Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(4): 4586–4602.
DOI: 10.3934/mbe.2021233
Received: 08 March 2021
Accepted: 19 May 2021
Published: 27 May 2021

Research article

Semi-supervised random forest regression model based on co-training and grouping with information entropy for evaluation of depression symptoms severity

Shengfu Lu[1,2,3], Xin Shi[1,2,3], Mi Li[1,2,3,4,*], Jinan Jiao[1,2,3], Lei Feng[5,6] and Gang Wang[5,6]

[1] Department of Automation, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2] The Beijing International Collaboration Base on Brain Informatics and Wisdom Services, Beijing 100124, China
[3] Engineering Research Center of Intelligent Perception and Autonomous Control, Ministry of Education, Beijing 100124, China
[4] Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China
[5] The National Clinical Research Center for Mental Disorders & Beijing Key Laboratory of Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing 100088, China
[6] The Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing 100088, China

* **Correspondence:** Email: limi@bjut.edu.cn.

Abstract: Semi-supervised learning has always been a hot topic in machine learning. It uses a large number of unlabeled data to improve the performance of the model. This paper combines the co-training strategy and random forest to propose a novel semi-supervised regression algorithm: semi-supervised random forest regression model based on co-training and grouping with information entropy (E-CoGRF), and applies it to the evaluation of depression symptoms severity. The algorithm inherits the ensemble characteristics of random forest, and combines well with co-training. In order to balance the accuracy and diversity of co-training random forests, the algorithm proposes a grouping strategy to decision trees. Moreover, the information entropy is used to measure the confidence, which avoids unnecessary repeated training and improves the efficiency of the model. In the practical application of evaluation of depression symptoms severity, we collect cognitive behavioral data of emotional conflict based on the depressive affective disorder. And on this basis, feature construction and normalization preprocessing are carried out. Finally, the test is conducted on 35 labeled and 80 unlabeled depression patients. The result shows that the proposed algorithm obtains MAE (Mean Absolute Error) = 3.63 and RMSE (Root Mean Squared Error) = 4.50, which is better than other semi-supervised regression algorithms. The proposed method effectively solves the modeling difficulties caused by insufficient labeled samples, and has important reference value for the diagnosis of depression symptoms severity.

Keywords: semi-supervised learning; E-CoGRF; depression; symptoms severity; emotional conflict

1. Introduction

In practical application, data tagging is limited by various factors, and sometimes even needs to pay a high price. Compared with labeled samples, the data acquisition of unlabeled samples is relatively simple. Semi-supervised learning is generated under the drive of practical application. It mainly studies how to make learners automatically use a large number of unlabeled data to assist the learning of a small number of labeled data under the premise of partial information missing in training data. Semi-supervised learning has been widely used in many fields because of its strong practical demand. Especially for psychiatric disorders research, there is a shortage of medical information talents and a low rate of clinical practice, so it is very expensive to obtain labeled samples. For example, in a recent study on bipolar disorder [1], the authors investigated the use of a semi-supervised learning method to process data obtained from patient-smartphone interactions, and considered the evolving time structure. The results shown that the semi-supervised model can effectively derive patient's status even if only 25% of the labeled data are available.

Depression is a psychiatric disorder characterized by significant and persistent loss of pleasure. To date, the pathogenesis of depression is still unclear. The diagnosis of depression is mainly carried out by psychiatrists through structured interviews based on diagnostic manuals [2]. Clinically, 17-item Hamilton depression rating scale (HAMD-17) [3] is considered as the gold standard for the evaluation of depression symptoms severity. However, the evaluation of the HAMD scores is time-consuming. On the one hand, the evaluation time depends on patients' symptoms and cooperation, and it will take longer if the patient is severely blocked. On the other hand, HAMD scores are evaluated independently by two trained evaluators through conversation and observation, which also requires a high level of experience and skill. Therefore, it is very difficult to obtain the HAMD scores, which results in less labeled (HAMD score) data and more unlabeled data in patients with depression. It may be difficult to train a learning system with strong generalization if only a small number of labeled data is used for supervised learning, while ignoring a large number of unlabeled data will cause a great waste of data resources. In this paper, the semi-supervised learning method is used to construct the evaluation model, which can solve the above problems well and achieve better performance.

In addition, machine learning methods based on objective biomarkers have become key technologies and effective means to diagnose depression. A large number of studies have shown that depression is related to the abnormalities in brain structure and function [4,5]. At present, the use of functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG) to record brain structure or brain activity has become an important means of identifying depression [6–8]. In addition, depression not only causes brain abnormalities, but also leads to behavioral disorders [9,10]. Therefore, in recent years, more and more attention has been paid to the correlation between depression and behavioral patterns such as facial expression, speech, language and body posture [11–13].

This study considers the basic characteristic of depression—affective disorder, and collects cognitive behavioral data derived from emotional conflict tasks. Cognitive behavioral data on conflict tasks are obtained by active attention and conscious cognitive. Compared with EEG/fMRI, or expression/speech/language data, it is purer, reliable, and more likely to show the depression mood of patients with depression [14,15].

Emotional conflict is caused by the processing conflict between the information arousing different emotions, which leads to the interference of irrelevant emotional stimulation on current emotional cognition. Researches based on cognitive psychology have shown that patients with depression show delayed reaction effect no matter in dealing with emotional conflict or non-conflict information. In particular, the delay is more significant when dealing with emotional conflict information [15]. The reason is that emotional disorder leads to negative emotional attention bias in patients with depression, which is manifested as excessive attention to sadness and insufficient attention to happiness [16,17]. Meanwhile, patients with depression have rumination. That is, they indulge and repeat negative emotions, which leads to slow thinking and movement, and difficult decision-making [18,19]. In addition, due to the sequence adjustment effect, the reaction time in dealing with an emotional conflict task is also affected by the previous task type [20]. Therefore, this study collects four kinds of cognitive behavior data: conflict/non-conflict/conflict monitoring/conflict resolution reaction time.

