

# A Predictive Model for Identifying Possible MCI to AD Conversions in the ADNI Database

Xiuyun Qu<sup>1</sup>, Bo Yuan<sup>1</sup>, Wenhuan Liu<sup>1</sup>

<sup>1</sup> : Graduated School at Shenzhen, Tsinghua University, Shenzhen, P.R. China  
e-mail: xiuyunqu@gmail.com, {yuanb, liuwh}@sz.tsinghua.edu.cn

**Abstract**—Alzheimer’s disease (AD) is one of the most common forms of dementia and has become a serious issue among the elderly in the aging society. Since AD is incurable and degenerative, early diagnosis is essential, which can give patients and their family more opportunities to arrange their lives. In the meantime, histopathologic studies have found that MCI (Mild Cognitive Impairment) subjects usually have intermediate levels of AD pathology. In this paper, a predictive model is developed for identifying possible conversions from MCI to AD based on the ADNI (Alzheimer’s Disease Neuroimaging Initiative) database. It is shown that, with the help of a range of advanced data mining techniques, the developed model can achieve promising performance with AUC around 0.88.

**Keywords**- Alzheimer’s Disease; ADNI; feature selection; imbalanced data

## 1. INTRODUCTION

Alzheimer’s disease (AD), also called Alzheimer disease, is a devastating neurodegenerative disease that impairs memory, thought, and behavior, reducing the quality of life and ultimately leading to death. As the development of the aging process in the world, AD has become a common disease among the elderly in the aging society. Over 4 million people in the US have been diagnosed with AD, and a very substantial number of people have other types of dementias. The cost to the US economy is well over \$100 billion per year and the incidence of dementia is expected to double during the next 20 years. No existing treatment has yet been shown to slow the progression of AD but a large number of potential treatments are under development. Once such treatments for patients with AD are approved, the next obvious step will be to perform prevention trials on those at high risk for AD, such as subjects with mild cognitive impairment (MCI), family histories of dementia, or genetic risk factors for AD [1].

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) aims to help researchers and clinicians develop effective treatments for Alzheimer’s disease by providing a large database of tests, measurements, and observations taken during the progression of MCI and AD. It is a large five-year (started on Oct. 1 2004) research project to study the change rates of cognition, function, brain structure and function, and biomarkers in around 200 elderly normal-level controls (NL), around 400 subjects with MCI, and around 200 subjects with AD. In accordance with the goal of collecting data along the timeline of AD and MCI progression, many of the examinations, measurements, and questionnaires repeated several times during the course of the study (mostly every six months). The large number of subjects and the quantity of recorded variables provide a desirable opportunity to search the ADNI database for clues about the progression of AD.

Since the database has only been made publicly available recently, systematic research based on it has been rare. Analysis of the data from conversion subjects in comparison to the data from subjects with consistent diagnoses provides a tool for identifying the patterns separating the two groups and for developing better early diagnosing techniques of AD.

In this paper, a predictive model for detecting MCI to AD conversions is built as the first step towards identifying a panel of biomarkers for the purpose of early diagnosis and intervention of AD. In face of the great challenges due to the properties of the ADNI dataset, a range of advanced data mining techniques are employed and the developed model is shown to be able to achieve promising results.

The rest part of this paper is organized as follows. Section 2 gives an overview of the existing research in AD. Section 3 focuses on the issues of ADNI data preprocessing and performance measurement. Experimental results are presented in Section 4 and this paper is concluded in Section 5.

## 2. AN OVERVIEW OF RESEARCH ON AD

In a living person, the diagnosis of possible or probable AD is based on the presence of cognitive deficits in two or more domains severe enough to interfere with normal daily functioning. The low specificity of the clinical criteria reflects the fact that AD shares many clinical features with other forms of dementia [2]. Nowadays, as the rapid development of medical neuroimaging technique, neuroimaging is playing a more and more important part in the diagnosis of AD [3] [4].

