The early prediction of AD evolution based on Trusted-LGBM

Pan Hao1*, Jiyun Li1
1School of Computer Science and Technology, Donghua University, Shanghai, 201620, China
*Corresponding author’s e-mail: 2191966@mail.dhu.edu.cn

Abstract. In order to achieve the purpose of early diagnosis of Alzheimer's disease, this study uses machine learning technology to process 14,659 original data from the public data set ADNI. First use the KNN algorithm to fill in the null value. Then a feature transformation method is proposed to transform regression problems into classification problems. Then this study calculates the trust level of each sample based on the feature correlation and the weight of null value, and proposes an improved LightGBM algorithm-Trusted-LGBM. Finally, the prediction results of the Trusted-LGBM algorithm are compared with SVM, XGBoost, and the unimproved LightGBM algorithm. The experimental results show that the Trusted-LGBM algorithm proposed by this research has a higher F1-score (0.784) and AUC (0.91). It can be seen that the method proposed in this study can more effectively support the early diagnosis of Alzheimer's disease.

1. Introduction

According to a survey by the International Alzheimer's Association[1]: In 2020, the number of Alzheimer's patients worldwide has reached 54 million, As of 2019, China has more than 10 million Alzheimer's patients, ranking first in the world, accounting for about 25% of the total number of Alzheimer's patients in the world[2], and it is expected to exceed 30 million by 2050. Alzheimer's disease is an irreversible neurodegenerative disease. Its development includes three stages: normal cognition (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). The early diagnosis of Alzheimer's disease is to detect the development trend of Alzheimer’s disease as early as possible, and to be able to carry out relevant treatment and intervention in time, which can prevent or slow down the development of AD. Therefore, the early auxiliary diagnosis for AD is particularly important.

In recent years, an increasing number of scholars in China and abroad have devoted themselves to the early diagnosis of AD, focusing on the analysis of neuroimaging or biomarkers of different modalities to assist in the diagnosis of AD [3-5]. Donohue et al. used the method of observing the cohort at different stages of the disease to study the slow development of the disease [6]; Sabuncu et al. proposed a method for statistical analysis of the correlation between longitudinal neuroimaging measurements and the time of related clinical events [7]; Schmidt-Richberg et al. proposed a method to estimate the progression of the disease [8]. They all use time series data to train their own models, but Alzheimer’s disease has a long development cycle and it is difficult to obtain a large amount of complete time series data. Using a small amount of time series data to train these models will lead to certain limitations. And it is easy to produce over-fitting, resulting in low accuracy of model prediction.

In order to make full use of the existing time series data in the ADNI data set, this paper proposes a Trusted-LGBM model based on the LightGBM algorithm. The sample trust level is calculated by using
the weights of null values obtained by the null-filling algorithm and the feature correlation obtained by the feature selection algorithm. The comparative experiment proves that the model can predict the trend of AD lesions more accurately in the early diagnosis.

2. Research methods

2.1. Question definition

The vertical data of AD subjects used in this paper are characterized by a small number of observations and uneven observation intervals. If a traditional regression model is used to predict their disease development, it will have an impact on the prediction results. Therefore, this paper proposes a feature transformation method to extract the association information before and after the original features as new features to generate a dataset that can be used for training of classification models, and to transform the regression problem of AD disease development into a classification problem, which can be modeled in the following form.

\[ t: S_{it}(C_i, V_{it}, y_{it}) \rightarrow t': S_{it'}(C_i, V_{it'}, y_{it'}) \]  \hspace{1cm} (1)

\( S_{it} \) represents the \( i \)-th subject at time \( t \). Its attributes contain the time-independent part \( C \) and the time-varying part \( V \). \( y_i \) denotes the type of subject at time \( t \). The transformed \( V_{it} \) is the change function of the variable part \( V \) with respect to time \( t \). Define the original dataset as \( D_{original} = \{X_i, y_i|i \in [1,n]\} \), \( X_i \) is the \( i \)-th feature vector in the dataset, and \( y_i \) is the corresponding label. The transformation process can be expressed as follows:

\[ D_{original} = \begin{bmatrix} x_{i1} & \ldots & x_{ip} & x_{i1+p+1} & \ldots & x_{im} & y_{i1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \ldots & x_{ip} & x_{i1+p+1} & \ldots & x_{im} & y_{i1} \\ x_{i+1} & \ldots & x_{ip} & x_{i+1+p+1} & \ldots & x_{i+m} & y_{i+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} & x_{n+p+1} & \ldots & x_{nm} & y_{n} \end{bmatrix} \]  \hspace{1cm} (2)