Although deep learning is widely studied in image processing, speech and text recognition [21,22], the emotional conflict data in this study is tabular data, which is not suitable for deep learning model. In the future, we will use deep learning to explore the evaluation method of depression severity based on facial expression, speech and language text data obtained under emotional stimulation.

The rest of this paper is arranged as follows: Section 2 is related work. Section 3 is data acquisition and processing. Section 4 is the proposed model. Section 5 is result and discussion. Section 6 is conclusion of this study.

2. Related works

At present, automatic depression evaluation methods based on machine learning have been widely studied. For example, Yoshida et al. (2017) adopted partial least squares regression for fMRI-based depression recognition. The Beck Depression Inventory-II (BDI-II) score was used to evaluate depression severity with a prediction error of RMSE = 9.56 [6]. Kang et al. (2017) presented a deep transformation learning (DTL) method for visual-based depression recognition, and the prediction error for BDI-II score was RMSE = 9.43 and MAE = 7.74 [11]. Haque et al. (2018) evaluated depression symptoms severity by using causal convolutional network (C-CNN) based on spoken language and 3D facial expressions. The final prediction error on the Patient Health Questionnaire-8 (PHQ-8) was MAE = 3.67 [12]. Muzammel et al. (2020) proposed a speech-based depression evaluation method through deep learning, with the evaluation error of PHQ-8 was RMSE = 3.22 [13].

Current studies on the evaluation of depression symptoms severity all use self-rating depression scales like PHQ-8 or BDI-II. However, these self-rating scales are too subjective to be used for clinical symptoms evaluation in patients with depression. HAMD depression scale is a clinical scale specially used to evaluate symptoms of patients with depression. In the evaluation process, two psychiatrists or two trained professionals are required to score independently. So the evaluation has a high accuracy and can be used as the standard for the selection of treatment methods or evaluation of treatment effects. Therefore, the HAMD scale is called the "gold standard" for clinical depression symptoms evaluation. The acquisition of HAMD-17 score is time-consuming and laborious, leading to the existence of a large number of samples without HAMD score labels in reality. Therefore, there is a lack of research on automatic evaluation methods of HAMD-17 score. The semi-supervised learning method introduced in this paper can solve this problem well.

At present, there are four main paradigms in semi-supervised learning: generative model, transductive SVM, graph-based algorithm and co-training algorithm [23]. In many branches of semi-supervised learning, co-training algorithm has the characteristics of simple, effective, stable and fast convergence, which has attracted the extensive attention of many scholars and achieved a lot of research results.

In 2005, Zhou and Li first studied semi-supervised regression and proposed co-training regressors (CoReg) [23]. This method generates two different KNN models based on different distance measures. Each regression model labels and selects unlabeled samples with high confidence to join the training set of another regression model. Lei et al. [24] proposed a semi-supervised regression algorithm based on support vector machine co-training. After that, Li et al. [25] proposed a semi-supervised regression algorithm based on co-training with SVR-KNN by combining SVR and KNN. In addition, some people [26,27] proposed the method of combining co-training strategy with partial least squares model (PLS).

In the above co-training methods, only one model is used to assist the training of the other model. That is, one model selects some high confidence samples from the unlabeled samples to label and add them to the training set of the other model. However, the prediction result of the single model is not reliable and it is easy to introduce noise data into the model. Therefore, Hady et al. [28] constructed a co-training algorithm based on ensemble method—CoBCReg. CoBCReg uses a committee of regressors (RBFNNs constructed by bagging) to predict unlabeled samples instead of a single regressor. The algorithm combines co-training and ensemble learning to generate a more powerful learning system.

Random forest is an ensemble learning method composed of decision trees, which is easy to implement in parallel. Moreover, randomness is introduced in both row and column directions, which can weaken over-fitting. Therefore, it is a powerful classification and regression method. Co-training of decision trees in random forest does not require the integration of different basic learners, but also ensures the diversity among members. However, there is a lack of research on co-training random forest in regression. Saitoh et al. [29] and Levatić et al. [30] only use random forest for self-training in regression. Therefore, this paper applies random forest to semi-supervised regression, and realizes the combination of ensemble learning and co-training.

There are still some problems in the semi-supervised random forest regression model based on co-training: 1) a small number of decision trees will lead to low prediction accuracy; a large number of decision trees will reduce the diversity. Therefore, in this paper, the random grouping strategy is used to group the decision trees to achieve the balance between prediction performance and model diversity. 2) The confidence measure method used in the previous research [23,28] has the problems of large amount of calculation and low generalization. Therefore, this paper proposes a confidence measure method based on information entropy.

3. Data acquisition and processing

3.1. Subjects

All 115 patients with depression in this study were from the clinic. All participants provided signed informed consent and this study was approved by the Ethics Committee at Beijing Anding Hospital, Capital Medical University, Beijing, China. And the experiments have been conducted in accordance with the Helsinki declaration.

In the decision-making tasks of emotional conflict, we collected the time-related features of all 115 patients with depression while performing the tasks. However, HAMD-17 scores were obtained from only 35 patients. When building the dataset using the acquired values of HAMD-17 and the time-related features of the decision-making tasks, HAMD-17 values for only 35 people are available, and those for the remaining 80 people are absent. Therefore, we adopted a semi-supervised machine learning approach to solve the problem of incomplete HAMD-17 labels in the samples.

Table 1. Distribution and depressive symptoms of HAMD-17 scores.

