Diagnosis of Dementia and Alzheimer's Disease Based on Classification Algorithms

Ci Song[1], Shuxian Zong[2]

1. School of Resources and Materials, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China.

2. Nanfang College Guangzhou, Guangzhou 510970, China.

Abstract: Alzheimer's disease is currently the most common kind of senile dementia. With the increasing aging degree of the global society, Alzheimer's disease will become an unavoidable social problem in an aging society. In order to improve this situation, artificial intelligence algorithms that are good at mining the internal laws of data are applied in the hope of more effectively diagnose this disease, which should be intervened as early as possible. After briefly restating the current situation of dementia and Alzheimer's disease, the diagnostic model for dementia is built using logistic regression, which achieves great accuracy despite the simplicity of the model. Then, two diagnostic models that can identify if the patient with dementia has Alzheimer's disease based on SVM and Random Forest are tested. Although both the algorithms perform poorly because of the sample imbalance, after processing the original data with SMOTE, their performances are largely improved.

Keywords: Machine Learning; Smote; Dementia Diagnosis

1. Introduction

Alzheimer's disease is a recognized public health problem for the GWHO. The number of Alzheimer's patients worldwide has exceeded 34 million in 2020, and is expected to reach 36.3 million in 2025. At present, the diagnostic method for Alzheimer's disease is mainly exclusion, the challenge of which is that the diseases of most patients are already advanced when explicitly diagnosed. researchers hope that new AI technology can change this situation by strengthening the effective data analysis of AD. Currently, machine learning and deep learning have been widely used in a variety of complex challenges, like medical, financial, industrial, and other fields, which may help improve the reliability, performance, predictability, and accuracy of diagnostic systems for Alzheimer's disease.

The motivation of this paper is that, considering traditional diagnostic methods tend to cause patients to miss out on optimal treatment time, new progress brought to diagnosing diseases by data analysis is crucial. With the development of algorithms that can homogenize samples, machine learning also obtain good accuracy for classification problems with sparse sample sizes such as diseases.

The main contribution of this paper is that, by first building a binary classification model based on logistic regression, an Intelligent diagnostic model for dementia can be obtained. Then, models that can distinguish whether a person with dementia has Alzheimer's disease are built based on SVM and random forest, whose performance are both poor. In order to improve poor accuracy caused by sample imbalance, SMOTE is applied to generate more scarce samples to strike a balance between two categories, which significantly improves the classification performance of SVM and random forest.

2. Overview of dementia

People with dementia and dementia tendencies are divided into several groups, namely cognitive normal elderly (CN), patients with subjective memory complaint (SMC), patients with early mild cognitive impairment (EMCI), patients with late mild cognitive impairment (LMCI) and patients with Alzheimer's disease (AD).

2.1 CN

Cognitive normal elderly can be intervened early on through psychotherapy, behavioral therapy, and drug therapy. The early manifestations of Alzheimer's disease are mainly short-term memory decline, learning ability decline and language ability decline, so for the elderly with normal cognition, prevention is the best choice.

2.2 SMC

For the older population with subjective memory complaint, subjective memory complaints predict an increase in mild cognitive impairment. As one of the manifestations of early and forgotten cognitive impairment in AD, Subjective Memory Complaint is defined as a self-memory impairment arising in the absence of both objective object signs and known pathological conditions. For SMC, the treatment is available through non-drug early intervention, which is a more feasible alternative treatment plan. This scheme can effectively release anxiety, depression and improve subjective memory, delayed memory, and overall cognitive function [1].

2.3 EMCI and LMCI

For EMCI and LMCI, which are the branches of MCI in time, are not easily distinguished from MCI in nature. Early intervention for EMCI and LMCI patients, mainly relies on nursing, and reasonable solutions should be provided for groups of different age.

For youth groups, nursing workers should intervene according to the risk factors in their age group to guide the youth group through the correct health.

In dealing with the intervention problem of MCI in the middle age population, the effective intervention of medical treatment and care are required to maintain MCI in a stable state through screening synchronization and risk factor assessment synchronization and dynamic monitoring, and reduce the tendency to shift to EMCI and LMCI.