Up to now, there has been no effect clinic treatment for AD. The many benefits to the patient, caregiver, and society are the motivating factors for establishing a diagnosis of AD as early in the course of the disease as possible. This goal can be obtained by watching for prodromal AD or MCI, which may occur before clinical AD [5]. There is now increasing evidence that the molecular pathomechanisms of AD become active several years before neurons start dying and cognitive deficits manifest [6]. During this stage, an effective treatment of AD would have the most impact because the cognitive function might be preserved at the highest level possible. Histopathologic studies have found that MCI subjects, as a group, usually have intermediate levels of AD pathology compared to healthy controls and subjects with probable or possible AD [7].

Researchers are often interested in finding determinants by statistic methodology. However, machine learning has played an effective part in learning from mass of data. As a result, the focus of this paper is on developing a systematic machine learning model for the MCI to AD prediction.

### 3. METHODOLOGY

#### 3.1 ADNI Database

The ADNI database consists of sixty-three tables, each of which relates to a particular set of clinical data organized into multiple columns. The dimension of the database is more than 2000, significantly higher than most common classification problems. In this paper, we focused on the analysis of MCI to AD conversions. In general, around 10% to 15% of MCI patients will change to AD per year [5]. As a result, the number of “MCI no change” patients is much higher than the number of “MCI to AD” patients, which means that the dataset is imbalanced. Also, in the process of data collection, there were some examinations in which only part of the subjects participated, resulting in an incomplete dataset.

In face of the above challenges, the key issues that need to be solved in order to build an effective predicative model are: (1) feature selection; (2) imbalanced data; (3) incomplete data. In this study, we focused on the first two topics.

#### 3.2 Data Preprocessing

In ADNI, the columns named DXCURREN, DXCONV, DXCONTYP, and DXREV contain the current diagnostic category, the occurrence of any conversions or reversions, the type of conversion (where applicable), and the type of reversion (where applicable) for each subject at each recorded visit respectively.

In our study, each row in the database (the examination record of a patient at a certain time) was marked as a positive sample if the patient had MCI at that time and turned into AD (“MCI to AD”) in the next visit (usually 6 months later). If the same patient still had MCI in the next visit, the record was marked as a negative sample (“MCI no change”). Note that records belonging to normal subjects or AD patients were not considered. In summary, our task was formulated as a binary classification problem aiming to predict whether a MCI patient will turn into AD within a period of 6 months.

The ADNI database contains a variety of data types such as numerical, character and text, making it very difficult to directly apply any existing classification algorithms. In order to simplify the problem at this stage, only numerical attributes were taken into account. Note that there are some tests in which only part of the subjects participated and the corresponding values were marked as “-1000” for the rest of the subjects. The resulting dataset is referred to as “Incomplete Data” in Table I. In the meantime, another dataset without missing values referred to as “Complete Data” in Table I was created by removing selected attributes and a small number of samples.

TABLE I. EXPERIMENTAL DATASETS

Name of Dataset	Number of Attributes	Positive Samples	Negative Samples
Incomplete Data (In_D)	562	72	650
Complete Data (C_D)	213	72	636

#### 3.3 Performances Measurement

Since the datasets to be classified are clearly imbalanced, the overall classification accuracy, which is the commonly used performance measure, is not

appropriate in this case. In this article, we adopted a popular performance measure in learning from imbalanced data: Receiver Operating Characteristic (ROC) analysis [8].

TABLE II. THE CONFUSION MATRIX

True class	Predicted Class		$\Sigma$
	Positive	Negative	
Positive	$TP$	$FN$	$P$
Negative	$FP$	$TN$	$N$

Table II is a confusion matrix for a binary classification problem. The ROC analysis generates a ROC curve (see Fig.1 for an example) to visualize the tradeoff between the false-positive ( $FP$ ) rate and the true-positive ( $TP$ ) rate, as in (1) and (2). The area under the ROC curve (AUC) can be used as a measurement of classifier performance and a good classifier should have a large AUC value.

