计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20170213
54(8):1736-1743,2017

基于 Storm 的在线序列极限学习机的气象预测模型

欧阳建权[1,2] 周 勇[1] 唐欢容[1,2]
[1](湘潭大学信息工程学院 湖南湘潭 411105)
[2](智能计算与信息处理教育部重点实验室(湘潭大学) 湖南湘潭 411105)
(oyjq@xtu.edu.cn)

A Meteorological Predication Model Based on Storm and Online Sequential Extreme Learning Machine

Ouyang Jianquan[1,2], Zhou Yong[1], and Tang Huanrong[1,2]
[1](*College of Information Engineering*, *Xiangtan University*, *Xiangtan*, *Hunan* 411105)
[2](*Key Laboratory of Intelligence Computing and Information Processing* (*Xiangtan University*), *Ministry of Education*, *Xiangtan*, *Hunan* 411105)

Abstract In order to improve the accuracy of meteorological forecasting, deal with frequent local meteorological disasters in real time, and have higher efficiency of dealing with massive data, this paper proposes a meteorological forecasting model using the Storm-based online sequential extreme learning machine. The model firstly initializes multiple online extreme learning machine. When new batches of data arrive, the model continually studies the new data samples based on the training results, and introduces the stochastic gradient descent method and the error weight adjustment method to give the error feedback for new prediction results and then update the error weight parameters in real time, and finally to improve prediction accuracy. In addition, the Storm flow processing framework is adopted to improve the proposed model in the aspect of parallelism in order to enhance the ability of dealing with massive high-dimensional data. The experimental results show that compared with the Hadoop-based parallel extreme learning machine (PELM), the proposed model has higher prediction accuracy and more excellent parallelism.

Key words Storm; extreme learning machine; meteorological predication; online sequence; machine learning

摘 要 为提高气象预测精度,实时应对频发的局域气象灾害,拥有更高的处理海量数据的效率,提出了一种基于 Storm 的在线序列的极限学习机气象预测模型.该模型首先初始化多个在线极限学习机,当新批次的数据不断到达时,模型能够在训练结果的基础上继续学习新样本,并引入随机梯度下降法和误差权值调整方法,对新的预测结果进行误差反馈,实时更新误差权值参数,以提高模型预测准确率.另外,采用 Storm 流式处理框架对提出的算法模型进行并行化改进,以提高处理海量高维数据的能力.实验结果表明:该模型与基于 Hadoop 的并行极限学习机算法(parallel extreme learning machine, PELM)相比,具有更高的预测精度和优异的并行性能.

关键词 Storm;极限学习机;气象预测;在线序列;机器学习

中图法分类号 TP391

收稿日期:2017-03-20;**修回日期**:2017-06-19

基金项目:国家自然科学基金项目(61672495);湖南省教育厅重点项目(16A208)

This work was supported by the National Natural Science Foundation of China (61672495) and the Key Projects of Hunan Provincial Department of Education (16A208).

随着信息技术的飞速发展,各个领域信息化程度不断加深,天气预报、工业生产、交通管理、图像识别、医疗诊断等人们生活中频繁接触的日常应用越来越依靠计算机系统来采集、存储和分析数据,而其中数据分析和处理的关键正是机器学习技术.随着机器学习技术的研究日益深入,也给气象行业带来了新的挑战.气象数据主要来自于地面观测、气象卫星遥感、天气雷达和数值预报产品.这 4 类数据占数据总量的 90%以上,直接应用于气象业务、天气预报、气候预测以及气象服务.

气象预报的发展使得气象数据积累速度迅速提高,因此对机器学习技术提出了更高的要求.传统的机器学习往往采用批量学习的方法,即所有的训练样本一次性学习完毕后,学习过程不再继续.但在实际应用中,训练样本空间的全部样本并不能一次全部得到,而往往是随着时间顺序得到.针对当前大部分机器学习算法无法在单个节点上处理的情况,研究者们通过并行处理的方式对大规模的数据进行学习,例如将学习过程分布到结点之间从而利用多核机器、计算结点集群甚至超级计算机的并行计算能力来完成机器学习的任务.虽然大规模的硬件资源能够在一定程度上缓解数据量大带来的问题,但是对新到达的数据进不能行快速处理学习并及时更新学习获得的知识[1].考虑到训练和预测的时空开销需求,能够在已有训练结果的基础上继续学习新样本,不断增强模型本身的识别能力,并且减少重复学习的时空开销的在线学习方法得到了广泛的关注[2].

