Knowledge-Based Systems 154 (2018) 1–16

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# An effective and efficient approach to classification with incomplete data

Cao Truong Tran[*,a,b], Mengjie Zhang[a], Peter Andreae[a], Bing Xue[a], Lam Thu Bui[b]

[a] School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
[b] Research Group of Computational Intelligence, Le Quy Don Technical University, 236 Hoang Quoc Viet St, Hanoi, Vietnam

ARTICLE INFO

*Keywords:*
Incomplete data
Missing data
Classification
Imputation
Feature selection
Ensemble learning

ABSTRACT

Many real-world datasets suffer from the unavoidable issue of missing values. Classification with incomplete data has to be carefully handled because inadequate treatment of missing values will cause large classification errors. Using imputation to transform incomplete data into complete data is a common approach to classification with incomplete data. However, simple imputation methods are often not accurate, and powerful imputation methods are usually computationally intensive. A recent approach to handling incomplete data constructs an ensemble of classifiers, each tailored to a known pattern of missing data. The main advantage of this approach is that it can classify new incomplete instances without requiring any imputation. This paper proposes an improvement on the ensemble approach by integrating imputation and genetic-based feature selection. The imputation creates higher quality training data. The feature selection reduces the number of missing patterns which increases the speed of classification, and greatly increases the fraction of new instances that can be classified by the ensemble. The results of experiments show that the proposed method is more accurate, and faster than previous common methods for classification with incomplete data.

## 1. Introduction

Classification is one of the most important tasks in machine learning and data mining [1]. Classification consists of two main processes: a training process and an application (test) process, where the training process builds a classifier which is then used to classify unseen instances in the application process. Classification has been successfully applied to many scientific domains such as face recognition, fingerprint, medical diagnosis and credit card fraud transaction. Many algorithms have been proposed to deal with classification problems, but the majority of them require complete data and cannot be directly applied to data with missing values. Even when some methods can be applied, missing values often lead to big classification error rates due to inadequate information for the training and application processes [2].

Unfortunately, missing values are a common issue in numerous real-world datasets. For example, 45% of the datasets in the UCI machine learning repository [3], which is one of the most popular benchmark databases for machine learning, contain missing values [2]. In an industrial experiment, results can be missing due to machine failure during the data collection process. Data collected from social surveys is often incomplete since respondents frequently ignore some questions. Medical datasets usually suffer from missing values because typically not all tests can be done for all patients [4,5]. Financial datasets also often contain missing values due to data change [6,7].

One of the most common approaches to classification with incomplete data is to use imputation methods to substitute missing values with plausible values [4,8,9]. For example, mean imputation replaces all missing values in a feature by the average of existing values in the same feature. Imputation can provide complete data which can then be used by any classification algorithm. Simple imputation methods such as mean imputation are often efficient but they are often not accurate enough. In contrast, powerful imputation methods such as multiple imputation [10] are usually more accurate, but are computationally expensive [11,12]. It is not straightforward to determine how to combine classification algorithms and imputation in a way that is both effective and efficient, particularly in the application process.

Ensemble learning is the process of constructing a set of classifiers instead of a single classifier for a classification task, and it has been proven to improve classification accuracy [13]. Ensemble learning also has been applied to classification with incomplete data by building multiple classifiers in the training process and then applicable classifiers are selected to classify each incomplete instance in the application process without requiring any imputation method [14–16]. However, existing ensemble methods for classification with incomplete data often cannot work well on datasets with numerous missing values [14,16]. Moreover, they usually have to build a large number of classifiers, which then require a lot of time to find applicable classifiers for each incomplete instance in the application process, especially when

[*] Corresponding author at: School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington, 6140, New Zealand.
*E-mail address:* cao.truong.tran@ecs.vuw.ac.nz (C.T. Tran).

https://doi.org/10.1016/j.knosys.2018.05.013
Received 17 January 2018; Received in revised form 8 May 2018; Accepted 10 May 2018

incomplete datasets contain a high proportion of missing values [14,15]. Therefore, how to construct a compact set of classifiers able to work well even on datasets with numerous missing values should be investigated.

Feature selection is the process of selecting relevant features from original features, and it has been widely used to improve classification with complete data [17]. Feature selection has also been investigated in incomplete data [18,19], but the existing methods typically still use imputation to estimate missing values in incomplete instances before classifying them. By removing redundant and irrelevant features, feature selection has the potential of reducing the number of incomplete instances, which could then improve accuracy and speed up classifying incomplete instances. However, this aspect of feature selection has not been investigated. This paper will show how to utilise feature selection to improve accuracy and speed up the application process for classification with incomplete data.