$$FP\ rate = FP / N \quad (1)$$

$$TP\ rate = TP / P \quad (2)$$

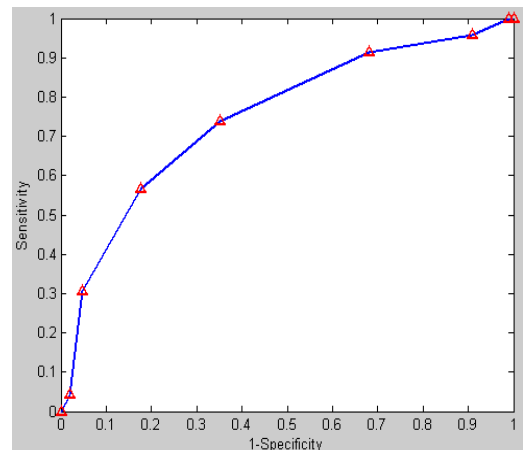


Figure 1. ROC curve

## 4. EXPERIMENTS

### 4.1 Feature Selection

Many classifiers perform badly on high-dimensional datasets with a large number of irrelevant or noisy features. In order to improve the prediction performance of the classifier, it is essential to conduct feature selection on the original dataset to reduce the dimensionality.

In this paper, we consulted an existing preliminary study on the statistical analysis of the numerical attributes and selected top 50 attributes (subset A) according to their  $p$ -values.

We also applied the CFS (correlation-based feature selection) method [9], which can generate a feature subset containing features that are highly correlated with the class and uncorrelated with each other. The feature subset evaluation function in CFS is shown in (3) where  $M_s$  is the heuristic “merit” of a feature subset  $S$  containing  $k$  features and  $\bar{r}_{cf}$  is the mean feature-class correlation ( $f \in S$ ), and  $\bar{r}_{ff}$  is the average feature-feature inter-correlation. Using this method, we selected 28 attributes (subset B) from

dataset C\_D and 30 attributes (subset C) from dataset In\_D, including 11 attributes with missing values.

$$M_s = k \bar{r}_{cf} / [k + k(k-1) \bar{r}_{ff}]^{1/2} \quad (3)$$

Next, the three feature sets were merged to create a new feature subset (without the attributes with missing values), which contained 33 attributes. A new dataset with these 33 attributes selected from C\_D was created, which is referred to as C\_F from now on.

#### 4.2 Imbalanced Data

Since the dataset of interest is imbalanced and this imbalance may lead to suboptimal classification performance, some special techniques for dealing with imbalanced data were employed [12].

At the dataset level, common solutions include various forms of re-sampling techniques, which can be grouped into two categories: over-sampling and under-sampling. These techniques directly manipulate the original dataset in order to alter the class distributions.

For over-sampling, we applied SMOTE [13] in which the minority class is over-sampled by taking each minority class sample  $x$  and introducing synthetic examples  $x_{new}$  along the line segments joining any/all of the  $k$  minority class nearest neighbors  $\bar{x}$ , as in (4). More general regions can be learned for the minority class samples rather than those being subsumed by the majority class samples around them. Fig. 2 shows an example of the effect of SMOTE. However, it is also argued that SMOTE may produce new data that are unauthentic in the medical domain.

$$x_{new} = x + rand(0, 1) * (\bar{x} - x) \quad (4)$$

For under-sampling, we applied random under-sampling, which randomly deletes some samples from the majority class.

At the algorithm level, we adopted the cost-sensitive learning method [15] and the Biased Minimax Probability Machine (BMPM) technique [14].

Cost-sensitive learning requires that a cost-matrix (Table III) is known for different types of errors. Usually, the costs of FP ( $c_{10}$ ) and FN ( $c_{01}$ ) are defined in advance and the costs of TP ( $c_{11}$ ) and TN ( $c_{00}$ ) are set to 0. In the ADNI dataset, the cost of labeling a positive sample (“MCI to AD”) incorrectly should be greater than the cost of labeling a negative sample (“MCI no Change”) incorrectly.

