Research Article

Journal of Scientific Enquiry
Vol: 3 (1) , Pages: 78-81

Computational Intelligence Based Prediction of Alzheimer Disease from Imbalanced Mild Cognitive Impairment Samples

[1], *Pradip Ghanty

[1]Department of Computer Science, Asansol Girls' College, Asansol, West Bengal, India

Received: 23.02.2023; Accepted: 22.06.2023; Published online: 30.06.2023

Abstract

The disease of Alzheimer is a progressive neurological disease. In Alzheimer's disease, as neurons are injured and die throughout the brain. Dementia is the most important and primary concern for Alzheimer's disease - there is deterioration in memory, thinking, behaviour and the ability to perform everyday activities. In the early stage of memory loss is the result of Mild cognitive impairment (MCI) and it also jeopardise the individual ability of performing his/her independently daily living activities. So, in later stage MCI patients may develop to Alzheimer disease (AD). In this paper a well-known Computational Intelligence tool called Artificial Neural Networks (ANNs) is used to predict AD from MCI samples. Two variants of ANNs - multilayer perceptron (MLP) and radial basis function (RBF) network are used for prediction. We have used 47 MCI samples (previously used by many researchers) for prediction problem. It has been observed that class distributions of MCI dataset are out of proportion i.e. lack of balance (imbalance). We have used oversampling algorithm with cross-validation for imbalanced MCI datasets to predict the developing of Alzheimer's or another dementia. In terms of prediction accuracy our findings are more relevant and it gives better results compared to previous studies which considered without imbalanced scenarios.

Keywords: Mild cognitive impairment, Alzheimer's disease, Imbalanced datasets, Artificial Neural Networks.

1. Introduction

The syndrome of dementia is loss of memory, deterioration of thinking ability and behaviour and inability to perform everyday activities. Dementia affects older people but ageing is not a normal part of dementia [8]. Around 55 million people are suffering from dementia and we find there are near about 11 million new cases in every year. Disability and dependency among the older people worldwide is the main symptoms of dementia. The impact of dementia is mainly psychological, physical, social, and economic.

It affects not only the suffering people of dementia, but also on their families, carers and society at large. Dementia is the most important and primary concern for Alzheimer's disease. In the early stage of memory loss is the result of Mild cognitive impairment (MCI) and it also jeopardise the individual ability of performing his/her independently daily living activities. So, in later stage MCI patients may develop to AD. Our objective is to predict develop of AD from MCI samples. Many researchers used computational intelligence tools to predict AD [1, 10-13]. Artificial Neural Networks (ANNs) are widely used for many prediction problems [1, 10] including prediction of AD. We have also used two types of ANNs variant - multilayer perceptron (MLP) and radial basis function (RBF) network.

In many datasets, it has been observed that class distributions are out of proportion i.e. lack of balance (imbalance). Normal approach of prediction with imbalance datasets suffers in prediction accuracy. Now a day's many approaches have been taken to handle imbalanced scenarios [14]. We have considered two oversampling algorithms viz Synthetic Majority Oversampling Technique (SMOTE) [4] and Adaptive Synthetic Sampling approach (ADASYN) [7].

Cross-validation method has been widely used for small datasets [1]. As number of samples in MCI dataset is very small, we have used cross-validation method to predict the developing of Alzheimer's or another dementia from imbalanced MCI dataset.

MCI Datasets

Ray et al. [8] collected 47 subjects diagnosed with MCI, 22 converted to AD within few (2–5) years (MCI -> AD), 8 converted to OD (MCI -> OD), whereas 17 were still diagnosed as MCI, 4–6 years later (MCI -> MCI). This is a three (3) class prediction problem. The percentage of data distribution among 3 classes are 47%, 17% and 36% respectively. Since the data distributions among 3 classes are not equal (ideal should be approx 33% in each class for balanced dataset), it is a prediction problem of imbalanced dataset. In Figure 1, the original distribution (imbalanced) and three scenarios when distributions could be balanced for MCI dataset are illustrated.

