Pattern Recognition 43 (2010) 2340–2350

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Selection–fusion approach for classification of datasets with missing values

Mostafa Ghannad-Rezaie [a,b,c], Hamid Soltanian-Zadeh [a,d,*], Hao Ying [b], Ming Dong [e]

[a] *Department of Diagnostic Radiology, Henry Ford Hospital, Detroit, MI 48202, USA*
[b] *Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202, USA*
[c] *Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48105, USA*
[d] *Control and Intelligent Processing Center of Excellence, Electrical and Computer Engineering Department, University of Tehran, Tehran 14395-515, Iran*
[e] *Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*

ARTICLE INFO

Article history:
Received 11 January 2009
Received in revised form
12 September 2009
Accepted 3 December 2009

Keywords:
Missing value management
Subspace classifiers
Ensemble classifiers
Multiple imputations
Pruning
Support vector machine (SVM)

ABSTRACT

This paper proposes a new approach based on missing value pattern discovery for classifying incomplete data. This approach is particularly designed for classification of datasets with a small number of samples and a high percentage of missing values where available missing value treatment approaches do not usually work well. Based on the pattern of the missing values, the proposed approach finds subsets of samples for which most of the features are available and trains a classifier for each subset. Then, it combines the outputs of the classifiers. Subset selection is translated into a clustering problem, allowing derivation of a mathematical framework for it. A trade off is established between the computational complexity (number of subsets) and the accuracy of the overall classifier. To deal with this trade off, a numerical criterion is proposed for the prediction of the overall performance. The proposed method is applied to seven datasets from the popular University of California, Irvine data mining archive and an epilepsy dataset from Henry Ford Hospital, Detroit, Michigan (total of eight datasets). Experimental results show that classification accuracy of the proposed method is superior to those of the widely used multiple imputations method and four other methods. They also show that the level of superiority depends on the pattern and percentage of missing values.

1. Introduction

Missing value is a common problem in real-world data processing and pattern recognition. Management of missing values becomes critical when the number of available samples is small [1]. Modifying an algorithm primarily designed to work on complete datasets to work on incomplete datasets is a challenge. In general, an appropriate strategy based on the ultimate processing goal may be developed. However, in the case of datasets with a small number of samples, not only the final goal but also the percentage and the distribution of missing values should be considered in algorithm development [2,3].

Traditional missing value management methods are based on the preprocessing of the data independent of the final goal and the associated processing scheme. In these methods, the missing values are estimated or the deficient samples are removed [1]. Although in this approach the data processing algorithm does not need to change, the data is not efficiently used, especially when a large portion of the samples have missing features. Modern missing value management methods are designed for specific applications and associated processing schemes where missing value management is integrated into the processing scheme [4]. These algorithms either apply multiple data processing stages, e.g., multiple imputations or somehow avoid the unknown values in the processing scheme, e.g., decision trees.

Although modern algorithms are shown to be successful in different applications, their proposed solutions are not designed to deal with a high percentage of missing features or a large number of systematic missing values that are frequent scenarios in some data categories such as medical datasets [1]. The main challenge arises from insufficient statistical power after the missing values are imputed. In this situation, the following questions arise:

- How to measure the complexity of the missing values?
- How to work with the missing values when imputation of the missing values is inappropriate?
- How to manage the missing values when the same features are missing in the test and training samples?

This paper proposes a new approach, named selection–fusion, based on the subspace classification method. In the proposed approach, missing value management is integrated not only in the training but also in the testing of the classifier. To this end, a set of

[*] Corresponding author at: Radiology Image Analysis Lab., One Ford Place, 2F, Detroit, Michigan 48202, USA. Tel.: +1 313 874 4482; fax: +1 313 874 4494.
E-mail address: hamids@rad.hfh.edu (H. Soltanian-Zadeh).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.12.003

classifiers are trained on the subspaces of the original feature space and then clustered using a distance metric. The best classifiers in each cluster, depending on the testing data, are combined to construct the overall classifier and estimate the final output.

The proposed approach is compared with the multiple imputations method as the most similar incomplete data processing method. Our major contributions can be summarized as follows:

- As part of the algorithm, we define a quantitative measure for the complexity of the missing values. Based on this measure, the usefulness of the algorithm for a particular dataset can be evaluated.
- We consider missing values in both of the training and the testing datasets without filling the missing values.
- We show that the proposed approach can be efficiently implemented for the support vector machine classifiers.

The rest of the paper is organized as follows. In the next section, we review the state-of-the-art for incomplete data processing. Details of the proposed selection–fusion method and its application to missing value management are described in Section 3. In this section, we address the above three challenges using multi-classifier fusion. We describe how each classifier is selected and how the results are combined to boost up the performance. The experimental results are presented in Section 4. We highlight the application areas of the new method and also discuss its limitations in Section 5 and conclude the paper in Section 6.

