

外类入侵度初始化参数的极限学习机

孔双双^{1,2}, 王开军^{1,2}, 林 崧^{1,2}

(1. 福建师范大学数学与信息学院, 福建 福州 350117)

(2. 福建师范大学数字福建环境监测物联网实验室, 福建 福州 350117)

[摘要] 针对经典极限学习机中输入权值随机初始化容易导致输出稳定性不够好进而影响分类性能的问题, 提出外类入侵度初始化参数的方法, 对极限学习机随机初始化的输入权值用样本的属性特征信息进行修正. 该方法对包含两个类别样本的数据集, 将其中一个类作为本类, 另一个类作为外类. 对于每个特征, 统计本类和外类样本重叠的区域占本类取值范围的比例, 也统计重叠区域中外类样本数目占重叠区域总样本数目的比例. 然后依据这两种占比值计算每个特征的外类入侵度. 再根据入侵度大小调整极限学习机模型中隐含层的输入权值. 在 10 个 UCI 数据集上进行的分类实验结果表明, 新方法的准确率比经典极限学习机提高了 1%~23%, 且泛化性能更稳定; 与另外两方法相比, 新方法的准确率稍高.

[关键词] 极限学习机, 重叠区域, 类入侵度, 权值修正

[中图分类号] TP183 **[文献标志码]** A **[文章编号]** 1672-1292(2019)03-0053-06

Extreme Learning Machine with Initialized Parameters Based on External Class Invasion Degree

Kong Shuangshuang^{1,2}, Wang Kaijun^{1,2}, Lin Song^{1,2}

(1. School of Mathematics and Information, Fujian Normal University, Fuzhou 350117, China)

(2. Digital Fujian Environmental Monitoring Internet of Things Laboratory, Fujian Normal University, Fuzhou 350117, China)

Abstract: Aiming at the problem that the random initialization of the input weights in the classical extreme learning machine which is easy to cause the output stability is not good enough and then affects the classification performance, an extreme learning machine with initialized parameters based on external class invasion degree is proposed by adjusting the weights initialized randomly of the extreme learning machine through the attribute characteristic information of samples. For the data set containing samples of two different categories, the method takes one class as this class and the other class as outer class. For each feature, the proportion of the overlapping areas between the samples of this class and outer class in the value range of this class is calculated, and the proportion of outer class samples in the total number of samples in the overlapping areas is also calculated. Then the invasion degree of each feature according to the two proportions is calculated. Finally, the input weights of the hidden layer of the extreme learning machine model is adjusted according to the degree of invasion. Experimental results on ten UCI data sets show that the accuracy of the new method is 1%~23% higher than that of the classical extreme learning machine, and the generalization performance is more stable; and that compared with other two methods, the accuracy of the new method is slightly higher.

Key words: extreme learning machine, overlapping area, external class invasion degree, weight correction

极限学习机(extreme learning machine, ELM)^[1-2]是一种单隐层前向神经网络(single-hidden layer feed-forward network, SLFN)的训练算法. 不同于传统的训练算法(如反向传播算法等), ELM 算法对输入层的权值和偏置进行随机赋值, 然后用求 Moore-Penrose 广义逆矩阵的方法直接解出隐含层到输出层的权值. ELM 的优势有: 需要手动设置的参数只有隐含层结点个数, 算法执行过程中不需要人工调整参数. 避免了传统训练算法反复迭代的过程, 快速收敛, 极大地减少了训练时间. 所得解是唯一最优解, 保证了网

收稿日期: 2019-07-05.

基金项目: 福建省自然科学基金(2018J01778)、国家自然科学基金(61772134)、博士后基金(2016M600494).

通讯联系人: 王开军, 博士, 副教授, 研究方向: 机器学习、数据挖掘等. E-mail: wkjwang@qq.com

络的泛化性能. 目前 ELM 已经广泛应用到各种回归和分类问题^[3-4].