目前,气象预测研究方面已经有前人的诸多成果,例如利用 SVM 和小波分解进行大气污染预测[3]、人工神经网络对水平面太阳辐射和风速的预测[4-5]、模糊神经网络对短期降雨量预测[6]、朴素贝叶斯预测和决策树方法对气温预测[7-8]、遗传算法和混合粒子群优化的 RBF 神经网络对降雨量的预测[9]和基于人工蜂群算法和遗传算法的混合分类器对降雨量的预测[10]等.

这些方法的不足主要表现在 2 个方面:1)采用离线分析气象数据,不能及时反映气象变化;2)随着气象预报要求不断提高,气象数据计算规模急剧膨胀,其处理数据的效率已经不能适应当前气象预测要求.

对于现阶段多层神经网络的深度学习,虽然其具有学习精度高、拟合能力强的优势,但是由于其多层复杂的神经网络训练以及大量的学习参数,使得深度学习极易陷入局部极小值和过拟合问题,并且需要花费大量的时间进行训练和消耗更多的硬件资源,不适合当前的实时在线学习的需求.黄广斌等人所提出来的极限学习机(extreme learning machine,ELM)[11],是一种求解单隐层神经网络的算法.ELM 最大的特点是相对于传统的神经网络,ELM 是单隐层前馈神经网络,它并不需要对所有的网络参数进行调整,输入权值和隐含层偏差在训练开始时随机给定,在训练过程中固定,而输出连接权值可通过求解线性方程组的最小二乘解来得到,具有泛化性能好的优点,在保证学习精度的前提下比传统的学习算法速度更快.

针对气象数据的实时计算与海量处理的需求,本文提出了一种基于 Storm 的在线序列的极限学习机气象预测模型.该模型首先初始化多个在线极限学习机,当新批次的数据不断到达时,模型能够在训练结果的基础上继续学习新样本,并引入随机梯度下降法和误差权值调整方法,对新的预测结果进行误差反馈,实时更新误差权值参数,以提高模型预测准确率.另外,采用 Storm 流式处理框架对提出的算法模型进行并行化改进,以提高处理海量高维数据的能力.实验结果表明,该模型与基于 Hadoop 的并行极限学习机算法(PELM)[12]相比,具有更高的预测精度和优异的并行性能.

1 ELM 算法分析

ELM 是一种单隐层前馈神经网络的学习算法,它并不需要对所有的网络参数进行调整,输入权值和隐含层偏差在训练开始时随机给定,在训练过程中固定,而输出连接权值可通过求解线性方程组的最小二乘解来得到.

虽然 ELM 在准确率、计算性能、执行时间方面都优于大部分机器学习算法,但 ELM 算法是一种批处理学习算法,在实际气象预测中,其并不完全适合气象预测场景,因此,在线序列优化是很有必要的.在线序列优化 ELM 中,当不断到达新的批次的数据时,能够在已有的训练结果上的基础上继续学习新样本,并引入随机梯度下降法和误差权值调整的思想,对新的预测结果进行误差反馈,实时更新误差权值参数,以提升模型预测准确率.另外采用 Storm 流式处理框架对提出的算法模型进行并行化改进,以提高处理海量高维数据的能力.

2 基于 Storm 的在线序列极限学习机模型

2.1 在线序列极限学习机模型

对于 N_0 个任意的训练样本$(\boldsymbol{x}_i,\boldsymbol{t}_i)$,令 $k=0$,

首先随机生成第 i 个隐层节点与输入节点的权值向量 $\boldsymbol{\omega}_i=[\omega_1,\omega_2,\cdots,\omega_n]^{\mathrm{T}}$ 和激励函数的偏置参数,利用极限学习机算法的思想,计算初始隐层输出矩阵 $\boldsymbol{H}_0$. 希望求得满足 $\|\boldsymbol{H}_0\boldsymbol{\beta}-\boldsymbol{T}_0\|$ 最小的 $\boldsymbol{\beta}^0$,则 $\boldsymbol{\beta}^0$ 由公式 $\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}\boldsymbol{T}$ 可以计算得到:

$$\boldsymbol{\beta}^0=\boldsymbol{K}_0^{-1}\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{T}_0, \tag{1}$$

其中,$\boldsymbol{K}_0=\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{H}_0$. 当一个新的训练数据进入系统时,假设为有 N_1 个样本进入模型,可求得:

$$\boldsymbol{\beta}^1=\boldsymbol{K}_1^{-1}\begin{bmatrix}\boldsymbol{H}_0\\ \boldsymbol{H}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{T}_0\\ \boldsymbol{T}_1\end{bmatrix}=$$

$$\boldsymbol{\beta}^0+\boldsymbol{K}_1^{-1}\boldsymbol{H}_1^{\mathrm{T}}(\boldsymbol{T}_1-\boldsymbol{H}_1\boldsymbol{\beta}^0). \tag{2}$$

