Neurocomputing 205 (2016) 152–164

Contents lists available at ScienceDirect

Neurocomputing

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

Missing data imputation using fuzzy-rough methods

Mehran Amiri [a,*], Richard Jensen [b]

[a] Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Kerman, Iran
[b] Department of Computer Science, Aberystwyth University, Ceredigion SY23 3DB, Wales, UK

ARTICLE INFO

Article history:
Received 13 July 2015
Received in revised form
12 February 2016
Accepted 18 April 2016
Available online 9 May 2016

Keywords:
Missing value imputation
Fuzzy-rough sets
Vaguely quantified rough sets
Ordered weighted average-based rough sets

ABSTRACT

Missing values exist in many generated datasets in science. Therefore, utilizing missing data imputation methods is a common and important practice. These methods are a kind of treatment for uncertainty and vagueness existing in datasets. On the other hand, methods based on fuzzy-rough sets provide excellent tools for dealing with uncertainty, possessing highly desirable properties such as robustness and noise tolerance. Furthermore, they can find minimal representations of data and do not need potentially erroneous user inputs. As a result, utilizing fuzzy-rough sets for imputation should be an effective approach. In this paper, we propose three missing value imputation methods based on fuzzy-rough sets and its recent extensions; namely, implicator/t-norm based fuzzy-rough sets, vaguely quantified rough sets and also ordered weighted average based rough sets. These methods are compared against 11 state-of-the-art imputation methods implemented in the KEEL data mining software on 27 benchmark datasets. The results show, via non-parametric statistical analysis, that the proposed methods exhibit excellent performance in general.

1. Introduction

In today's world, the understanding of behaviors of phenomena can be achieved by analyzing relevant datasets. These interpretations could be for classification, regression or time series data. Perhaps the most beneficial part of this work is the prediction of parameter values which are of high importance. Thus, gathering relevant data related to a phenomenon is widely carried out. Many different datasets are generated everyday in most fields of science and are constructed in different ways. There are a number of factors that can affect the data gathered; for example, power system failures, noise, environmental factors (such as humidity, temperature, etc), a lack of response in scientific experiments, human error in measurements, problems of data transfer in digital systems or respondents' unwillingness to respond to survey questions, low quality of sensors and many more [1–4]. Hence, the existence of missing values in gathered datasets is somewhat inevitable. However, the presence of missing values could dramatically degrade the results of interpretation of datasets which are usually carried out with the aid of machine learning techniques [5]. Therefore, dealing with missing values is an important issue in data mining and machine learning communities [5–9].

There are several ways to deal with missing values in datasets. Deleting or ignoring them are the simplest approaches. Replacing the missing values with zero or the mean of the attributes is another option. These treatment methods have a major drawback in that they degrade the quality of estimations by removing some information present in instances containing missing values. This could potentially bias the results of estimation. Hence, another option is to deal with this problem by estimating the missing values. These methods are usually called *imputation* methods.

Based on the process undergone to produce a dataset and also the nature of the data itself, there could be three types of missing values. These are: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) [10]. When missing values are considered to be of type MCAR, this means that the missing values are independent of other variables. If the values are considered to be of type MAR, then these can be estimated using other values (i.e. the mechanism by which values are missing can be ignored). And finally, if the values are of type NMAR then these values depend on other missing variables and the mechanism by which values are missing will need to be modeled for effective imputation. Hence the missing values cannot be estimated from existing variables. For most approaches, the missing values are assumed to be of type MAR.

* Corresponding author. Postal address: No.1, Block 20, Tabatabaie Alley, South Sardar-e-Jangal Blvd, Poonak, Tehran, Iran. Tel.: +98 918 359 1069, +98 935 374 1942.
E-mail address: m.amiri.ac@gmail.com (M. Amiri).

http://dx.doi.org/10.1016/j.neucom.2016.04.015
0925-2312/

Missing value imputation methods could be categorized based on the technique used for approximating missing values. They could be classified into two different types [11]. First, there are approaches that use mathematical or statistical methods to predict missing values. These consist of very simple methods which simply replace missing values with the mean or mode value of the features and also more sophisticated methods that are based on more advanced statistical techniques. Second, there are methods that use machine learning techniques to impute missing values. Methods belonging to this category build a model or a combination of models based on information available in the dataset. Afterwards, missing values are predicted using this model. Sometimes missing values are predicted during the training phase and they are iteratively amended to achieve best values. This category itself can be divided into some subcategories, such as methods using Neural Networks [12–16], Support Vector Machines [17], Nearest Neighbor based methods [18,19], and also methods based on unsupervised learning, e.g. clustering [7,8].

Methods which are based on Nearest Neighbors simply predict missing values based on complete instances which are located in the proximity of the instances with missing values. These methods are accurate, but the main reason to propose and use them is that they are intuitive and simple. Besides, they have some drawbacks such as a need to provide the number of neighbors by the user, a need to compare all instances to find NNs which results in a high time complexity and also the problem of local optima due to their local nature.