HAMD-17 score	Symptoms severity	Samples size
0–7	none	0
8–17	mild	7
18–24	moderate	19
> 25	severe	9

3.2. Data acquisition

3.2.1. Acquisition method

In order to obtain the cognitive behavioral data on emotional conflict, we require participants to complete the word-face Stroop tasks [14,31]. Each Stroop task is composed of an emotional face (happy face or sad face) and an emotional word ("HAPPY" or "SAD") in red ink. Emotional face pictures are from the International Standard Expressions Library (NimStim Set of Facial Expressions [32]), including 40 happy face pictures (22 males and 18 females) and 40 sad face pictures (22 males and 18 females). Photoshop is used to process all pictures in a unified way, so that the size of all face pictures are consistent.

According to whether the emotions expressed by the face picture and the word are consistent, the experimental tasks are divided into two categories: 1) conflict tasks or incongruent tasks (happy face + "SAD" and sad face + "HAPPY"); 2) non-conflict tasks or congruent tasks (happy face + "HAPPY" and sad face + "SAD"), as shown in Figure 1.

In addition, due to the sequence adjustment effect, the reaction time in dealing with an emotional conflict task is also affected by the previous task type. Therefore, conflict tasks can be further divided into two types: 1) conflict monitoring tasks (Figure 1a): a non-conflict task activates a non-conflict expectation mechanism, resulting in the weakening of conflict resolution in the next conflict task (high conflict monitoring and low conflict resolution). In other words, after completing a non-conflict task, the participants will make a decision relatively slowly on the next conflict task. 2) conflict resolution tasks (Figure 1b): a conflict expectation mechanism activated by a conflict task enhances the conflict resolution in the next conflict task (high conflict resolution and low conflict monitoring). In other words, the participants will have conflict awareness after completing a conflict task, so that they will make a decision relatively quickly on the next conflict task.

In the end, it constitutes a total of 80 conflict tasks and 80 non-conflict tasks. Because of the different types of the previous task, the 80 conflict tasks are divided into 40 conflict monitoring tasks and 40 conflict resolution tasks.

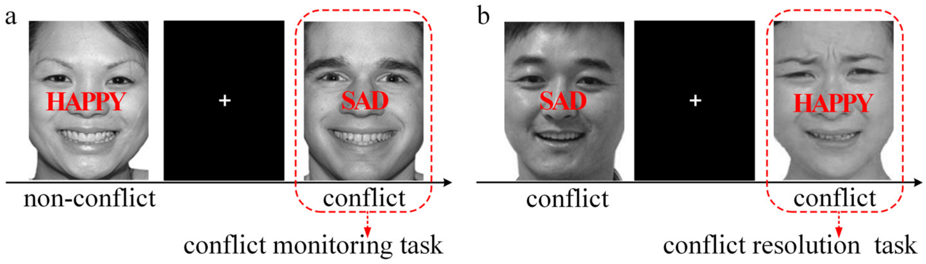

Figure 1. Experimental tasks.

3.2.2. Acquisition process

The data acquisition process is shown in Figure 2. The 160 emotional decision-making tasks (80 conflict tasks and 80 non-conflict tasks) are divided into two sessions, and each session contains 80 trials. At the beginning of each trial, a "+" is displayed at the center of the screen for 3–5 s, followed by an emotional decision-making task for 1s. Participants are required to ignore the word, and quickly recognize the emotional valence (happy or sad) of the face and press the key (left key for happy and right key for sad). The next trial starts automatically after the key or task ends.

When the participants carry out each task, the acquisition program automatically records the start time and end time, and calculates the reaction time.

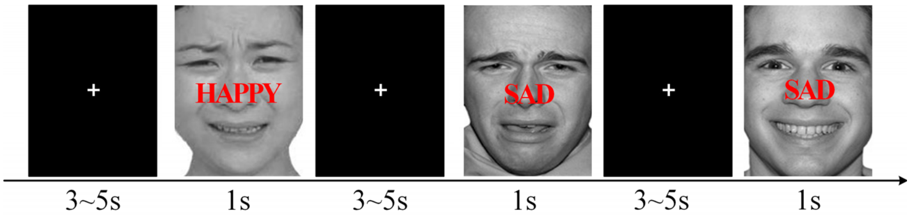

Figure 2. Data acquisition process.

3.2.3. Depression symptoms severity and conflict decision-making

Table 2 shows the mean reaction time of patients with different depression symptoms in conflict or non-conflict decision-making tasks. It can be seen that the more severe depression symptoms, the longer the reaction time required for decision-making. The result shows that the decision-making ability decreases with the increase of depression symptoms severity. It also illustrates that the decision-making reaction time under emotional conflict can characterize depression symptoms severity. Since the HAMD scale is a rating scale for depression symptoms severity, if the HAMD values are used as the labels of the samples, we can use the decision-making information on emotional conflict to predict the HAMD value and thus evaluate depression symptoms severity.

Table 2. Depression symptoms severity and conflict decision-making.

Symptoms severity	Mean of reaction time (*ms*)	
	conflict	non-conflict
mild	752.25	702.48
moderate	758.15	709.32
severe	771.84	725.07

3.3. Feature construction

In the process of data acquisition, each participant acquired 80 non-conflict reaction time, 80 conflict reaction time, 40 conflict monitoring reaction time and 40 conflict resolution reaction time.

From the perspective of statistical distribution, we extract 9 features from each kind of reaction time data of each participant, which are: 4 central tendency features (mean, median, 1st quartile, 3rd quartile), 3 dispersion tendency features (minimum, maximum, standard deviation) and 2 distribution pattern features (skewness coefficient, kurtosis coefficient).

Therefore, a total of 36 features are constructed in this study, as shown in Table 3. The value of each feature is the mean of samples (*ms*).

Table 3. Feature construction (Q1 = 1st quartile, Q3 = 3rd quartile).

