Expert Systems With Applications 89 (2017) 52–65

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers

Unai Garciarena*, Roberto Santana

Faculty of Informatics, University of the Basque Country. Paseo Manuel Lardizabal, 1 - 20018 Donostia-San Sebastián, Gipuzkoa, Spain

ARTICLE INFO

Article history:
Received 14 November 2016
Revised 23 June 2017
Accepted 15 July 2017
Available online 17 July 2017

Keywords:
Missing data
Imputation methods
Supervised classifiers
Machine learning

ABSTRACT

When applying data-mining techniques to real-world data, we often find ourselves facing observations that have no value recorded for some attributes. This can be caused by several phenomena, such as a machine's incapability to record certain characteristics or a person refusing to answer a question in a poll. Depending on that motivation, values gone missing may follow one kind of pattern or another, or describe no regularity at all. One approach to palliate the effect of missing data on machine learning tasks is to replace the missing observations. Imputation algorithms attempt to calculate a value for a missing gap, using information associated with it, i.e., the attribute and/or other values in the same observation. While several imputation methods have been proposed in the literature, few works have addressed the question of the relationship between the type of missing data, the choice of the imputation method, and the effectiveness of classification algorithms that used the imputed data. In this paper we address the relationship among these three factors. By constructing a benchmark of hundreds of databases containing different types of missing data, and applying several imputation methods and classification algorithms, we empirically show that an interaction between imputation methods and supervised classification can be deduced. Besides, differences in terms of classification performance for the same imputation method in different missing data patterns have been found. This points to the convenience of considering the combined choice of the imputation method and the classifier algorithm according to the missing data type.

1. Introduction

Missing values are ubiquitous in almost every type of real-world datasets. They can be particularly detrimental for certain applications of the datasets, especially when the distribution of the missing data (MD) is not uniform and a possible mechanism that could explain the lost values is unknown. Perhaps the most used among the non-trivial alternatives to deal with MD are imputation methods (IMs). These methods replace the missing values by estimates that can be taken from the database (DB), derived from statistics of known values (e.g., the mean of a given variable), or obtained using more sophisticated algorithms.

There is consensus on the importance of the application of IMs, especially when DBs with MD are used as a basis for learning supervised classifiers. However, the choice of the IM, and its impact on the classifier performance can be very dependent on the MD type. For example, an improper choice of the IM can bias the learned classifier, producing a low classification quality on test data.

When the problem of supervised classification is considered, these three elements are strongly intertwined. In this paper we analyze this relationship by investigating problems with different types of MD, addressed using a set of IMs with the final goal of supervised classification by means of different types of classifiers. Our aim is to determine to what extent there is a relationship between the choice of the IM and the precision of the classifiers when considering DBs that exhibit different types of MD.

Previous work (Batista & Monard, 2003; Luengo, García, & Herrera, 2012) has analyzed the relationship between the IMs used for treating MD and classifiers. Batista and Monard (2003) evaluated four IMs for two different classifiers concluding that the choice of the IM influences the performance of the classifiers. A more in-depth study on the relationship between IMs and classifiers was presented by Luengo et al. (2012). The authors conducted an extensive evaluation of classifiers and IMs on real-world DBs and concluded that the choice of the IM should indeed be conditioned on the type of classification method used.

* Corresponding author.
E-mail addresses: unai.garciarena@ehu.es (U. Garciarena), roberto.santana@ehu.es (R. Santana).

http://dx.doi.org/10.1016/j.eswa.2017.07.026
0957-4174/

In this paper we go beyond the analysis of the relationship between IMs and classification algorithms, and consider as another factor the particular characteristics of the MD. We hypothesize that the three previously mentioned factors can influence the classification results and should be considered in their interaction. We investigate this hypothesis by devising procedures that generate DBs with different types of MD, and using them as a benchmark, we evaluate the effect of the MD type and the IMs on the performance provided by the classifier. Another contribution of our work is the simultaneous use of real-world DBs, which are used as a basis to construct the benchmark, with an artificially generated MD type which is introduced in the original DB. Following this strategy, we can control the characteristics of the MD and evaluate the effect on the other factors analyzed. In our investigation we also evaluate an extensive number of classifiers, including many of those investigated in previous works and some other more recent classification approaches.