*Email: pradipg@agc.ac.in

DOI: https://doi.org/10.54280/jse.233114

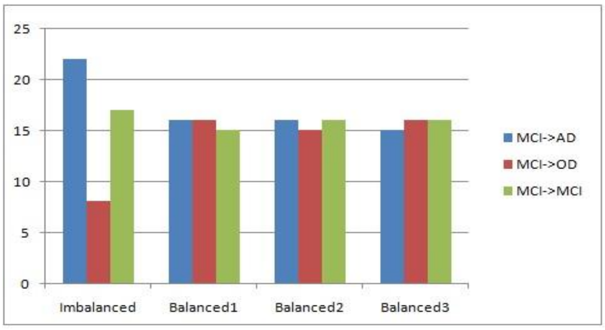


Figure 1: Class Distributions of MCI Dataset – Imbalanced (Original) and Possibilities of Three Balanced Scenarios

2. Method

A well known Computational Intelligence tool called Artificial Neural Networks (ANNs) is used to predict Alzheimer disease [1]. There are two types of ANNs variant - multilayer perceptron (MLP) and radial basis function (RBF) network are used for prediction.

The MLP is a layered network with at least three layers: the input layer, hidden layer and the output layer. There could be more than one hidden layer. The successive layers are completely connected but there is no connection between the nodes within a layer. Every node in the hidden and output layers computes the weighted sum of its inputs and applies an activation function to compute its output. The output is then transmitted to the nodes in the next layer [9]. During training, find the connection weights so that the error between the network output and the target output is the minimum on the training data. Typically the back propagation learning algorithm is used to minimise the sum squared error between the desired output and the computed output.

The RBF network is also layered but it consists of exactly three layers: input layer, basis function layer and output layer. Unlike MLP, the activation functions of the hidden nodes are not sigmoidal type, rather each hidden node represents a radial basis function. The transformation from the input space to the hidden space is non-linear but each node in the output layer computes just the weighted sum of the outputs of the previous layer, i.e. output layer nodes makes a linear transformation. The basic theory of RBF network is rooted in interpolation where for each data point one basis function is used [9]. The learning of RBF network is usually performed in two phases. An unsupervised learning method is applied to estimate the basis function parameters. Then a supervised learning method such as gradient descent or least square error estimate is applied to tune the network weights between the hidden layer and output layer.

We can address imbalanced scenario by two main approaches viz undersampling and oversampling. Former removes majority samples and the latter depicts the minority samples. We have used oversampling procedures since they are capable of maintaining class distributions accepting critically potential majority samples.

For oversampling algorithm [5, 6], Synthetic Minority Oversampling Technique (SMOTE) [4] is used as a benchmark. Another important oversampling technique is Adaptive Synthetic Sampling approach (ADASYN). It develops the learning about the samples distribution in an efficient manner [7]. Synthetic Majority Oversampling Technique (SMOTE) helps to produce synthesis minority samples with the line segments (to produce synthetic sample based on the distance between the minority sample and the closest minority sample therefore the new synthetic sample will be formed between the two minority samples), joining randomly which is chosen P minority examples and their k-nearest minority class neighbours. P is known as the number of minority samples to oversample. In order to obtain the expected balancing ratio between the classes by mentioning similar samples to the existing minority points, SMOTE works larger and less specific boundaries that enhance the generalization abilities of classifiers. It increases their performance. In lieu of generating an equal number of synthesis minority samples for each minority instance, the ADASYN algorithm, indicates that minority instances harder to learn are given a greater importance, being oversampled more often. The oversampling technique is the same as SMOTE; only difference is that harder minority instances are replicated more often. We have used SMOTE and ADASYN as oversampling technique for imbalanced MCI dataset.

In Cross-validation (CV), the dataset is primarily partitioned into k folds, where k-1 folds help to train the classifier and the left-out fold is used for testing. Then all the rotated folds are used for training and testing the classifier. Final performance metrics are averaged across the k estimates of each test fold. This technique assures the k independent sets are used to test the classifier, simulating unseen data. During the training of the model the test set is never found, to avoid overfitting the data.

3. Performance Metrics for Imbalanced Scenarios

Accuracy (ACC) = (TP+TN) / (TP+FN+FP+TN), where
TP is the true positives,
TN is the true negatives,
FP is the false positives,
FN is the false negatives,
given that ACC is biased towards the majority class. A true positive is correctly considered as the class of interest. Similarly, a true negative is correctly

considered as not the class of interest. A false positive is incorrectly restricted as the class of interest. And a false negative is incorrectly restricted as not the class of interest. For the multi-class problem, the most common performance measures consider the classifier ability to discriminate one class versus all others. The class of interest is called the positive class and all others are called negative class.

SENS = TP/(TP+FN) helps to calculate the sensitivity (SENS) and also it helps to measure the percentage of positive samples which are correctly classified, while Specificity (SPEC) refers to the percentage of negative samples correctly classified and can be computed as SPEC = TN /(TN+FP).