尽管 Huang 等学者已证明了在整个训练阶段,随机输入权值不变的 SLFN 具有较好的逼近能力^[5],但输入权值问题也引起了众多研究者的关注^[6]. 文献[7]认为随机赋值的输入权重会降低算法的有效性,同时也会导致 ELM 的输出不太稳定进而影响算法的性能. 为了解决这个问题,一些学者提出了改进方法来提高 ELM 泛化性能. 文献[8]中提供了基于投票机制的极限学习机 V-ELM,通过训练多个独立且有相同结构的极限学习机并使用投票方式整合各 ELM 的结构方法来避免随机初始化的隐含层输入权值和偏置对分类结果造成的不稳定的影响. 但为使其正常工作需要一个足够大的独立训练样本数,同时随着训练样本数的增加其消耗时间也会成倍增加. 文献[9]中提出了一种通过使用限制玻尔兹曼机来确定输入权重和偏置的方法 RBM-ELM,实验结果表明其算法性能较为稳定但时间消耗较大. 文献[10]提出的 RO-ELM 算法中使用蒙特卡洛算法和随机正交投影的方法生成 ELM 输入权值的样本结构保持能力,并利用这些投影方法分析了 ELM 的学习性能;但其主要研究的是输入权重全局样本结构保持性能,而未考虑到局部样本结构保持性能. 文献[11]中提出在计算输出层权值之前通过调整输入层的权值和偏置,使得隐含层的输出矩阵达到满秩条件,进而提高网络的分类准确性和鲁棒性;但其仅在数据集较小的情况下分类和回归性能优于传统 ELM. 可以看出,以上方法比较复杂且计算耗时多或者仅适用于小数据集;尚未由依据数据集中样本的属性特征来设置 ELM 输入权值的方法.

本文提出了一种用样本特征修正权值的外类入侵度初始化参数的极限学习机(extreme learning machine with initialized parameters based on external class invasion degree, EID-ELM). 这种方法通过计算外类入侵度来判断各特征对于区分不同种类的能力,并用于设置输入层到隐含层的权值,以提高 ELM 的分类性能,改善随机初始化参数导致的 ELM 泛化性能的不稳定性.

1 背景知识

ELM 是一种利用解析解而非标准梯度下降算法来确定单隐层前向 SLFN 输出权值的有效算法. 通过随机初始化输入权值矩阵,并使用广义逆运算求解输出权值矩阵,克服了利用梯度下降算法不断迭代更新权重的运行速度慢的问题.

设有 n 个输入训练样本 $(x_i, t_i), i=1, 2, \dots, N$, 其中 $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T, \mathbf{t}_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T$. 含有 L 个隐含层神经元以及激活函数为 $g(x)$ 的 SLFN 的输出为:

$$\sum_{i=1}^L \beta_i g(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) = o_j, j=1, \dots, N. \quad (1)$$

式中, $\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$ 为输入层和隐含层的连接矩阵, $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$ 为隐含层和输出层的连接矩阵, b_i 为第 i 个神经元的偏置.

这样的 SLFN 可以无限逼近这 N 个样本: $\sum_{j=1}^N \|o_j - t_j\| = 0$, 于是可将式(1)简写为矩阵形式 $\mathbf{H}\boldsymbol{\beta} = \mathbf{O} = \mathbf{T}$. 通过广义逆运算可计算出输出层权值矩阵 $\boldsymbol{\beta}^{[12]}$ (其中 \mathbf{H}^\dagger 为 \mathbf{H} 的广义逆矩阵):

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \rightarrow \boldsymbol{\beta} = \mathbf{H}^\dagger \mathbf{T}. \quad (2)$$

2 外类入侵度初始化参数的极限学习机

2.1 方法的相关定义

设数据集由 n 个 m 维(特征)样本组成,一个特征上的二类数据集的值若有重叠,则该特征区分二类样本的能力变弱. 本文将设计体现二类数据集在一个特征上值域重叠程度的入侵度指标,来描述第 2 类(也称为外类)样本混入(也称为入侵)第 1 类(也称为本类)样本值域的程度.

定义 1 (二类数据集) 设二类数据集 D 包含 A 类样本集 X 和 B 类样本集 Y , A 类样本集 $X = \{x_1, x_2, \dots, x_{n_1}\}$ 具有 n_1 个样本,每个样本 $\mathbf{x}_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$ 具有 m 个分量(特征),特征集为 $\mathbf{F} = \{f_1, f_2, \dots, f_m\}$; B 类样本集 $Y = \{y_1, y_2, \dots, y_{n_2}\}$ 具有 n_2 个样本,每个样本 $\mathbf{y}_i = \{y_{i1}, y_{i2}, \dots, y_{im}\}$ 具有 m 个分量(特征).

