IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 16, 2018, accepted August 28, 2018, date of publication September 6, 2018, date of current version October 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2868998

A Hierarchical Extreme Learning Machine Algorithm for Advertisement Click-Through Rate Prediction

SEN ZHANG, ZHENG LIU, AND WENDONG XIAO, (Senior Member, IEEE)

Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Sen Zhang (zhangsen@ustb.edu.cn)

This work was supported in part by the Beijing Natural Science Foundation under Grant 4182039 and in part by the National Natural Science Foundation of China under Grants 61673056 and 61673055.

ABSTRACT Click-through rate (CTR) prediction plays a predominant role in the online advertisements. CTR prediction is a problem of binary classification with imbalanced data. Many existing approaches for imbalance learning only focus on over-sampling and under-sampling, but these methods definitely ignore some vital information of the original data. In this paper, we first propose a weighted output extreme learning machine (WO-ELM) to learn the imbalanced data. A hierarchical extreme learning machine (H-C-ELM) is proposed based on the proposed WO-ELM and the weighted extreme learning machine (W-ELM). The H-C-ELM has two levels in its structure. In the first level, the WO-ELM and the W-ELM are trained on different combined fields of the CTR (each field has some attributes). The two extreme learning machines (ELMs) output their predicted scores of the corresponding combined fields of the CTR. The WO-ELM and the W-ELM have different predicted results on the same combined fields because of the difference of the two ELMs. Therefore, in the second level, another ELM is applied based on the outputs of the two ELMs in the first level and the actual outputs in order to improve the prediction accuracy. The experimental results demonstrate that the proposed H-C-ELM method has better performance for the binary classification with imbalanced data than the other related algorithms on CTR prediction, such as the WO-ELM, the W-ELM, and the stacked autoencoder-logistic regression.

INDEX TERMS Imbalance learning, CTR prediction, ELM, WO-ELM, H-C-ELM.

I. INTRODUCTION

With the dramatic development of Internet, the advertising industry pays more attention to the online advertisements rather than the traditional advertisements such as newspapers and TVs. One of the fundamental technologies of the online advertisements is the prediction of Click-Through Rate (CTR) [1], which not only heightens the advertising companies' reputation and earnings, but also helps the advertisers to optimize the advertising budgets.

Nowadays, CTR prediction has attracted much more attention, and many different approaches have been proposed. Shan *et al.* [2] proposed two dimensional matrix factorization model which simultaneously took users, ads and publishers into consideration because of the complicated interaction among the three factors. Due to the biased information of search engine click logs, Dupret and Piwowarski [3] provided a unbiased estimate of the document relevance through a set of assumptions on user browsing behavior. Fang *et al.* [4] established a bayesian network model which was selected as a framework for representing and inferring dependencies and uncertainties among variables to predict the CTRs of new ads. Kumar *et al.* [5] proposed and established a model to predict the CTR by adopting Logistic Regression (LR). Ma *et al.* [6] presented a useful CTR prediction model for ads of abundant history data by using Logistic Regression. Wang *et al.* [7] used the ensemble method for reference and proposed a feature selection algorithm based on the gradient boosting. Wang *et al.* [8] adopted multiple criteria linear programming regression model for the CTR prediction. Shan *et al.* [9] proposed a feature-based fully coupled interaction tensor factorization to predict CTR. Li *et al.* [10] proposed a model of CTR prediction based on a convolution neural network.

An embedding layer was proposed to learn a distributed representation of categorical data in CTR prediction [11]–[13]. Zhu *et al.* [14] proposed a Softmax-based ensemble model for CTR estimation in Real Time Bidding(RTB) advertising. As a matter of fact, the experimental results in [2]–[14] were got in an ideal condition where the data was approximately balanced. Shi and Jin-Ji [15] adopted a balanced sampling strategy to the CTR prediction, but this method will definitely lose the fully information of the original data. Some methods about the prediction of CTR focus on the representation of attributes. For instance, Jiang *et al.* [17] used Deep Belief Network (DBN) [16] to extract the complex attributes of the original dataset and then logistic regression was trained on the extracted features; in the model of stacked autoencoder-logistic regression(SAE-LR) [18], the outputs of autoencoder [19] were regarded as the inputs of logistic regression. However, these two methods are a little time-consuming because lots of time is needed to train the DBN and the autoencoder.