Precision (PREC) corresponds to the percentage of positive samples correctly classified, considering the set of all the samples classified as positive, PREC = TP/(TP+FP).

4. Experiments

We have used selected 29 plasma signalling proteins [1] important for predicting future AD from MCI plasma samples [8]. We have used MLP and RBF neural networks as classifiers for 3 class problem with 29 features. Since size of the MCI dataset is small, we have used 10-fold CV for prediction of AD from MCI samples. We have used cross-validation during oversampling as in cross-validation after oversampling it is possible that copies of the same patterns appear in both the training and test sets, making the design subjected to overoptimism. The hyper-parameters of the classifiers are chosen by double cross-validation [15]. We have used 2500 iterations to train MLP and RBF neural networks. The detailed steps of the experiment are illustrated in the algorithm below.

Algorithm: Prediction of AD from MCI samples

Input: 47 MCI Sample with 20 features

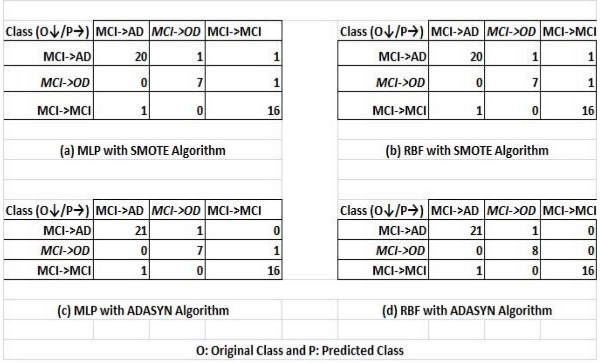
Output: Accuracy (ACC), Sensitivity (SENS), Specificity (SPEC) and Precision (PREC)

Steps:

1. **Divide the samples into 10-folds**
2. **For each fold k=1 to 10**
3. **Prepare training data from 9 folds (All samples – Sample of kth fold)**
4. **Using oversampling algorithm (SMOTE/ADASYN) prepare balanced**
5. **Set classifier (MLP/RBF) hyper-parameters by double cross-validation of prepared training data in step 4.**
6. **Train the classifier for 2500 iterations with selected hyper-parameters in step 5.**
7. **Test the classifier of the kth fold test data with trained classifier in step 6.**
8. **Calculate the TP. TN, FP and FN for kth fold test data .**
9. **Calculate ACC, SENS, SPEC and PREC for kth fold test data by calculated TP, TN, FP and FN in step 8.**
10. **EndFor**
11. **Final ACC, SENS, SPEC and PREC are calculated by averaging on 10-folds.**

5. Results

In Figure 2, the confusion matrices are shown conducting 10-folds cross-validation during oversampling. From Figure 2, it is observed that change of classifiers and/or oversampling algorithms no much effect on MCI imbalanced dataset. Though it is performed better in terms of prediction accuracy (95.75%) with RBF neural networks and ADASYN oversampling algorithm. The accuracy, sensitivity, specificity and precision are illustrated in Table 1 for each class (MCI->AD, MCI->OD and MCI->MCI) taking RBF classifier and ADASYN oversampling algorithm. The comparison results with (oversampled dataset) and without (original dataset) imbalanced scenario is shown in Figure 3. From Figure 3, it is clearly observed that prediction matrices are better compared to prediction of MCI samples after oversampled than original imbalanced MCI dataset.



Class (O↓/P→)	MCI->AD	*MCI->OD*	MCI->MCI
MCI->AD	20	1	1
MCI->OD	0	7	1
MCI->MCI	1	0	16

(a) MLP with SMOTE Algorithm

Class (O↓/P→)	MCI->AD	*MCI->OD*	MCI->MCI
MCI->AD	20	1	1
MCI->OD	0	7	1
MCI->MCI	1	0	16

(b) RBF with SMOTE Algorithm

Class (O↓/P→)	MCI->AD	*MCI->OD*	MCI->MCI
MCI->AD	21	1	0
MCI->OD	0	7	1
MCI->MCI	1	0	16

(c) MLP with ADASYN Algorithm

Class (O↓/P→)	MCI->AD	*MCI->OD*	MCI->MCI
MCI->AD	21	1	0
MCI->OD	0	8	0
MCI->MCI	1	0	16

(d) RBF with ADASYN Algorithm

O: Original Class and P: Predicted Class

Figure 2: Confusion Matrices with MLP & RBF neural networks and SMOTE & ADASYN oversampling algorithms