当有 k 个样本进入模型,可以得到在线序列极限学习机算法的输出权重 $\boldsymbol{\beta}$ 的递推公式:

$$\boldsymbol{\beta}^{k+1}=\boldsymbol{\beta}^k+\boldsymbol{K}_{k+1}^{-1}\boldsymbol{H}_{k+1}^{\mathrm{T}}(\boldsymbol{T}_{k+1}-\boldsymbol{H}_{k+1}\boldsymbol{\beta}^k). \tag{3}$$

2.2 误差权值的调整方法

为了使多个并行节点的在线序列极限学习机[13]模型拥有更高的预测准确率,本文采用随机梯度下降算法(stochastic gradient descent, SGD)[14]和加权平均的方法,动态调整集群中多个不同节点预测输出结果的误差权值. 预测准确率高的节点被赋予更高的权重. 最终预测值 $\bar{y}$ 通过各个节点输出结果和误差权值加权平均得到:

$$\bar{y}_j=\sum_{i=1}^{k}\sigma_{ji}y_{ji}\Big/\sum_{i=1}^{k}\sigma_{ji}, \tag{4}$$

其中,σ_{ji} 为第 i 个学习机节点的误差权值,y_{ji} 为第 i 个学习机节点的输出值,j 为预测阶段的第 j 批次预测. 误差权值通过误差函数 E 计算:

$$E=\frac{1}{2}(y_{ji}-\bar{y}_j)^2, \tag{5}$$

要使误差函数达到最小值,对误差函数求导:

$$\Delta E(\sigma_{ji})=\frac{\partial E}{\partial\sigma_{ji}}=\frac{\partial E}{\partial\bar{y}_j}\frac{\partial\bar{y}_j}{\partial\sigma_{ji}}=-(y_{ji}-\bar{y}_j)\frac{\partial\bar{y}_j}{\partial\sigma_{ji}}=$$

$$-(y_{ji}-\bar{y}_j)\left[\frac{y_i\sum_{i=1}^{k}\sigma_{ji}-\sum_{i=1}^{k}\sigma_{ji}y_{ji}}{\left(\sum_{i=1}^{k}\sigma_{ji}\right)^2}\right]. \tag{6}$$

使用随机梯度下降法,可得到预测权重更新式(7)和式(8),其中,η 为学习速率,这里设置为 $\eta=0.1$.

$$\Delta\sigma_{ji}=-\eta\frac{\partial E}{\partial\sigma_{ji}}=\eta(y_{ji}-\bar{y}_j)\times$$

$$\left[\frac{y_i\sum_{i=1}^{k}\sigma_{ji}-\sum_{i=1}^{k}\sigma_{ji}y_{ji}}{\left(\sum_{i=1}^{k}\sigma_{ji}\right)^2}\right], \tag{7}$$

由于是求最小化误差函数,所以按每个参数 σ 的梯度负方向来更新每个 σ,得到:

$$\sigma_{j,i+1}=\sigma_{ji}-\Delta\sigma_{ji}. \tag{8}$$

2.3 基于 Storm 的在线序列极限学习机流程

基于 Storm 的在线序列极限学习机气象预测模型流程示意图如图 1 所示:

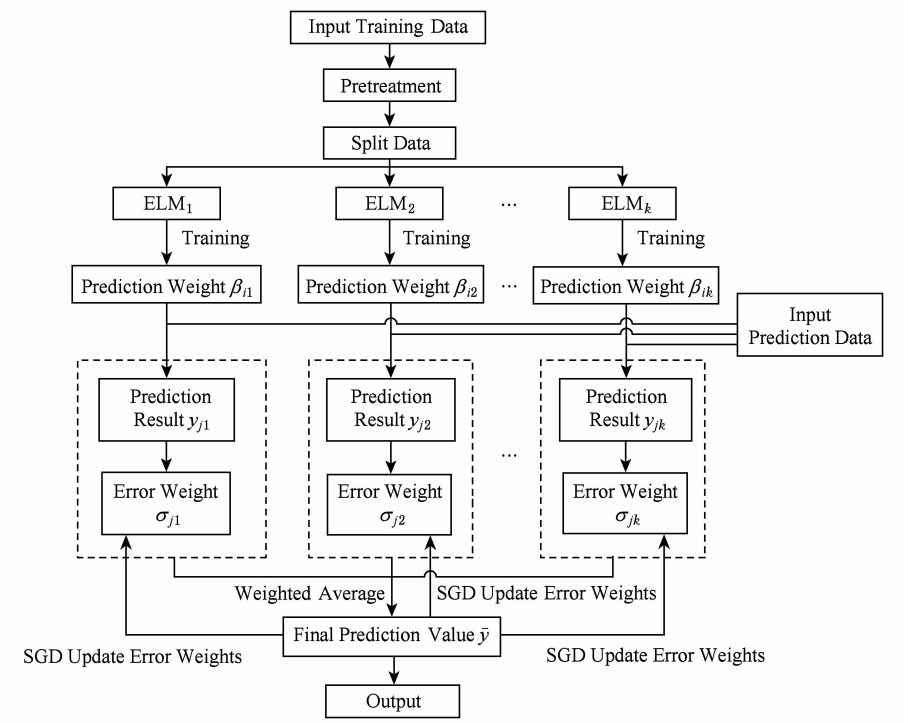

Fig. 1 The flow chart of the meteorological prediction model using S-OSELM

图 1 S-OSELM 气象预测模型流程图

1) 初始化阶段

首先输入训练数据,数据经过关联分析和离散化的预处理后,之后采用 Hash 值求余的方式对数据均匀分割,通过 Kafka 分布式消息队列机制,将数据发送到 Storm 集群中 k 个 ELM 节点,并将每个误差权值 σ 初始化为 1.

2) 训练阶段

通过在线序列极限学习机的的训练方法,对不断传送过来的数据进行分布式训练得到 k 个输出权重向量 $\boldsymbol{\beta}$,每传送过来一批数据集则将输出权重向量 $\boldsymbol{\beta}$ 更新一次,不断增强模型识别能力.

3) 预测阶段

该阶段分为 2 个步骤:

① 输入预处理后的用于预测的第 j 批次数据集,由 $\boldsymbol{H\beta}=\boldsymbol{T}$ 得到 k 个预测结果 $y_{j1},y_{j2},\cdots,y_{jk}$;

② 将 k 个预测结果和预测结果对应的误差权值用加权平均得到最终预测值 $\bar{y}_j$.

4) 调整误差权值阶段

通过 SGD 求解预测结果和最终预测值 $\bar{y}_j$ 得到误差权值 σ_j,并及时更新当前的误差权值.

5) 预测结果输出阶段

输出最终预测值 $\bar{y}_j$. 按照在线序列学习机制以及长期降雨量预测要求,返回步骤 3)预测下一阶段的气象数据.

3 实验分析

3.1 实验环境

本文实验环境基于 Storm 集群,采用完全分布式模式搭建 9 个节点,其中 1 台主节点(Nimbus),其余 8 台为从节点(Supervisor). 每个节点机器配置为 2.60 GHz 四核 CPU,4 GB 内存,操作系统为 Ubuntu-Server Linux14.04,网络带宽为 100 Mbps,Storm 版本为 0.9.2.

3.2 实验样本

在诸多气象指标预测中,降雨量是防灾减灾的重要参量,很大程度反映灾害发生趋势,降雨量对农业生产、水土流式和工程应用等有着重要的影响,对一个地区的降雨量进行准确预测,可以帮助农业、水利部门提高防治旱涝灾害的能力,将危害降低到最低.

实验样本选用英国 Met Office① 发布的气象数据. 本文使用的是该网站提供的华南某地区城市 2005 年至 2016 年真实的气象数据. 气象预测目标是通过以上数据样本预测某时段的的降雨量.

这些数据的属性有 28 项之多,如大气压、平均气温、湿度、风速、风向、土壤温度等. 为了提高算法预测速率和准确率,本文对该数据采用相关性分析和离散化方法进行预处理,预处理的结果将作为训练集和测试集样本.

1) 相关性分析

首先剔除不完整和格式不正确的数据. 然后选取与预测目标相关性大的气象属性,并剔除相关性小的气象属性,达到降维的目的. 最后计算它们与降雨量之间的相关系数 γ_{xy}[15]:

$$\gamma_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},\quad(9)$$

其中,$\bar{x}=\sum_{i=1}^{n}x_i/n,\bar{y}=\sum_{i=1}^{n}y_i/n$.