Single hidden layer feedforward networks (SLFNs) can work as universal approximators, which was proved in an incremental constructive method. Extreme Learning Machine (ELM), as a training method of SLFNs, has some exceptional features such as fast learning speed, good generation, etc. Inspired by the relationship between ELM network and its subnetworks [20], we give separate analysis on the minority class and the majority class. A novel cost-sensitive method WO-ELM, in this paper, is proposed to solve the issue of the imbalanced data. The WO-ELM and the W-ELM [25] will give their predicted scores on different combined fields. It's hard for us to figure out the contributions of different scores by experience. Therefore, we design a model structure of hierarchical ELM (H-C-ELM) to find effective representation between the predicted scores of two ELMs and the real outputs.

The paper is organized as follows: Section II gives the background of CTR prediction. Evaluation metric of CTR prediction is introduced in Section III. Section IV has a brief review of ELM theory. The details of the proposed WO-ELM and H-C-ELM including the difference between the W-ELM and the WO-ELM are described in Section V. Section VI gives the experimental results. Section 7 draws the conclusions.

II. THE BACKGROUND OF THE CTR PREDICTION

The online advertising can be divided into two categories. One is the searching advertising which means that the search engine uses the inputting keywords of the users to orient the advertising content and the position. The other one is real time bidding (RTB) advertising. RTB advertising has three important parts: Supply Side Platform (SSP), Ad Exchange (ADX) and Demand Side Platform (DSP). DSP, ADX, SSP and advertisers are included in the transaction model of RTB. Advertisers will put their demand of the advertisement on DSP. Through SSP, media websites will put their ad impression on the corresponding ADX. The advertisers will begin the bid competition through DSPs and ADX. Finally, the ADX will choose the ads for impression which has the highest bidding price [1].

CTR can be expressed as

$$CTR = \frac{Click\_num}{Impression\_num}. \tag{1}$$

where *Click_num* and *Impression_num* are the number of the clicks and the number of the impressions.

Currently, the popular advertising model is Cost Per Click (CPC). That is to say, the advertiser need to pay for every click. The specific computing way is shown as

$$value = CPC \times CTR. \tag{2}$$

where *value* is the profit of the advertising company.

In the equation above, CPC usually is a constant. If the advertising companies expect to make a healthy profit, they should spare no effort to increasing the value of CTR as much as possible.

III. EVALUATION METRIC

Accuracy is not the ideal evaluation metric for imbalance learning. In this paper, we select Area Under Curve (AUC) [21] as the evaluation metric of binary classification with imbalanced data. AUC, which is the abbreviation of the Area Under ROC Curve [22], is based on confusion matrix (see Table 1), where the true positive, the false negative, the false positive and the true negative are denoted as TP, FN, FP and TN respectively.

TABLE 1. Confusion matrix.

	Predicted positive examples	Predicted negative examples
Actual positive examples	TP	FN
Actual negative examples	FP	TN

ROC curve can be obtained by the horizontal coordinate FPR and the vertical coordinate TPR. The results of the classification are excellent when the value of AUC is close to 1. We can compute the FPR and the TPR as

$$FPR = \frac{FP}{FP + TN}. \tag{3}$$

$$TPR = \frac{TP}{TP + FN}. \tag{4}$$

IV. A BRIEF INTRODUCTION OF ELM

As displayed in Fig. 1, ELM [23] is a learning algorithm for single-hidden layer feedforward neural networks. Given N training samples $\{(X, T)|X = [x_1, x_2, \ldots, x_N]^T, T = [t_1, t_2, \ldots, t_N]^T\}$ and the hidden layer input (with L nodes) matrix X, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^n$ and $t_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \in R^m$, ELM is represented as (5).

$$\sum_{i=1}^{L} \beta_i \cdot G(a_i.x_j + b_i) = t_j, \quad j = 1, 2, \ldots, N. \tag{5}$$

