[J]. IEEE transactions on visualization and computer graphics, 2014, 20(12): 2654-2663.

[16] CLEVELAND U R B, CLEVELAND U W S, MCRAE U J E, et al. STL: a seasonal-trend decomposition procedure based on loess[J]. Journal of official statistics, 1990, 6(1): 3-73.

[17] SCHULTE J A. Wavelet analysis for non-stationary, nonlinear time series[J]. Nonlinear processes in geophysics, 2016, 23(4): 257-267.

[18] RILLING G, FLANDRIN P, GONCALVES P. On empirical mode decomposition and its algorithms[C]//Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing. Piscataway: IEEE, 2003: 8-11.

[19] CHENG H T, KOC L, HARMSEN J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. New York: ACM, 2016: 7-10.

Base Station Network Traffic Prediction Method Based on Wide & Deep Learning

CHEN Haojie[1,2], HUANG Jin[1,2], ZUO Xingquan[1,2], HAN Jing[3], ZHANG Baisheng[3]

(1.School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2.Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China; 3.Zhongxing Telecommunication Equipment Corporation, Shenzhen 518057, China)

Abstract: Aiming at the long-term prediction problem of wireless communication network traffic, a base station network traffic prediction method was proposed based on Wide & Deep learning. Firstly, S-H-ESD (seasonal hybrid extreme studentized deviate test) algorithm and data smoothing method were used to preprocess the network traffic data, and to reduce the impact of noise data on the prediction. Then, the network flow was input into the deep part (neural network) of the Wide & Deep model, the radio resource control (RRC) and physical resource block (PRB) were input into the wide part (linear model) of the Wide & Deep model, and the deep and wide parts were combined to predict the network traffic. The method established one prediction model for the network traffic of all base stations. The root mean squared logarithmic error (*RMSLE*) of prediction results was 0.985, which was significantly better than that of the traditional seasonal autoregressive integrated moving average model (*RMSLE* was 2.095) and that of the long short-term memory network model (*RMSLE* was 3.281). Experimental results showed that the Wide & Deep model could better solve the problem of long-term prediction of wireless network traffic via combining the memory ability of the linear model and the generalization ability of the depth model.

Keywords: Wide & Deep model; deep learning; base station network traffic; traffic prediction; time series prediction; neural network

文章编号:1671-6833(2022)01-0014-06

改进的协同训练半监督 SVM 在油层识别中的应用

潘用科, 贺紫平, 夏克文, 牛文佳

(河北工业大学 电子信息工程学院,天津 300401)

摘 要: 实际石油测井中有标签数据获取代价昂贵,而大量低廉的无标签数据未被使用,如何利用有限的有标签样本及大量的无标签样本获取准确的油层分布有待解决。半监督学习方法因能同时利用少量有标签样本及大量无标签样本便可获取良好的分类模型而被广泛应用。因此,基于半监督支持向量机(S3VM),提出一种改进的基于量子行为粒子群优化(QPSO)的协同训练 S3VM 油层识别算法(QPSO-CS3VM)。首先引入多视图的协同训练策略,构造 2 个独立的初始分类器提高识别精度;然后为提高初始分类精度,引入了量子行为粒子群算法以优化 S3VM;最后引入一种改进的近邻数据剪辑方法用于预测无标签样本伪标签的置信度,从而避免因错分样本导致的模型性能恶化的问题。通过对具有代表性的两口井的测井数据进行油层识别,结果表明:改进的协同训练半监督 SVM 相较于传统的协同训练算法在两口井中的识别率分别提升了 5.00 百分点和 3.12 百分点。所提算法油层识别精度较高,有一定的实际应用意义。

关键词: 半监督支持向量机; 协同训练; 量子行为粒子群优化; 数据剪辑; 油层识别

中图分类号: TP18 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.01.001

0 引言

在传统的石油钻探及测井识别中,油层识别扮演着重要的角色。由于传统的油层识别技术已经无法满足现代化石油工业的需求,因此,研究者们将现代的数据挖掘方法如二阶锥优化的多核相关向量机(multiple-kernel relevance vector machine on second order cone programming, SOCP-MKRVM)[1]、多核相关向量机(multi-core support vector machine, MKRVM)[2]、随机森林(random forest, RF)[3]等方法用在石油测井识别中,并且获得了不错的测井识别效果。但在石油测井中,有标签的数据往往很难获得,而大量的无标签的数据却没有被利用。半监督算法因能够同时利用未标记和有标记样本来进行训练和改善分类器的性能,所以本文将半监督支持向量机(semi-supervised support vector machine,S3VM)思想引入到油层识别中,以提高油层预测精度及减少样本获取代价。