计算得到结果如表 1 所示. 当 $|\gamma_{xy}|=0$ 时,称 x 与 y 不相关;当 $|\gamma_{xy}|=1$ 时,此时 x、y 之间具有线性相关性. $|\gamma_{xy}|$ 的值越大,表示 x 与 y 相关性越高. 在实验组选取 $|\gamma_{xy}|>0.4$ 的气象属性作为预测属性,它们分别是相对湿度、总云量、露点温度、平均气温、大气压和风速.

Table 1 The Correlation Coefficient Between Different Weather Attributes and Rainfall

表 1 各气象属性与降雨量之间的相关系数

Number	Attribute Name	Correlation Coefficient
1	Relative Humidity	0.971 452
2	Average Temperature	−0.402 308
3	Dew-point Temperature	0.452 773
4	Soil Surface Temperature	0.081 465
5	Sunshine Hours	−0.291 572
6	Atmosphere Pressure	−0.737 522
7	Wind Speed	0.723 267
8	Wind Direction	0.098 411
9	Total Amount of Cloud	0.865 124
10	Horizontal Visibility	−0.018 393 3
⋮	⋮	⋮

① http://rp5.ru/

2) 数据离散化处理

采用 PKID 算法[16] (proportional k-interval discretization)离散处理,最终得到样本属性如表 2 所示.输入样本是 7 维属性向量,分别为相对湿度、总云量、露点温度、平均气温、大气压、风速和降雨量.

Table 2 The Sample Attributes Table

表 2 样本属性表

Number	Attribute Name	Data Type
1	Relative Humidity	int
2	Total Amount of Cloud	int
3	Dew-point Temperature	int
4	Average Temperature	int
5	Atmosphere Pressure	int
6	Wind Speed	int
7	Rainfall	double

3.3 实验结果分析

本文采用精度和性能对实验结果进行评估.

3.3.1 精度评估

计算正确率:

$$R_{\text{accuracy}}=\frac{S_{\text{correct}}}{S_{\text{total}}}. \tag{10}$$

本实验将支持向量机分类算法(SVM)、BP 神经网络算法(BPNN)、朴素贝叶斯分类算法(NB)、极限学习机算法(ELM)、并行 ELM 算法(PELM)与本文的算法(S-OSELM)算法对气象数据预测结果进行比较.主要比较它们的训练精度、预测精度.其中,SVM 选用高斯核函数,其中核函数参数 γ 和分类器惩罚参数 C 的取值通过十折交叉验证法来确定.对于 BP 神经网络算法、极限学习机算法、PELM 算法和 S-OSELM 算法,隐藏层的激活函数选用 sigmoid 函数.

表 3 给出了支持向量机算法(SVM)、BP 神经网络算法(BPNN)、朴素贝叶斯算法(NB)、极限学习机算法(ELM)、并行的极限学习机算法(PELM)和本文提出的算法(S-OSELM)六种分类算法的实验结果.从表 2 中可以看出,在 6 种分类算法中 S-OSELM 对降雨量预测效果最好,预测精度达到 90.68%.

支持向量机算法(SVM)理论提供了一种避开高维空间的复杂性,利用在线性可分情况下的求解方法直接求解对应的高维空间的决策问题.当核函数已知,可以简化高维空间问题的求解难度,即 SVM 算法适合于小样本预测分类,相比神经网络具有较好的泛化能力.但是对于大规模气象数据预测分类,SVM 算法在求解问题分类时,涉及到求解二次规划的 m 阶矩阵的计算,如此一来将耗费大量机器内存和运算时间,并且对缺失的数据敏感,间接影响了分类精度.

Table 3 The Classification Results of the Six Algorithms

表 3 6 种算法的分类结果

Accuracy/%	SVM	BPNN	NB	ELM	PELM	S-OSELM
Training	78.51	85.46	81.12	84.67	88.36	**91.23**
Prediction	77.80	85.09	80.62	84.18	88.75	**90.68**

BP 神经网络算法(BPNN)具有预测分类精度高、非线性映射能力强等特点,但是该算法收敛速度慢.在大规模气象数据预测分类中,其存在预测能力和训练能力的矛盾的问题.一般情况下,训练能力差时,预测能力也差,并且一定程度上,随着训练能力的提高,预测能力会得到提高.当达到一定值时,随着训练能力的提高,预测能力反而会下降,也即出现所谓"过拟合"现象.出现该现象的原因是网络学习了过多的样本细节导致.

朴素贝叶斯算法(NB)在通过计算概率来进行分类,可以处理多分类问题,同时在小规模数据训练分类表现良好,但是对于大规模气象预测分类方面,存在着一些准确率上的损失,需要计算先验概率,分类决策上存在错误率.