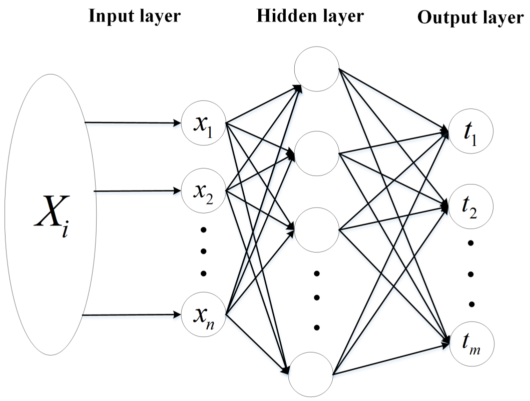

FIGURE 1. The model structure of ELM.

where $G(ax + b)$ is a activation function. $a_i = [a_{i1}, a_{i2}, \ldots, a_{in}]^T$ is the weight vector bridging the input layer and the i-th hidden neuron and n is the number of the input neurons. b_i is the bias of the i-th hidden neuron. $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}]^T$ is the connecting weight vector between the output layer and the i-th hidden neuron, m is the number of the output neurons.

The output matrix of the hidden layer is

$$H = \begin{pmatrix} G(a_1 \cdot x_1 + b_1) & \ldots & G(a_L \cdot x_1 + b_L) \\ \vdots & \ddots & \vdots \\ G(a_1 \cdot x_N + b_1) & \cdots & G(a_L \cdot x_N + b_L) \end{pmatrix}_{N \times L} = \begin{pmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{pmatrix}. \quad (6)$$

β is calculated in the way of (7).

$$\beta = H^+ T. \quad (7)$$

where H^+ is the Moore−Penrose generalized inverse of the matrix H.

The optimization problem can be written as
Minimize:

$$Loss = \frac{1}{2}\|\beta\|^2 + \frac{C}{2}\sum_{i=1}^{N} \xi^2. \quad (8)$$

Subject to:

$$h(x_i)\beta = t_i - \xi_i, \quad i = 1, 2, \ldots, N. \quad (9)$$

where ξ_i is the training error vector of the m-th output node with respect to the training sample x_i. C is the regularization parameter.

According to KKT theorem, we can get
if $N < L$,

$$\beta = H^T(I/C + HH^T)^{-1}T. \quad (10)$$

if $N > L$,

$$\beta = (I/C + H^T H)^{-1} H^T T. \quad (11)$$

where I is the unit matrix.

V. THE PROPOSED APPROACH

Given the target vector $T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}_{N \times 1}$, where $T_1 = [-1, -1, \ldots, -1]^T \in R^{N_1}$, $T_2 = [1, 1, \ldots, 1]^T \in R^{N_2}$, $N_1 + N_2 = N$, X^- and X^+ are input matrixs of the two classes, we propose the weighted output ELM(WO-ELM) for the binary classification with imbalanced data. Furthermore, a hierarchical ELM based on the outputs of the WO-ELM and the W-ELM [25] is given in subsection V.C.

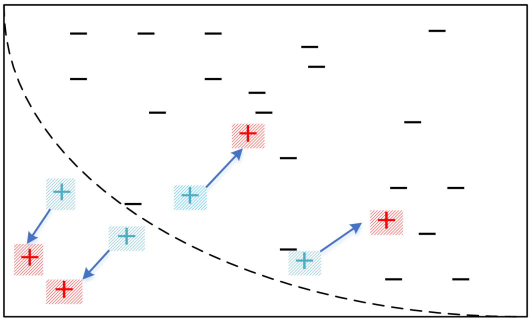

FIGURE 2. WO-ELM for imbalance learning.

A. THE WEIGHTED OUTPUT EXTREME LEARNING MACHINE (WO-ELM)

As shown in Fig. 2, the WO-ELM amplifies the outputs of the positive samples by moving $h(x_i)\beta$ which is denoted as the blue plus sign to the place that is denoted as the red plus sign. Intuitively speaking, this method maintains a long distance between $h(x_i)\beta$ and the separating boundary and obtains two similar errors of the classes indirectly. From (14), we can find that if a instance x_i belonging to the positive class is misclassified as the negative class, its error ζ_i is enlarged by parameter ε, compared to (9).