where \( x_{i1}, \ldots, x_{ip} \) is the time-independent part \( C_i \), \( x_{ip+1}, \ldots, x_{im} \) is the time-varying \( V_{it} \). Let \( X_i, X_{i+1}, X_{i+2} \) be a set of feature vectors with correlations (i.e., vertical data belonging to the same subject in the dataset), \( t \) be a time-point feature, \( j \in [p+1, m], j \in [i, i+1] \). Then, for variable features and labels there is the following transformation relation:

\[ \Delta x_{kj} = \frac{x_{kj} - x_{kj}}{t_{k+1} - t_k} \]  \hspace{1cm} (3)

\[ y_k' = \begin{cases} 0, & y \text{the last in group } < 2 \\ 1, & y \text{the last in group } \geq 2 \end{cases} \]  \hspace{1cm} (4)

This set of feature vectors is calculated according to equations (3) and (4) to obtain the new vector \( X'_i, X'_{i+1} \), thus transforming the original dataset into a classification-friendly dataset \( D_{new} \):

\[ \begin{bmatrix} X_i \\ X_{i+1} \\ X_{i+2} \end{bmatrix} \rightarrow \begin{bmatrix} X'_i \\ X'_{i+1} \\ X'_{i+2} \end{bmatrix} = \begin{bmatrix} x_{i1} & \ldots & x_{ip} & x_{i1+p+1} & \ldots & x_{im} & \Delta x_{i,p+1} & \ldots & \Delta x_{i,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \ldots & x_{ip} & x_{i1+p+1} & \ldots & x_{im} & \Delta x_{i+1,p+1} & \ldots & \Delta x_{i+1,m} \end{bmatrix} \]  \hspace{1cm} (5)

\[ D_{new} = \begin{bmatrix} X'_i \\ X'_{i+1} \\ \vdots \\ X'_{i+2} \end{bmatrix} = \begin{bmatrix} y'_i \\ y'_{i+1} \\ \vdots \\ y'_{i+2} \end{bmatrix} \]  \hspace{1cm} (6)

The feature transformation method generates a new dataset based on the changes in attributes based on the original vertical data, which retains the original diagnostic results and also expresses the changes in each examination index explicitly, making it an important discriminative feature in the classification model.
2.2. Improved LightGBM model

2.2.1 LightGBM algorithm principle. LightGBM (Light Gradient Boosting Machine) is a landed implementation of the GBDT (Gradient Boosting Decision Tree) algorithm released by Microsoft Research Asia in 2016\[9,10\]. The main idea of GBDT is to use CART (Classification And Regression Trees) as a weak learner to continuously iterate training to get the optimal model, so GBDT can be represented in the form of an additive model of decision tree, let the gradient boosting tree contain $n$ CART trees and $\theta_i$ denotes the parameter of the ith decision tree, then the following expression is given:

$$f_i(x) = \sum_{i=1}^{n} F(x; \theta_i)$$

(7)

According to the forward iteration algorithm, the model undergoes the ith iteration can be expressed as:

$$f_i(x) = f_{i-1}(x) + F(x; \theta_i)$$

(8)

If the loss function is taken to be the squared loss and let $y_j$ be the label value of sample $x_j$, then we have the following loss function:

$$L(f_i(x_j); y_j) = \frac{(y_j - f_i(x_j))^2}{2}$$

(9)

Optimizing equation (9) yields the parameters of the ith decision tree:

$$\theta'_i = \arg\min_{\theta_i} \sum_{j=1}^{n} L(f_{i-1}(x) + T(x; \theta_i); y_j)$$

(10)

The final classification model is obtained by updating the parameters $\theta_i$ of each regression tree in continuous iterations.

2.2.2. Improved model based on sample trust level. The ADNI dataset contains null values. In order to utilize as much as possible the test results obtained from each visit of the subject, we used the KNN algorithm to fill the null values.