半监督学习(semi-supervised learning, SSL)[4]早在 20 世纪 70 年代首次将无标签的样本用于自训练(self-training, SL)方法,但由于该方法的学习性能完全依靠内部的 SL 方法,会导致模型的性能下降。因此,学者们提出了直推式学习,该方法是基于 SL 方法的改进来预测训练集和测试集中的无标签样本的类标签。再后来,协同训练(co-training)[5]和直推式支持向量机(transductive support vector machine, TSVM)[6-7]等方法被陆续提出。协同训练算法核心思想是利用充分冗余的视图训练出两个具有差异性的学习机,以提高无标签样本的预测标签置信度[8]。由于传统协同训练存在初始分类器精度不高的问题,弱分类器很容易受到另一个弱分类器错误预测的无标记样本及其对应的错误标签的影响,因此协同训练算法的性能通常是不稳定的,容易导致错误的累积[9]。为此,本文提出了一种改进的基于量子行为粒子群优化[10]的协同训练 S3VM 油层识别算法。该算法首先采用协同训练的策略,并同时采用了文献[11]的方法以避免传统协同训练的训练过程中的错误累积。通过建立两个独立的初始分类器,然后两个分类器互相交换高置信度的无标签样本来达到提高本身性能的目的。其次采用量子行为粒子群算法优化 S3VM,提高初始

收稿日期:2021-06-21;**修订日期**:2021-07-27

基金项目:国家自然科学基金资助项目(42075129);河北省重点研发计划项目(19210404D,20351802D)

通信作者:夏克文(1965—),男,湖南武冈人,河北工业大学教授,博士,主要从事人工智能、智能信息处理、无线通信等研究,E-mail:kwxia@ hebut.edu.cn。

无标签样本的分类精度。又考虑到错分类的无标签样本进入循环从而导致模型总体性能下降的问题,使用了一种改进的近邻数据剪辑方法来预测无标签样本伪标签的置信度。最后,将该改进的算法模型应用于实际测井数据挖掘的油层识别,并验证其在石油实际应用中的有效性。

1　协同训练半监督 SVM 算法

1.1　半监督 SVM 算法

半监督 SVM 算法的基本原理是首先设有标签训练样本集为$\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$,$y_i\in\{-1,+1\}$ 和无标签样本集$\{x_1^*,x_2^*,\cdots,x_k^*\}$,然后对无标签样本集进行预测,得到其标签为$\{y_1^*,y_2^*,\cdots,y_k^*\}$。当无标签样本集线性不可分时,为使最优超平面方程$\boldsymbol{\omega}\cdot\boldsymbol{x}+\boldsymbol{b}=0$的分类间隔最大,则以上问题可表示为以下形式:

$$\begin{aligned}&\min_{\boldsymbol{\omega},\boldsymbol{b},y,\xi}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{i=l+1}^{m}\xi_i^*,\\&\text{s.t.}\quad\forall_{i=l+1}^{m}:y_i^*(\boldsymbol{\omega}\cdot x_j^*+\boldsymbol{b})\geqslant 1-\xi_i^*,\\&i=l+1,l+2,\cdots,m,\\&\forall_{i=1}^{l}:y_i(\boldsymbol{\omega}\cdot x_i+\boldsymbol{b})\geqslant 1-\xi_i,i=1,2,\cdots,l,\\&\xi_i\geqslant 0,i=1,2,\cdots,m。\end{aligned}\tag{1}$$

式中:$\boldsymbol{\omega}$ 为对无标签样本进行分类时的最优超平面;$\boldsymbol{b}$ 为超平面偏移量;C 和 C^* 分别为有标签样本与无标签样本的影响因子;ξ 和 ξ^* 分别为有标签样本与无标签样本的惩罚因子。

另外,本文选择径向基核函数(RBF)作为核函数:$K(x,x_i)=\exp(-\gamma\|x-x_i\|^2)$,核参数为 γ,惩罚系数为 C_1。其中核参数和惩罚系数需人为指定,且不同的数值组合会影响模型的分类性能。

1.2　量子行为粒子群优化算法

粒子群算法[12]是模仿鸟群捕食行为的智能优化算法。量子行为粒子群优化算法针对粒子群算法易陷入局部最小值的问题[13],引入了量子的概念,提升了粒子的随机性,从而提升了粒子群算法的全局搜索能力。量子行为粒子群优化算法的原理如下。