	Central tendency				Dispersion tendency			Distribution pattern	
	Mean	Median	Q1	Q3	Min	Max	Std.	Skewness	Kurtosis
conflict	743.40	722.71	643.29	819.20	487.50	1226.53	146.84	0.97	4.37
non-conflict	703.12	675.71	606.55	764.43	467.13	1185.40	141.53	1.14	4.89
conflict monitoring	744.22	722.02	641.90	824.34	502.63	1161.20	146.64	0.91	4.16
conflict resolution	742.57	719.04	642.84	821.61	512.10	1162.20	144.96	0.86	3.94

3.4. Feature normalization

As shown in Table 3, the values of skewness feature and kurtosis feature are much smaller than other features, so their role in model construction will be ignored. In order for all features to participate in decision-making equally, it is necessary to perform feature normalization.

In this paper, the min-max conversion function is used for feature normalization, and the feature data is converted to the interval [0, 1]. For each feature, the normalization method is shown in formula (1).

$$x_i' = \frac{x_i - min(X)}{max(X) - min(X)} \tag{1}$$

where, $X = \{x_i\}, i = 1, 2, \ldots, N$. X is the dataset (N is the number of samples), x_i' is the normalized value of x_i. $max(X)$、 $min(X)$ are the maximum and minimum of X, respectively.

4. Proposed model

In this study, a semi-supervised regression method based on co-training is used to evaluate HAMD scores of patients with depression. The idea of co-training algorithm is as follows: First, train at least two models on the initial labeled sample set; Then, one of them is selected as the main model in turn, and the others constitute the auxiliary model of the main model; The auxiliary model predicts the unlabeled samples and provides the samples with high confidence for the main model; The main model retrains on the updated labeled sample set. Keep training in this way until convergence.

This research combines co-training algorithm and random forest (CoRF), and on this basis, introduces a grouping strategy and a confidence measure method using information entropy.

4.1. Grouping strategy

Co-training is effective only when members are diverse. Obviously, if the regression models are similar or even identical, there is no more information to be transmitted between the members, and the co-training is meaningless.

CoRF: For a random forest with *N* decision trees, one of the decision trees is selected as the main model by turns in each iteration, and the other *N*-1 decision trees constitute its auxiliary model. Only one decision tree is different between the auxiliary models of any two main models, and the diversity is 1/ (*N*-1). The larger *N* is, the smaller the diversity between the two auxiliary models. The diversity between the two new labeled sample sets predicted and selected by the two auxiliary models will also be small, leading to the similarity of the two main models after retraining. Similarly, any two models will become similar to each other. After many iterations, the diversity between models is destroyed.

CoGRF: According to the above description, when there are many decision trees, it is difficult for CoRF to ensure its diversity; when there are few decision trees, it is impossible to guarantee the accuracy. Therefore, a grouping strategy is introduced in this paper. *N* decision trees are divided into *m* groups, and there are *N*/*m* decision trees in each group. In each iteration, one group is taken as the main model, and the other *m*-1 groups are used as its auxiliary model. One group is different between the auxiliary models of any two main models, including *N*/*m* decision trees, and the diversity is 1/ (m-1). The grouping strategy greatly improves the diversity of the co-training model and ensures the number of decision trees.

4.2. Confidence measure method based on information entropy

An important factor affecting the performance of co-training algorithm is how to measure the confidence of new labeled samples. Incorrect confidence measure method will lead to the selection and addition of samples with wrong labels, which will have a negative impact on the algorithm.

In previous studies, RMSE difference was often used to measure the reliability of new labeled samples [23,28]. That is, the most reliable new labeled sample should be the one that reduces the regression model's error on the labeled sample set the most, as shown in formula (2).

$$\Delta_{x_u} = RMSE - RMSE^* \tag{2}$$

where x_u is the unlabeled sample. RMSE is the prediction error of the original regression model on the validation set. $RMSE^*$ is the prediction error of the new regression model on the validation set. The new regression model is obtained by adding x_u to the original training set and retraining. Finally, the sample with the largest Δ_{x_u} is regarded as the sample with the highest confidence.

This method has a significant disadvantage. That is, in each iteration, it has to repeated training model and calculate RMSE after adding each sample of the new labeled sample set. This method is computationally intensive and time-consuming. Therefore, this study proposes a simple and effective method to measure the confidence of samples: information entropy.

In the actual system, the probability of each possible case in the system may not be the same. So information entropy is used to measure the uncertainty of the whole system. It is also a measure of the consistency of all possible cases in the system. The calculation formula of information entropy is as follows:

$$H = -\sum_{i=1}^{n} p_i * logp_i \tag{3}$$

where, *n* indicates that there are *n* possible cases in the system. p_i is the probability of each possible case. The larger H is, the higher the system uncertainty is and the higher the consistency of all possible cases is.

In the auxiliary model, the predicted values of different decision trees are different. Therefore, we use information entropy to measure the consistency of predictions, that is, the similarity degree of the predicted values of all decision trees. For sample x_u, The confidence measure method follows formula (3), where p_i is defined as follows:

$$p_i = T_i(x_u) / \sum_{i=1}^{n} T_i(x_u) \tag{4}$$

where *n* is the number of decision trees in the auxiliary model. $T_i(x_u)$ is the predicted value of the *i-th* decision tree to sample x_u. p_i represents the proportion of predicted value of the *i-th* decision tree to predicted values of all decision trees in the auxiliary model.

The larger the information entropy of the sample is, the higher the consistency of the predictions of the auxiliary decision trees is, and the higher the confidence of the sample is. The confidence measure method based on information entropy, on the one hand, greatly reduces the training process and the time complexity. On the other hand, considering the difference of predictions of all auxiliary decision trees, the selection of samples with the highest consistency is conducive to improving the generalization of the model.

4.3. E-CoGRF algorithm process

The algorithm process of semi-supervised random forest regression model based on co-training and grouping with information entropy (E-CoGRF) is shown in Figure 3.