According to the analysis above, the optimization problem of the WO-ELM can be written as
Minimize:

$$Loss = \frac{1}{2}\|\beta\|^2 + \frac{C}{2}\left(\sum_{i=1}^{N_1} {\xi_i}^2 + \sum_{i=1}^{N_2} {\zeta_i}^2\right). \quad (12)$$

Subject to:

$$\xi_i = t_i - h(x_i)\beta, \quad i = 1, \ldots, N_1. \quad (13)$$

$$\zeta_i = t_i - \varepsilon h(x_i)\beta, \quad i = 1, \ldots, N_2. \quad (14)$$

where ξ_i is the error of i-th negative sample and ζ_i is the error of i-th positive sample.

According to KKT theorem, the equivalent dual optimization problem is

$$L = \frac{1}{2}\|\beta\|^2 + \frac{C}{2}\left(\sum_{i=1}^{N_1} {\xi_i}^2 + \sum_{i=1}^{N_2} {\zeta_i}^2\right) - \sum_{i=1}^{N_1} \alpha_{1i}(h(x_i)\beta - t_i + \xi_i) - \sum_{i=1}^{N_2} \alpha_{2i}(\varepsilon h(x_i)\beta - t_i + \zeta_i). \quad (15)$$

According to the KKT conditions, we can get

$$\frac{\partial L}{\partial \beta} = 0 \rightarrow \beta = H^T \alpha. \tag{16}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_{1i} = C\xi_i. \tag{17}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_{2i} = C\zeta_i. \tag{18}$$

$$\frac{\partial L}{\partial \alpha_{1i}} = 0 \rightarrow h(x_i)\beta - t_i + \xi_i = 0. \tag{19}$$

$$\frac{\partial L}{\partial \alpha_{2i}} = 0 \rightarrow \varepsilon h(x_i)\beta - t_i + \zeta_i = 0. \tag{20}$$

H is written as

$$H = \begin{pmatrix} h(X_1^-) \\ \vdots \\ h(X_{N_1}^-) \\ h(X_1^+) \\ \vdots \\ h(X_{N_2}^+) \end{pmatrix} = \begin{pmatrix} H_- \\ H_+ \end{pmatrix}. \tag{21}$$

We can compute β as follows.

If $N > L$,

$$\beta = \left({}^{I}/_{C} + H_-^T H_- + \varepsilon^2 H_+^T H_+\right)^{-1} \left(H_-^T T_1 + \varepsilon H_+^T T_2\right). \tag{22}$$

If $N < L$,

$$\beta = \left(H_-^T, \varepsilon H_+^T\right) W^{-1} \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}. \tag{23}$$

where $W = {}^{I}/_{C} + \begin{pmatrix} H_- H_-^T & \varepsilon H_- H_+^T \\ \varepsilon H_+ H_-^T & \varepsilon^2 H_+ H_+^T \end{pmatrix}$.

Finally, the WO-ELM classifier is obtained as the following equation for the binary classification.

$$f(x_i) = sign(h(x_i)\beta). \tag{24}$$

B. THE DIFFERENCE BETWEEN THE WO-ELM AND THE W-ELM

The optimization problem of the W-ELM [25], weighting scheme W_1, weighting scheme W_2 and the calculating way are shown as the following equations.

Minimize:

$$L = \frac{1}{2}\|\beta\|^2 + CW\frac{1}{2}\sum_{i=1}^{N}\left\|\xi_i^2\right\|. \tag{25}$$

Subject to :

$$h(x_i)\beta = t_i - \xi_i, \quad i = 1, 2, \ldots, N. \tag{26}$$

Weighting Scheme W_1:

$$W_1 : W_{ii} = 1/\#(t_i). \tag{27}$$

Weighting Scheme W_2:

$$W_2 \begin{cases} W_{ii} = 0.618/\#(t_i), & t_i > AVG(t_i) \\ W_{ii} = 1/\#(t_i), & t_i \leq AVG(t_i). \end{cases} \tag{28}$$

β is written as

$$\beta = \begin{cases} H^T\left({}^{I}/_{C} + WHH^T\right)^{-1} WT, N < L \\ \left({}^{I}/_{C} + H^T WH\right)^{-1} H^T WT, N > L. \end{cases} \tag{29}$$

where $\#(t_i)$ is the number of samples belonging to a specific class, the weight matrix $W = diag\{W_{ij}\}$ is obtained by (27) and (28).