极限学习机算法(ELM)随机产生输入层与隐含层间的连接权值及隐含层神经元的偏置,且在训练过程中无需调整,只需设置隐含层神经元的个数,便可获得唯一的最优解,与传统的 BP 神经网络算法相比,ELM 方法学习速度快、泛化性能好.在大规模气象数据预测分类方面产生了较好的实验结果.

基于 Hadoop 的并行的极限学习机算法(PELM)是采用 MapReduce 的框架对极限学习机进行并行化优化的算法,对于大规模气象数据来说,ELM 算法计算过程中最复杂的部分是大规模矩阵乘法和大规模矩阵转置的运算,根据矩阵乘法每个元素的计算彼此间不存在依赖关系,采用并行计算,把大规模矩阵乘法转换成向量点乘和向量求和 2 个过程.通过合理设定元素(key,$value$)键值对,实现大规模矩阵的转置,该算法具有较好的分类精度和并行效率.

基于 Storm 的在线序列极限学习机算法(S-OSELM)是一种基于在线序列极限学习机的气象预

测算法,该算法利用 Storm 流式处理框架,对多个 ELM 进行并行训练,并引入在线序列,从而每次训练样本只需要迭代处理一个传输过来的样本,而不需要对整个样本重新训练,提高了训练和预测效率.同时对多个预测结果采用随机梯度下降法进行误差权值反馈更新,最后用加权平均对分类结果进行整合,在预测精度得到了较大的提高.

3.3.2 性能评估

本文采用运行时间和加速比来测试 S-OSELM 算法的并行性能.加速比是衡量并行系统或者程序并行化的性能指标,加速比 γ_{speedup} 如式(11)所示:

$$\gamma_{\text{speedup}}=\frac{T_{\text{single}}}{T_{\text{cluster}}}, \tag{11}$$

其中,T_{single} 是单机运行的时间,T_{cluster} 是集群运行的时间.

在性能评估的实验中,本文的 S-OSELM 算法与 PELM 算法在运行时间和加速比上做对比.表 4 给出了实验对比参数和范围.在每组实验中,改变一个参数,同时设置剩余参数为默认值.

Table 4 The Experimental Parameter of Performance Estimation

表 4 性能评估实验参数

Parameter	Range and Default
Number of Hidden Nodes	80,140,**200**,260,320
Number of Records	**1×10^8(1.6 GB)**,2×10^8(3.2 GB),3×10^8(4.8 GB),4×10^8(6.4 GB),5×10^8(8.0 GB)
Number of Working Nodes	1,2,3,4,5,6,7,**8**

首先,在不同隐层节点数的对比实验中,如图 2(a)所示,样本的训练时间随着隐层节点数的增加而增加.ELM 的隐层节点数的增加,会使得 $\boldsymbol{H}$ 隐层输出矩阵的变大,那么 S-OSELM 和 PELM 下的中间结果增大,同时也增加了数据在集群中间的传输时间.S-OSELM 算法是基于 Storm 流式处理框架,PELM 算法是基于 MapReduce 批量处理框架.Storm 是直接在内存中计算和传递数据,而 Hadoop 是使用 HDFS 进行磁盘读写,因此,S-OSELM 在处理时延上要比 PELM 算法快.图 2(b)表示不同隐层节点数下集群运行的加速比.在相同隐层节点数下,S-OSELM 算法的并行系统的加速比优于 PELM 算法的加速比.理论上,并行系统的加速比是线性增加.但在实际应用中,随着隐层节点数增加,节点间的网络传输消耗也不断增加,即理想的线性加速比是非常难以达到的.

另外,在不同学习样本量的对比实验中,由图 3(a)得知,随学习样本量的增加,实验运行时间也相应地增加.S-OSELM 通过分布式消息队列将样本数据分发到各个 Storm 集群节点,数据在内存中快速计算并返回最后的运算结果再进行磁盘存储.然而,PELM 算法中 Map 和 Reduce 的任务都是在磁盘上进行读写.如图 3(a)所示,随着数据量越大,S-OSELM 算法优势越明显.图 3(b)中,相同学习样本量的 S-OSELM 算法下系统加速比优于 PELM 算法下系统的加速比.

最后,图 4(a)(b)分别表示不同工作节点数下的 2 种算法的运行时间和加速比.随着工作节点数的增加,2 种算法运行时间减少,系统并行性能增加,同时也增加集群节点之间传输成本,因此加速比也越小于理论值.由图 4 的实验结果表明,S-OSELM 算法运行效率和系统加速比优于 PELM 算法.