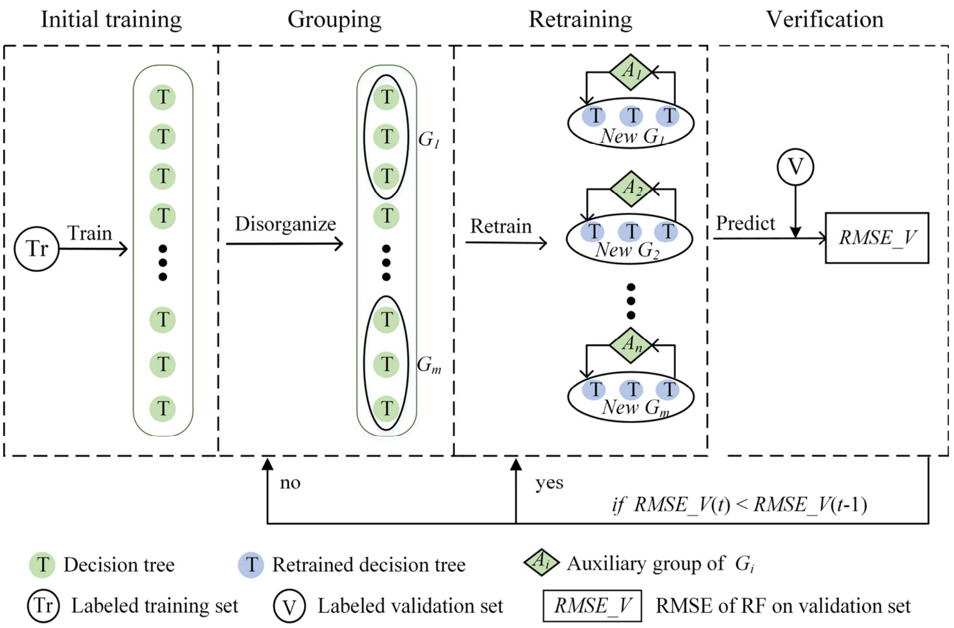

Figure 3. E-CoGRF algorithm process.

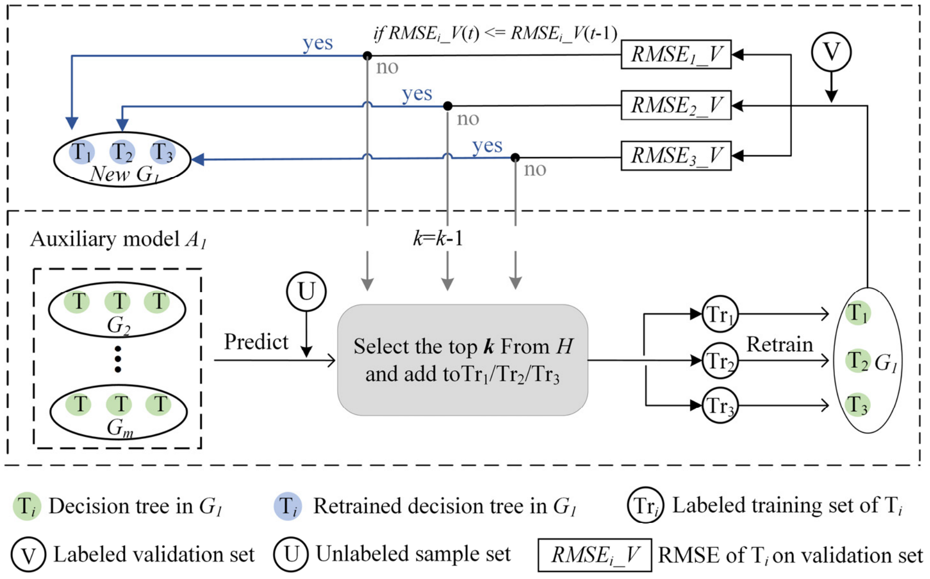
if $RMSE_i\_V(t) <= RMSE_i\_V(t-1)$
yes
no
$RMSE_1\_V$
$RMSE_2\_V$
$RMSE_3\_V$
V
New G_1
Auxiliary model A_1
$k=k-1$
G_2
G_m
U
Predict
Select the top k From H and add toTr_1/Tr_2/Tr_3
Retrain
Tr_1
Tr_2
Tr_3
G_1
T_i Decision tree in G_1
T_i Retrained decision tree in G_1
Tr_i Labeled training set of T_i
Labeled validation set
Unlabeled sample set
$RMSE_i\_V$ RMSE of T_i on validation set

Figure 4. Retraining process of G_1.

After the random forest is trained and randomly grouped, each group is retrained with the help of its auxiliary model (Figure 4). Verify whether the prediction error of the retrained random forest on the validation set decreases, that is, whether $RMSE\_V(t) < RMSE\_V(t-1)$ is correct. If it is true, repeat the Retraining-Verification process until the max iteration is reached; otherwise, regroup and then perform the Retraining-Verification process.

Figure 4 describes the retraining process of G_1 in Figure 3 in detail. In this process, G_1 is the main model, and other groups constitute the auxiliary model A_1. All auxiliary decision trees in A_1 predict and measure the confidence of all unlabeled samples. The k samples with the highest confidence are selected and added to the training set of all decision trees in G_1. Verify whether the addition of k samples reduces the prediction error of the main decision tree on the verification set. That is, for the main decision tree T_i, judge whether $RMSE_i\_V(t) < RMSE_i\_V(t-1)$ is correct. If not, execute k-1, retrain and verify. In the end, *New* G_1 is obtained with the assistance of A_1. See Table 4 for the specific algorithm.

