Tutorial on practical tips of the most influential data preprocessing algorithms in data mining

Salvador García[1], Julián Luengo[1], Francisco Herrera[1]

[1]Dept. of Computer Science and Artificial Intelligence, University of Granada, Granada, 18071, Spain.
salvagl@decsai.ugr.es, julianlm@decsai.ugr.es, herrera@decsai.ugr.es

Abstract. This is a summary of our article published in Knowledge-Based Systems [2] to be part of the MultiConference CAEPIA'16 Key-Works.

Keywords: Data preprocessing, data reduction, missing values imputation, noise filtering, dimensionality reduction, instance reduction, discretization, data mining

1 Summary

Data preprocessing for Data Mining (DM) [1] focuses on one of the most meaningful issues within the famous Knowledge Discovery from Data process [3]. Data will likely have inconsistencies, errors, out of range values, impossible data combinations, missing values or most substantially, data is not suitable to start a DM process. The growing amount of data in current business applications, science, industry and academia, also demands more complex mechanisms to analyze it. With data preprocessing, we can adapt the data to accomplish the input requirements of each DM algorithm, thus converting the impractical into possible.

In an effort to identify some of the most influential data preprocessing algorithms that have been widely used in the DM community, we enumerate them according to their usage, popularity and extensions proposed in the research community. The selection of the algorithms is based entirely on our criteria and expertise, especially after the composition of a recent book on this topic [1]. The selection criteria is based on (i) *usage* (the algorithm is widely used); (ii) *referable* (is published); (iii) *popularity* (is highly cited); (iv) *standarization* (is the baseline for extensions); (v) *smart* (incorporates a smart procedure); and (vi) *variability* (each preprocessing family has enough representatives).

The categorization of the data preprocessing families obeys a taxonomy that directly draws from the specialized literature. Three main categories are considered: imperfect data, data reduction and imbalanced data preprocessing. Among these categories, several subfamilies are established and the nominated algorithms are selected. They are as summarized in Figure 1.

Once the nomination is made, we provide a practical study including numerous tips that it is divided in two parts. First, by using the banana data set, we

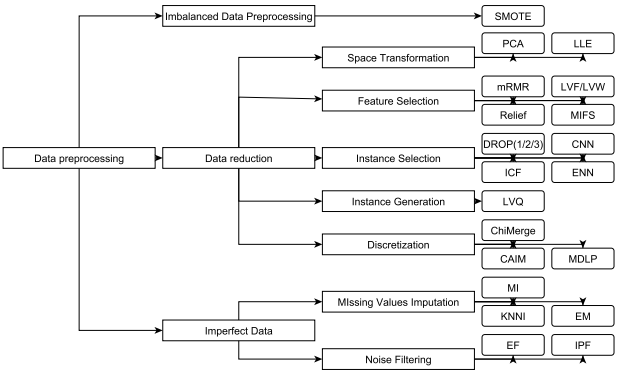

Fig. 1. Most influential data preprocessing algorithms classified by type

illustrate the results of applying the reviewed techniques in the paper. In the case of the dimensionality reduction techniques (Feature Selection and Space Transformation), the sonar data set is considered instead as it contains a large number of features. The effects on the data and a comparison among related techniques for these data sets are thus easily attainable, as the effect on the data is visually depicted for every technique.

We also consider a real-world problem from the ECDBL'14 Big Data competition related to bioinformatics. Thus, an appropriate setting is posed for showing the importance of appropriately using several preprocessing techniques. We selected one technique of each type and combined them to solve the different problems present in the data, also showing benefits in the model obtained by the C4.5 classifier. Variations in the order used to apply the data preprocessing techniques, illustrating the possibilities and tips that can be drawn from the differences among them, highlighting the major importance that falls on the correct arrangement of data preprocessing algorithms and in the adjustment of parameters.

References

1. S. García, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*. Springer, 2015.
2. S. García, J. Luengo, and F. Herrera. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98:1–29, 2016.
3. M. J. Zaki and W. Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY, USA, 2014.