The symptom of MCI in the old age stage is mainly reflected in the poor lifestyle, chronic disease, or the previous genetic characteristics, which can aggravate the degree of cognitive impairment. This is more likely to be LMCI, where mainly drug intervention as well as non-drug intervention are used to increase the plasticity of brain function and delay the development of AD. the diagnosis of EMCI and LMCI is like the diagnosis of MCI, which is mainly divided into the following five diagnostic criteria:

1.Relevant information vouchers provided by Informed persons such as family members.

2.Impairment of cognitive functions or other cognitive functions inconsistent with age and education level shown in objective tests.

3.Overall cognitive function is relatively intact.

4.Whether the function of daily life is affected.

4.Whether it meets the diagnostic criteria for dementia [2,3].

2.4 AD

For AD, the early intervention measures are mainly based on the premise of early detection and early treatment through the careful discovery of community doctors and the close relationship between doctors and patients. As for early intervention, community hospitals are very effective in detecting early AD, based on which early psychosocial intervention promotes a patient-family-centered model transformation to provide or change existing treatment options for patients. Nutritional interventions in the elderly can also reduce the risk of AD. Among the criteria for the diagnosis of AD, the best time to diagnosis is within 1-2 years of the early onset, and the early psychological and pharmacological intervention in AD patients can greatly reduce the morbidity [4].

3. Diagnostic model for dementia based on Logistic Regression

3.1 Data description: Alzheimer Features

The data for this section is collected from Kaggle [5], The dataset contains features, namely Gender, Age, Years of Education, Socioeconomic Status, Mini Mental State Examination, Clinical Dementia Rating, Estimated Total Intracranial Volume, Normalize Whole Brain Volume and Atlas Scaling Factor. The label is Demented or Nondemented, which are represented 1 and 0 in this section. The whole dataset is split into training set and test set by 7:3.

3.2 Diagnostic model based on Logistic Regression

Logistic Regression is a classification model and is commonly used for binary classification, which is popular for its simplicity, parallelizability, and interpretability.

The mathematical model is based on sigmoid function, which is of the form:

$$y = g(x) = \frac{1}{1+e^{x}}, for\ x \in (- \infty, + \infty) \qquad (1)$$

The feature of this function is that, when $x \geq 0, y \geq 0.5$. Assume y is a label from 0-1 classification, when $x > 0$, it means the data point is more likely to be classified into the category labeled 1, which gives logistic regression mathematical meaning in classification problems. The logistic regression in this case is defined as:

$$\hat{y} = f(x_1, x_2, ..., x_9) = g(\theta_0 + \theta_1 x_1 + ... + \theta_9 x_9) = g(\Theta) \qquad (2)$$

The loss function is defined as:

$$J(\Theta) = -\frac{1}{n}\sum_{i=1}^{n} (y_i \ln \hat{y}_i + (1 - y_i) \ln (1 - \hat{y}_i)) \qquad (3)$$

Which can be minimized by Gradient descent, where the update is:

$$\theta_{i+1} = \theta_i - \alpha \frac{\partial J(\Theta)}{\partial \theta} \qquad (4)$$

α is the learning rate, which is an empirically set hyperparameters [6]. Use gradient descent to iterate until convergence, then the best decision boundary in binary classification problems is obtained.

To evaluate the performance of the classification model, confusion matrix is first introduced, the form of which is shown in table 1:

Table 1: Confusion matrix in binary classification

TP	*FP*
FN	*TN*

TP and TN are the numbers of positive and negative samples that are predicted correctly, while FP and FN are the other way around. Using confusion matrix, the following indicators can be calculated:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The Intelligent diagnostic model is first trained on the training set and then tested on the test set, the confusion matrices and classification reports achieved by which on both sets are shown below:

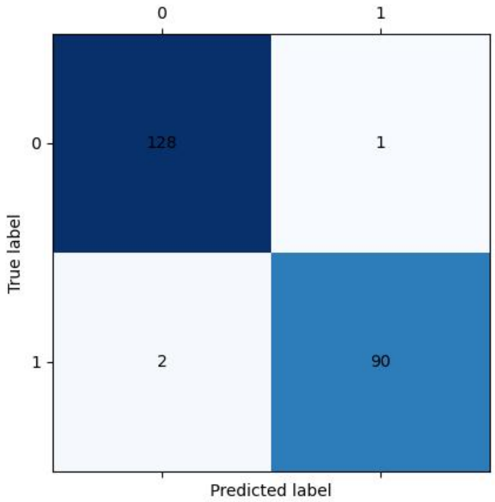


Figure 1: Confusion matrix achieved by logistic regression on training set

Table 2: Classification report of logistic regression on training set

Logistic Regression	0	1
Precision	0.98	0.99
Recall	0.99	0.98
F1-score	0.99	0.98
Accuracy	0.99	

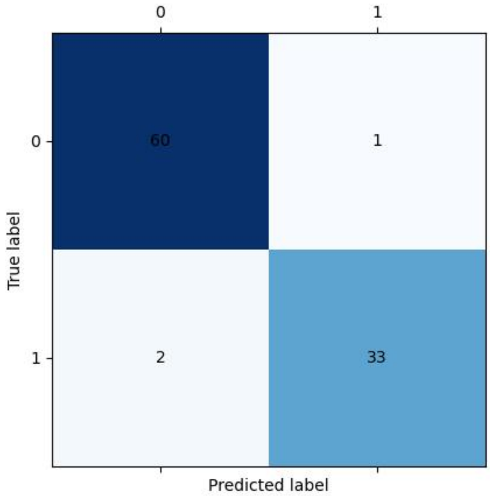


Figure 2: Confusion matrix achieved by logistic regression on test set

Logistic Regression	0	1
Precision	0.97	0.97
Recall	0.98	0.94
F1-score	0.98	0.96
Accuracy	0.97	

Table 3: Classification report of logistic regression on test set

It can be seen from the charts that the indicators obtained by logistic regression on both sets are excellent, which means no overfitting or underfitting occurs and the intelligent diagnostic model has good generalization ability.

3.3 Diagnostic models for Alzheimer's disease based on SVM and Random Forest

In this section, two binary classifier, Support Vector Machine (SVM) and Random Forest are both applied as diagnostic models for Alzheimer's disease.

Data description: Alzheimer's clinical data

The dataset is again from Kaggle [7], which contains only 5 features, namely Gender, Age, MMSE, CDR, Memory level. The labels are Uncertain dementia and Dementia, which are again represented by 0 and 1. The train-test split proportion is still 7:3.

3.4 Diagnostic model based on SVM

SVC is rather popular in binary classification; the core idea is to solve the following optimization problem:

$$minimize\ \frac{\|\vec{w}\|^2}{2}$$

$$subject\ to: g_i(\vec{w},b)=y_i\cdot(\vec{w}\cdot\vec{x}_i+b)-1\geq 0$$

$$for\ i=1,2,\ldots,n$$

Where $\vec{w}$ is the weight vector of the hyperplane equation that separates two categories and $g_i(\vec{w},b)\geq 0$ is the constraint. To solve such problem, the following Lagrange equation can be constructed:

$$L(\vec{w},b,C_i,p_i)=\frac{\|\vec{w}\|^2}{2}-\sum_{i=1}^{n}C_i\,(y_i(\vec{w}\cdot\vec{x}+b)-1-p_i^2) \qquad (5)$$

where C is the penalty factor and p_i is applied to transform constraints into equation. This equation can be solved by KKT conditions:

$$\begin{cases} \vec{w}-\sum\limits_{i=1}^{n}C_i y_i\vec{x}_i=0 \\ \sum\limits_{i=1}^{n}C_i y_i=0 \\ y_i(\vec{w}\cdot\vec{x}_i+b)-1-p_i=0 \\ C_i p_i^2=0 \\ C_i\geq 0 \end{cases}$$