Table 4. The E-CoGRF pseudo-code.

```
Algorithm: E-CoGRF

1: Initialize: the labeled training set Tr, the labeled validation set V, the unlabeled sample set U, max iteration $T$, the number of decision trees $N$, the number of groups $m$, the number of decision trees within a group $n=N/m$.
2: Train: Construct a Random Forest consisting $N$ Decision Tree, $\{T_1, T_2, \dots, T_N\}$
3: Group: Randomly divide all trees into m groups equally, $\{G_1, G_2, \dots, G_m\}$
Construct Auxiliary groups for each group, $\{A_1, A_2, \dots, A_m\}$
4:     Repeat for $T$ round:
5:         for $i \in \{1,2,\dots,m\}$
6:             for each $x \in U$ do
7:                 for each $T_j \in A_i$, where $j \in \{1,2,\dots,N-n\}$, do
8:                     predict: $\hat{y}_j \leftarrow T_j(x)$
9:                 end of for
10:                Mean: $\hat{y}_u = \sum_{j=1}^{N-n} \hat{y}_j / (N-n)$
11:                Confidence: $p_j = \hat{y}_j / \sum_{j=1}^{N-n} \hat{y}_j$
12:                        $H(x) = -\sum_{j=1}^{N-n} p_j * logp_j$
13:            end of for
14:             Sort $H$ in descending order
15:             Select top $k$, $X=\{x_1, x_2, \dots, x_k\}$, $Y=\{\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_k}\}$
16:             Add: $Tr_i' \leftarrow Tr_i \cup (X, Y)$
17:             Retrain: $T_i' \leftarrow learnTree(Tr_i')$
18:             Verify: $RMSE_i\_V(t) \leftarrow T_i'(V)$
19:                 if $RMSE_i\_V(t)$ >= $RMSE_i\_V(t-1)$
20:                     $k=k$-1, repeat step 15-19
21:           end of for
22:           Verify: $RMSE\_V(t) \leftarrow RF(V)$
23:               if $RMSE\_V(t)$ >= $RMSE\_V(t$-1)
24:                   Regroup and Reconstruct the auxiliary model
25:      end of Repeat
26: Output: Regressor $RF^*(x) = \frac{1}{n}(T_1(x) + T_2(x) + \cdots + T_N(x))$
```

5. Results and discussion

5.1. Model training strategy

The total number of samples used in the study is 115, including 35 labeled (HAMD score) samples and 80 unlabeled samples. In order to obtain general results, we perform 5 fold cross-validation on 35 labeled samples. One of the folds (7 samples) is used as the test set to evaluate the model each time. The rest are used as the training set to train the model. In the iterative training process, 80 unlabeled samples are reused to achieve semi-supervised learning. Finally, five evaluation models are obtained, and the evaluation results of the five models are averaged to obtain the final evaluation value.

In E-CoGRF algorithm, the number of decision trees in random forest is set to 12, and the max number of iterations is set to 100.

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to evaluate the results of the model.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y^{(i)} - \hat{y}^{(i)}\right| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2} \tag{6}$$

where, *m* is the sample size, $y^{(i)}$ is the actual value of the *i-th* sample, $\hat{y}^{(i)}$ is the predicted value of the *i-th* sample.

5.2. Comparison of different number of groups

In order to obtain the best grouping effect, we compare the prediction accuracy of the proposed E-CoGRF algorithm for HAMD scores when the grouping number is 2, 3, 4, and 6, as shown in Table 5 below. Among them, when the number of groups is 4, the lowest prediction error and the largest improvement range are obtained. Therefore, the optimal number of groups of 12 decision trees is 4.

Table 5. Performance comparison of E-CoGRF with different number of groups.

groups	MAE			RMSE		
	Initial	Final	Improved	Initial	Final	Improved
2	3.97 ± 0.64	3.82 ± 0.63	3.78%	4.97 ± 0.91	4.73 ± 0.88	4.83%
3	3.88 ± 0.72	3.7 ± 0.65	4.64%	4.91 ± 0.79	4.63 ± 0.75	5.70%
4	**4.15 ± 0.89**	**3.63 ± 0.65**	**12.53%**	**5.12 ± 1.04**	**4.50 ± 0.97**	**12.11%**
6	4.09 ± 0.51	3.78 ± 0.44	7.58%	5.06 ± 0.77	4.71 ± 0.73	6.92%

Note: Initial: the prediction error of the model trained on the initial labeled sample set.

Final: the prediction error of the model after 100 iterations.

Improved: the improvement rate of 'Final' compared to 'Initial'.

5.3. Comparison of CoRF and E-CoGRF

The following experimental analysis is performed with 12 decision trees divided into 4 groups, and the number of decision trees within each group is 3.

E-CoGRF is compared with CoRF, a co-training random forest regression algorithm, as shown in Table 6. It can be seen from the table that both the CoRF and E-CoGRF improve the evaluation accuracy of HAMD scores, which shows that they can effectively use unlabeled samples for regression estimation. Further analysis shows that compared with CoRF, E-CoGRF has lower error and shorter runtime. This indicates that introduction of the grouping strategy and the information entropy strategy effectively improves the performance of the model.

Figure 5 draws the error curves of the two algorithms in iteration process. It can be intuitively seen that E-CoGRF achieves the best convergence accuracy for both MAE and RMSE.

Table 6. Performance comparison between CoRF and E-CoGRF.

	MAE			RMSE			Runtime(*s*)
	Initial	Final	Improved	Initial	Final	Improved	
CoRF	4.22 ± 0.74	3.76 ± 0.60	10.90%	5.34 ± 0.91	4.82 ± 0.84	9.74%	7.77×10^2
E-CoGRF	**4.15 ± 0.89**	**3.63 ± 0.65**	**12.53%**	**5.12 ± 1.04**	**4.50 ± 0.97**	**12.11%**	$\mathbf{1.12 \times 10^2}$

Note: Initial: the prediction error of the model trained on the initial labeled sample set.

Final: the prediction error of the model after 100 iterations.

Improved: the improvement rate of 'Final' compared to 'Initial'.