But, it also introduces another problem: when the model is trained, the original data and training samples containing non-original data computed according to the KNN algorithm are treated equivalently by the classifier, which will affect the uncertainty of the model prediction results. In Bayesian linear regression, for example, the exponent can be multiplied by the trust level of the sample $w$:

$$p(y|x, \theta, \sigma) \propto e^{-\frac{1}{2\sigma^2}(y - f(x))^2}$$

(11)

Here $\sigma$ is the uncertainty inherent in the objective function. $w = 1$ indicates a normal level of trust and the model can trust this sample completely, and a lower $w$ value characterizes an increase in uncertainty. So, an example with a lower sample trust level $w$ setting would result in a wider distribution due to uncertainty about the target value. If this information is fed back into the model, a corrected loss function is obtained:

$$loss = \sum w_i * L(x_i, y_i)$$

(12)

Here, $w_i$ is the trust level of the ith sample. Based on the above theory and the tree building process of LightGBM, this paper proposes a method to enhance the LightGBM algorithm by introducing the sample plausibility during model training. Using the weighted training property of the LightGBM algorithm, the trust level of the samples is finally transformed into a weighting of the loss function. By setting the trust level $w_i (w_i \in [0,1])$ for each sample to be trained and using the gradient of $w_i$ times the loss function and the hessian (second order derivative) to obtain a new gradient with hessian to influence the splitting of the nodes during the tree construction \[9\]. This allows the model to perceive the difference in uncertainty between samples. The sample trust level is mainly influenced by two factors: the number of null-valued features contained in the original sample and the importance of that null-valued feature. Let the number of samples in the dataset be $n$, and each sample $X_i$ contains $m$ features, the following algorithm is used in this paper to calculate the sample trust level $w_i$:

$$e_{ij} = \begin{cases} 1, & x_{ij} is empty \\ 0, & x_{ij} is not empty \end{cases}$$

$$w_i = 1 - \frac{\sum_{j=1}^{m} (e_{ij} * l_{ij})}{m}$$

(13)

(14)
where $i \in [1, n], j \in [1, m]$, $e_{ij}$ indicates whether this sample feature $x_{ij}$ is empty before the null filling using the KNN algorithm, and $I_{ij}$ is the normalized value of the feature importance given by the random forest algorithm model.

3. Experiment and Analysis

The data of AD subjects are obtained from the public database ADNI (Alzheimer's Disease Neuroimaging Initiative), and the data set to be trained is obtained through data processing operations such as extraction, cleaning and transformation. The prediction results were evaluated in comparison with other algorithms.

3.1. Data pre-processing

3.1.1. Data cleaning and feature selection. For the purpose of early prediction, subjects with baseline diagnosis of AD and those who had only one diagnosis in the original dataset should be removed first, and the resulting dataset contains 848 subjects with a total of 3061 examination records. Since more of these examination items have null values, this paper uses the KNN(k-Nearest Neighbors) algorithm in machine learning to fill the null values.

In this paper, a feature selection method based on Random Forest algorithm was used to extract the top 10 features with high correlation with Alzheimer’s disease condition, and the importance of each feature after data normalization is shown in Table 1.

| Features name             | Features’ importance | rank |
|---------------------------|----------------------|------|
| MMSE                     | 0.0653               | 1    |
| CDRSB                    | 0.0646               | 2    |
| Hippocampus/WholeBrain   | 0.0645               | 3    |
| MOCA                     | 0.0633               | 4    |
| AGE                      | 0.0620               | 5    |
| APOE4                    | 0.0618               | 6    |
| PTEDUCAT                 | 0.0614               | 7    |
| TAU                      | 0.0602               | 8    |
| PTAU                     | 0.0601               | 9    |
| ABETA                    | 0.0591               | 10   |

3.1.2. Feature transformation.

The vertical data of AD subjects are grouped by subject’s id and feature transformed using the method proposed in Section 2.1 in this paper. The steps are as follows. The demographic features of the newly generated dataset are shown in Table 2.

| Label                             | Range of age | Years of education | Gender |
|-----------------------------------|--------------|--------------------|--------|
| Eventual deterioration to AD      | 55–91        | 8–20               | 415    | 658    |
| Stable or returning to normal condition | 55–90     | 6–20               | 573    | 848    |

3.2. Evaluation Indicators

In this paper, $F1$ – score and ROC are used to evaluate the prediction results of early AD lesions obtained from different models. $F1$ – score is the summed average of the precision and recall rates, and the horizontal coordinate of the ROC is the false positive rate ($FPR$) and the vertical coordinate is the true rate ($TPR$), which is defined as follows:
\[ F1 - score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (15) \]
\[ \text{precision} = \frac{TP}{TP + FN}, \text{recall} = \frac{TP}{TP + FN} \]
\[ FPR = \frac{FP}{TP + FN}, TPR = \frac{TP}{TP + FN} \quad (16) \]

where \( TP \) denotes samples predicted positive by the model, \( FP \) denotes negative samples predicted positive by the model, \( FN \) denotes positive samples predicted negative by the model, and \( TN \) denotes negative samples predicted negative by the model.