In addition, SVM classification model must also use Kernel function to map data points to high-dimensional space in order to find the plane of segmentation. This paper chooses Gaussian Radial Kernel Function as the Kernel Function, which is of the form:

$$K(x_i,x_j)=\exp\{-\frac{\|x_i-x_j\|^2}{2\sigma^2}\} \qquad (6)$$

where $\gamma=\frac{1}{2\sigma^2}$ is a hyperparameter [8]. Thus, to build an SVC model, the value of C and γ are crucial. After some attempts, it turns out that SVM achieves the best performance when $C=100$ and $\gamma=5$, the confusion matrices and classification reports on both training and test set are as follow:

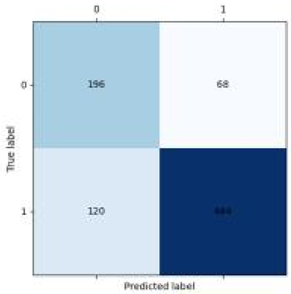


Figure 3: Confusion matrix achieved by SVM on training set

Table 4: Classification report of SVM on training set

SVM	0	1
Precision	0.62	0.87
Recall	0.74	0.79
F1-score	0.68	0.83
Accuracy	0.77	

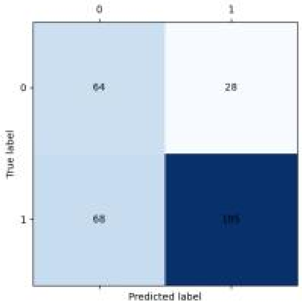


Figure 4: Confusion matrix achieved by SVM on test set

Table 5: Classification report of SVM on test set

SVM	0	1
Precision	0.48	0.87
Recall	0.70	0.74
F1-score	0.57	0.80
Accuracy	0.73	

Obviously, the classification performance achieved by SVM is not ideal enough, especially in terms of the recognition of the samples from the category labeled 0.

3.5 Diagnostic model based on Random Forest

Random forest is an extended variant of the ensemble learning method Bagging. By sampling the original training sample and selecting the feature nodes, many different trees can be obtained. The random forest is the integrated version of the decision tree. It has two main features:

1. randomness: The Bagging algorithm randomly selects n training samples based on the self-service sampling method, and each n training samples is used to train a base learner. A total of t sample sets is sampled to construct t base learner. In addition, in the process of base learner training, the randomness of attribute selection is applied, which is the extension of random forest to Bagging.

2. Forest: The base learner of random forest is CART decision tree, which integrates the learning results of multiple decision trees to determine the result of the model, which is called "forest" [9].

The visual diagram in figure 5 shows the structure of Random Forest

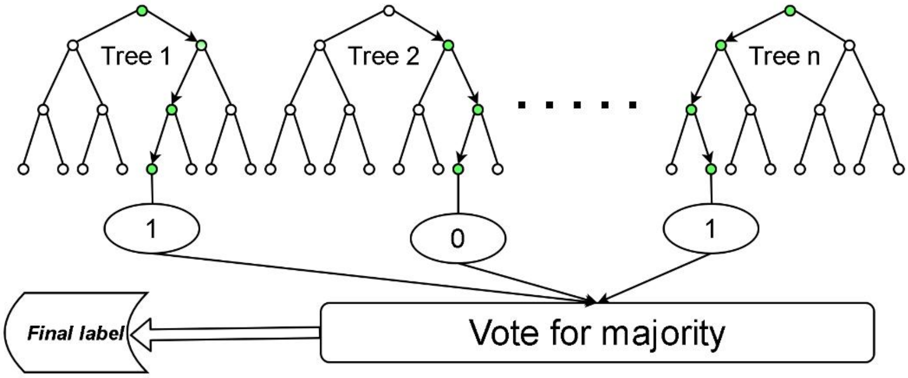


Figure 5: Structure of Random Forest

As can be noticed from the figure, the number of trees is a parameter, which is set as 25 in this case. The confusion matrices and classification reports achieved by Random Forest on training and test are:

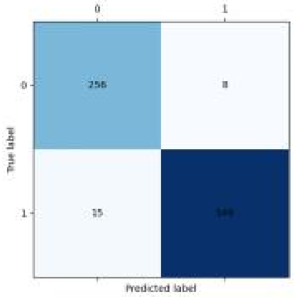
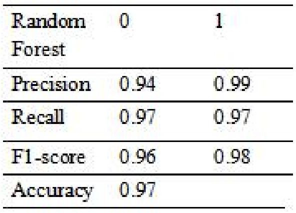