The use of simple statistical models can cause the resulting imputed data to become biased. Thus, models built using the data will also be biased. On the other hand, methods based on the statistical learning are typically more complex. They cannot directly deal with missing values, thus at the very beginning all of them need an initial guess to be optimized later. This initial guess is usually the mean of the feature values. Most of them use eigenvectors to describe data and they usually ignore some of the less important eigenvectors. This causes data loss and the final outcome of this could be degradation in accuracy of predictions. They also need the user to determine the number of the most important eigenvectors as an input. This might require expert knowledge in order to produce good results. Since these methods are iterative, they also need a predefined threshold to stop. This is usually given to the algorithm by the user. Furthermore, their performance is sensitive to the type of data being analyzed. For instance, SVDI [19] usually performs better on time series data. Most of them are essentially a form of linear regression.

Methods based on Neural Networks are another option to impute missing values. This family of methods usually defines an error and iteratively tries to minimize this. A main drawback of this set of algorithms is their time complexity; for example, Multi Task Learning (MTL) neural networks [20] use a quadratic error for minimization. They also need a predefined threshold to stop. Methods based on clustering are alternatives, but have some drawbacks: they have higher time complexities and they need user specified thresholds to terminate. Some of these methods need a user-specified number of clusters at the start. Most of them usually converge to a local minimum. To converge to a global minimum, several repetitions should be undertaken, which is very computationally expensive.

In contrast, fuzzy-rough set theory provides an excellent framework to deal with uncertainty and has some desirable characteristics which makes it a good choice to be applied to the problem of imputation. Fuzzy rough methods are not essentially optimization problems; thus, they do not iterate through algorithm steps. This is important because they do not need a stopping criterion when finding a good one is usually hard. They also do not need user-specified parameter values which could be erroneous. Another reason to use fuzzy-rough techniques for this domain is their simplicity and understandability. They simply calculate the fuzzy similarities of instances and make decisions based on these. They can easily and effectively work in presence of noise and also can deal with missing data (as demonstrated in this paper). Hence, they do not need initial guesses for missing values. Furthermore, they can easily deal with imprecise data. The missing value imputation domain needs methods which can deal with imprecise data easily and also effectively. The mentioned reasons make fuzzy-rough techniques suitable for imputation of missing values.

The nearest neighbor algorithm has proven itself as a very accurate method, yet simple. In this paper we have introduced several methods to impute missing values based on fuzzy-rough sets combined with the nearest neighbor algorithm. This way, one can benefit from both the simplicity and accuracy of nearest neighbor prediction along with the robustness and noise tolerance of fuzzy-rough sets. The accuracy of fuzzy-rough nearest neighbor methods has already been demonstrated in the classification domain [21]. We have used three types of fuzzy rough sets; namely, implicator/t-norm based fuzzy-rough sets, vaguely quantified rough sets and also ordered weighted average-based fuzzy-rough sets. The two latter approaches are proven to be more robust in the presence of noise.

The rest of the paper is as follows: In the next section we review the literature in the area and focus on the major approaches. Section 3 describes the theoretical background necessary to understand the proposed methods. Afterwards, the proposed algorithms are introduced. These methods are then applied to benchmark data and evaluated in Section 5 using non-parametric statistical analysis. The last section concludes the paper and outlines future work.

2. Literature review

There are many imputation methods in the literature based on different approaches. Although many such algorithms exist, they are often proposed to be used in specific domains or even for specific datasets, e.g. transportation [22], meteorology [23] and others [24]. Hence, they are not publicly used. In contrast, there are several imputation methods which are used in many domains. In this section, we are going to focus mainly on these general algorithms, as the focus of this paper is a general approach to missing value imputation.

In [18], an intuitive method for the imputation of missing values is introduced, called KNNI. KNNI finds the nearest neighbors of instances and replaces the missing value with the mean value of the specific feature of its neighbors. The number of neighbors for this algorithm must be found empirically. Weighted nearest neighbor imputation (WKNNI) [19] is another imputation method based on nearest neighbors. This method uses weights based on the Euclidean distance in order to better estimate the missing values. Again this method needs the number of neighbors to be determined empirically. In [19] another imputation method is introduced, SVDI, which is developed to deal with missing gene expression values in DNA arrays, using singular value decomposition to find gene expression patterns which could be linearly combined to approximate the missing values. Since SVD cannot deal with missing values, the missing values of the dataset are initially replaced by the corresponding average. After this process, an expectation maximization algorithm is utilized to reach a final approximation for the missing values. The algorithm firstly finds all eigenvectors of a dataset, which are called eigengenes. Their relative eigenvalues are found also. Afterwards, eigengenes are sorted based on their corresponding eigenvalues. The k most significant eigengenes are selected consequently. To find the missing