集成在线贯序超限学习机的分类算法

王麗雯[1,2] 闫羿成[1,2] 包洋[1,2]

（1.盐城师范学院信息工程学院，江苏 盐城 224002；2.南京工业大学，江苏 南京 211816）

摘 要： 为了有效解决数据流的分类问题，提出一种集成在线贯序超限学习机算法（A-OSELM）。为了解决现有分类算法精度不高的问题，本文将经典的集成算法AdaBoost与在线贯序超限学习机相结合，生成一种集成在线贯序超限学习机(A-OSELM)算法，形成一个强分类器，在有效提升分类精度的同时，也不会造成训练速度过慢的问题，最终实验结果证明这一算法的有效性。

关键词： 在线学习；在线贯序超限学习机；数据流分类；集成算法

0 引言

随着时代的发展，数据对我们的生活产生了巨大的影响，所以数据挖掘已经成为各行业的重要话题[1]。其中数据分类是数据挖掘中的一重要课题，受到公众的广泛关注。且随着大数据的快速发展，互联网行业每天在源源不断地生成海量数据，我们所接触的信息也都是以流式数据形式出现，具有快速、连续、实时、海量的特点[2]。故需要用在线的方式去解决，所以在线学习应运而生，许多学者对数据流的分类进行深入研究，常见的分类算法有随机森林、在线支持向量机等。目前的算法都存在计算量过大或者是分类精度不高的问题，如何有效兼顾训练速度与分类精度一直是学者们所要解决的问题。

1 OSELM算法

超限学习机（ELM）是OSELM的应用基础，是基于神经网络思想提出的一种新型单隐层前馈神经网络（Single-hidden Layer Feedforward Neural Network，SLFNs）的学习方法。通过随机产生输入权值与隐含层参数，通过最小二乘法计算得到输出权重，在延续神经网络泛化性能好的同时还具有操作简单、运行速度较快等特点。超限学习机的基本思想如下：假设有N个不同的训练样本$(\boldsymbol{x}_j,\boldsymbol{t}_j)\in R^n\times R^m$。其中：$\boldsymbol{x}_j$是$n\times 1$维的输入向量，$\boldsymbol{t}_j$是$m\times 1$维的目标向量。若该神经网络包含$N$个隐层节点，则ELM可以表示为：

$$f_N(\boldsymbol{x}_j)=\sum_{i=1}^{N}\boldsymbol{\beta}_i G(\boldsymbol{a}_i,\boldsymbol{b}_i,\boldsymbol{x}_j)=\boldsymbol{t}_j,\quad j=1,2,\ldots,N. \tag{1}$$

其中：$G(\boldsymbol{a}_i,\boldsymbol{b}_i,\boldsymbol{x}_j)$是第$i$个隐层节点对于输入向量$\boldsymbol{x}_j$的激活函数，$\boldsymbol{a}_i$和$\boldsymbol{b}_i$是关于输入节点与第$i$个隐藏节点随机生成的连接权值向量与偏置，$\boldsymbol{\beta}_i\in R^m$是连接输出节点的与第$i$个隐藏节点权值向量。其中式（1）可写成矩阵形式：

$$\boldsymbol{T}=\boldsymbol{H}\boldsymbol{\beta} \tag{2}$$

其中$\boldsymbol{H}$是隐藏层的输出矩阵，$\boldsymbol{\beta}$为输出权重矩阵，$\boldsymbol{T}$为目标矩阵，可以表示为：

$$\boldsymbol{H}=\begin{bmatrix} G(\boldsymbol{a}_1,\boldsymbol{b}_1,\boldsymbol{x}_1) & \cdots & G(\boldsymbol{a}_N,\boldsymbol{b}_N,\boldsymbol{x}_1)\\ \vdots & \ddots & \vdots \\ G(\boldsymbol{a}_1,\boldsymbol{b}_1,\boldsymbol{x}_N) & \cdots & G(\boldsymbol{a}_N,\boldsymbol{b}_N,\boldsymbol{x}_N)\end{bmatrix}_{N\times N} \tag{3}$$

$$\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1^T\\ \vdots\\ \boldsymbol{\beta}_N^T\end{bmatrix}_{N\times m},\quad \boldsymbol{T}=\begin{bmatrix}\boldsymbol{t}_1^T\\ \vdots\\ \boldsymbol{t}_N^T\end{bmatrix}_{N\times m}$$

而后求得输出权重矩阵$\boldsymbol{\beta}$的最小二乘解为：

$$\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}\boldsymbol{Y}=\left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{Y} \tag{4}$$