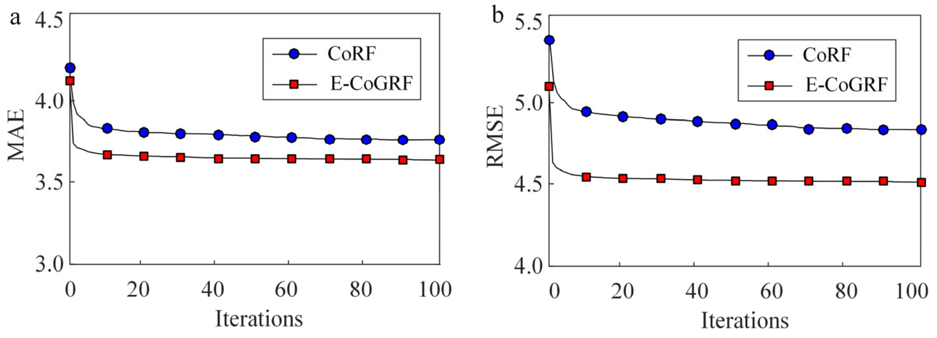

Figure 5. Error convergence curves in iteration process.

5.4. Comparison with other methods

In order to further verify the performance of the proposed E-CoGRF, five semi-supervised regression algorithms are applied to evaluate HAMD scores of depression. The comparative algorithms include co-training methods based on different basic learners (#1–#4) and a self-training method based on random forest (#5). As shown in Table 7.

Among the various co-training methods, CoBCReg achieves the best accuracy, which reveals the important role of ensemble learning in semi-supervised co-training regression. Compared with RBFNN (Radial Basis Function Neural Networks), KNN (K-Nearest Neighbor), SVR (Support Vector Regression), PLS (Partial Least Squares Regression) and other basic learners, RF shows better performance (#5). The proposed algorithm combines RF with co-training, and achieves better performance than the other five algorithms.

Table 7. Comparison of evaluation results of several semi-supervised regression methods.

#	Method	Regressor	MAE	RMSE
1	CoBCReg (2009) [28]	RBFNN	4.07	4.68
2	SVR-KNN CoREG (2014) [25]	SVR and KNN	4.63	5.51
3	COSVR (2019) [33]	SVR	5.22	5.85
4	Co-training PLS (2020) [27]	PLS	4.13	4.81
5	Self-training RF (2017) [30]	RF	3.97	4.54
6	**Proposed**	**RF**	**3.63**	**4.50**

6. Conclusions

In this study, a semi-supervised random forest regression model based on co-training and grouping with information entropy is proposed and applied to evaluate depression symptoms severity. Finally, the evaluation error of MAE = 3.63 and RMSE = 4.50 is obtained on 35 labeled (HAMD-17 scores) and 80 unlabeled samples.

Firstly, the conflict/non-conflict decision-making tasks of two different emotions, happiness and sadness, are constructed based on the basic characteristics of depression affective disorders. Then four kinds of decision-making behavior data (reaction time) of conflict/non-conflict/conflict monitoring/conflict resolution are collected. On this basis, 36 features are constructed according to the statistical distribution. And then the features are normalized.

Secondly, random forest is introduced into semi-supervised co-training algorithm. On this basis, a grouping strategy is proposed to balance the accuracy and diversity of the model, and a confidence measure method based on information entropy is proposed to reduce unnecessary repeated training in the model. Compared with CoRF, E-CoGRF achieves the lowest error and the shortest runtime.

In addition, compared with other semi-supervised regression algorithms in recent years, E-CoGRF achieves better evaluation performance. This paper provides an effective method to solve the problem of less labeled samples and more unlabeled samples, and provides technical support for the evaluation of depression symptoms severity.

The automatic evaluation method of HAMD proposed in this study not only solves the problem of low efficiency of manual evaluation in clinical, but also improves the evaluation accuracy.

The work only focuses on the evaluation of depression symptoms severity in patients with depression, but does not involve an automatic evaluation method for depression risk. Moreover, due to the limitations of data type, deep learning method is not applied. Based on this, we will further explore this issue by using facial expression, speech, language and other features acquired under emotional stimulation and deep learning methods in the future.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61602017), the National Basic Research Programme of China (2014CB744600), “RixinScientist” Foundation of

Beijing University of Technology (2017-RX(1)-03), the Beijing Natural Science Foundation (4164080), the Beijing Outstanding Talent Training Foundation (2014000020124G039), the National Natural Science Foundation of China (61420106005), the International Science & Technology Cooperation Program of China (2013DFA32180), the Special Fund of Beijing Municipal Science and Technology Commission (Z171100000117004), Beijing Hospitals Authority Youth Programme (QML20181904), Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support (ZYLX201607) and Beijing Municipal Administration of Hospitals' Ascent Plan (DFL20151801).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. G Casalino, G Castellano, F Galetta, K. Kaczmarek-Majer, Dynamic incremental semi-supervised fuzzy clustering for bipolar disorder episode prediction, in *International Conference on Discovery Science*, Springer, Cham, (2020), 79–93.
2. J. C. Wakefield and S. Demazeux, *Introduction: Depression, one and many*, Sadness or Depression?, Netherlands, Springer, 2016, 1–15.
3. M. E. Gerbasi, A. Eldar-Lissai, S. Acaster, M. Fridman, V. Bonthapally, P. Hodgkins, et al., Associations between commonly used patient-reported outcome tools in postpartum depression clinical practice and the Hamilton Rating Scale for Depression, *Arch. Women's Mental Health*, **23** (2020), 727–735.
4. C. L. Allan, C. E. Sexton, N. Filippini, A. Topiwala, A. Mahmood, E. Zsoldos, et al., Sub-threshold depressive symptoms and brain structure: A magnetic resonance imaging study within the Whitehall II cohort, *J. Affective Disord.*, **204** (2016), 219–225.
5. X. Li, Z. Jing, B. Hu, J. Zhu, N. Zhong, M. Li, et al., A resting-state brain functional network study in MDD based on minimum spanning tree analysis and the hierarchical clustering, *Complexity*, **2017** (2017), 9514369.
6. K. Yoshida, Y. Shimizu, J. Yoshimoto, M. Takamura, G. Okada, Y. Okamoto, et al., Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional MRI data with partial least squares regression, *Plos One*, **12** (2017), e0179638.
7. S. Sun, X. Li, J. Zhu, Y. Wang, R. La, X. Zhang, et al., Graph theory analysis of functional connectivity in major depression disorder with high-density resting state EEG data, *IEEE Trans. Neural Syst. Rehabil. Eng.*, **27** (2019), 429–439.
8. U. R. Acharya, S. L. Oh, Y Hagiwara, J. Tan, H. Adeli, D. P. Subha, Automated EEG-based screening of depression using deep convolutional neural network, *Comput. Methods Prog. Biomed.*, **161** (2018), 103–113.
9. R. W. Lam, S. H. Kennedy, R. S. McIntyre, A. Khullar, Cognitive dysfunction in major depressive disorder: effects on psychosocial functioning and implications for treatment, *Can. J. Psychiatry*, **59** (2014), 649–654.