### 1.1. Goals

To deal with the issues stated above, this paper aims to develop an effective and efficient approach for classification with incomplete data, which uses three powerful techniques: imputation, feature selection and ensemble learning. Imputation is used to transform incomplete training data to complete training data which is then further enhanced by feature selection. After that, the proposed method builds a set of specialised classifiers which can classify new incomplete instances without the need of imputation. The proposed method is compared with other common approaches for classification with incomplete data to investigate the following main objectives:

1. How to effectively and efficiently use imputation for classification with incomplete data; and
2. How to use feature selection for classification with incomplete data to not only improve classification accuracy but also speed up classifying new instances; and
3. How to build a set of classifiers which can effectively and efficiently classify incomplete instances without the need of imputation; and
4. Whether the proposed method can be more accurate and faster than using imputation both in the training process and the application process; and
5. Whether the proposed method can be more accurate and faster than the existing ensemble methods.

### 1.2. Organisation

The rest of this paper is organised as follows. Section 2 presents a survey of related work. The proposed method is described in Section 3. Section 4 explains experiment design. The results and analysis are presented and discussed in Section 5. Section 6 states conclusions and future work.

## 2. Related work

This section firstly introduces traditional approaches to classification with incomplete data. It then discusses ensemble learning for classification with incomplete data. Finally, it presents typical work on feature selection.

### 2.1. Traditional approaches to classification with incomplete data

There are several traditional approaches to classification with incomplete data. The deletion approach simply deletes all instances containing missing values. This approach is limited to datasets with only a few missing values in the training data and no missing values in the application process [20]. A second approach is to use one of classifies such as C4.5 which can directly classify incomplete datasets using a probabilistic approach [21]. However, their accuracy is limited when there are a lot of missing values [22].

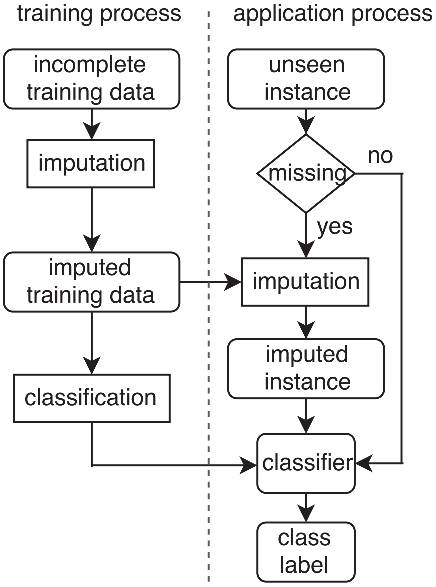

**Fig. 1.** A common approach to using imputation for classification with incomplete data.

The most used approach to classification with incomplete data is to use imputation methods to transform incomplete data into complete data before building a classifier in the training process or classifying a new incomplete instance in the application process. This approach has the advantage that the imputed complete data can be used by any classification algorithm. This approach also can deal with incomplete datasets with a large number of missing values [8,23].

Fig. 1 shows the main steps using imputation for classification with incomplete data. In the training process, imputation is used to estimate missing values for incomplete training data. After that, imputed training data is put into a classification algorithm to build a classifier. In the application process, complete instances are directly classified by the classifier. With each incomplete instance, its missing values are first replaced by plausible values by using the imputation to generate a complete instance which is then classified by the classifier.

There are two classes of imputation: single imputation and multiple imputation.

#### 2.1.1. Single imputation

Single imputation estimates a single value for each missing value. Mean imputation is an example of single imputation methods which fills all missing values in each feature by the average of all existing values in the same feature.

kNN-based imputation is one of the most powerful single imputation methods [22]. To estimate missing values in an incomplete instance, it first searches for its $k$ nearest neighbour instances. After that, it replaces missing values of the instance with the average of existing values in the $k$ instances. kNN-based imputation is often more accurate than mean imputation [22]. However, kNN-based imputation is more computationally expensive than mean imputation because it takes time to find the nearest instances, especially with datasets containing a large number of instances and a large value of $k$ [22].

Single imputation has been widely used to estimate missing values for classification with incomplete data [8,20,22,23]. In [20] and [22], kNN-based imputation is shown to outperform C4.5, mean imputation and mode imputation. In [23], the impact of six imputation methods on classification accuracy for six classification algorithms are investigated. Results show imputation on average can improve classification accuracy when compared to without using imputation. However, in [8],