其中$\boldsymbol{H}^{\dagger}$为H的Moore-Penrose广义逆，若$(\boldsymbol{H}^T\boldsymbol{H})$为非奇异，可通过正交投影法表示得到式（4）。

为了从大量连续到达的数据流中挖掘出所需知识，Liang提出的OSELM以学习增量数据的方式进行在线学习。该算法分为两部分，第一部分为初始化，第二部分为在线序列学习。具体过程如下：

Step1：初始阶段：

（1）从给定的数据集$D=\{(\boldsymbol{x}_i,\boldsymbol{y}_i),i=1,2,\cdots N\}$中选取部分数据集$D_0=\{(\boldsymbol{x}_i,\boldsymbol{y}_i),i=1,2,\cdots N_0\}$进行初始训练，其中$N_0\geq N$；（2）根据随机生成的输入权值$\boldsymbol{a}_i$和$\boldsymbol{b}_i\,(i=1,2,\cdots N)$，求得隐藏层输出矩阵$\boldsymbol{H}_0$；（3）根据式（4）求得初始输出权值$\boldsymbol{\beta}^0=\boldsymbol{P}_0\boldsymbol{H}_0^T\boldsymbol{T}_0$，其中，$\boldsymbol{P}_0=\left(\boldsymbol{H}_0^T\boldsymbol{H}_0\right)^{-1}$。

Step2：在线阶段：

假设$D_{k+1}=\left\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\,|\,i=\left(\sum_{l=0}^{k}N_l\right)+1,\cdots\sum_{l=0}^{k+1}N_l\right\}$新数据到来，$N_{k+1}$表示第$K+1$个数据块中的样本个数，首先计算出对应的隐藏层输出矩阵$\boldsymbol{H}_{k+1}$，再更新输出权重矩阵$\boldsymbol{\beta}^{k+1}$：

$$\boldsymbol{P}_{k+1}=\boldsymbol{P}_k-\boldsymbol{P}_k\boldsymbol{H}_{k+1}^T\left(\boldsymbol{I}+\boldsymbol{H}_{k+1}\boldsymbol{P}_k\boldsymbol{H}_{k+1}^T\right)^{-1}\boldsymbol{H}_{k+1}\boldsymbol{P}_k \tag{5}$$

$$\boldsymbol{\beta}^{k+1}=\boldsymbol{\beta}^k+\boldsymbol{P}_{k+1}\boldsymbol{H}_{k+1}^T\left(\boldsymbol{T}_{k+1}-\boldsymbol{H}_{k+1}\boldsymbol{\beta}^k\right) \tag{6}$$

2 AdaBoost集成算法

Boosting是集成的一种，个体学习器间存在强依赖关系、必须串行生成的序列化方法。而AdaBoost则是其代表方法之一，通过不断迭代地训练原始样本，用于自适应地改变训练样本的分布，每次迭代根据上一次的训练结果修改原始样本的分布，使得基分类器关注在那些容易误分的样本上，最终各基分类器之间进行集成，形成一个强分类器[3]。

3 集成在线贯序超限学习机算法

首先将每个样本设置相同的初始权重进行训练，再根据迭代函数来调整每个样本的权重分布，将新的权值发送给下层分类器进行新一轮的训练。对于错分样本，再次加大其权重，使得在下次的训练中其更多地关注错分样本，以提高整体分类精度。迭代完成后进行组合，最终生成强分类器。算法描述如下：

输入数据$\{x_i,y_i\}$，$x_i\in R^d$，$y_i\in R^q$，其中$i\in 1,\ldots,N$。T为迭代次数，g为基分类器OSELM。

样本权重$D_1=(W_{11},W_{12},\cdots,W_{1N})$初始化，其中$W_{1i}=1/N,i=1,2,3,\cdots,N$。

For $t=1,\ldots\ldots T$ do

根据训练数据的权值分布，D_t的训练基分类器$g_t(x_i)$。

计算$g_t(x_i)$在训练集上的分类误差$e_t=\sum_{i=1}^{N}w_{ti}I(g_t(x_i)\neq y_i)$。

计算弱分类器在最终分类器中所占的权重：

$\alpha_t=\frac{1}{2}\ln\left(\frac{1-e_t}{e_t}\right)$。

更新样本的权值分布：$D_{t+1}=\frac{D_t(i)\exp(-\alpha_t y_i g_t(x_i))}{Z_t}$，其中$Z_t$为归一化因子。