### 3.3. Experimental results and analysis

In order to demonstrate that the improvement method proposed in this paper for LightGBM on AD early prediction problem makes the model more robust and superior in classification effect, this paper uses SVM algorithm model and XGBoost algorithm model on the same dataset for comparison experiments. The results are shown in Table 3.

| Model       | \( E(F1 - score) \) |
|-------------|---------------------|
| SVM         | 0.698               |
| XGBoost     | 0.727               |
| LightGBM    | 0.731               |
| Trusted-LGBM| 0.784               |

The experimental results from Table 4 show that the \( E(F1 - score) \) of the Trusted-LGBM algorithm using the method in this paper is 0.784, which is better than the remaining three control groups and improves by 7.25 percentage points compared to LightGBM. The ROC curves of the four models were plotted according to the definitions in Section 3.2, see Figure 1. It can be seen from the figure that the Trusted-LGBM model has a better classification performance with an AUC value of 0.91, which is 0.1 improvement over the LightGBM model.

The analysis of the experimental results shows that the Trusted-LGBM algorithm, which is improved for early AD lesion prediction, outperforms other classification algorithms in terms of prediction performance and evaluation results, and proves the effectiveness and feasibility of the sample trust level calculation method proposed in this paper.
4. Conclusions
Early diagnosis of Alzheimer's disease has been a major problem in the medical community, and its etiology is complex and hidden. In this paper, we analyze the longitudinal data of AD subjects through machine learning techniques to filter out the features most closely related to Alzheimer's disease. And this study proposes an improved LightGBM algorithm model, which introduces sample trust level in the training process, thereby enhancing the model’s tolerance to sample null values and improving the model’s classification accuracy. Finally, a series of comparison experiments designed in this paper showed that the method can effectively predict the early lesion trend of Alzheimer's disease. With the increasing data of AD subjects, incremental training of the model with supporting incremental input will be the focus of future research in this paper.

References
[1] Alzheimer’s Disease International. (2018) World Alzheimer Report 2018. https://www.alzint.org/resource/world-alzheimer-report-2018/
[2] Nichols E., Szoeke C., Vollset S., et al. (2019) A systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology.,18(1):88-106.
[3] Davatzikos C., Bhatt P., Shaw L.M., et al. (2013) Prediction of MCI toAD conversion, viaMRI, CSF biomarkers, and pattern classification. J. Neurobiol Aging., 32(12): 19-27.
[4] GARCIA-RIBAS G., ARBIzu J., CARRiO I., et al. (2017) PET biomarkers: use of imaging techniques in Alzheimer disease and neurodegenerationclinical diagnosis. J. Neurologia.,32(5):275-277.
[5] Ruan Q., D'onofrio G., Sancarlo D., et al. (2016) Potential neuroimaging biomarkers of pathologic brain changes in mildcognitive impairment and Alzheimer's disease: a systematic review. J. BMC Geriatr.,16(1):104.
[6] Michael C.D., Hélène J.G., Mélanie L.G., Ronald G.T., et al. (2014) Estimating long-term multivariate progression from short-term data. J. Alzheimer's Dementia: The Journal of the Alzheimer's Association., 10(5):64-78.
[7] Mert R.S., Jorge L.B., Martin R., et al. (2014) Event time analysis of longitudinal neuroimage data. J. NeuroImage., 97(2):9-18.
[8] Schmidt-Richberg A, Ledig C, Guerrero R, Molina-Abril H, Frangi A, Rueckert D. (2016) Learning Biomarker Models for Progression Estimation of Alzheimer's Disease. J. Plos One., 11(4):0153040.
[9] Yin D., Nobata C., Langlois J.M., et al. (2016) Ranking Relevance in Yahoo Search. In: Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. pp. 178-187