定义 2 (入侵深度比) 在数据集 D 的任一个特征上,外类和本类样本取值范围的重叠区域占本类总取值范围长度的比值,称为外类在该特征上对本类的入侵深度比.

定义 3 (入侵个数比)在数据集 D 的任一个特征上,外类和本类样本取值范围的重叠区域中外类样本的个数与该重叠区域中两类样本总个数的比值,称为外类在该特征上对本类的入侵个数比。

定义 4 (外类入侵度 R)在数据集 D 的任一个特征上,外类对本类的入侵深度比和入侵个数比按各占一半的比例相加得到外类对本类的入侵度。

外类入侵度可以描述一个特征区分本类和外类样本的能力。外类对本类的入侵度取值范围在 $[0, 1]$ 内,值越大说明特征的二类区分能力越差,值为 0 说明在该特征上二类样本完全分开,没有重叠。

2.2 极限学习机的初始化参数

经典的极限学习机输入层到隐含层的参数是随机生成的,因而会导致使用不同的初始化参数训练出的极限学习机泛化性能有差异;同时随机初始化的参数也没有考虑到不同特征对于区分不同类别样本的能力。因此将外类入侵度用于比例缩放极限学习机模型中隐含层输入权值的初始化方法,以保持网络的稳定性。

对包含二类样本的数值型数据集 $D_{(n_1+n_2) \times m}$,假设 A 类样本类中心小于 B 类样本,若要求出特征 f_j 上 A 类对 B 类的外入侵度 $R(f_j, A, B)$,首先应判断在特征 f_j 上两类之间的重叠区域,其可通过两类样本取值范围的交叉区域求得。 A 、 B 两类样本在特征 f_j 上的取值如图 1 所示。

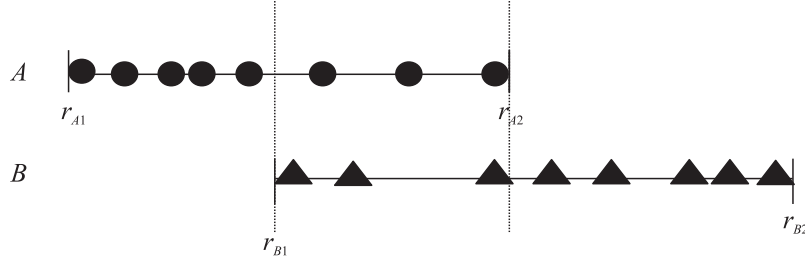


图 1 A 、 B 两类样本在特征 f_j 上的取值区间及重叠区域

Fig. 1 The value range and overlap area of A and B samples on feature f_j

其中在特征 f_j 上 A 类样本的取值范围为 $[r_{A1}, r_{A2}]$, B 类样本的取值范围为 $[r_{B1}, r_{B2}]$, 且 $r_{A2} > r_{B1}$, 则重叠区域为 $S = [r_{B1}, r_{A2}]$, 其长度为 $|S| = r_{A2} - r_{B1}$ 。

特征 f_j 上的入侵深度比,可根据不同类别样本取值范围的重叠区域占各自总取值范围长度的比值计算求得。设 $R_{\text{depth}}(f_j, A, B)$ 表示在特征 f_j 上 A 类对 B 类的入侵深度比,则有:

$$R_{\text{depth}}(f_j, A, B) = \frac{|S|}{r_{B2} - r_{B1}}. \quad (3)$$

特征 f_j 上的入侵个数比,可根据不同类别样本取值范围的重叠区域中入侵的样本数占该区域内总样本数的比值计算求得,则在特征 f_j 上 A 类对 B 类的入侵个数比 $R_{\text{num}}(f_j, A, B)$ 有:

$$R_{\text{num}}(f_j, A, B) = \frac{|I(A, S)|}{|I(A, S)| + |I(B, S)|}. \quad (4)$$

式中, $|I(A, S)|$ 、 $|I(B, S)|$ 分别代表 A 、 B 两类样本在重叠区域 S 内的样本个数。据此,可计算出在特征 f_j 上 A 类对 B 类的类入侵度 $R(f_j, A, B)$:

$$R(f_j, A, B) = \frac{R_{\text{depth}}(f_j, A, B) + R_{\text{num}}(f_j, A, B)}{2}. \quad (5)$$