The paper is organized as follows. In the next section, some essential background on the main concepts covered in the paper is given. Related work, emphasizing the connection with our proposal, is discussed in Section 3. Section 4 gives a formal presentation of the methods used to generate the different types of MD. This section also describes the databases selected to evaluate the relationships between the methods and algorithms. Sections 5 and 6 respectively present the imputation and classification methods investigated. In Section 7 we describe the experimental framework, the results of the experiments, and discuss some of our findings. Section 8 concludes the paper and presents some lines for future research.

2. Background

2.1. Missing data types

Many different reasons can cause MD in real-world databases. Identifying any pattern in the MD is a key aspect when conceiving methods to deal with the missing observations. In particular, the type of MD can directly impact the quality of the predictions of the classification methods applied to the data. Therefore, several works have been devoted to characterizing the types of MD, and suggesting algorithms for imputation. In this section we review the most common accepted classes of MD and their expected effect on the behavior of supervised classification techniques (Batista & Monard, 2003; Blomberg & Ruiz, 2013; Gelman & Hill, 2006; Hernández-Pereira, Álvarez-Estévez, & Moret-Bonillo, 2015; Luengo et al., 2012):

- Missing Completely At Random (MCAR): When the database's measurement failures occur randomly, there is no specific pattern to be identified. The impact of MCAR on a classification algorithm will depend on the MD distribution over the data. The more uniform the distribution of the MD is, the less bias is expected to be introduced in the database.
- Missing At Random (MAR): MD is cataloged as MAR when a pattern can be identified, i.e., we can find a common factor in all the observations with missing values. For example, we find that when a certain variable (with no MD) takes extreme values for an observation, two other variables tend to be missing for that same observation.
- Missing Not At Random (MNAR): This MD type is similar to MAR. However, in this case the values *causing* others to be missing are not known, this can have two origins:
  - Missingness depending on unobserved Variables (MuOV): One of the reasons these values are not known can be that simply they were not observed.
  - Missingness depending on its Value Itself (MIV): An element can be missing depending on its value itself. This could happen when a variable takes a value out of its representation range.

In general, it is not possible to identify the MCAR type of MD, since in real databases there is no way to track the cause of this MD. MCAR can be caused by a huge variety of reasons, from data loss during an information transference, to a person's refusal to provide personal data in a poll, etc. As stated above, assuming that the missing values are uniformly distributed, the dataset will not experience a considerable loss of information. As long as the amount of missing values is not significant, even discarding observations containing MD will not necessarily have an impact on posterior classification. However, even though the amount of information lost regarding missing observations may be small, the quality of the IM could be as important as in other MD types.

The MAR kind of MD is not as common as MCAR, but it is easier to infer its origin by studying other variables of the dataset. For example, in a situation in which people are asked about their habits and health, some information about sedentary lifestyle might be available. However, while some subjects may be open to share information about their weight, other subjects (more likely those with an overweight condition) might be more reluctant to disclose this type of information. This example illustrates situations in which a cause for MAR can be inferred from an analysis of the characteristics of the database.

The MAR type of MD can be a potential source of problems for the performance offered by the classification algorithms. Since in this case there is an underlying reason for the MD, it is likely that observations containing MD will be similar to each other and will be tagged in the same class. This could lead to an unbalanced database that will potentially affect the classification. In this case, discarding data is not an advisable option, and the use of IMs is a requirement.

Finally, MNAR presents a considerably more difficult situation. Following the previous example, the task would become much more tedious if we had not asked about other medical and lifestyle parameters (MuOV). Another scenario needs to be addressed when an individual is ashamed and refuses to disclose, for example, information about the amount of money he or she spends on drugs. This variable most likely depends only on itself. In this second case we would have MIV. These two situations are the most problematic ones, since it can be as harmful as MAR for the data, but it can be easily misidentified as MCAR, as a result of the impossibility identifying a pattern in the unobserved variables.

2.2. Imputation methods

There are two main ways of handling MD. Ignoring observations with missing data is generally a good choice when the cardinality of missing values is relatively small and the MD is homogeneously distributed. Nevertheless, when these measurement failures are concentrated in a single variable, or when ignoring them would suppose a big loss of information, we consider techniques to fill the gaps. This is essentially what imputation does. Several strategies have been proposed for this purpose and they can exhibit important differences in terms of complexity, and output quality (Batista & Monard, 2003; Brownstone & Valletta, 2001; Lakshminarayan, Harp, & Samad, 1999; Liu & Brown, 2013). Section 5 will present a number of imputation methods relevant for our work.

2.3. Classification problems and supervised classification algorithms

Classification is the task of learning a target function that maps an attribute collection to a predefined class. Algorithms that solve