Evaluating Imputation Techniques for Missing Data in ADNI: A Patient Classification Study

Sergio Campos[1(✉)], Luis Pizarro[3], Carlos Valle[1], Katherine R. Gray[2], Daniel Rueckert[2], and Héctor Allende[1]

[1] Departamento de Informática,
Universidad Técnica Federico Santa María, Valparaíso, Chile
sergio0.1@gmail.com
[2] Department of Computer Science, University College London, London, UK
[3] Department of Computing, Imperial College London, London, UK

Abstract. In real-world applications it is common to find data sets whose records contain missing values. As many data analysis algorithms are not designed to work with missing data, all variables associated with such records are generally removed from the analysis. A better alternative is to employ data imputation techniques to estimate the missing values using statistical relationships among the variables. In this work, we test the most common imputation methods used in the literature for filling missing records in the ADNI (Alzheimer's Disease Neuroimaging Initiative) data set, which affects about 80% of the patients–making unwise the removal of most of the data. We measure the imputation error of the different techniques and then evaluate their impact on classification performance. We train support vector machine and random forest classifiers using all the imputed data as opposed to a reduced set of samples having complete records, for the task of discriminating among different stages of the Alzheimer's disease. Our results show the importance of using imputation procedures to achieve higher accuracy and robustness in the classification.

Keywords: Missing data · Imputation · Classification · ADNI · Alzheimer

1 Introduction

Alzheimer's disease (AD) is the most common type of dementia in the elderly, representing about 80% of all dementia patients and the sixth cause of death in the USA. 26.6 million people worldwide were estimated to suffer from some degree of dementia in 2006, and 100 million impaired people are expected by 2050 [1]. Unfortunately, no drug treatment reducing the risk of developing AD or delaying its progression has been discovered so far. The Alzheimer's Disease Neuroimaging Initiative[1] (ADNI), launched in 2004, contributes to the development of biomarkers for the early detection (diagnostic) and tracking (prognostic)

[1] http://adni.loni.usc.edu/

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 3–10, 2015.
DOI: 10.1007/978-3-319-25751-8_1

of AD using longitudinal clinical, imaging, genetic and biochemical data from patients with AD, mild cognitive impairment, and healthy controls. Its major achievements have been reviewed in [2]. Pattern recognition techniques have been instrumental in identifying disease patterns. Tasks such as classification, prediction, feature extraction and selection, multimodal data fusion, dimensionality reduction, among others, are at the core of this ongoing multidisciplinary research initiative. Pattern analysis is, however, hampered by *missing data* in the ADNI dataset, i.e. patients with incomplete records, cases where the different data modalities are partially or fully absent due to several reasons: high measurement cost, equipment failure, unsatisfactory data quality, patients missing appointments or dropping out of the study, and unwillingness to undergo invasive procedures. The missing data problem can be handled in two ways. Firstly, all samples having a missing record are removed before any analysis takes place. This is a reasonable approach when the percentage of removed samples is low so that a possible bias in the study can be discarded. Secondly, the missing values can be estimated from the incomplete measured data. This approach is known as *imputation* [3] and is recommended when the adopted data analysis techniques are not designed to work with missing entries. About 80% of the ADNI patients have missing records. Despite this, such patients are discarded in the vast majority of ADNI studies, which is a disuse of valuable incomplete information. Only recently, pattern recognition and machine learning techniques that can cope with missing entries or perform data imputation have been investigated. This article focuses on the task of patient classification into clinical groups. In particular, we conduct a comparative study of different imputation techniques and evaluate their impact on classification performance. We train support vector machine and random forest classifiers using all the imputed data as opposed to a reduced set of samples having complete records, for the task of discriminating among different stages of AD. We show the importance of including imputation and data analysis procedures to achieve more accurate and robust classification results.

In section 2 we provide further background on the classification task for ADNI patients and briefly describe the imputation and classification methods we used in this study. Section 3 details the experimental settings on which we tested the different methods against imputation error and classification performance, discussing our findings. Final remarks and future work are examined in section 4.

2 Methods

2.1 Classification with Incomplete Data

The ADNI study provides a database of multimodal entries for 819 subjects: 229 participants with normal cognition as healthy controls (HC), 397 with mild cognitive impairment (MCI), and 193 with mild Alzheimer's disease. Individuals with MCI are divided into two groups: those who remained in a *stable* condition (sMCI) and those who later *progressed* to AD (pMCI). It is therefore crucial to diagnose the patients into these clinical categories correctly in order to choose an

appropriate treatment and further monitoring the disease. This task is especially difficult when approximately 80% of the participants have missing observations.