As shown in (27) and (28), two distinct weighting schemes are introduced to assign different weights to the error of samples from different classes in W-ELM. Specifically speaking, a larger weight will be assigned to the error of minority class.W-ELM only focus on the errors of the classification in the outer manner. In other words, they ignore the ways that the errors are generated concretely. Besides, the two weighting schemes may not give much attention to minority samples. Therefore, W-ELM may not have a obvious advantage to unravel the problem of imbalanced data. A new way is adapted to process the errors of two classes in WO-ELM. The difference between WO-ELM and W-ELM is the way to get similar errors of two classes.

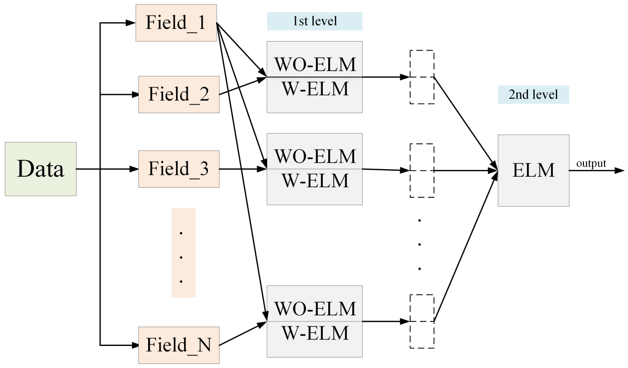

FIGURE 3. The model structure of H-C-ELM.

C. THE PROPOSED HIERARCHICAL ELM (H-C-ELM)

There are some fields with many attributes in the dataset of advertisement click-through rate. We usually use all attributes as inputs to train the CTR prediction model. In this way, all fields with some attributes decide the predicted results directly. Several algorithms aiming at resolving the problem of the imbalanced data failed to explore the relationship between the combined fields in the CTR and the results of the classification, such as Adaboost W-ELM [24], W-ELM [25]. In this subsection, we incorporate the advantages of the WO-ELM and the W-ELM in the two layers H-C-ELM to explore the contribution of different combined fields of the CTR. The model structure of H-C-ELM is shown in Fig. 3. In the first level of H-C-ELM, the WO-ELM and the W-ELM will give their scores on a specific combined fields. We denote the outputs of the WO-ELM and the W-ELM as o_{wo}^i and o_w^i on the i-th combined fields of the CTR. Nonetheless, o_{wo}^i and o_w^i are different because of the difference of the two ELM variants. For a specific sample, all the predicted scores of the first level are denoted as $O = [o_{wo}^1, o_w^1, o_{wo}^2, o_w^2, \cdots, o_{wo}^n, o_w^n]$, where n is the number of the combined fields of the CTR. If we assign weights vectors $W_1 = [0, 0, 0, 0, \cdots, 0, 1]$ and $W_2 = [0, 0, 0, 0, \cdots, 1, 0]$ to the predicted scores, $O \cdot W_1^T$ and $O \cdot W_2^T$ are equivalent to the W-ELM and the WO-ELM respectively. Each combination of different fields has a big or small contribution on the predicted results. But it is hard to define the contributions of different scores by experience. Therefore, in this model structure, all outputs of the WO-ELM and the W-ELM are concatenated as the inputs of ELM, which is expected to explore the relationship between the real outputs and the scores of the WO-ELM and the W-ELM.

TABLE 2. The detailed description of the four fields.

Field's name	Data type	Attributes' number
MD	0 or 1	6
APD	0 or 1	11
AF1	Double	23
AF2	Double	19

VI. PERFORMANCE EVALUATION

In this section, we will evaluate the performance of H-C-ELM. 5-fold cross validation is applied in the experiments. Five datasets with different imbalance ratios are used in these experiments to verify the effectiveness of the WO-ELM and the H-C-ELM. These are four fields in these datasets. They are Media ID (MD), Advertising Position ID (APD) and Anonymous Fields (AF1 and AF2). The detailed description of the four fields is given in Table 2. 2^6 is chosen as the value of C. A grid search of ε on {2,4,6,…,300} and L on {50,100,150,…,350} is conducted to find the optimal value of the related models. $sigmoid = \frac{1}{1+e^{-x}}$ is selected as the activation function of ELM.The imbalance ratio is defined as

$$IR = \frac{\#positive}{\#negative}. \tag{30}$$

where $\#positive$ is the number of the positive samples, and $\#negative$ is the number of the negative samples.

