计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):163−168 [doi: 10.15888/j.cnki.csa.006723]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 GDBN 网络的文本情感倾向分类算法①

陈颖熙[1], 廖晓东[1,2,3], 苏例月[1], 陶　状[1]

[1](福建师范大学 光电与信息工程学院, 福州 350007)
[2](福建师范大学 医学光电科学与技术教育部重点实验室 福建省光子技术重点实验室, 福州 350007)
[3](福建师范大学 福建省先进光电传感与智能信息应用工程技术研究中心, 福州 350007)

摘　要: 情感倾向性分类是自然语言处理领域中的热门话题, 它的一个重要应用是挖掘线上评论中的重要信息, 掌握网络舆论走向, 因此本文提出一种基于 GDBN 网络的文本情感倾向性分类算法. 该算法通过引入遗传算法来改进深度置信网络模型中的隐层, 使模型自行对隐单元个数寻优, 取得当前模型的适宜值, 并以此模型进行深层建模与特征提取. 最后通过反向传播网络对提取到的特征进行情感倾向性分类. 在多个文本数据集上进行实验验证, 验证结果表明了本文算法的有效性.

关键词: 情感倾向性分类; 寻优搜索; 遗传算法; 深度置信网络

引用格式: 陈颖熙,廖晓东,苏例月,陶状.基于 GDBN 网络的文本情感倾向分类算法.计算机系统应用,2019,28(1):163–168. http://www.c-s-a.org.cn/1003-3254/6723.html

Text Sentiment Classification Based on GDBN Neural Network

CHEN Ying-Xi[1], LIAO Xiao-Dong[1,2,3], SU Li-Yue[1], TAO Zhuang[1]

[1](College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, China)
[2](Key Laboratory of Optoelectronic Science and Technology for Medicine (Ministry of Education) Cum. Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou 350007, China)
[3](Fujian Provincial Engineering Research Center for Optoelectronic Sensors and Intelligent Information, Fuzhou 350007, China)

Abstract: Text sentiment classification is a hot topic in the field of natural language processing. One of its important applications is to dig out important information from online comments and grasp the trend of public opinion on the Internet. Therefore, this study proposes a method of text sentiment classification based on GDBN neural network. The algorithm improves the hidden layer in the DBN neural network by introducing genetic algorithm, which is of powerful global searching ability, and the algorithm optimizes the number of hidden units and obtains the appropriate value of the current model, then the modeling and feature extraction of this model. Finally, we can classify the extracted features of the BP neural network. By testing multiple data, the results show that the proposed algorithm is effective.

Key words: text sentiment classification; optimization search; genetic algorithm; deep belief networks

1　引言

近年来, 随着互联网信息技术的高速发展, 各种社交平台和电子商务平台的兴起使得门户网站上的评论信息呈指数增长, 用户通过移动网络可以方便、自由的对人或事进行评价与分析, 表达自己的看法、观点以及情感倾向[1]. 面对线上各大平台的大量无规律的评论词语和文本内容, 有必要利用自然语言处理技术建立一种智能高效的文本情感分类模型对文本所表达的

① 基金项目: 省科技厅区域科技重大项目 (2015H4007); 中央引导地方科技发展专项 (2017L3009)
Foundation item: Regional Key Science and Technology Program of Fujian Provincial Science and Technology Bureau (2015H4007); Special Fund of Central Government for Local Science and Technology Development (2017L3009)
收稿时间: 2018-07-02; 修改时间: 2018-07-27; 采用时间: 2018-08-08; csa 在线出版时间: 2018-12-26

情感倾向 (正向、负向、中立) 进行分析判断, 从海量无规律的文本数据中提取重要的信息.

目前, 互联网上的信息大多以短文本的形式存在, 例如淘宝商品评论、搜索引擎的搜索结果、微博、豆瓣、文档文献摘要等. 其中在微博评论中就有明确规定字数必须限制在 140 字以内. 由于短文本具有特征稀疏性、实时性、动态性、交错性、不规则性等特点[2], 传统的文本情感分类方法对其分类的准确率较低, 无法达到理想的结果.

短文本在搜索引擎、论坛信息交流等方面具有重要作用, 因此对短文本情感分类的研究具有一定的实用价值并且得到了广泛的关注. 近些年国内外学者们提出了许多在文本情感倾向性分类的有效的方法, 大致可分为三大类, 即基于规则的方法、基于机器学习的方法和深度学习方法.