Let $X \in \mathbb{R}^{n \times p}$ be an incomplete matrix with n samples (subjects) and p variables (features). X can be seen as two matrices, one containing the observed data X_o, and the other one representing the missing data to be estimated X_m. Most classification methods from the literature discard all samples having at least one missing value. Only a few works that use all the available data exist, such as [4–7] where direct data imputation is avoided, and [8] where a subset of the missing data is estimated based on variable and sample selection. It is then important to investigate the different causes of the missing data to evaluate the utilisation of adequate imputation methods. Little and Rubin [3] define three missing data mechanisms: *i*) missing completely at random, MCAR: missing values are independent of both observed and unobserved data; *ii*) missing at random, MAR: given the observed data, missing values are independent of unobserved data; and *iii*) missing not at random, MNAR: missing values depend on the unobserved data. A recent longitudinal study [9] found that missing data in ADNI are not MCAR, but rather conditional to other features in addition to cognitive function. Moreover, the authors found evidence of different missing data mechanisms between different biomarkers and clinical groups.

2.2 Imputation Methods

Efforts to define a taxonomy of imputation methods have been reported in [3,10]. In this work we compare some common techniques used in the literature.

1. *Zero.* This method consists of imputing missing data with 0 (zero) values.
2. *Mean.* Missing values filled with the mean of the observed values per variable.
3. *Median.* Missing values filled with the median of the observed values per variable. The median is more robust against outliers than the mean. It tolerates up to 50% of outliers [11].
4. *Winsorised mean.* Provides a more robust estimate for the mean, which is calculated after replacing a given percentage (α) of the largest and smallest values with the closest observations to them. We used $\alpha = 10\%$. This method also controls the effect of outliers.
5. *k-nearest neighbours (kNN).* Missing values filled with the mean of the k-nearest observed samples based on the Euclidean distance. We use a modified cross-validation approach [12] to find the parameter k in the range $[1, \sqrt{n_{obs}}]$.
6. *Regularised expectation maximisation (RegEM).* Proposed by Schneider [13], this imputation method makes two important assumptions: the data follow a Normal distribution, and the missing values are generated by a MAR process. The missing entries are estimated by the linear regression model $x_m = \mu_m + (x_o - \mu_o)B + e$, where $x_o \in \mathbb{R}^{1 \times p_o}$ and $x_m \in \mathbb{R}^{1 \times p_m}$ are row vectors of the observed data matrix X_o and the estimated missing data matrix X_m, respectively; μ_o and μ_m are their corresponding means; $B \in \mathbb{R}^{p_o \times p_m}$ is the regression coefficients matrix, and $e \in \mathbb{R}^{1 \times p_m}$ is a zero-mean random residual vector with unknown covariance matrix $C \in \mathbb{R}^{p_m \times p_m}$. Initially, the algorithm

estimates the missing data with the Mean method, followed by i) E-step: compute the expected mean μ and covariance matrix Σ of X, ii) M-step: compute the maximum likelihood estimates of the regression parameters B and C, conditional to the estimates (μ, Σ), and iii) impute missing values using the regression model. These three steps are iterated until convergence, i.e., until the estimates (μ, Σ) stabilise. We run Schneider's implementation[2] using the individual ridge regression model.

3 Experimental Results

3.1 Data

In this work, we consider three baseline ADNI modalities: cerebrospinal fluid (CSF), magnetic resonance imaging (MRI) and positron emission tomography (PET). The modalities were preprocessed according to [14], with 43 out of 819 subjects excluded for not passing the quality control. The CSF source contains 3 variables that measure the levels of some proteins and amino acids that are crucially involved in AD. The MRI source provides volumetric features of 83 brain anatomical regions. The PET source (with FDG radiotracer) provides the average brain function, in terms of the rate of cerebral glucose metabolism, within the 83 anatomical regions. Hence, each subject consists of 169 features.

3.2 Imputation

In this section we work with the 147 subjects who have complete records: 35 HC, 75 MCI and 37 AD. We synthesise different patterns of missing data, considering the individual modalities and pairs of them: CSF, MRI, PET, CSF-MRI, CSF-PET and MRI-PET. For each pattern we removed such features from a given percentage {10, 20, 30, 40, 50}% of subjects that were chosen randomly. The performance of the different imputation methods is assessed between three clinically relevant pairs of diagnostic groups: AD/HC, MCI/HC and pMCI/sMCI.