最终分类器：$G(x)=sign\left(\sum_{t=1}^{T}\alpha_t g_t(x_i)\right)$。

4 实验操作及结果分析

本次实验的所有仿真软件均在同一系统环境下运行，系统运行环境为PC（i5-7200U CPU@ 2.50 GHz，4.00 GB RAM），操作系统为Windows10，仿真软件为Matlab R2017a。实验中，将所有数据归一化到[-1,1]区间中。ELM网络的激活函数均为Sigmoid函数。

4.1 数据集

为了验证所提算法的有效性，本文采用五个常见的二分类数据集进行对比实验，将每个数据集分成训练集与验证集，为了排除偶然性，每组数据随机进行10次实验，最终结果取10次实验结果的平均值（见表1）。根据算法设定，隐层节点的节点个数$N<N_0$。

4.2 实验结果及分析

由表2可知，所提算法A-OSELM的测试精度相较于OSELM有明显的提升，这一结果验证了集成在线贯序超限学习机的有效性，在实验过程中，训练时间并没有明显地增加，且每次的训练结果都比较稳定，未出现明显的波动。这是因为集成训练时分类器重点关注于错分样本，使得其在下次的训练中加大错分权重，直至分类正确。

表 1 二分类数据集描述

dataset	No. Of attributes	No. of training samples	No. of testing samples
Pima	8	500	268
Yeast1	8	1000	484
diabetes	8	576	192
Vehicle0	18	600	246
iris0	4	100	50

表 2 隐层节点个数、初始块、序列块大小与测试精度

dataset	No. Of hidden neurons	Initialization set size	Chunk size	测试精度	
				OSELM	A–OSELM
Pima	8	500	268	0.5037	0.6041
Yeast1	8	1000	484	0.5072	0.6529
diabetes	8	576	192	0.6729	0.7313
Vehicle0	18	600	246	0.4671	0.7093
iris0	4	100	50	1.0000	1.0000

5 结语

本文针对数据流分类时测试精度不高的问题，提出集成在线贯序超限学习机算法（A-OSELM），选择OSELM为基分类器，结合Adaboost集成思想，不断调整各基分类器的权重进行迭代训练，重点关注错分样本，直至生成强分类器，以保证分类结果的精度有所提升。先对算法进行理论分析，再选取5组数据进行对比实验，最终结果表明A-OSELM算法的有效性。但实验都是在假设数据流未出现波动的情况下进行的，接下来的工作则需要考虑在线情况下数据流出现波动等问题，这会对数据分布以及分类结果产生怎样的影响。

参考文献

[1] 黄树东 . 协同聚类及集成的关键技术研究 [D]. 西南交通大学, 2015

[2] 吕光春 , 秦斌 , 祝兴星 . 基于极限学习机的高铁永磁直驱电机转子位置预测 [J]. 电子元器件与信息技术 ,2018(12):82–84.

[3] 董学辉 . 逻辑回归算法及其 GPU 并行实现研究 [D]. 哈尔滨工业大学 ,2016.

（上接第234页）

施工围蔽阻断了建设六马路的过境交通并影响片区的对外、内部交通。过境交通分流使环市东路东行流量增加527pcu/h，服务水平从C级降为D级；先烈南路南行流量增加357pcu/h；农林下路南行流量增加170pcu/h；东风东路流量变化较小服务水平不变。交叉口方面，农林下路-东风东路交叉口北进口最大排队长度与现状相比增加98.9m；先烈南路-东风东路交叉口东北进口、东进口排队长度均增加近20m。

5 结语

本研究以复杂环境下地铁施工引起的交通疏解为研究对象，以广州十三号线建设六马路站为例探讨交通疏解需要注意的要点，分析溯源车流运行趋势，通过车流仿真模型为手段，提出具有可行性的方案，为维持地铁施工期间的道路服务水平具有实际意义。

参考文献

[1] 张旭丹 . 复杂环境下地下车站施工方法的研究 [J]. 广东建材 ,2012,28(7):77–79.

[2] 安应选 , 杨硕 , 余璞 . 复杂环境下地铁区间风井设计方案研究 [J]. 现代城市轨道交通 ,2021(8):55–59.

[3] 王琳颖 . 主干道施工交通疏解策略——以南京建宁路为例 [J]. 商品与质量 ,2020(11):265–266.

[4] 杨青 . 我国城市轨道交通运营管理存在的问题探寻 [J].IT 经理世界 ,2021(3):139,141.

[5] 孙洪岩 , 杨平 , 张彦红 , 等 . 地铁车站交通疏解钢便桥数值分析研究 [J]. 地下空间与工程学报 ,2010,6(3):472–476.