根据上述方法计算各特征的入侵度后,将传统的 ELM 隐含层中随机初始化的输入权值根据入侵度值的大小进行调整,并将其作为新的输入权重加入到极限学习机中进行分类任务,以此可保证经入侵度赋值的隐含层神经元的分类能力更强,进而使得网络的性能更好。

本文设计的方法是在计算输出层神经元的输出 $g(\cdot)$ 前使用类入侵度将输入权值 W 进行调整,调整方法如下:

输入样本 $D_{(n_1+n_2) \times m}$ 通过类入侵度计算方法可得到类入侵度 $R_{m \times 1}$; 对于输入层权值 W , 先将其随机初始化,再将隐含层输入权值用 $R_{m \times 1}$ 进行缩放:

$$W_{pi} = W_{pi} \times (1 - R_p)^2, \quad p = 1, 2, \dots, m, i = 1, 2, \dots, L. \quad (6)$$

根据 Bartlett 理论^[13]可知,对于需要训练误差越小的前馈神经网络来说,权重越小,网络越稳定.为此,对隐含层输入权值再乘以小于 1 的 $(1-R)^2$,有利于保证网络的稳定性.

3 实验及结果分析

本节首先说明实验数据集和性能指标,然后给出实验结果,最后对算法的有效性进行对比分析.

3.1 实验数据和性能指标

本实验使用 python3.6 实现算法编码,选用 UCI 数据集^[14]和 2012 年 2 月的 Lending Club 数据集对算法进行测试.数据集的相关信息如表 1 所示.

表 1 实验数据集信息
Table 1 Experimental data sets information

数据集	样本数	特征数	类标签数	各类样本数量比
zoo	101	16	7	41:20:5:13:4:8:10
StatlogHeart	270	13	2	150:120
ionosphere	351	34	2	126:225
ForestTypes	523	27	4	159:86:83:195
australian	690	14	2	383:307
Lending Club	710	20	3	439:193:78
Germman	1 000	20	2	700:300
mfeat	2 000	649	10	200:200:200:200:200:200:200:200:200:200
CTG	2 126	20	3	165:1115:846
segment	2 310	19	7	330:330:330:330:330:330:330
waveform	5 000	40	3	1647:1696:1657

其中,对于 Lending Club 数据集,保留 Loan Purpose、Amount Requested、Interest Rate、APR、Amount Funded、Number of Lenders、Home Ownership、Monthly Income、Debt-To-Income Ratio、FICO Range、Open CREDIT Lines、Total CREDIT Lines、Revolving CREDIT Balance、Revolving Line Utilization、Inquiries in the Last 6 Months、Accounts Now Delinquent、Delinquent Amount、Delinquencies(Last 2 yrs)、Public Records On File、Employment Length 共 20 个特征,并以 CREDIT Rating 进行归类,分别将 A、B 等级归为第 1 类,C、D 等级归为第 2 类,E、F、G 等级归为第 3 类.

本小节选用经典 ELM、RBM-ELM、RO-ELM 以及本文的 EID-ELM 4 种方法分别对分类问题进行性能比较,其中经典 ELM 算法的代码来源于文献[1]作者^[15],RBM-ELM 算法和 RO-ELM 算法的代码来源于文献[9]作者^[16].

本实验采用准确率和 F1-score 作为评价指标^[17].准确率为正确分类的样本数占总样本数的比值,准确率越大,说明分类性能越好,定义准确率为

$$\frac{1}{n} \sum_{i=1}^n I(g(x_i)=y_i), \quad (7)$$

F1-score 为精确率和召回率的调和平均,其中精确率是精确性的度量(即预测为正类的样本实际也为正类的百分比),召回率是完全性的度量(即正类样本被预测为正类的百分比).本实验中精确率指某类样本中正确分类的样本数占所有被预测为该类的样本的比重,召回率指某类样本中正确分类的样本数占该类样本的比重.

$$F = \frac{2RP}{R+P} \quad (8)$$

$$P = \frac{N_{A-\text{correct}}}{N_{A-\text{predict}}} \quad (9)$$

$$R = \frac{N_{A-\text{correct}}}{N_A} \quad (10)$$

上式中 F 为 R 和 P 的调和平均,其中 F 为 F1-score, R 为召回率, P 为精确率, N 为样本总数, N_A 表示 A 类样本的数量, $N_{A-\text{correct}}$ 表示 A 类样本中被正确分类的数量, $N_{A-\text{predict}}$ 表示所有样本中被预测为 A 类样本的数量.