A. THE EXPERIMENTAL RESULTS

In real CTR application, AUC is used to evaluate the performance of the predicted results in the fields of advertisement CTR prediction. There are five datasets with different imbalanced ratios in the following experiments. Table 3 describes the detailed distribution of the five datasets.

TABLE 3. The detailed description of the five datasets.

IR	1:5	1:20	1:100	1:150	1:3000
Samples' number	4440	15540	74740	111740	318455

As shown in Fig. 4, in view of the issue of complexity, the suitable numbers of hidden neurons of H-C-ELM, W-ELM and WO-ELM are 50, 100,100 respectively. The detailed results of different fields are described in Table 4. According to the table, we can find that each combination of different fields makes a little contribution to the performance of classification. The W-ELM and the WO-ELM, as two cost-sensitive methods, have a better performance to train the imbalanced data, compared to the original ELM. In these combined fields, the W-ELM and the WO-ELM give their different scores. However, the scores only represent the potential relation between output and the combined fields. Assimilating different scores of different combined fields, H-C-ELM bridges a complex connection between the scores and the outputs, where the highest AUC (0.731) was obtained. Furthermore, in order to test the advantage of H-C-ELM, some related algorithms for advertisement CTR prediction are listed to make a comparison, as illustrated in Table 5–Table 9. These algorithms are trained with the whole attributes, without taking the interaction of fields

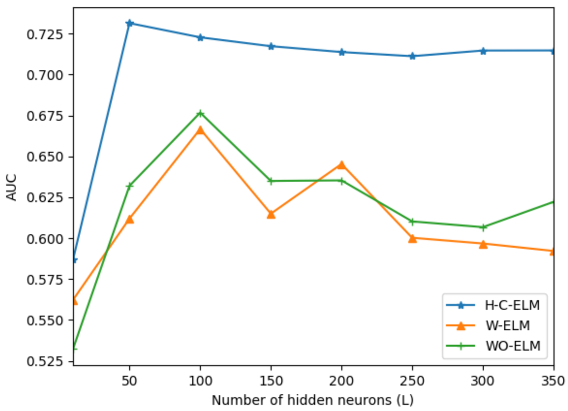

FIGURE 4. The effect of the number of the hidden neurons on AUC where IR is 3:1000.

TABLE 4. Detailed results of the H-C-ELM model where IR is 3:1000.

Combined Fields	W-ELM (AUC)	WO-ELM (AUC)	H-C-ELM (AUC)
MD+APD	0.533	0.539	0.731
MD+AF1	0.560	0.582	
MD+AF2	0.588	0.585	
APD+AF1	0.578	0.605	
APD+AF2	0.602	0.611	
AF1+AF2	0.627	0.626	
MD+APD+AF1	0.593	0.600	
MD+APD+AF2	0.618	0.617	
APD+AF1+AF2	0.641	0.647	
MD+APD+AF1+AF2	0.667	0.677	

TABLE 5. The experimental results of five algorithms where IR is 3:1000.

Method	The number of hidden neurons	AUC
Logistic regression [5]	/	0.619
SAE-LR [18]	150	0.682
WO-ELM	100	0.677
W-ELM[25]	100	0.667
H-C-ELM	50	0.724

TABLE 6. The experimental results of five algorithms where IR is 1:5.

Method	The number of hidden neurons	AUC
Logistic regression [5]	/	0.832
SAE-LR [18]	300	0.877
WO-ELM	400	0.848
W-ELM[25]	300	0.845
H-C-ELM	150	0.881

TABLE 7. The experimental results of five algorithms where IR is 1:20.

Method	The number of hidden neurons	AUC
Logistic regression [5]	/	0.815
SAE-LR [18]	300	0.845
WO-ELM	100	0.840
W-ELM[25]	100	0.841
H-C-ELM	200	0.856

TABLE 8. The experimental results of five algorithms where IR is 1:100.

Method	The number of hidden neurons	AUC
Logistic regression [5]	/	0.753
SAE-LR [18]	150	0.801
WO-ELM	150	0.786
W-ELM[25]	450	0.779
H-C-ELM	100	0.810

TABLE 9. The experimental results of five algorithms where IR is 1:150.