TABLE III. THE COST MATRIX

True class	Predicted Class	
	Positive	Negative
Positive	$c_{11}$	$c_{01}$
Negative	$c_{10}$	$c_{00}$

Given the reliable estimation of the mean and covariance of data  $\{x, \Sigma_x\}, \{y, \Sigma_y\}$ , BMPM constructs a decision boundary  $a^T z = b$  ( $z \in R^n, b \in R$ ) where samples satisfying the condition of  $a^T z > b$  are classified as class  $x$  and samples satisfying the condition of  $a^T z < b$  are classified as class  $y$ . During training, the classification accuracy ( $\alpha$ ) of the minority class of data ( $x$ ) is maximized

under the condition that the accuracy of the majority class of data ( $y$ ) are not below a pre-specified acceptable level  $\beta_0$ , as in (5)–(7). Thus, it provides a systematic and rigorous treatment for skewed data.

$$\max_{\alpha, \beta, b, a \neq 0} \alpha \text{ s.t. } \inf_{x \in \{\bar{x}, \Sigma_x\}} \Pr\{a^T x \geq b\} \geq \alpha \quad (5)$$

$$\inf_{y \in \{\bar{y}, \Sigma_y\}} \Pr\{a^T y \leq b\} \geq \beta \quad (6)$$

$$\beta \geq \beta_0 \quad (7)$$

A set of experiments were carried on dataset C\_F (the complete dataset C\_D after feature selection) where 2/3 of the dataset was used as the training set and the rest part was used as the test set. The AUC value of the BMPM classifier was 0.75 and the effect of sampling and cost-sensitive learning on various classifiers is shown in Table IV. It is clear that the performance of these classifiers varied significantly with different combinations of techniques for imbalanced datasets.

TABLE IV. THE EFFECT OF SAMPLING AND COST-SENSITIVE LEARNING

	AUC <sup>a</sup>		
	C_F <sup>b</sup>	Random Delete <sup>c</sup>	SMOTE <sup>d</sup>
Naïve Bayes [10]	0.876	0.872	0.879
C-libsvm <sup>e</sup> [16]	0.704	0.756	0.758
C4.5 [17]	0.613	0.613	0.540
C-C4.5	0.591	0.680	0.572
KNN [18]	0.685	0.673	0.659
C-KNN <sup>f</sup>	0.685	0.733	0.768
AdaboostM1 [11]	0.748	0.669	0.768

- a. Test set Positive/All 23/233  
b. Original training set Positive/All 49/476  
c. Random deleted training set Positive/All 49/257  
d. SMOTE sampling training set Positive/All 98/525  
e. The prefix ‘c-’ means cost-sensitive learning and the cost ratio was 4:1.  
f. The  $k$  of KNN in this experiment was 3.

## 5. CONCLUSIONS

The major contribution of our work is to introduce a principled and systematic machine learning approach to the problem of predicting MCI to AD conversions in the ADNI dataset, which has potentially significant practical meaning for the early diagnosis of AD. Experimental results showed that Naïve Bayes can predict the conversion with a reasonably good AUC value after feature selection. Various sampling and cost-sensitive learning techniques can also help improve the performance of a number of other classifiers.

Certainly, there is still a lot of room for improvement in order to make the predictive model more effective. For example, only the complete dataset was considered in the current work while there may be some discriminative attributes with missing data and ignoring these attributes could lead to the loss of valuable information. As a result, how to deal with attributes with missing values is a very

important challenge in the analysis of the ADNI dataset. Note that, in medical datasets, the existence of missing values is often not due to random factors as in other datasets but may be due to different treatments based on the specific conditions of each patient.