Due to space limitations, Fig. 1 only shows the results for the experiment AD/HC (72 subjects) with the CSF-PET missing data pattern. 95% confidence intervals were computed for the Pearson correlation (PC) and the relative error (RE) over 100 runs. As expected, we observed that the PC of the imputed variables decreases with the amount of missing data. It is noteworthy that the PC for the Zero method is the lowest because this technique does not consider any additional information for estimating the data. Moreover, the RE for each method seems rather constant. Since it is computed as $\text{RE} = |x_o - x_m|/x_o$, the Zero method will always produce RE = 1. Filling CSF data produces an error of about 45% for the Median, Winsorised mean and kNN methods, which outperform the Mean and EM methods. Filling PET data produces an error of about 13% for most techniques, except for the Zero method. This low error can be explained by inspecting the actual PET values. Fig. 2 shows the histograms

[2] www.clidyn.ethz.ch/imputation

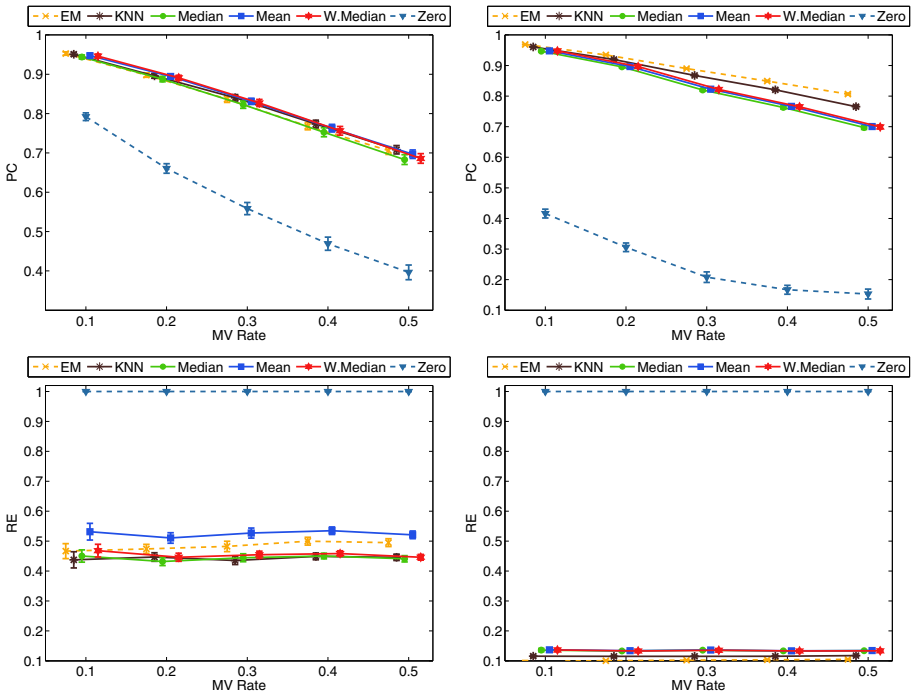

Fig. 1. Imputation performance. Pearson correlation (*bottom*) and relative error (*top*) for the imputation of missing values (MV) in CSF (*left panels*) and PET (*right panels*) for the CSF-PET missing data pattern.

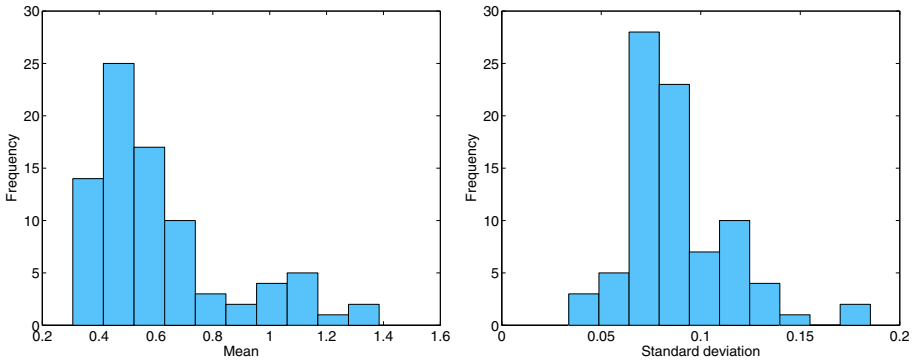

Fig. 2. Mean (*left*) and standard deviation (*right*) histograms over all 83 PET variables.

of the mean and standard deviation over all 83 PET variables. These small quantities indicate that the PET values are bunched up close to the mean. For this reason, the methods tend to provide estimates around this value even if they do not directly impute using the mean.

Table 1. AD/HC multi-modality classification accuracy (acc.), area under the curve (AUC), sensitivity (sens.), specificity (spec.), and F-measure (F) based on filling missing data with different imputation methods before training a support vector machine (SVM) and a random forest (RF) classifiers. Results are expressed as mean (standard deviation).