Method	The number of hidden neurons	AUC
Logistic regression [5]	/	0.718
SAE-LR [18]	200	0.771
WO-ELM	300	0.730
W-ELM[25]	200	0.728
H-C-ELM	100	0.796

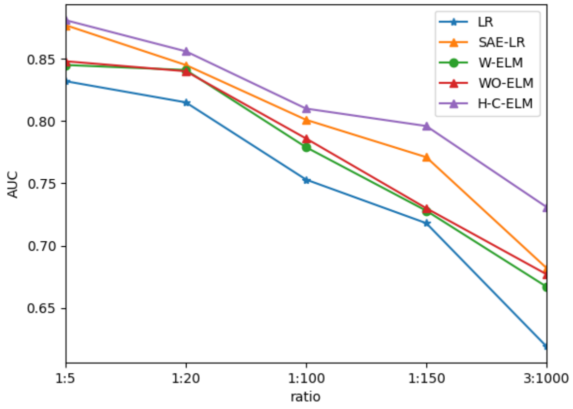

FIGURE 5. The influence of IR on AUC.

into consideration. In Fig. 5, we can find that 1) by assigning a big weight in a new way to the error of minority sample, the performance of the WO-ELM is slightly better than the W-ELM; 2) when IR is small (from 1:5 to 1:150), compared with the LR, the W-ELM and the WO-ELM, the SAE-LR [18] has an obvious advantage to tackle the problem of the imbalanced data, but the SAE-LR has a similar performance with the WO-ELM and the W-ELM when IR is 3:1000; 3) among the five algorithms, the H-C-ELM's performance is the best. From the statistic above, we can draw a conclusion that the model structure H-C-ELM have explored the potential representation of the combined fields and the output.

VII. CONCLUSIONS

In this paper, we firstly propose a weighted output extreme learning machine (WO-ELM) to learn the imbalanced data. According to the real problem of the advertisement CTR, the model structure of H-C-ELM is proposed to explore the contributions of distinct combined fields in the CTR. H-C-ELM has two levels. In the first level of H-C-ELM, the WO-ELM and the W-ELM are used to give their scores on each combined fields. ELM is selected as the final classifier in the second level of the model. Experimental results disclose that the proposed WO-ELM has a good performance on the imbalance learning, and H-C-ELM has better performance compared to the WO-ELM. Furthermore, the H-C-ELM offers a significant guide to the display of ads. However, some deep relationship between the attributes may not be fully explored. The future work may include the use of deep neutral network such as CNN [26] and RBM [27] to explore the interaction between attributes in CTR prediction.

REFERENCES

[1] A. Y. Zhou, "Computational advertising: A data-centric comprehensive Web application," *Chin. J. Comput.*, vol. 34, no. 10, pp. 1805–1819, 2011.

[2] L. Shan, L. Lin, D. Shao, and X. Wang, "CTR prediction for DSP with improved cube factorization model from historical bidding log," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Springer, 2014, pp. 17–24.

[3] G. E. Dupret and B. Piwowarski, "A user browsing model to predict search engine click data from past observations," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2008, pp. 331–338.

[4] Z. Fang, K. Yue, J. Zhang, D. Zhang, and W. Liu, "Predicting click-through rates of new advertisements based on the Bayesian network," *Math. Problems Eng.*, vol. 2014, no. 4, pp. 1–9, 2014.

[5] R. Kumar, S. M. Naik, V. D. Naik, S. Shiralli, S. V. G, and M. Husain, "Predicting clicks: CTR estimation of advertisements using logistic regression classifier," in *Proc. Adv. Comput. Conf.*, 2015, pp. 1134–1138.

[6] J. Ma, X. Chen, Y. Lu, and K. Zhang, "A click-through rate prediction model and its applications to sponsored search advertising," in *Proc. Int. Conf. Cyberspace Technol.*, 2014, pp. 500–503.

[7] Z. Wang, Q. Yu, C. Shen, and W. Hu, "Feature selection in click-through rate prediction based on gradient boosting," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.*, 2016, pp. 134–142.

[8] F. Wang, W. Suphamitmongkol, and B. Wang, "Advertisement click-through rate prediction using multiple criteria linear programming regression model," *Procedia Comput. Sci.*, vol. 17, pp. 803–811, Jan. 2013.

[9] L. Shan, L. Lin, C. Sun, and X. Wang, "Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization," *Electron. Commerce Res. Appl.*, vol. 16, pp. 30–42, Mar. 2016.