假设有 m 个粒子组成的种群在 N 维的求解空间中,第 i 个粒子 t 时刻的位置为$\boldsymbol{X}^{(t)}(i)=[\boldsymbol{X}^{(t)}(i,1),\boldsymbol{X}^{(t)}(i,2),\cdots,\boldsymbol{X}^{(t)}(i,d)]$,第 i 个粒子的历史最好位置为$\boldsymbol{G}^{(t)}(i)=[\boldsymbol{G}^{(t)}(i,1),\boldsymbol{G}^{(t)}(i,2),\cdots,\boldsymbol{G}^{(t)}(i,d)]$,种群的全局最优位置为 $\boldsymbol{G}^{(t)}(g)$。

粒子更新位置的公式为

$$\begin{cases}\boldsymbol{X}^{(t+1)}(i)=\boldsymbol{G}(i)+\alpha\cdot|\boldsymbol{M}_{\text{best}}-\boldsymbol{X}^{(t)}(i)|\cdot\\\quad\ln\dfrac{1}{u},\quad\text{当 }k\geqslant 0.5;\\\boldsymbol{X}^{(t+1)}(i)=\boldsymbol{G}(i)-\alpha\cdot|\boldsymbol{M}_{\text{best}}-\boldsymbol{X}^{(t)}(i)|\cdot\\\quad\ln\dfrac{1}{u},\quad\text{当 }k<0.5。\end{cases}\tag{2}$$

式中:α 为收缩膨胀系数控制算法的收敛速度;u 和 k 为[0,1]随机数;$\boldsymbol{M}_{\text{best}}$为历史个体平均最优位置;$\boldsymbol{G}(i)$为局部吸引器用以保证算法的收敛性。

$$\boldsymbol{M}_{\text{best}}^{(t)}=\frac{1}{m}\sum_{i=1}^{N}\boldsymbol{G}^{(t)}(i);\tag{3}$$

$$\boldsymbol{G}(i)=\phi\boldsymbol{G}^{(t)}(i)+(1-\phi)\boldsymbol{G}^{(t)}(g)。\tag{4}$$

量子行为粒子群优化算法的应用范围广泛,其中应用比较多、效果比较好的在于支持向量机中的核函数和惩罚系数的参数寻优[14-15]。所以本文引入量子行为粒子群算法对 S3VM 的核参数和惩罚参数进行寻优。

1.3　协同训练算法

协同训练算法需要满足以下 2 个条件:

(1)必须有足够的数据集在所有属性集中分别训练出强分类器;

(2)如果数据集中的标签样本是已知的,那么其中样本中的属性集合需要各自独立。

以上条件意味着样本集必须拥有冗余且充分的视图。若上述条件都满足时,协同训练算法如下所示。

给定有标记数据集 X 与无标记数据集 U,数据集 X^1、X^2 为有标记数据集 X 的两个独立的属性视图。利用数据集 X^1 和 X^2 分别训练出两个不同的分类器,然后让每个分类器分别将无标记数据集 U 置信度最高的样本赋予伪标记,同时提供给另一个分类器最新增加的有标记样本用于训练更新,如此循环往复直至达到预先设定的迭代次数或者分类器不再变化。无标签样本未标记的置信度的计算式如下所示:

$$\Delta_u=\frac{1}{|X|}\sum_{x_i\in X}(y_i-f(x_i))^2-\frac{1}{|X|}\sum_{x_i\in X}(y_i-f'(x_i))^2。\tag{5}$$

式中:$f(x_i)$ 为当前分类器;$f'(x_i)$ 为加入标记过的无标记样本训练得到的分类器;y_i 为标签。

2　改进的协同训练半监督 SVM 模型

2.1　模型的提出

协同训练主要依赖于多视图的“相容互补性”。若数据包含 2 个充分、冗余视图,可在每个视图下,利用有标签样本训练得到一个分类

器[16];然后使用各个分类器分别对无标签样本标记进行预测,从而得到无标签样本的标记;最后在每个分类器中依据置信度估计方法,将预测标记置信度最高的无标记样本及其标签放置到另一个分类器中。循环此过程,直到达到最大循环次数或分类器都不再变化。

协同训练关键的步骤在于如何选择合适的初始分类器,以及如何计算无标签样本置信度[17]。为此,本文提出了改进的协同训练半监督SVM算法来解决此问题:首先利用QPSO优化的S3VM作为协同训练的初始分类器,提高初始分类器分类精度;然后将数据剪辑方法用于对无标签样本伪标记置信度的估计,减少分类器错分的概率。