Classifier	Imputation	Acc. (%)	AUC (%)	Sens. (%)	Spec. (%)	F (%)
SVM	*none*	83.5 (10.7)	92.4 (8.0)	81.7 (15.7)	86.1 (12.4)	82.2 (12.2)
	Zero	**88.7** (3.3)	93.9 (3.0)	**89.0** (4.7)	88.6 (5.2)	**89.5** (3.1)
	Mean	86.6 (2.3)	92.2 (2.4)	85.9 (5.6)	87.7 (4.7)	87.7 (2.4)
	Median	88.5 (3.7)	93.7 (3.1)	88.3 (3.9)	88.5 (6.7)	89.3 (3.3)
	Winsor m.	88.4 (3.4)	**94.3** (2.4)	88.6 (4.6)	**88.9** (5.5)	89.2 (3.2)
	*k*NN	88.5 (3.0)	93.5 (2.5)	88.3 (3.6)	88.8 (4.7)	89.1 (3.2)
	EM	88.1 (4.0)	93.7 (2.8)	87.9 (5.6)	88.3 (4.3)	88.9 (3.8)
RF	*none*	84.8 (9.0)	93.2 (5.3)	85.9 (13.2)	85.5 (11.3)	84.3(9.6)
	Zero	86.2 (3.5)	93.2 (3.1)	87.1 (4.9)	85.4 (5.4)	87.1 (3.2)
	Mean	86.6 (2.8)	92.7 (2.5)	87.2 (5.5)	86.1 (5.3)	87.5 (2.7)
	Median	86.3 (3.5)	93.6 (3.2)	87.2 (3.8)	85.2 (6.7)	87.3 (3.1)
	Winsor m.	**88.4** (3.1)	**94.3** (2.0)	**89.1** (4.2)	**87.8** (4.1)	**89.1** (2.8)
	*k*NN	85.3 (3.3)	92.5 (2.3)	86.5 (4.7)	84.2 (4.8)	85.9 (3.4)
	EM	86.3 (4.2)	93.1 (2.6)	87.1 (5.1)	85.6 (5.9)	87.1 (4.0)

Table 2. MCI/HC multi-modality classification results.

Classifier	Imputation	Acc. (%)	AUC (%)	Sens. (%)	Spec. (%)	F (%)
SVM	*none*	69.1 (11.8)	69.5 (14.6)	48.3 (21.5)	74.9 (13.4)	44.3 (22.8)
	Zero	70.9 (3.3)	74.4 (11.1)	59.8 (5.6)	76.3 (4.3)	56.6 (5.4)
	Mean	71.9 (3.8)	76.1 (12.2)	62.5 (6.8)	76.7 (4.0)	58.8 (6.0)
	Median	71.0 (3.9)	76.3 (4.1)	62.4 (6.1)	75.8 (4.9)	59.1 (6.5)
	Winsor m.	72.0 (3.6)	77.9 (3.7)	**63.4** (5.6)	76.2 (4.6)	59.0 (5.1)
	*k*NN	72.6 (3.8)	78.8 (3.6)	63.3 (6.3)	77.3 (5.6)	58.9 (6.7)
	EM	**73.1** (4.2)	**78.9** (2.8)	62.8 (7.1)	**78.6** (5.1)	**59.5** (7.3)
RF	*none*	71.1 (8.3)	75.3 (10.6)	65.0 (21.9)	74.3 (10.2)	42.6 (14.9)
	Zero	**73.6** (3.2)	78.3 (3.8)	**67.1** (8.0)	76.1 (4.2)	56.7 (7.1)
	Mean	71.9 (3.3)	77.6 (5.0)	65.0 (8.0)	75.0 (2.8)	56.1 (4.5)
	Median	71.4 (3.8)	76.5 (4.7)	66.1 (7.3)	73.5 (3.8)	55.7 (6.3)
	Winsor m.	72.2 (3.8)	77.5 (3.6)	66.9 (4.2)	74.3 (4.9)	55.5 (6.1)
	*k*NN	73.2 (4.2)	**78.6** (4.6)	65.5 (8.0)	**76.7** (5.4)	**58.5** (6.1)
	EM	72.6 (3.7)	77.6 (3.8)	63.2 (7.4)	**76.7** (3.9)	56.8 (5.3)

3.3 Classification

We now use all 776 subjects to assess the impact of the different imputation methods on patient classification. The whole data set has 33% of missing values, from which 97% correspond to PET values. The remaining 3% are CSF values, while the MRI source is complete. We consider two experiments: AD/HC with

395 subjects (185 AD and 210 HC) and MCI/HC with 591 subjects (381 MCI and 210 HC). In each experiment we used 75% of the data to train two classifiers, namely a ν-Support Vector Machine (ν-SVM) and a Random Forest (RF), evaluated over 25 runs. The other 25% of the data was used for testing. We employed the implementations found in the scikit-learn library[3]. The ν and σ parameters for ν-SVM and the number of trees and number of features for RF were tuned using 5-fold CV.