基于规则的方法最早是由麻省理工媒体实验室的 Picard 教授提出[3], 它通过将文本中表达情感倾向的词语与已建立的情感词典对比然后进行评估打分, 进而通过计算分数实现文本情感倾向性分类. 由于该方法过分依赖于人工构建的词典, 所以存在一系列缺点, 如词典覆盖面窄、易丢失部分有挖掘价值的文本数据、易受到一词多义的影响等,并且该方法难以捕捉到深层次特征.

基于深度学习的文本情感分类方法是近几年的研究热点, 它广泛应用于计算机视觉领域和音频领域, 近几年才被引用到自然语言处理领域中, 其中深度置信网络 (Deep Belief Networks, DBN)[4]是最经典的深度学习神经网络之一, 它弥补了机器学习方法的局限性, 可以通过网络模型自动地学习提取文本的深层次特征, 但是存在隐层单元个数的选择问题. 深度置信网络的隐层单元个数通常依据经验进行认为选择, 且一旦选定则无法修改. 当隐层单元数超过所需个数时, 多余的隐层单元会增加网络的复杂度, 使得计算量变大从而导致训练时间呈指数增长; 当隐层单元数低于所需个数时, 由于网络无法满足训练所需规模, 从而导致达不到理性的训练结果. 因此, 本文提出了 GDBN 网络 (Genetic Deep Belief Networks), 通过利用遗传算法 (Genetic Algorithm, GA)[5]的全局快速寻优的能力对 DBN 的隐层单元个数自动进行设定. 实验结果表明, 本文所提出的 GDBN 网络在文本情感倾向性分类中能取得较好的分类效果.

2 相关工作

2.1 深度置信网络

深度置信网络 (Deep Belief Networks, DBN) 最初是由 Hinton 等学者于 2006 年提出的一种由多层 RBMs 堆叠和一层反向传播 (Back Propagation) 网络组成的深度学习神经网络[4]. DBN 的主要任务是实现对数据从底层到高层的特征提取, 帮助系统将数据分类成不同的类别. 其网络结构如图 1 所示[6].

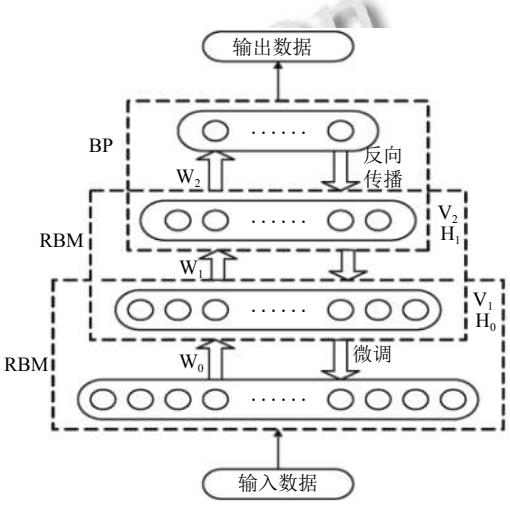

图 1 DBN 网络结构图

DBN 的训练步骤分为两步: 第一步为预训练, 对网络中 RBMs 采用逐层无监督的方法来学习各层参数, 使得每层 RBM 达到最佳特征表示; 第二步为微调, 将 BP 网络输出数据和标准标注信息进行对比, 对从下往上的认知权重 w 和从上往下的生成权重进行反向微调, 以得到更好的生成模型.

近些年来学者们在 DBN 模型上提出了一系列的改进, 使得改进后的模型能够更高效的应用于文本检测. 例如, Mleczko 等[7]在 DBN 模型的基础上引入粗糙集理论 (RDBN), RDBN 模型主要用于识别与分类具有缺失文字的文本信息. Jiang 等[8]提出将采用不同参数优化算法的 Softmax 分类器与 DBN 模型结合, 利用分类器对 DBN 所提取到的文本数据特征进行分类, 该模型能有效地提高分类精度.

2.2 RBM 预训练过程

受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM)[9]是以玻尔兹曼机为基础的改进算法, 它是一种具有快速学习和简单网络结构的无监督训练特征提取器. 其结构模型如图 2 所示.