2.2 初始分类器的选择

为了提高协同训练初始分类器对样本标注的准确率及训练速度,采用量子行为粒子群优化算法[10]来优化半监督支持向量机,获得一个强S3VM分类器作为初始分类器。基于QPSO-S3VM构建两个初始分类器,并引入半监督学习的思想。通过QPSO算法对S3VM的惩罚系数 C_1 和核参数 γ 这两个参数进行快速寻优,减少迭代训练的时间,提高样本标注的准确率,其次由于半监督学习思想的引进,减少了算法模型对有标签样本的依赖程度,在实际应用中大大降低了提取样本信息的成本代价。

2.3 基于近邻数据剪辑技术的置信度

在协同训练中,由于初始已标记数据集规模很小,以及初始分类器分类能力不强,在协同训练过程中噪声样本不断地引入,会导致模型分类能力低下。因此,数据剪辑(data editing)技术[11]被应用到协同训练中,切边权重统计(cut edge weight statistic)方法就是其中的一种。

通过一组有标记样本 $\boldsymbol{L}$ 构造一个无定向的近邻图 $\boldsymbol{G}_L$,探索近邻图 $\boldsymbol{G}_L$ 上的结构信息判断样本点 x_p 的标签 y_p 是否正确。在此基础上,每个样本 x_p 及其标签 y_p 的置信度可由切边权重统计估计为

$$J_p = \sum_{x_p \in C_p} w_{pq} I_{pq}。 \tag{6}$$

式中:C_p 为在近邻图 $\boldsymbol{G}_L$ 与 x_p 相关的所有的样本总集;$w_{pq} \in [0,1]$ 为近邻图中的权重,$w_{pq} = (1 + d(x_p, x_q))^{-1}$,$d(x_p, x_q)$ 可由欧式距离求得;每个 I_{pq} 对应一个独立同分布的伯努利随机变量,当 y_q 与 y_p 的标签不同时,I_{pq} 为1,通常,$p_r(I_{pq} = 1) = 1 - p_r(y = y_p)$。

当 C_p 的样本集充分大时,根据中心极限定理,J_p 可由正态分布的平均值 $\frac{\mu_p}{H_0}$ 及方差 $\frac{\sigma_p}{H_0}$ 近似建模为

$$J_p^s = \frac{\left(J_p - \frac{\mu_p}{H_0}\right)}{\frac{\sigma_p}{H_0}}; \tag{7}$$

$$\frac{\mu_p}{H_0} = (1 - p_r(y = y_p)) \sum_{x_p \in C_p} w_{pq}; \tag{8}$$

$$\frac{\sigma_p}{H_0} = p_r(y = y_p)(1 - p_r(y = y_p)) \sum_{x_p \in C_p} w_{pq}^2。 \tag{9}$$

可定义样本 (x_p, y_p) 置信度为

$$CF_Z(x_p, y_p) = 1 - \Phi(J_p^s)。 \tag{10}$$

式中:$\Phi(J_p^s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{J_p^s} e^{-\frac{t^2}{2}} dt$。

本文对于任何的标签样本 (x, y) 都可以采用基于近邻数据剪辑技术 $CF_Z(x_p, y_p)$ 来估计样本标签置信度。

2.4 量子行为粒子群优化的协同训练半监督SVM算法

2.4.1 算法描述

给出有标签样本集 $L = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$,且属性集为 X,$y_i \in \{+1, -1\}$,$i = 1, 2, \cdots, N$。属性集 X 由两个独立同分布的属性集 X^1、X^2 表示,在属性集 X^1、X^2 上将有标记样本集 L 划分为 L_1,L_2,然后利用基于 X^1、X^2 的属性集的有标签样本集 L_1 和 L_2 分别构造出两个存在差异性的QPSO-S3VM分类模型,用于对无标签样本数据的预测,最后选择置信度最高的无标签样本为伪标签,并将之放置于另一个分类器的有标记样本子集中,如此反复,直至两个分类器都不再发生变化,或达到了预先的迭代次数。

2.4.2 算法步骤

改进的协同训练半监督SVM算法如下所示。

算法1 协同训练算法。

输入:基于 X^1 的有标签样本集 L_1,基于 X^2 的有标签样本集 L_2,基于 X^1 的无标签样本集 U_1,基于 X^2 的无标签样本集 U_2;