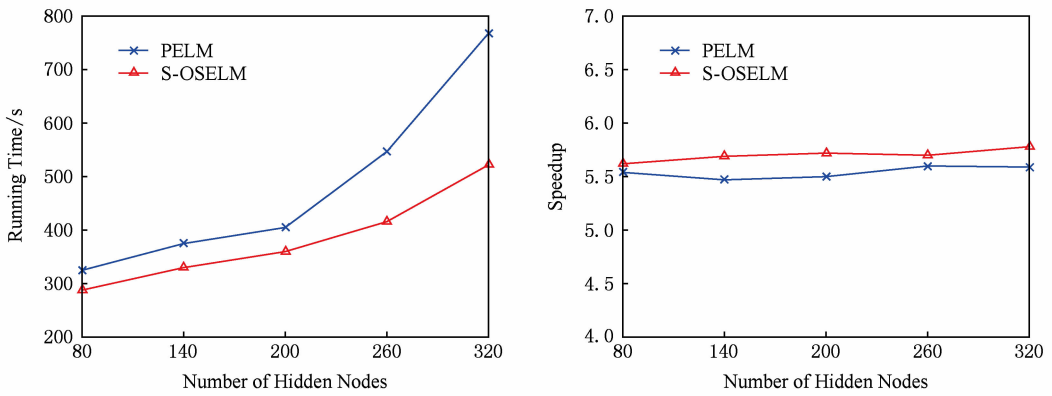

Fig. 2 The experimental results of different number of hidden nodes

图 2 不同隐层节点数的实验结果

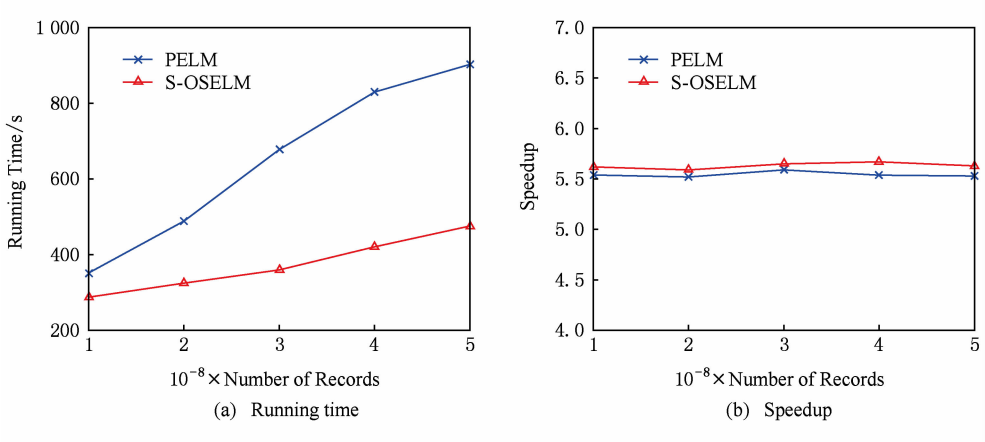

Fig. 3 The experimental results of different volume of learning samples

图 3 不同学习样本量的实验结果

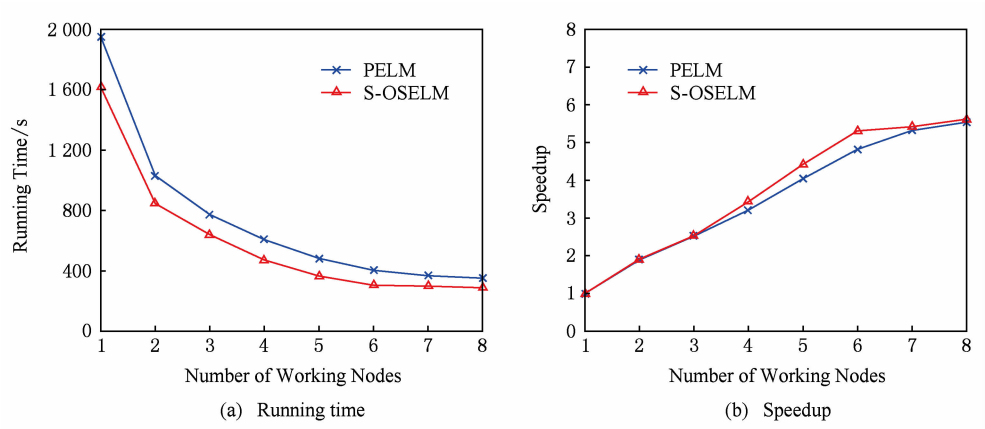

Fig. 4 The experimental results of different number of working nodes

图 4 不同工作节点数的实验结果

综上所述,S-OSELM 算法具有速度快、可扩展性好的特点,它是应对大规模数据在线学习的一个有效工具,具有广泛实际应用前景.