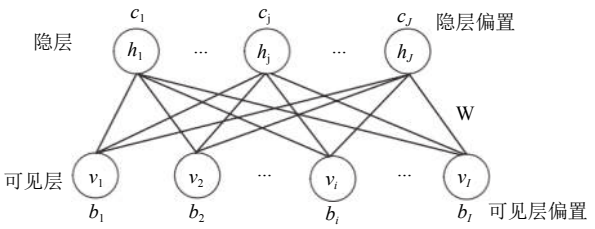

图 2 RBM 结构模型图

图 2 中偏置b_i、c_j和连接权重$w_{i\times j}$构成模型内部参数向量, 表示为$\theta=(w,b,c)$. RBM 是基于能量概率模型[10], 对于每一个(v,h)可得到联合概率分布:

$$P_\theta(v,h)=\frac{1}{Z_\theta}e^{-E_\theta(v,h)} \tag{1}$$

其中, Z_θ为归一化因子, $Z_\theta=\sum_{v,h}e^{-E_\theta(v,h)}$, 其能量函数定义为:

$$\begin{aligned}E_\theta(v,h)&=-b^{\mathrm{T}}v-c^{\mathrm{T}}h-hw\\&=-\sum_{i=1}^{I}b_iv_i-\sum_{j=1}^{J}c_jh_j-\sum_{i=1}^{I}\sum_{j=1}^{J}v_ih_jw_{ij}\end{aligned} \tag{2}$$

式中, v_i为可见层单元, h_j为隐层单元, 具有{0,1}两种状态, 即“激活”和“未激活”状态, 且状态的取值只根据概率统计来计算. RBM 的训练步骤分为两步[11]:

(1) 初始化. 随机初始化θ, $w_{i\times j}$初始化为来自正态分布的随机数, c_j初始化为 0, b_i按式进行初始化;

$$b_i=\log\frac{P_i}{1-P_i} \tag{3}$$

式中, P_i表示训练样本中第 i 个特征处于{1}状态所占的比例.

(2) Gibbs 采样. 通过 Gibbs 采样得到v_i和h_j, 并且重构v_i, 更新得到最佳权值.

训练时, 采用逐层无监督的方法来学习参数. 进而完成 DBN 的预训练过程.

2.3 BP 网络微调过程

RBM 训练中无监督学习方法只能使得该层单元状态达到局部最优, 然而并不能使模型整体效果最优, 因此, 采用 BP 网络[12]对整个网络的参数进行微调. 在 RBM 完成预训练后, 将 RBM 训练好的数据正向传播, 做为 BP 网络的输入, 当输出数据和标准标注信息有误差时, 利用 BP 网络的误差反向传播的特性, 对从下往上的认知权重 w 和从上往下的生成权重以及偏置进行微调, 让整个网络的单元状态达到全局最优, 以得到更好的生成模型.

3 GDBN 情感分类算法

本文提出的基于 GDBN 网络的文本情感倾向性分类算法的主要工作有: 首先通过网络爬虫程序从微博平台上采集实验所需文本数据, 之后对文本数据进行预处理, 然后通过遗传算法来改进深度置信网络模型, 并以此模型进行深层建模与特征提取, 最后通过反向传播网络对提取到的特征进行情感倾向性分类.

3.1 GDBN 理论基础

遗传深度置信网络 (GDBN) 是结合遗传算法 (Genetic Algorithm, GA)[5]和深度置信网络 (Deep Belief Networks, DBN)[4]的学习方法, 它利用遗传算法的全局寻优搜索能力对 DBN 的隐层单元个数进行自动寻优, 结合 DBN 强大的数据特征提取和处理高复杂度的非线性数据的能力, 使网络模型效果更接近于其上限. GA 具有较强全局寻优搜索能力, 然而它最大的缺点就是易出现“早熟”现象, 即容易陷入局部极值, 导致神经网络参数质量不高, 所以在设计 GDBN 算法的遗传操作中, 增大交叉率和变异率. GDBN 算法设计如下:

(1) 编码

在确定可见单元v_i后, 由于隐层各单元之间相互独立, 所以模型性能只与隐层单元个数相关, 因此采用实数编码方式对其个数进行编码.