输出:最终分类器 f,最优参数组合 $\{C_1, \gamma\}$。

Step 1 油层数据样本预处理。

Step 2 初始化粒子群。初始化粒子群 (C_1, γ),确定群体模型,设定粒子群参数及最大迭代次数 $T_{\max}$,每个粒子的个体最优解 $pbest_i$ 初始值

为 x_i 的初始值, $gbest_i$ 为全局最优解。

Step 3　评价各粒子适应度(*fitness*)。首先用 QPSO-S3VM 分别对有标签样本集 L_1、L_2 进行训练,得到两个初始分类器 f_1、f_2。其次将分类器 f_1 和 f_2 分别对无标签样本集 U_1 和 U_2 进行测试,用于预测无标签样本的标签。然后用无标签样本及其预测标签构造近邻图,并利用式(10)估计无标签样本预测标签的置信度。分别从 U_1 和 U_2 中选择最优的一组无标签样本及其预测标签放置对方的有标签样本训练集中。最后更新有标签样本集 L_1、L_2。在更新后的样本集 L_1、L_2 上重新训练,得到新的分类器模型。再对测试样本进行预测,采用 K 折交叉验证法计算的平均准确率 α_{k-cv} 计算每个样本的粒子适应度。

Step 4　对每个粒子,比较当前适应度 $f(x_i)$ 和历史最好位置适应度 $f(pbest_i)$,如果 $f(x_i) < f(pbest_i)$,那么 $pbest_i = x_i$;比较群体所有粒子当前适应度 $f(x_i)$ 和群体最好位置适应度 $f(gbest_i)$,如果 $f(x_i) < f(gbest_i)$,那么全局最优解 $gbest_i = x_i$。

Step 5　使用式(2)~(4)更新粒子的位置,产生新种群 $\boldsymbol{X}^{(t+1)}$。

Step 6　检查结束条件,若满足,则结束寻优,返回当前最优个体为结果,否则 $t = t+1$,转至 Step 3。设定结束条件为无标签样本集 U_1、U_2 为空且寻优达到最大迭代次数 $T_{\max}$ 或评价值小于给定精度。

Step 7　输出参数寻优结果。当满足结束条件时,最大适应度函数所对应的 C_1 和 γ 即为最优组合 $\{C_1,\gamma\}$。

Step 8　输出最终模型。当无标签样本集 U_1、U_2 为空,或达到最大迭代次数时,合并两个训练集 L_1、L_2 形成最终训练集 L,重新训练得到最终分类器 f,即识别模型。

3　油层识别应用

3.1　油层识别基本模型

基于量子行为粒子群优化的协同训练半监督 SVM(QPSO-CS3VM)油层识别模型如图 1 所示。油层识别一共有 5 个步骤。

(1)油层数据样本的预处理。对数据进行归一化的处理。

(2)油层属性的离散化。使用 0 表示无油层,1 表示油层,所以决策属性为 $D = \{d\}$, $d = \{d_i = i, i = 0,1\}$。

(3)对油层数据样本中的属性进行约简。常规的测井数据中不仅有多余的无效属性,且至少含有 15 种以上的测井信息。因此,本文采用基于属性重要性的约简算法[18]将属性集 X 分为 X^1、X^2 两部分。

(4)协同训练半监督 SVM 算法建模。在协同训练半监督 SVM 模型中,输入经属性约简后的样本信息,采用 QPSO-S3VM 算法进行训练,对无标签样本进行标注,同时利用数据剪辑技术以减少分类器的错分,提高样本标注准确率,最后得到 QPSO-CS3VM 分类模型。

(5)油层数据识别。使用 QPSO-CS3VM 模型对油层数据进行识别得到最终结果。

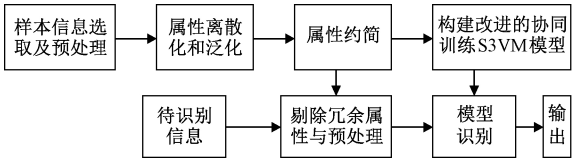

图 1　QPSO-CS3VM 油层识别模型

Figure 1　Oil layer recognition of QPSO-CS3VM

3.2　实际应用

本文选取了具有代表性的两口井(Q1 和 Q2)的实际测井数据进行训练和测试,用于证明本文提出的改进协同训练半监督 SVM 油层识别模型的应用效果。Q1 井的数据如表 1 所示,Q2 井的数据如表 2 所示。

表 1　Q1 井的基本数据

Table 1　Basic data of Q1 well

数据集种类	深度/m	样本数据点个数	油层数据个数	无油层数据点个数
训练集	1 220~1 290	281	53	228
测试集	1 180~1 320	2 537	478	2 059

