Extended Abstract of the Doctoral Thesis:
Intelligent techniques to improve data pre-processing

Ogerta Elezaj
ogertae@stud.ntnu.no

1 Introduction

Pre-processing of large scale datasets in order to ensure data quality is a very important task in data mining. One of the serious threats to data quality is the lack of data collected during field experiments, which negatively affects the data quality. The missing data usually have significant effects in many real-life pattern classification scenarios, especially when it leads to biased parameter estimates but also disqualify for analysis purposes. The process of filling in the missing data based on other valid values of rest of the variables of a data set is known as the imputation process.

In this research, we present a new data-driven machine learning approach for imputing the missing data. Even though Machine Learning methods are used in order to impute missing data in the literature, it is difficult to decide on a single method to apply on a given data set for imputation. This is because imputation process is not considered as science but as art that focuses on choosing the best method with the least biased value. For this reason, we compare different machine learning methods, such as decision tree (J48), Bayesian network, clustering algorithm and artificial neural networks in this research. The comparison of the algorithms indicates that, for predicting categorical and numerical missing information in large survey data sets, clustering method is the most efficient out of the others methods found in literature. A hybrid method is introduced which combines unsupervised learning methods with supervised ones based on the missing ratio, for achieving a data imputation with higher accuracy.

Additionally some statistical imputation methods such as Mean\Mode, Hot-Deck have been applied emphasizing their limitations in large scale datasets in comparison to the machine learning methods. A comparison of all above mentioned methods, traditional statistical methods and machine learning methods has been made and conclusions are drawn for achieving data imputation with higher accuracy in data sets of large scale survey. Also, another objective of the research is to discover the effect of balancing the training data set in the performance of classifiers. All methods are tested by applying them to imputed artificially created missing data to a large real world data set, population and housing census.

Also, handling sensitive data it is important that the data and the classifier remain private through identifying a set of core operations over encrypted data that underlie many classification protocols. Using encryption scheme, it is possible to delegate the execution of machine learning algorithms to a computing service while retaining confidentiality of the training and test data. Our main result is a privacy-preserving data imputation protocol for databases that are horizontally partitioned between two parties. Protocols using machine learning algorithm allows

either party to compute missing values without requiring the parties to share any information about their data and without revealing the traversed path to either party.

2 Hypotheses

There are a wide range of research issues to be addressed in this project. These can be summarized at a high level as the following set of overarching hypothesis:

a) The integration of supervised and unsupervised machine learning methods for imputing missing values in large scale surveys data sets will result in values that are as close as or closer to the actual values than those found by traditional statistical methods.

b) No universal method seems to be superior for a particular dataset type problem. Even if one methodology works well with one type of dataset, the results often cannot be repeated on similar datasets. This is due to the underlying distributions in the datasets, correlations of attributes, the amount of missing values and the sample size.

c) Privacy-preserving protocol for filling in missing values using machine learning imputation algorithm for data that is horizontally partitioned between two parties makes possible the process of imputation without requiring the parties to share any information among them

3 Research methodology

The proposed machine learning method methods areapproach is implemented in Weka 3.8 and executed in a PC with Intel® Core i5 processor with 2.7 GHz speed and 8 GB of RAM. SAS software, Version 9.2 has also been used for processing. Hot-Deck imputation is done using CONCORD JAVA (CONtrollo e CORrezionedeiDati version with Java interface) software developed for data editing and imputation, and IDEA ((Indices for Data Editing Assessment) software is used for calculating similarity indexes. These are open source software developed by Italian National Institute of Statistics (ISTAT). The principle of analyses is shown below.

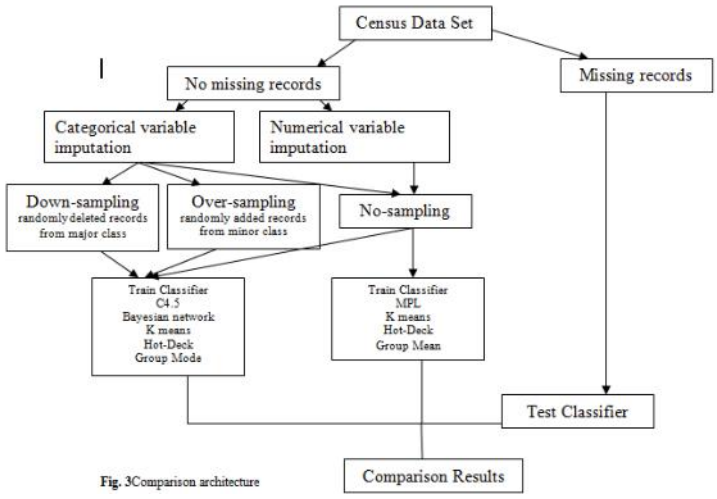

Fig. 3Comparison architecture