2. Related work

In a missing value problem, considerable portions of the data fields may be incomplete. To describe the seriousness of the data deficiency, the primary question in a typical missing value problem is “the missing value pattern.” For example, in Pneumonia data described in [6], on average 6.3% of the feature values are missing while one individual feature is missing for 61% of the cases. On the other hand, in C-Section problem [6], only 1.2% of the feature are missing, while 27.9% of the cases have at least one missing feature. However, these figures do not provide clear ideas about the complexity of these problems. In fact, despite a smaller percentage of the missing values, the second problem is more complicated than the first.

To describe the complexity of a missing value pattern, some statistical models are used in the literature. *Missing completely at random* (MCAR) and *missed at random* (MAR) are the models most frequently used in the database literature. Although due to their simplicity, they are not always realistic models for the real-world problems, they provide relative measures of complexity. The missing value for a random variable X is MCAR if the missing probability is independent of the actual value of X or the values of the other features. The missing value is called MAR if the missing probability is independent of the value of X after controlling the other variables. Missing values due to equipment malfunction is an example of the MCAR well-described pattern. However, in many real-world applications, MAR is a more realistic model than MCAR [2].

Generally speaking, there are five classes of well-established strategies to deal with the missing values: (1) discard the incomplete samples (e.g., *pairwise deletion* [2]); (2) avoid the missing features by dynamic decisions (e.g., *decision trees such as CART* [7]); (3) recover unknown values from the similar samples (e.g., *Expectation Maximization* (EM) [8]); (4) insert possible values for the missing features, classify after each insertion and combine the classification results (e.g., *Multiple Imputations* (MI) [9]); and (5) design multiple classifiers on the subsets of the data and combine the classification results (e.g., *ensemble classifiers* [17]).

Discarding the incomplete samples and filling the missing values are very simple but undesirable methods for a dataset with a small number of samples and a large percentage of missing values. The former approach may discard significant amount of information when the number of samples is limited and the latter approach may add considerable distortion to the data when the percentage of the missing values is high.

Recovering the missing values form the other samples, also called single imputation, is the traditional approach for the treatment of incomplete datasets with a small number of samples. Many single imputation methods have been proposed over the years. Decision tree imputation [7], nearest neighbor imputation [10], and mean value substitution [11,12] are examples of classical imputation methods. These methods are only valid under specific assumptions such as MCAR assumption for the mean value substitution approach or dense sampling for the nearest neighbor imputation approach. Bayesian missing value treatment is a modern approach that replaces the missing values with the most probable values [8].

From the classification point of view, there is a common problem in all traditional missing value treatment methods: they provide a solution independent of the ultimate goal. Multiple imputations (MI) method [1,9] is an alternative solution that uses Monte Carlo simulation to generate more than one imputation of the missing values. However, the MI usually implies several assumptions on the data distribution such as joint normality [13] and regression relationships [14]. Application of MI is particularly favorable when the number of samples is relatively small (100 cases or less). Markov Chain Monte Carlo (MCMC) method is a successful MI method for datasets with a small number of samples [13–15].

Recently, ensemble classifiers technique has been shown to be a valuable tool for missing value management. In this approach, the results of multiple weak classifiers are combined to boost-up the performance. Different groups have shown effectiveness of this approach for general classification problems [16,17]. Recently, it has also been applied to the missing value problem [18]. Despite its advantages, this approach suffers from two major limitations in its application to the missing value problem: (1) lack of mathematical framework for the selection of the weak classifiers and (2) handling of the missing values in the testing data. In this paper, we overcome both of these limitations.

From the performance point of view, the most effective ensemble approach in the literature utilizes fusion. In this approach, outputs of a s et of inaccurate classifiers are combined to generate highly accurate classification results. A simple implementation of this idea, also known as selection–fusion (SF), trains each classifier on a random subset of data [19]. This implementation is shown to be effective for small datasets and improve the performance compared with traditional methods. However, in the large datasets, since the number of possible classifiers increases quickly, this implementation of fusion would not work well. A systematic method to find an optimal set of classifiers, as proposed in the paper, solves the problem using a manageable number of classifiers. In addition, when there are missing values in the data, as is the case in this paper, random selection of the subsets is inapplicable.

In general, both of the testing and training datasets may have missing values. When a feature is missing in the testing data, filling the missing value is the most common approach [19,20]. The advantage of the filling method has been mostly discussed under certain conditions like the MCAR model and a sufficient sample size. Apparently, the performance degrades if these