表 2　Q2 井的基本数据

Table 2　Basic data of Q2 well

数据集种类	深度/m	样本数据点个数	油层数据个数	无油层数据点个数
训练集	1 230~1 270	160	47	113
测试集	1 190~1 290	641	192	449

(1)油层数据样本的预处理。Q1 井具有 11 个条件属性,分别为:AC、CALI、GR、NG、RA2、RA4、RI、RM、RT、RXO、SP。决策属性 $D = \{d\}$, $d = \{d_i = i, i = 0,1\}$, 0 和 1 分别表示无油层和油层。

Q2 井具有 28 个条件属性,分别为:AC、CNL、DEN、GR、RT、RI、RXO、SP、R2M、R025、BZSP、RA2、C1、C2、CALI、RINC、PORT、VCL、VMA1、VMA6、RHOG、SW、VO、WO、PORE、VXO、VW、

AC1。决策属性 $D=\{d\}$，$d=\{d_i=i,i=0,1\}$。0和1分别表示无油层和油层。

(2)对测井数据中的信息属性进行约简。Q1井的数据经过属性约简后，属性 X^1 由以下4个属性组成：AC、NG、RI、SP，属性 X^2 由CALI、GR、RA2、RA4、RM、RT、RXO这7个属性组成。

Q2井的数据经过属性约简后，属性 X^1 由以下5个属性组成：AC、GR、RT、RXO、SP，属性 X^2 由CNL、DEN、RI、R2M、R025、BZSP、RA2、C1、C2、CALI、RINC、PORT、VCL、VMA1、VMA6、RHOG、SW、VO、WO、PORE、VXO、VW、AC1这23个属性组成。

最后，对约简后的数据进行归一化处理以便模型进行油层识别，其中，Q1井数据属性 X^1 的归一化图如图2所示。

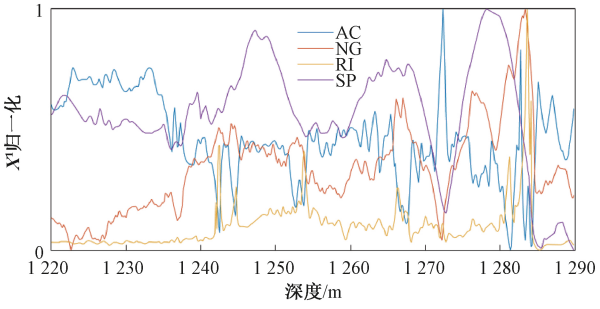

图2 属性 X^1 归一化处理

Figure 2 Normalization attribute of X^1

(3)识别结果及比较。将在Q1井的训练集上训练好的预测模型对Q1井1 180~1 320 m的2 537个样本进行油层识别。同时将在Q2井的训练集上训练好的预测模型对Q2井1 190~1 290 m的641个样本进行油层识别。最后，将本文提出的油层识别模型与PSO优化的SVM模型、QPSO优化的SVM模型[15]和传统协同训练半监督SVM(CS3VM)模型相比较，Q1井上测得的性能指标见表3，Q2井上测得的性能指标见表4，图3表示Q1和Q2井的真实油层分布及其预测油层分布。其中，运行时间是在CPU为Intel Core i7，内存为8 GB的计算机上的运行时间。

表3 Q1井的油层识别结果

Table 3 Oil layer recognition results of Q1 well

识别模型	训练时间/s	识别率/%
QPSO-SVM	17.79	93.50
PSO-SVM	12.06	92.00
CS3VM	0.37	90.00
本文算法	4.92	95.00

表4 Q2井的油层识别结果

Table 4 Oil layer recognition results of Q2 well

识别模型	训练时间/s	识别率/%
QPSO-SVM	10.63	93.91
PSO-SVM	7.03	91.89
CS3VM	0.35	90.95
本文算法	3.85	94.07