3.2 实验性能对比

实验时首先对数据进行归一化处理,使样本在各特征下的取值分布在 $[0,1]$ 内.对于 Lending Club 数据集,分别对 Loan Purpose、Home Ownership 两个特征的数据用正整数序数 1、2、3 等替换;FICO Range 评分区间取其平均数;在 Employment Length 中大于 10 年时取 10,小于 10 年时取 0.5,n/a 取 0;并对百分数的数据转为小数制并保留四位小数.

参照文献[1]中实验设置方式,本文中的实验都随机抽取数据集中 25% 和 75% 的样本分别作为训练集和测试集,且对每个数据集分别进行 50 次独立实验并取其平均值作为实验结果.实验中经典 ELM、RBM-ELM、RO-ELM 和 EID-ELM 算法的隐含层神经元个数均设置为 20 个,且隐含层激活函数设置均为 sigmoid 函数.

RBM-ELM 算法中根据文献[9]中作者的设置方式其参数 maxIter、lr、wc、iMom、fMom、cdIter、batchSize、freqPrint 分别设置为 200、0.001、0.0002、0.5、0.9、1、100 和 10. EID-ELM 方法按式(3)、(4)设定隐含层的输入权值缩小相应随机初始化权值的 $(1-R)^2$ 倍.

各种方法给出测试集上的实验结果如表 2、表 3 所示:

表 2 各种方法在各数据集上的准确率

Table 2 Prediction accuracy of several algorithms on data sets

	经典 ELM	RBM-ELM	RO-ELM	EID-ELM
zoo	0.732 8	0.771 5	0.773 3	0.796 8
StatlogHeart	0.779 7	0.789 8	0.785 6	0.796 9
ionosphere	0.821 7	0.808 4	0.816 4	0.828 4
ForestTypes	0.819 7	0.816 8	0.821 2	0.850 4
australian	0.848 3	0.853 8	0.858 8	0.858 1
Lending Club	0.795 0	0.836 5	0.787 9	0.856 1
Germman	0.729 4	0.739 7	0.718 8	0.740 5
mfeat	0.625 2	0.745 0	0.743 8	0.858 9
CTG	0.783 5	0.786 6	0.728 3	0.784 1
segment	0.859 0	0.855 4	0.869 7	0.877 4
waveform	0.815 4	0.836 9	0.830 0	0.833 2

表 3 各种方法在各数据集上的 F1-score

Table 3 F1-score of several algorithms on data sets

	经典 ELM	RBM-ELM	RO-ELM	EID-ELM
zoo	0.727 5	0.737 3	0.735 2	0.793 3
StatlogHeart	0.776 2	0.781 4	0.781 4	0.793 3
ionosphere	0.818 9	0.800 5	0.818 2	0.831 7
ForestTypes	0.811 5	0.818 3	0.819 7	0.815 5
australian	0.850 6	0.856 2	0.856 3	0.856 2
Lending Club	0.799 8	0.837 1	0.788 6	0.862 6
Germman	0.728 1	0.738 8	0.717 0	0.739 5
mfeat	0.619 6	0.750 1	0.739 3	0.853 7
CTG	0.782 0	0.786 8	0.733 0	0.780 9
segment	0.860 6	0.853 5	0.870 3	0.876 8
waveform	0.815 4	0.836 9	0.830 0	0.833 2

对于 ionosphere、StatlogHeart、australian、CTG 和 waveform 数据集,由于计算出的类入侵度大小相差较小,说明不同特征的二类区分性能相差不大,因此调整后的权值与调整前仅在量级上不同,而权值间的关系没有太大改变,所以新方法在训练集和测试集上的准确率仅略高于其他方法或与其他方法相当.对于 mfeat 数据集,其计算出的入侵度总体较小,而具有较大入侵度的特征通过权值调整弱化了其分类作用,使得分类性能好的特征作用增强,因此新方法的准确率高出其他算法且更优.对于 ForestTypes、Lending Club、Germman、segment 和 zoo 数据集,其计算出的类入侵度值均较大,即有较多区分能力差的特征存在,调整后的权值也弱化了其区分能力差的特征的分类作用,因此准确率高出其他方法且效果更优.通过以上结果分析表明,当计算出的类入侵度有部分为较大或较小值时,通过入侵度调整输入层权值的 EID-ELM 方法整体表现均优于经典 ELM,且与对比方法 RBM-ELM 和 RO-ELM 相当或稍好.