10. R. S. McIntyre, D. S. Cha, J. K. Soczynska, H. O. Woldeyohannes, L. A. Gallaugher, P. Kudlow, et al., Cognitive deficits and functional outcomes in major depressive disorder: determinants, substrates, and treatment interventions, *Depression Anxiety*, **30** (2013), 515–527.
11. Y. Kang, X. Jiang, Y. Yin, Y. Shang, X. Zhou, Deep transformation learning for depression diagnosis from facial images, in *Chinese Conference on Biometric Recognition,* Springer, Cham, (2017), 13–22.
12. A. Haque, M. Guo, A. S. Miner, F. Li, Measuring depression symptom severity from spoken language and 3D facial expressions, preprint, arXiv:1811.08592.
13. M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, A. Othmani, AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis, *Mach. Learn. Appl.*, **2** (2020), 100005.
14. J. Zhu, J. Li, X. Li, J. Rao, Y. Hao, Z. Ding, et al., Neural basis of the emotional conflict processing in major depression: ERPs and source localization analysis on the N450 and P300 components, *Front. Human Neurosci.*, **12** (2018), 214.
15. B. W. Haas, K. Omura, R. T. Constable, T. Canli, Interference produced by emotional conflict associated with anterior cingulate activation, *Cognit. Affective Behav. Neurosci.,* **6** (2006), 152–156.
16. T. Armstrong, B. O. Olatunji, Eye tracking of attention in the affective disorders: a meta-analytic review and synthesis, *Clin. Psychol. Rev.*, **32** (2012), 704–723.
17. A. Duque, C. Vázquez, Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study, *J Behav. Ther. Exp. Psychiatry*, **46** (2015), 107–114.
18. S. P. Karparova, A. Kersting, T. Suslow, Disengagement of attention from facial emotion in unipolar depression, *Psychiatry Clin. Neurosci.*, **59** (2005), 723–729.
19. M. P. Caligiuri, J. Ellwanger, Motor and cognitive aspects of motor retardation in depression, *J. Affective Disord.*, **57** (2000), 83–93.
20. A. Etkin, T. Egner, D. M. Peraza, E. R. Kandel, J. Hirsch, Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala, *Neuron*, **51** (2006), 871–882.
21. K Mohan, A Seal, O Krejcar, A. Yazidi, FER-net: facial expression recognition using deep neural net, *Neural Comput. Appl.*, (2021), 1–12.
22. K Mohan, A Seal, O Krejcar, A. Yazidi, Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks, *IEEE Trans. Instrum. Meas.*, **70** (2020), 1–12.
23. Z. Zhou, M. Li, Semi-supervised regression with co-training, in *IJCAI*, (2005), 908–913.
24. M. A. Lei, W. Xili, Semi-supervised regression based on support vector machine co-training, *Comput. Eng. Appl.*, **47** (2011), 177–180.
25. Y. Q. Li, M. Tian, A semi-supervised regression algorithm based on co-training with SVR-KNN, in *Advanced Materials Research,* Trans Tech Publications Ltd, (2014), 2914–2918.
26. L. Bao, X. Yuan, Z. Ge, Co-training partial least squares model for semi-supervised soft sensor development, *Chemom. Intell. Lab. Syst.*, **147** (2015), 75–85.
27. D. Li, Y. Liu, D. Huang, Development of semi-supervised multiple-output soft-sensors with Co-training and tri-training MPLS and MRVM, *Chemom. Intell. Lab. Syst.*, **199** (2020), 103970.

28. M. F. A. Hady, F. Schwenker and G. Palm, Semi-supervised learning for regression with co-training by committee, in *International Conference on Artificial Neural Networks,* Springer, Berlin, Heidelberg, (2009), 121–130.
29. F. Saitoh, Predictive modeling of corporate credit ratings using a semi-supervised random forest regression, *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM*), IEEE, (2016), 429–433.
30. J. Levatić, M. Ceci, D. Kocev, S. Džeroski, Self-training for multi-target regression with tree ensembles, *Knowl. Based Syst.*, **123** (2017), 41–60.
31. S. Xue, S. Wang, X. Kong, J. Qiu, Abnormal neural basis of emotional conflict control in treatment-resistant depression: An event-related potential study, *Clin. EEG Neurosci.*, **48** (2017), 103–110.
32. N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. A. Hare, et al., The NimStim set of facial expressions: Judgments from untrained research participants, *Psychiatry Res.*, **168** (2009), 242–249.
33. M. Lei, J. Yang, S. Wang, L. Zhao, P. Xia, G. Jiang, et al., Semi-supervised modeling and compensation for the thermal error of precision feed axes, *Int. J. Adv. Manuf. Technol.*, **104** (2019), 4629–4640.