Figure 6: Confusion matrix achieved by random forest on training set

Table 6: Classification report of random forest on training set



Random Forest	0	1
Precision	0.94	0.99
Recall	0.97	0.97
F1-score	0.96	0.98
Accuracy	0.97	

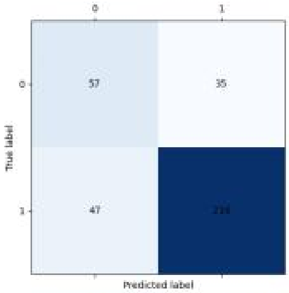
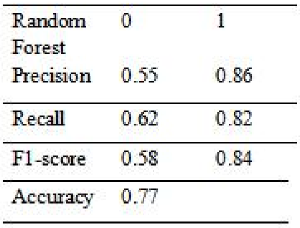


Figure 7: Confusion matrix achieved by random forest on test set

Table 7: Classification report of random forest on training set



Random Forest	0	1
Precision	0.55	0.86
Recall	0.62	0.82
F1-score	0.58	0.84
Accuracy	0.77	

It can be concluded from the charts above that Random Forest performs absolutely well on the training set but far from ideal on test set, which is no better than SVM. Thus, it is likely that the error comes primarily from the distribution of the data rather than models.

3.6 Sample set homogenization using SMOTE

By observing the data, it is easy to acknowledge that the sample is unbalanced, where the size of Category 1 is much larger than Category 0, to solve which an algorithm that can make samples more evenly distributed called SMOTE is now introduced.

Synthetic Minority Over-Sampling Technique (SMOTE) strikes a balance in the entire sample set by artificially synthesizing minority samples, the process of which is shown in Figure 8.

First, a sample point from a minority class is chosen randomly and it selects another sample point at random from the K points closest to it. Then, a new artificial sample point can be generated randomly at the line segment that connects two points, whose location coordinates can be described as:

$$x_{newly\ generated} = x_1 + \beta \cdot |x_1 - x_2| \qquad (7)$$

for $\beta \in (0,1)$, where x_1 and x_2 is the coordinates of the first and second chosen points.

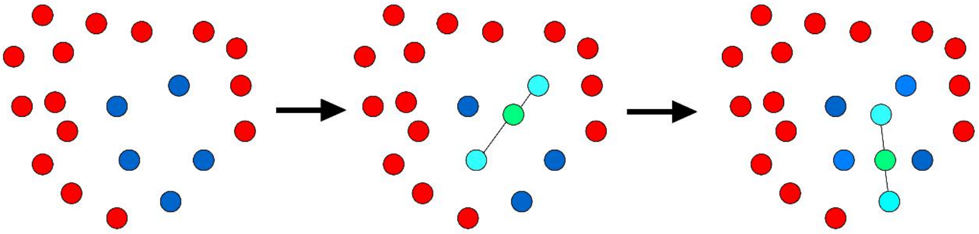


Figure 8: Process of oversampling by SMOTE

After homogenizing the samples using SMOTE, SVM and Random Forest are again deployed to perform classification task, whose results on training and test sets are shown in table 8 and 9 along with the results achieved without SMOTE for comparison.

Table 8: Results on training set

Algorithms		SVM	SVM(SMOTE)	RandomForest	RandomForest(Smote)
Precision	0	0.62	0.72	0.94	0.97
	1	0.87	0.96	0.99	0.99
Recall	0	0.74	0.98	0.97	0.99
	1	0.79	0.62	0.97	0.97
F1 score	0	0.68	0.83	0.96	0.98
	1	0.83	0.76	0.98	0.98
Accuracy		0.77	0.80	0.97	0.98

Table 9: Results on test set

Algorithms		SVM	SVM(SMOTE)	RandomForest	RandomForest(Smote)
Precision	0	0.48	0.72	0.55	0.78
	1	0.87	0.95	0.86	0.81
Recall	0	0.70	0.97	0.62	0.82
	1	0.74	0.64	0.82	0.78
F1 score	0	0.57	0.83	0.58	0.80
	1	0.80	0.76	0.84	0.80
Accuracy		0.73	0.80	0.77	0.80