[10] S. Q. Li, L. Lin, and C. Sun, "Click-through rate prediction for search advertising based on convolution neural network," *Intell. Comput. Appl.*, vol. 5, pp. 22–25, May 2015.

[11] W. Zhang, T. Du, and J. Wang, "Deep learning over multi-field categorical data," in *Proc. Eur. Conf. Inf. Retr.*, 2016, pp. 45–57.

[12] Y. Qu *et al.*, "Product-based neural networks for user response prediction," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2017, pp. 1149–1154.

[13] Q. Liu, F. Yu, S. Wu, and L. Wang, "A convolutional click prediction model," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1743–1746.

[14] W.-Y. Zhu, C.-H. Wang, W.-Y. Shih, W.-C. Peng, and J.-L. Huang, "SEM: A Softmax-based ensemble model for CTR estimation in real-time bidding advertising," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, Feb. 2017, pp. 5–12.

[15] M. Y. Shi and G. U. Jin-Ji, "Balance-sampling based light-weighted advertisement CTR prediction method," *Appl. Res. Comput.*, vol. 31, no. 1, pp. 33–36, Jan. 2014.

[16] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang, "Deep belief network-based approaches for link prediction in signed social networks," *Entropy*, vol. 17, no. 4, pp. 2140–2169, 2015.

[17] Z. Jiang, S. Gao, and W. Dai, "Research on CTR prediction for contextual advertising based on deep architecture model," *Control Eng. Appl. Inform.*, vol. 18, no. 1, pp. 11–19, 2016.

[18] Z. Jiang, S. Gao, and W. Dai, "A CTR prediction approach for text advertising based on the SAE-LR deep neural network," *J. Inf. Process. Syst.*, vol. 13, no. 5, pp. 1052–1070, 2017.

[19] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2012, pp. 4153–4156.

[20] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, S. Mao, G.-B. Huang, "A theoretical study of the relationship between an ELM network and its subnetworks," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 1794–1801.

[21] D. Brzezinski and J. Stefanowski, "Prequential AUC: Properties of the area under the ROC curve for data streams with concept drift," *Knowl. Inf. Syst.*, vol. 52, no. 2, pp. 531–562, 2017.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[23] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. Int. Joint Conf. Neural Netw*, vol. 2, 2004, pp. 985–990.

[24] S. Zhang, Q. Fu, and W. Xiao, "Advertisement click-through rate prediction based on the weighted-ELM and adaboost algorithm," *Sci. Program.*, vol. 2017, no. 1, pp. 1–8, 2017.

[25] W. Zong, G. B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, no. 3, pp. 229–242, 2013.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[27] R. D. Hjelm, V. D. Calhoun, R. Salakhutdinov, E. A. Allen, T. Adali, and S. M. Plis, "Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks," *Neuroimage*, vol. 96, no. 8, pp. 245–260, 2014.

SEN ZHANG received the Ph.D. degree in electrical engineering from Nanyang Technological University in 2005. She has been a Post-Doctoral Research Fellow with the National University of Singapore and a Lecturer in Charge with Singapore Polytechnic. She is currently an associate professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing. Her research interests include extreme learning machine, target tracking, and estimation theory.

ZHENG LIU is currently pursuing the M.S. degree with the Department of Control Science and Engineering, University of Science and Technology Beijing. His research interests include machine learning and its application to advertisement click-through rate prediction.

WENDONG XIAO (M'01–SM'09) received the Ph.D. degree from Northeastern University, China, in 1995. His previous appointments include the Scientist III with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, from 2004 to 2012; a Research Fellow with Nanyang Technological University, Singapore, from 2001 to 2004; an Associate Professor with Northeastern University from 1999 to 2001; and a Post-Doctorate Research Fellow with the POSCO Technical Research Laboratories, South Korea, from 1996 to 1999. He is currently a Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. His current research focuses on wireless localization and tracking, energy-harvesting-based network resource management, wearable computing for healthcare, big data processing, wireless sensor networks, and Internet of Things. He has authored about 150 papers in journals and conferences and has been participating in a number of research and industrial projects in the related areas. He is actively participating in the organizations for more than 70 international conferences and is a reviewer for many top international journals.