Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 269 (2015) 30–46

FUZZY sets and systems

www.elsevier.com/locate/fss

Rough subspace neuro-fuzzy system

Krzysztof Simiński

Institute of Informatics, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland

Received 3 January 2014; received in revised form 3 July 2014; accepted 6 July 2014

Available online 11 July 2014

Abstract

The missing values can be an important obstacle and challenging problem in data analysis. The paper presents the neuro-fuzzy system that handles incomplete data. The system is complete: it can extract the fuzzy rule base from both complete and incomplete data and can elaborate answers for complete and incomplete data. The second major feature of the system is the assignment of weights to attributes in fuzzy rules. The weights are assigned locally: each fuzzy rule has its own weights for attributes. This feature may improve the precision of answers elaborated by the system and may reveal relations between attributes in the data set.

The paper is accompanied by experimental results. The results show that the subspace technique is advantageous in handling data set with missing values. The results also reveal that for approximation of complete data it is better to apply techniques without subspace approach.

Keywords: Missing values; Incomplete data; Subspace analysis; Neuro-fuzzy system; Rough fuzzy approach

1. Introduction

The real life data are sometimes far from being ideal. They may suffer from two problems: (1) the diversity of importance of dimensions (attributes) and (2) incompleteness of data.

Some attributes (features, dimensions) may be of minor importance or even represent nothing but noise. Many methods for removing of attributes have been proposed. The global removal of attributes by feature transformation (e.g. principal component analysis or singular value decomposition) may deteriorate the interpretability of elaborated models. Some data may exist in certain subspaces of the domain. It is possible that clusters of data exist each in its own subspace defined by different attributes (data features). In such cases global removal of attributes is not recommended. This is why subspace data analysis algorithm has been proposed. The popular data analysis is clustering. There are two kinds of subspace clustering: bottom-up and top-down [44]. The bottom-up approach first applies the grid partition and then analyses the density of data in each grid cell. Simultaneously the important dimensions are selected [10,5,24]. The top-down approach starts with full dimensional clusters and tries to remove the less important attributes

E-mail address: krzysztof.siminski@polsl.pl.

http://dx.doi.org/10.1016/j.fss.2014.07.003
0165-0114/

[3,22,21,4,63]. The common feature of all these algorithms is the assignment of the value 1 when the attribute is valid in the subspace of the cluster and 0 when the cluster's space is not described by this attribute.

The second problem mentioned at the beginning is the lack of values in data tuples. The reasons are various [56]: errors in answer acquisition, failure of sensors, impossibility to get data (e.g. patient has died), the refusal to answer some questions in the questionnaire, inapplicability of questions, random noise, impossible values, retrospective usage of data i.e. the data were gathered for other purpose then the research needs. In [47] the medical example is given where only 1 patient in 55 had all blood tests done. Overall 9.2% of blood test results are missing. In [33] the real life data set is presented with more than 50% of values missing. The paper [35] presents three classes of missing data types:

MCAR – missing completely at random: The probability of a tuple having a missing value for an attribute depends neither on the known values nor on the missing data, e.g. the data values missing from the whole data set.

MAR – missing at random: The probability that the tuple have a missing value for an attribute may depend on the known values, but not on the value of the missing data itself, e.g. the data miss at random only from one cluster;

NMAR – not missing at random: The probability of an instance having missing value for an attribute could depend on the value of that attribute, e.g. the values greater than 100 miss with probability 0.1.

The paper [2] defines four classes of difficulty in function of the ratio of missing values: (1) trivial data with less than 1% of missing values, (2) manageable data with 1–5% of missing values, (3) data requiring some sophisticated approach (5–15% of missing value) and (4) data with more than 15% missing value that "severely impact any kind of interpretation".

The methods for handling incomplete data can gathered into three classes: (1) marginalisation (whole data strategy): the data tuples [60,29] or attributes [12] with missing values are deleted from the data set; (2) imputation: the missing values are filled in with some values [14,23,47,61,65,67]; (3) applying rough sets [40,27,25].

Imputation and marginalisation are commonly used for their simplicity. The marginalisation of attributes with missing values leads to reduction of dimensionality. Imputation is more frequently used than marginalisation [30]. The review of imputation methods can be found in [36,37].

Various techniques of imputation have been proposed: imputation with zeros, random numbers [61], most common value of the attribute [11], mean values over all data set [38], mean value over the class of the incomplete tuple [26], median imputation (over all data, over the class). Missing categorical value can be imputed with all possible value [28]. The techniques for elaboration of missing values based on more sophisticated methods have been proposed: deduction of values from other values of the incomplete data vector, regression [8], k-Nearest Neighbour imputation [6], nearest neighbourhood imputation [65,67], partial imputation (parimputation) [66], imputation by utilising information within incomplete data items [68], expectation-maximisation (EM) [14], imputation based on fuzzy k-means clustering (FKMI) [2].

The imputation has some disadvantages: (1) the imputed value cannot be fully trusted [60,17], (2) the imputed value may not exist or have no sense in the task domain [61,48,19,16].

In general the neuro-fuzzy systems are not designed to handle data sets with missing values. Some systems have been proposed for classification [40–43,32] but they are not full systems and require the fuzzy rule base to be delivered. The paper [20] incorporates the missing values into fuzzy rules. The firing strength for missing value is substituted with 1.

In the paper we present the neuro-fuzzy system that handles data with missing values and assigns weights to the attributes [58,55]. The system is an extension the neuro-rough-fuzzy system for data with missing values [56]. In the system each fuzzy rule exists in its own subspace. This is a complete system designed for regression: the system creates the fuzzy model basing on full or missing value data sets and can elaborate the answer for both full tuples or tuples with missing values. The values in the train data set can miss from all attributes.

The paper is organised as follows: Section 2 describes the rough subspace neuro-fuzzy system: system's architecture (Sections 2.1, 2.2), creation of model (Sections 2.3, 2.4, 2.5, 2.10), elaboration of system's answer (Sections 2.6, 2.7, 2.8, 2.9). Section 3 describes the data sets (Section 3.1) and experiments (Section 3.2) and finally Section 4 summarises the paper.

In the paper the blackboard bold characters ($\mathbb{A}$) are used to denote the sets, bolds ($\mathbf{a}$) – matrices and vectors, upper case italics (A) – the cardinality of sets, lower case italics (a) – scalars and set elements. The overline $\bar{a}$ denotes upper