(2) 适应度函数

GDBN 网络模型中可见层和隐层之间表现为层内无连接, 层间全连接, 隐单元的状态只与可见单元v_i有关, 所以在函数设计时不但要考虑样本的似然程度还要考虑v_i维度对模型训练的影响.

本文采用重构误差[13]的方法来评价样本的似然程度, 所谓重构误差就是通过 Gibbs 采样重构的单元与训练样本原始数据的平方差, 其具体流程如下:

① 误差初始化, 即令$Error=0$;

② 对所有$v^{(t)}$进行采样, $hP(\cdot|v^{(t)})$为隐层采样; $vP(\cdot|h)$为可见层采样, 其采样公式如下:

$$P(h_j=1|v)=\sigma(c_j+\sum_i v_iw_{ij}) \tag{4}$$

$$P(v_i=1|h)=\sigma(b_i+\sum_j h_jw_{ij}) \tag{5}$$

③ 累计当前误差, 即$Error=Error+\|v-v^{(t)}\|$, 并将其返回总误差. 综上, 定义 GDBN 模型中的适应度函数如下:

$$Fit(k)=1-\frac{Error}{I+S} \tag{6}$$

式中, I 为可见单元个数, S 为样本维度, 根据适应度的大小对个体进行选择, 当适应度值越大时, 则个体越好, 即该个体对应的 GDBN 模型似然度最高.

(3) 遗传操作

在遗传算法 (GA) 改进网络模型后, 进一步优化精调真个模型, 其算法流程如图 3 所示.

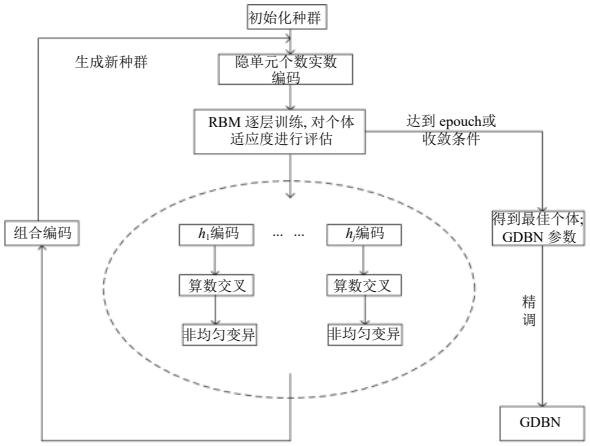

图 3 算法流程

3.2 框架实现

(1) 文本预处理: 将通过爬虫得到的数据内容进行处理, 将其中涉及到个人隐私、url 链接或敏感信息的内容删除.

(2) 分词、去停用词: 由于中文评论无法像英文评论一样直接通过空格来分隔单词, 所以本文采用 Jieba 工具, 进行中文分词, 并去掉停用词, 如“的”、“和”等一些出现频率高但无情感意义的词, 为特征提取提供较为准确的基元.

(3) 特征提取: 通过 GDBN 网络模型进行深层建模与特征提取.

(4) 情感分类: BP 网络对提取到的特征进行情感倾向性分类.

4 实验验证及结果分析

4.1 实验环境与数据

本文具体实验环境如表 1 所示.

为了验证本文所提出的分类算法的有效性, 本文基于三个中文文本数据集进行实验验证. (1) 使用中科院谭松波教授的酒店评论语料 (D1), 该语料采集于携程网, 规模为 10 000 篇, 被整理成 4 个子集, 1、ChnSentiCorp-Htl-ba-2000: 平衡语料, 正负类各 2k; 2、ChnSentiCorp-Htl-ba-4000:平衡语料, 正负类各 4k; 3、ChnSentiCorp-Htl-ba-6000: 平衡语料, 正负类各 3k; 4、ChnSentiCorp-Htl-ba-10000: 非平衡语料, 其中正类为 7k. (2) 使用 COAE2014 微博观点数据集, 在该数据集中随机抽取 30 000 条作为实验数据集, 对其中部分训练数据进行不同情感倾向的人工标注, 主要情感有开心、愤怒、厌恶、低落四个类别. (3) 通过网络爬虫程序从微博平台上采集的 50 000 条微博数据 (D3), 其中标注的积极微博有 25 000 条, 消极微博有 20 000 条, 中性微博有 5000 条. 考虑到其中部分能容可能含有用户隐私, 删除了数据集中的 url 链接等信息.