As is clearly shown in the tables, after SMOTE, the performances achieved by both SVM and Random Forest are both upgraded to a large degree. By comparing the results of SVM and Random Forest, both after SMOTE, it can be found that SVM can accurately identify patients with uncertain dementia and random forests perform equally in the identification of both categories, which are Intuitively visible in their confusion matrix on test set below:

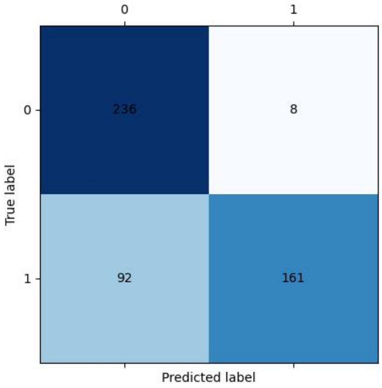


Figure 9: Confusion matrix achieved by Random Forest after SMOTE

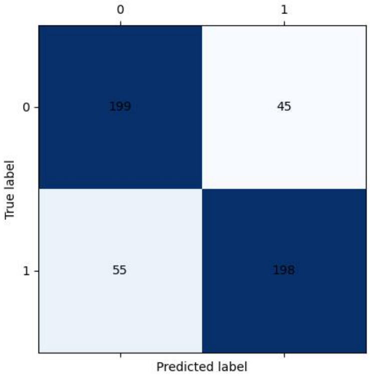


Figure 10: Confusion matrix achieved by SVM after SMOTE

Conclusion

It can be seen by summarizing the early intervention and evaluation criteria for dementia and Alzheimer's disease that it is necessary to use accurate intelligent diagnostic systems to identify the type of disease early. After providing the Gender, Age, Years of Education, Socioeconomic Status, Mini Mental State Examination, Clinical Dementia Rating, Estimated Total Intracranial Volume, Normalize Whole Brain Volume and Atlas Scaling Factor of the patient, the diagnostic model based on logistic regression can precisely diagnose whether a patient has dementia or not. In the case of diagnosing whether the type of dementia is Alzheimer's disease, this paper discusses the diagnostic capabilities of the models based on SVM and Random Forest before and after SMOTE, which turns out SMOTE can significantly improve the diagnostic performance. Also, the difference in SVM and Random Forest shows that each of the algorithm has its own advantages when acting as intelligent diagnostic model for the types of dementia.

References

[1] Zhao Q, Chen P, Zhu SQ, et al. Effects of nonpharmacological interventions on elderly people with subjective memory complaints: a systematic review[J]. Chinese General Practice, 2020, 23(29): 3719-3728.

[2] Li N, Zhao Y. Research progress on early recognition of mild cognitive impairment and related theoretical models[J]. Chinese Journal of Nursing, 2018, 53(5):6.

[3] Whitehouse P, Price D, Struble R, et al. Alzheimer's dementia: loss of neurons in the basal forebrain[J]. Annals of Neurology, 1982, 10:122-126.

[4] Yan An. Community Health Service in Early Diagnosis and Adjuvant Therapy of Alzheimer Disease[J]. Continuing Medical Education, 2015, 000(007):131-133.

[5] Alzheimer Features, Available from: https://www.kaggle.com/datasets/brsdincer/ alzheimer-features.

[6] Liao JG, Chin KV. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. [J]. Bioinformatics, 2007, 23(15):1945-1951.

[7] Alzheimer's clinical data, Available from: https://www.kaggle.com/datasets/ legendahmed/alzheimers-clinical-data.

[8] Vapnik, VladimirN. An Overview of Statistical Learning Theory. [J]. IEEE Transactions on Neural Networks, 1999.

[9] Long BT. Network Video Customer Churn Prediction and Analysis Based on Random Forest and K-means Algorithm[J]. Journal of Hubei Minzu University (Natural Science Edition), 2022,40(02):202-207.

[10] Hui H, Wang WY, Mao BH. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[J]. Lecture Notes in Computer Science, 2005.