4 总结与展望

本文提出了一种基于在线序列极限学习机和 Storm 云平台结合处理大规模气象数据的方法,该方法能实时在线对气象数据进行分析预测,并具有较高的准确率以及并行性能. 该方法在大多数流数据的实际应用场景具有重要的参考价值,例如在视频流中关键帧的抽取、实时股票走向预测、实时分析用户状态并为用户个性化推荐等等. 今后的工作是将该模型应用智能交通,实时感知道路状态,并分析预测流量情况,以便有效进行指挥和调度.

参 考 文 献

[1] Zhao Qiangli. The research on ensemble pruning and its application in on-line machine learning [D]. Changsha: National University of Defense Technology, 2010 (in Chinese)
(赵强利. 基于选择性集成的在线机器学习关键技术研究[D]. 长沙: 国防科学技术大学, 2010)

[2] Wang Aiping, Wan Guowei, Cheng Zhiquan. Incremental learning extremely random forest classifier for online learning [J]. Journal of Software, 2011, 22(9): 2059-2074 (in Chinese)
(王爱平, 万国伟, 程志全. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报, 2011, 22(9): 2059-2074)

[3] Osowski S, Garanty K. Forecasting of the daily meteorological pollution using wavelets and support vector machine [J]. Engineering Applications of Artificial Intelligence, 2007, 20(6): 745-755

[4] Behranga M A, Assareha E. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data [J]. Solar Energy, 2010, 84(8): 1468-1480

[5] Bilgili M, Sahin B, Yasar A. Application of artificial neural networks for the wind speed prediction of target station using reference stations data [J]. Renewable Energy, 2007, 32(14): 2350-2360

[6] Jin Long, Jin Jian, Yao Cai. A short-term climate prediction model based on a modular fuzzy neural network [J]. Advances in Atmospheric Sciences, 2005, 22(3): 428-435

[7] Zhang H, Su Jiang. Naive bayes for optimal ranking [J]. Journal of Experimental & TheoreticalArtificial Intelligence, 2008, 20(2): 79-93

[8] Jiang Wenrui, Wang Yuying, Hao Xiaoqi, et al. Application of decision tree in temperature prediction [J]. Computer Applications and Software, 2012, 29(8): 141-144) (in Chinese)
(姜文瑞, 王玉英, 郝小琪, 等. 决策树方法在气温预测中的应用[J]. 计算机应用与软件, 2012, 29(8): 141-144)

[9] Wu Jiansheng, Long Jin, Liu Mingzhe. Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm [J]. Neurocomputing, 2015, 148(2): 136-142

[10] KavithaRani B, Govardhan A. Effective features and hybrid classifier for rainfall prediction [J]. International Journal of Computational Intelligence Systems, 2014, 7(5): 937-951

[11] Huang Guangbin, Zhu Qinyu, Siew C K. Extreme learning machine: A new learning scheme of feedforward neural networks [C] //Proc of the IEEE Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2004: 985-990

[12] He Qing, Shang Tianfeng, Zhuang Fuzhen, et al. Parallel extreme learning machine for regression based on MapReduce [J]. Neurocomputing, 2013, 102(2): 52-58

[13] Liang Nanying, Huang Guangbin, Saratchandran P, et al. AS fast and accurate on-line sequential learning algorithm for feedforward networks [J]. IEEE Trans on Neural Networks, 2006, 17(6): 1411-1423

[14] Bottou L. Large-scale machine learning with stochastic gradient descent [C] //Proc of COMPSTAT'2010. Paris: Physica-Verlag HD, 2010: 177-186

[15] Rodgers J, Nicewander W. Thirteen ways to look at the correlation coefficient [J]. The American Statistician, 1988, 42(1): 59-66

[16] Yang Ying, Webb G I. Weighted proportional k-interval discretization for naive bayes classifiers [C] //Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 501-512

Ouyang Jianquan, born in 1973. Professor, PhD supervisor, visiting scholar in the Department of Computer Science, University of Georgia, USA. Member of CCF. His main research interests include machine learning and multimedia analysis and retrieval.

Zhou Yong, born in 1990. Master. Student member of CCF. His main research interests include machine learning and data mining.

Tang Huanrong, born in 1976. Associate professor. Member of CCF. Her main research interests include multi-objective evolutionary computation, information security and video image analysis.