表 1 实验环境

项目	配置
操作系统	Windows7
CPU	Inter(R)Core(TM) i7-7700 @3.60 GHz
内存	8 GB
硬盘	1 TB
开发平台	JetBrains PyCharm Community Edition
开发语言	Python3.6

4.2 实验设计

实验方案总体过程如图 4 所示.

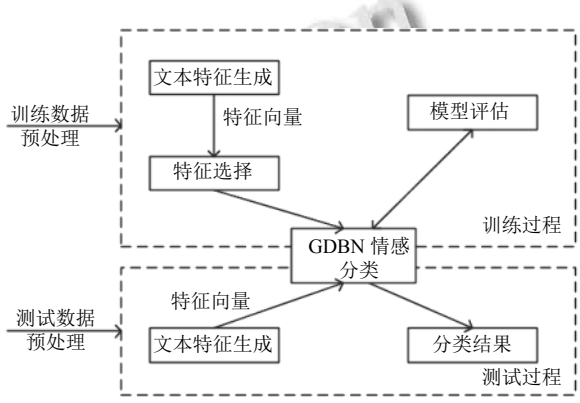

图 4 实验方案

首先对训练数据进行预处理, 生成文本特征向量, 然后将训练后的 GDBN 情感分类模型用于测试数据分类并检验分类效果.

4.3 性能评估

本文采用准确率 P(precision)、召回率 R(recall)、F_1 值($F_1-Score$)作为评价指标, 通过 ROC 曲线图来评估文本情感倾向性分类模型的性能.

precision主要体现模型对负样本的区分能力, 通常用 P 表示, 设 TP 为分类正确的文本数, N 为样本总数, 其计算公式如下:

$$P = \frac{TP}{N} \tag{7}$$

recall主要体现模型对正样本的识别能力, 通常用 R 表示, 设 N_+为某一类的样本总数, 其计算公式如下:

$$R = \frac{TP}{N_+} \tag{8}$$

F_1 值为两者的综合, 当 F_1 值越高时证明模型越好. 其计算方法如下:

$$F_1 = \frac{2*P*R}{P+R} \tag{9}$$

4.4 实验结果与分析

为了验证本文提出的基于 GDBN 网络的文本情感倾向性分类算法的有效性, 将 SVM、DBN 与本文算法进行对比, 其对比实验结果如表 2 所示. 且作出 GDBN 算法用于三个中文文本数据集 (D1、D2、D3) 的迭代曲线图如图 5 所示, 其结果表明, GDBN 算法较于 DBN 和 SVM 算法更能有效的对文本情感倾向进行分类.

表 2 实验结果对比 (单位: %)

算法	D1			D2			D3		
	P	R	F_1	P	R	F_1	P	R	F_1
SVM	86.3	68.8	76.6	83.1	67.3	74.4	82.5	65.8	73.2
DBN	80.4	79.7	80.0	81.6	71.3	76.1	86.3	67.8	75.9
GDBN	84.3	82.1	83.3	84.7	78.2	81.9	87.1	79.7	83.2

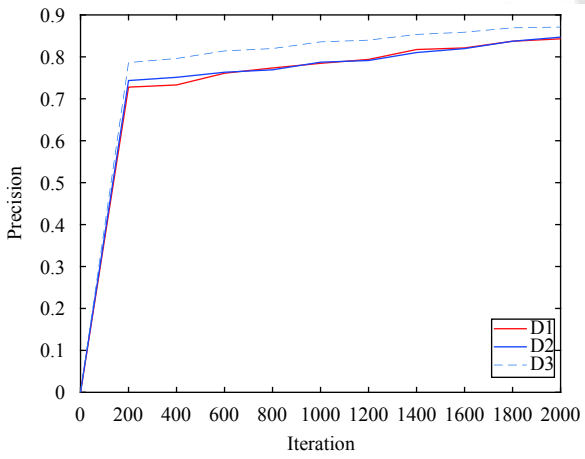

图 5 GDBN 迭代曲线图

本文对三种分类算法做 ROC 曲线进行模型评估, 如图 6 所示. ROC 曲线下面积越大代表模型性能越好, 由图 6 可知基于 GDBN 算法的文本情感分类模型具有更高的分类性能.