Tables 1 and 2 show the classification results for the experiments AD/HC and MCI/HC, respectively. We juxtapose both classifiers, SVM and RF, for the different imputation methods. For completeness, we include the results when the classifiers are trained solely with the reduced set of subjects having complete records and thus no imputation is needed. It can be noticed that the classification improves considerably when the full data set is used, imputing the missing values. This clearly provides more information to discriminate among the different diagnostic groups. These experiments suggest that the Winsorised mean, the kNN and the EM methods should be preferred as imputation methods as they provide more stable performance. The Zero method seems competitive, which is explained again by the fact that the most of the missing data come from the PET source which presents low dispersion values close to zero. Both classifiers present similar performances in each scenario. A remarkable point, is that their robustness (low variance) is increased in cases with imputation.

4 Conclusions and Future Work

We have seen how imputation techniques allow for the utilisation of additional information, that would otherwise be discarded, to better distinguish between different diagnostic groups. The development of biomarkers using more evidence could result in more accurate diagnosis and prognosis of Alzheimer's patients. Our results showed that training classifiers with imputed data is better than constructing a predictive model with a reduced number of subjects with complete records. This is supported by the fact that all imputation techniques increase both performance metrics and robustness of the classifiers. An apparently unexpected finding is that the Zero method is competitive with the other methods, according to the performance metrics used in this article. It is expected that more sophisticated methods such as kNN and EM would deliver better results. However, as we stated before, possibly more relevant than the quality of the imputation algorithms is the nature of the data, which plays an important role in the performance as we have seen.

Future work includes studying other imputation and classification techniques, as well as exploring multi-class extensions and alternative ways of treating the feature space to handle data-dependent imputation pitfalls. There is interest in comparing imputation methods with methods that can internally handle the missing values, as Artificial Neural Networks (ANN) [15] and SVM [16].

[3] scikit-learn.org/stable

References

1. Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M.: Forecasting the global burden of Alzheimer's disease. Alzheimer's & Dementia **3**(3), 186–191 (2007)
2. Weiner, M.W., et al.: The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. Alzheimer's & Dementia **9**(5), 111–194 (2013)
3. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley-Interscience (2002)
4. Wang, C., Liao, X., Carin, L., Dunson, D.B.: Classification with incomplete data using Dirichlet process priors. JMLR **11**, 3269–3311 (2010)
5. Ingalhalikar, M., Parker, W.A., Bloy, L., Roberts, T.P.L., Verma, R.: Using multiparametric data with missing features for learning patterns of pathology. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 468–475. Springer, Heidelberg (2012)
6. Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J.: Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage **61**(3), 622–632 (2012)
7. Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J.: Bi-level multi-source learning for heterogeneous block-wise missing data. NeuroImage **102**, Part 1, 192–206 (2014)
8. Thung, K.-H., Wee, C.-Y., Yap, P.-T., Shen, D.: Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. NeuroImage **91**, 386–400 (2014)
9. Lo, R.Y., Jagust, W.J.: Predicting missing biomarker data in a longitudinal study of Alzheimer disease. Neurology **78**, 1376–1382 (2012)
10. García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R.: Pattern classification with missing data: A review. Neural Computing and Applications **19**(2), 263–282 (2010)
11. Maronna, R.A., Martin, D.R., Yohai, V.J.: Robust Statistics: Theory and Methods. John Wiley and Sons, New York (2006)
12. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statistics Surveys **4**, 40–79 (2010)
13. Schneider, T.: Analysis of incomplete climate data: Estimation of mean valuesand covariance matrices and imputation of missing values. Journal of Climate **14**, 853–871 (2001)
14. Gray, K., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D.: Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage **65**, 167–175 (2013)
15. Báez, P.G., Araujo, C.P.S., Viadero, C.F., García, J.R.: Automatic prognostic determination and evolution of cognitive decline using artificial neural networks. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 898–907. Springer, Heidelberg (2007)
16. Pelckmans, K., Brabanter, J.D., Suykens, J.A.K., Moor, B.D.: Handling missing values in support vector machine classifiers. Neural Networks **18**(5–6), 684–692 (2005)