## REFERENCES

- [1] Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Website, available at <http://www.loni.ucla.edu/ADNI/>.
- [2] D. S. Knopman, S. T. DeKosky, J. L. Cummings, et al., "Practice parameter: diagnosis of dementia (an evidence based review)," *Neurology*, vol. 56, pp. 1143–1153, 2001.
- [3] A. K. Demetriades, "Functional neuroimaging in Alzheimer's type dementia," *Journal of the Neurological Sciences*, vol 203, pp. 247–251, 2002.
- [4] A. C. Burggren and S. Y. Bookheimer, "Structural and functional neuroimaging in Alzheimer's disease: an update," *Current Topics in Medicinal Chemistry*, vol. 2, pp. 385–393, April 2002.
- [5] B. P. Leifer, "Early diagnosis of Alzheimer's disease: clinical and economic benefits," *Journal of the American Geriatrics Society*, vol. 51, pp. 281–288, 2003.
- [6] S. T. DeKosky and K. Marek, "Looking backward to move forward: early detection of neurodegenerative disorders," *Science*, vol. 302, Oct. 2003, pp. 830–834, doi: 10.1126/science.1090349.
- [7] D. A. Bennett, J. A. Schneider, J. L. Bienias, D. A. Evans and R. S. Wilson, "Mild cognitive impairment is related to Alzheimer disease pathology and cerebral infarctions," *Neurology*, vol. 64, pp. 834–841, 2005.
- [8] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, July 1997.
- [9] M. A. Hall, Correlation-based Feature Subset Selection for Machine Learning, Ph. D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998, unpublished.
- [10] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," *Proc. the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, Morgan Kaufmann, pp. 338–345, 1995.
- [11] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," *Proc. the Thirteenth International Conference on Machine Learning (ICML'96)*, Bari, Italy, Morgan Kaufmann, pp.148–156, 1996.
- [12] N. V. Chawla, N. Japkowicz and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, pp. 1–6, June 2004.
- [13] N. V. Chawla, Ke. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [14] K. Huang, H. Yang, I. King and M. R. Lyu, "Learning classifiers from imbalanced data based on biased minimax probability machine," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, D. C., USA, vol. 2, pp. 558–563, 2004.
- [15] C. Elkan, "The foundations of cost-sensitive learning," *Proc. the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, Washington, D. C., USA, pp. 973–978, 2001.
- [16] C. Chang and C. Lin, LIBSVM: A Library for Support Vector Machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [17] J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann Publishers, 1993.
- [18] D. W. Aha, D. Kibler and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

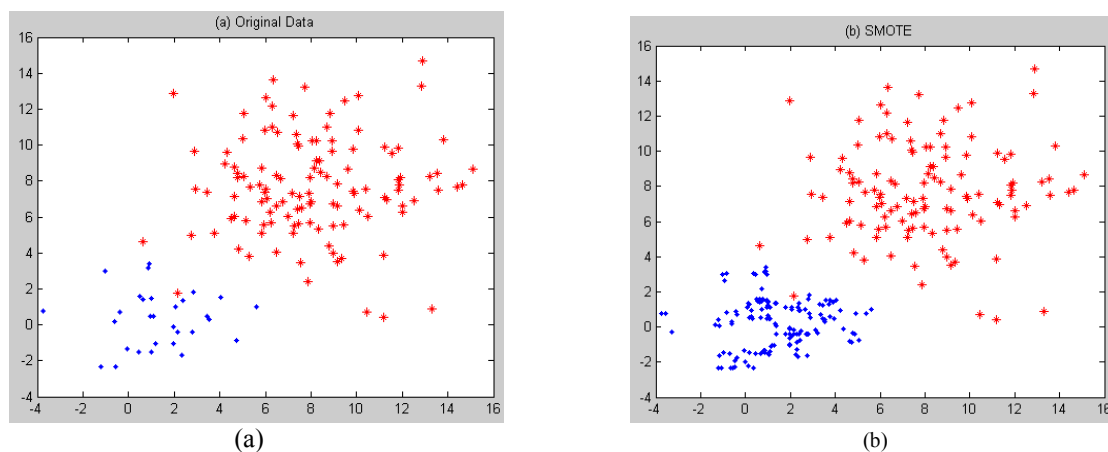


Figure 2. An example of SMOTE: (a) original data; (b) after re-sampling.