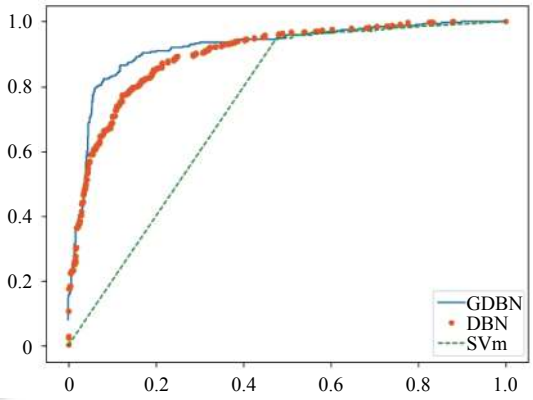

图 6 ROC 曲线图

5 结语

为了更好的解决中文文本情感分类问题, 本文基于深度学习算法构建了一个 GDBN 网络模型, 针对 DBN 网络人工进行隐层单元个数选择从而导致模型性能存在极大不确定性的问题, 引入具有强大全局寻优搜索能力的遗传算法, 根据实验输入数据自行对隐单元个数寻优, 取得当前模型的适宜值. 经实验验证可得, 本文所提方法在分类准确性和降低模型复杂性上均有提升, 能取得良好的效果, 但仍存在不足. 在今后的工作中, 将继续改进本文算法, 比如在对提取到的特征进行分类时候, 针对 BP 网络存在的网络“震荡”等问题, 采用 XGBoost 算法来进行分类, 进一步提高模型情感分类的精度.

参考文献

1 Somasundaran S, Wilson T, Wiebe J, *et al*. QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. Proceedings of the International Conference on Weblogs and Social Media. Boulder, CO, USA. 2007.

2 胡雯雯, 高俊波, 施志伟, 等. 基于词性特征的特征权重计算方法. 计算机系统应用, 2018, 27(1): 92–97. [doi: 10.15888/j.cnki.csa.006127]

3 Picard RW. Affective Computing. Cambridge: MIT Press, 1997.

4 Sarikaya R, Hinton GE, Deoras A. Application of Deep Belief Networks for natural language understanding.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(4): 778–784. [doi: 10.1109/TASLP.2014.2303296]

5 Uysal AK, Gunal S. Text classification using genetic algorithm oriented latent semantic features. Expert Systems with Applications, 2014, 41(13): 5938–5947. [doi: 10.1016/j.eswa.2014.03.041]

6 张翔, 石力, 尚勃, 等. 深度置信网络的 Spark 并行化在微博情感分类中的应用研究. 计算机应用与软件, 2018, 35(2): 48–53. [doi: 10.3969/j.issn.1000-386x.2018.02.008]

7 Mleczko WK, Kapuściński T, Nowicki RK. Rough deep belief network—application to incomplete handwritten digits pattern classification. In: Dregvaite G, Damasevicius R, eds. Information and Software Technologies. Cham: Springer, 2015. 400–411.

8 Jiang MY, Liang YC, Feng XY, *et al*. Text classification based on deep belief network and Softmax regression. Neural Computing and Applications, 2018, 29(1): 61–70. [doi: 10.1007/s00521-016-2401-x]

9 Chen CLP, Zhang CY, Chen L, *et al*. Fuzzy restricted Boltzmann machine for the enhancement of deep learning. IEEE Transactions on Fuzzy Systems, 2015, 23(6): 2163–2173. [doi: 10.1109/TFUZZ.2015.2406889]

10 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]

11 Alexandre E, Cuadra L, Nieto-Borge JC, *et al*. A hybrid genetic algorithm—extreme learning machine approach for accurate significant wave height reconstruction. Ocean Modelling, 2015, 92: 115–123. [doi: 10.1016/j.ocemod.2015.06.010]

12 Li J, Cheng JH, Shi JY, *et al*. Brief introduction of back propagation (BP) neural network algorithm and its improvement. In: Jin D, Lin S, eds. Advances in Computer Science and Information Engineering, Volume 2. Berlin Heidelberg. Springer. 2012. 553–558.

13 Lyu C, Lu YN, Ji DH, *et al*. Deep learning for textual entailment recognition. Proceedings of the 27th IEEE International Conference on TOOLS with Artificial Intelligence. Vietri sul Mare, Italy. 2016. 154–161.