Measurement ↓ **Class→**	**MCI->AD**	**MCI->OD**	**MCI->MCI**

True Positive (TP)	21	08	16
True Negative (TN)	24	38	30
False Positive (FP)	01	01	00
False Negative (FN)	01	00	01
Accuracy (ACC)	95.75%	97.87%	97.87%
Sensitivity (SENS)	95.45%	100.0%	94.18%
Specificity (SPEC)	96.00%	97.44%	100.0%
Precision (PREC)	95.45%	88.89%	100.0%

Table 1: Accuracy, sensitivity, specificity and precision of each class

6. Conclusion

It is not necessary that the set of 29 proteins are the minimal set of plasma proteins. All these proteins are working in carry AD specific signature among the MCI patients observed by Agarwal et al. [1]. In future study we can use feature selection methods to find minimal set of plasma proteins from oversamples MCI datasets. We can use other oversampling algorithms for better prediction accuracy.

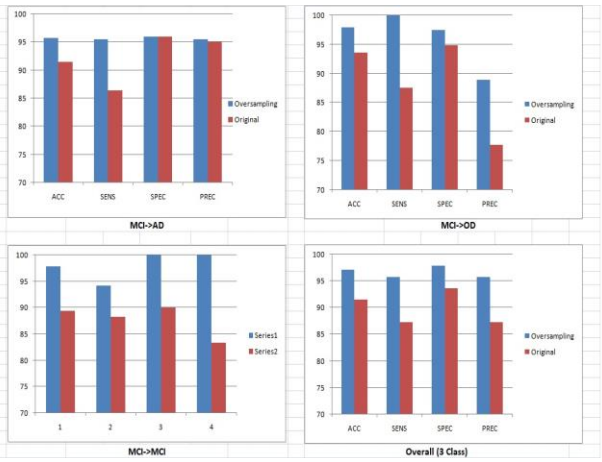


Figure 3: Comparison of prediction matrices with (Oversampling) vs without (Original) imbalance scenario

References

[1] S. Agarwal, P. Ghanty, N. R. Pal, “Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer’s disease”, Bioinformatics (2015) 31:2505–13.

[2] Z. Zheng, Y. Cai, Y. Li, "Oversampling method for imbalanced classification", Computing and Informatics, (January 2015), 34(5): 1017-37.

[3] N. V. Chawla et al., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research (2002): 321-357.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002.

[5] J. Wang, M. Xu, H. Wang, and J. Zhang, “Classification of imbalanced data by using the smote algorithm and locally linear embedding,” in Proc. 8th Int. Conf. Signal Process., vol. 3. 2006, pp. 1–4.

[6] C. S. Ertekin, “Adaptive oversampling for imbalanced data classification,” in Proc. 28th Int. Symp. Comput. Inf. Sci., vol. 264. Sep. 2013, pp. 261–269.

[7] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell., Jun. 2008, pp. 1322–1328.

[8] S. Ray et al. (2007) Classification and prediction of clinical Alzheimer’s diagnosis based on plasma signaling proteins. Nature Medicine, 13, 1359–1362.

[9] S. Haykin, Neural networks: a comprehensive foundation, 2002, PHI.

[10] Luís Costa et al., Application of Machine Learning in Postural Control Kinematics for the Diagnosis of Alzheimer’s Disease. Computational Intelligence and Neuroscience, vol. 2016, Article ID 3891253, 15 pages, 2016.

[11] D. Horn, E. Ruppin, M. Usher, M. Herrmann; Neural Network Modeling of Memory Deterioration in Alzheimer's Disease. *Neural Comput* 1993; 5 (5): 736–749.

[12] A. Savio A et al., Classification Results of Artificial Neural Networks for Alzheimer’s Disease Detection. In: Corchado E., Yin H. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2009. IDEAL 2009. Lecture Notes in Computer Science, vol 5788. Springer, Berlin, Heidelberg.

[13] G. S. Babu, S. Suresh and B. S. Mahanand, Meta-cognitive q-Gaussian RBF network for binary classification: Application to mild cognitive impairment (MCI), *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1-8,

[14] Y. Sun, A.K. C. Wong and M. S. Kamel, Classification of Imbalanced Data: A Review, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 23, No. 04, pp. 687-719, 2009.

[15] P. Ghanty, S. Paul and N.R. Pal, NEUROSVM: An Architecture to Reduce the Effect of the Choice of Kernel on the Performance of SVM, Journal of Machine Learning Research 10 (2009) 591-622.