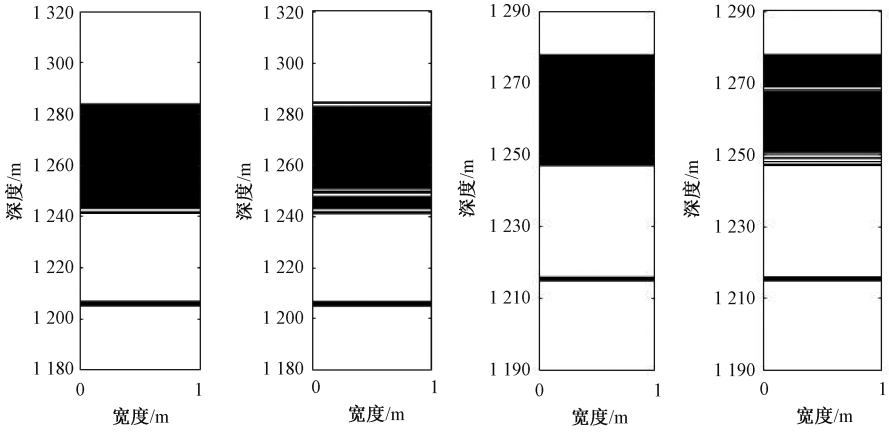

图3 Q1井与Q2井真实与预测油层分布对比

Figure 3 Comparison of actual and predicted oil layer distribution of Q1 and Q2 wells

由表3和表4可知，在相同条件下，在识别率方面，本文提出的改进的协同训练半监督SVM识别模型明显优于协同训练半监督SVM模型，主要是因为由QPSO优化的半监督SVM模型的分类效果要强于标准的半监督SVM识别模型。与全监督算法相比，本文算法在Q1井的测试集中得到了比PSO优化的SVM与QPSO优化的SVM模型更高的识别率，并且本文所提算法的识别率相较于基于S3VM的协同训练算法提高了5.00百分点。在Q2井的测试集训练结果中，本文算法

与其他 3 种算法相比,取得了 94.07%的最高识别率,并且相较于基于 S3VM 的协同训练算法,本文算法的识别率提高了 3.12 百分点,说明本文改进算法应用效果十分显著。

由图 3 可知,无论是对 Q1 井还是 Q2 井,本文提出模型预测的油层分布与真实油层分布十分接近,表现出了优异的油层识别性能。由此可验证本文提出的 QPSO 优化的协同训练半监督学习在油层识别中的有效性。

4　结论

针对传统 S3VM 算法分类精度较低,分类效果差的问题,本文采用了协同训练的思想,构建了两个分类器互相学习协同合作从而提高彼此分类精度。其次,为提高两个初始分类器的分类效果,引入了 QPSO 算法来优化 S3VM,以获得一个较好的初始分类结果,从而达到提高最终总体模型分类效果的目的。最后,使用一种改进的近邻数据剪辑方法预测无标签样本伪标签置信度,进而提高无标签样本预测精度,避免错分类样本进入循环而导致模型性能恶化。此方法应用于油层识别时,实验结果表明该改进模型分类效果优异,并在仅使用少量有标签样本的条件下,相对于其他对比算法,本文模型识别精度高,从而减少了获取有标签样本的代价,体现了半监督思想的优异性和有效性,具有很好的应用前景。

参考文献:

[1] ZHU C, XIA K, ZU B, et al. Multiple-kernel relevance vector machine based on SOCP[J]. Journal of computational information systems, 2014, 10(24): 10831-10838.

[2] LIU L, XIA X, XIA K, et al. Oil layer recognition by support vector machine based on semi-definite programming[J]. Journal of computational information systems, 2014, 10(6): 2579-2586.

[3] ZHANG J N, XIA K W, HE Z P, et al. Dynamic multi-swarm differential learning quantum bird swarm algorithm and its application in random forest classification model [J]. Computational intelligence and neuroscience, 2020, 2020: 1-24.

[4] 韩嵩,韩秋弘.半监督学习研究的述评[J].计算机工程与应用,2020,56(6):19-27.

[5] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 1998 11th Annual Conference on Computational Learning Theory. New York: ACM, 1998: 92-100.

[6] BIE T D, CRISTIANINI N. Convex methods for transduction[J]. Advances in neural information processing systems, 2004, 16: 1-8.

[7] XU P C, LU W K, WANG B F. A semi-supervised learning framework for gas chimney detection based on sparse autoencoder and TSVM [J]. Journal of geophysics and engineering, 2019, 16(1): 52-61.

[8] DAI D, LI H X, JIA X Y, et al. A co-training approach for sequential three-way decisions [J]. International journal of machine learning and cybernetics, 2020, 11(5): 1129-1139.

[9] TARIQ S, AKHTAR N, AFZAL H, et al. A novel co-training-based approach for the classification of mental illnesses using social media posts [J]. IEEE access, 2019, 7: 166165-166172.

[10] SUN J, XU W B, FENG B. A global search strategy of quantum-behaved particle swarm optimization [C]// IEEE Conference on Cybernetics and Intelligent Systems, 2004. Piscataway: IEEE, 2004: 111-116.