新方法通过计算出的外类入侵度来调整隐含层的输入权值,即具有较好区分能力特征的分类作用加

强,而具有较差区分能力特征的分类作用减弱;但当数据集本身各特征的区分能力相当时,其效果会与经典 ELM 相当.

4 结语

本文设计的类入侵度,为数据集的每个特征用于区分不同类别的样本提供了依据,可用于分类任务中的特征选择工作和改进分类模型;并据此设计了将类入侵度用于改进极限学习机模型中隐含层输入权值的初始化方法.实验结果表明,新方法在分类方面的准确率总体比经典 ELM 方法更优,与对比方法 RBM-ELM 和 RO-ELM 相当或稍好,解决了经典极限学习机中输入权值随机初始化容易导致输出性能不稳定的问题.

[参考文献](References)

- [1] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1): 489-501.
- [2] HUANG G B, ZHOU H M, DING X J, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE transactions on systems, man and cybernetics, 2011, 42(2): 513-529, 2012.
- [3] LIU B, YAN S, YOU H L, et al. Road surface temperature prediction based on gradient extreme learning machine boosting[J]. Computers in industry, 2018, 99: 294-302.
- [4] 裘日辉, 刘康玲, 梁军. 基于极限学习机的分类算法及在故障识别中的应用[J]. 浙江大学学报(工学版), 2016, 50(10): 1965-1972.
QIU R H, LIU K L, LIANG J. Classification algorithm based on extreme learning machine and its application in fault identification of Tennessee Eastman process[J]. Engineering of journal of Zhejiang university, 2016, 50(10): 1965-1972. (in Chinese)
- [5] HUANG G B, LEI C, SIEW C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE transactions on neural networks, 2006, 17(4): 879-892.
- [6] WANG R, SAN K, WANG X Z. A study on random weights between input and hidden layers in extreme learning machine[J]. Soft computing, 2012, 16(9): 1465-1475.
- [7] ZHANG X F, LIN X L, ASHFAQ. Impact of different random initializations on generalization performance of extreme learning machine[J]. Journal of computers, 2018, 13(7): 805-821.
- [8] CAO J W, LIN Z P, HUANG G B, et al. Voting based extreme learning machine[J]. Information sciences, 2012, 185(1): 66-77.
- [9] PACHECO A G C, KROHLING R A, SILVA C A S, et al. Restricted Boltzmann machine to determine the input weights for extreme learning machines[J]. Expert system with application, 2018, 96: 77-85.
- [10] WANG W H, LIU X Y. The selection of input weights of extreme learning machine: a sample structure preserving point of view[J]. Neurocomputing, 2017, 261: 28-36.
- [11] WANG Y G, CAO F L, YUAN Y B. A study on effectiveness of extreme learning machine[J]. Neurocomputing, 2011, 74(16): 1483-1490.
- [12] RAKHA, MEDHAT A. On the Moore-Penrose generalized inverse matrix[J]. Applied mathematics and computation, 2004, 158(1): 185-200.
- [13] BARTLETT, PETER L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network[J]. IEEE transactions on information theory, 1998, 44(2): 525-536.
- [14] BLAKE C L, MERZ C J. UCI Repository of machine learning databases[EB/OL]. [2019-3-20]: <https://archive.ics.uci.edu/ml/index.php>.
- [15] HUANG G B. ELM source codes[EB/OL]. [2019-3-20] <http://www.ntu.edu.sg/home/egbhuang/>.
- [16] PACHECO A G C. RBM-ELM source codes[EB/OL]. [2019-3-20] <http://github.com/paaatcha/RBM-ELM>.
- [17] HAN J W, MICHELIN K, JIAN P, 等. 数据挖掘: 概念与技术[M]. 北京: 机械工业出版社, 2012.
HAN J W, MICHELIN K, JIAN P, et al. Data mining: concepts and techniques[M]. Beijing: China Machine Press, 2012. (in Chinese)

[责任编辑: 陈 庆]