[11] ZHANG M L, ZHOU Z H. CoTrade: confident co-training with data editing[J]. IEEE transactions on systems, man, and cybernetics, part B (cybernetics), 2011, 41(6): 1612-1626.

[12] 高岳林,武少华.基于自适应粒子群算法的机器人路径规划[J].郑州大学学报(工学版),2020,41(4):46-51.

[13] 章昕,张飞,肖雄,等.基于量子粒子群算法-支持向量机的冷连轧断带故障诊断[J].冶金自动化,2020,44(6):17-24.

[14] LUO J, WANG X Y, XU Y G. Vibration fault diagnosis for hydroelectric generating unit based on generalized S-transform and QPSO-SVM[C]//2019 IEEE Sustainable Power and Energy Conference (iSPEC). Piscataway: IEEE, 2019: 2133-2137.

[15] LIU L. Oil layer recognition by support vector machine based on quantum-behaved particle swarm optimization [J]. Journal of information and computational science, 2014, 11(5): 1511-1518.

[16] 周志华.基于分歧的半监督学习[J].自动化学报,2013,39(11):1871-1878.

[17] 鲜焱,吕佳.基于核均值漂移聚类的改进局部协同训练算法[J].重庆师范大学学报(自然科学版),2020,37(4):106-113.

[18] BAI J C, XIA K W, LIN Y L, et al. Attribute reduction based on consistent covering rough set and its application[J]. Complexity, 2017, 2017: 1-9.

(下转第 26 页)

An Improved Ghost-YOLOv5 Infrared Target Detection Algorithm Based on Feature Distillation

LI Beiming[1], JIN Ronglu[1,2], XU Zhaofei[2], LIU Qing[2], WANG Shuigen[2]

(1.College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China; 2.Yantai IRay Technology Co., Ltd., Yantai 264000, China)

Abstract: Infrared target detection algorithms suffered from problems such as poor adaptability and high computational complexity. An improved Ghost-YOLOv5 infrared target detection algorithm was proposed based on feature distillation to solve the above problems. Firstly, GhostNet block was used for backbone pruning. Secondly, two effective data enhancement methods including Mosaic and Copy-paste were used, together with feature distillation to improve the accuracy in object detection. Furthermore, an infrared image dataset that contained a variety of scenarios with pedestrians, motor vehicles, and non motorized targets was constructed. The test experimental results on the above dataset showed that the model parameters obtained by the algorithm proposed in this paper using GhostNet module were only 1.9M, and the accuracy of the small model on the infrared dataset were improved by 6.6% through feature distillation and data enhancement. And the overall *mAP* value reached 90.1%. The detection speed of the model could reach 25 frames per second and the average detection accuracy could reach 90.2% when measured empirically on Hisi, all achieving higher detection accuracy compared to a variety of common models portable to this platform.

Keywords: infrared target detection; data enhancement; model pruning; feature distillation; Hisi platform

---

(上接第19页)

Improved Co-training Semi-supervised SVM and Its Application in Oil Layer Recognition

PAN Yongke, HE Ziping, XIA Kewen, NIU Wenjia

(School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: It is expensive to obtain labeled data in actual oil logging, and a large amount of cheap unlabeled data are not used. How to use limited labeled samples and a large number of unlabeled samples to obtain accurate oil layer distribution remains to be solved. The semi-supervised learning methods were widely used because they could obtain good classification models using both a small number of labeled samples and a large number of unlabeled samples. Therefore, based on a semi-supervised support vector machine (S3VM), an improved semi-supervised support vector machine based on co-training and quantum-behaved particle swarm optimization algorithm (QPSO-CS3VM) was proposed for oil layer recognition. Firstly, the multi-view-based co-training strategy combined with S3VM was used to construct two independent initial classifiers, and then exchanged and labelled unlabeled samples to improve the overall oil layer recognition accuracy. Secondly, in order to improve the initial classification accuracy of original classifiers, the quantum behavioral particle swarm algorithm was introduced to optimize S3VM. Finally, a newly nearest neighbor data editing approach was used to predict the confidence of the pseudo-labelling of unlabeled data to reduce the deterioration of model performance caused by misclassification of data. The improved co-training semi-supervised SVM proposed in this paper improved the classification accuracy by 5.00% and 3.12% compared to the traditional co-training algorithm by performing oil layer recognition on the logging data of the two wells. The algorithm proposed in this paper had high accuracy in oil layer recognition and had practical application.

Keywords: semi-supervised support vector machine; co-training; QPSO; data editing; oil layer recognition