Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
©2023 ESREL2023 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P726-cd

Importance analysis in the evaluation of input attributes of classifiers

Elena Zaitseva

Department of Informatics, University of Zilina, Slovakia. E-mail: elena.zaitseva@fri.uniza.sk

Vitaly Levashenko

Department of Informatics, University of Zilina, Slovakia. E-mail: vitally.levashenko@fri.uniza.sk

Sergey Stankevich

Scientific Centre for Aerospace Research of the Earth, NAS of Ukraine, Ukraine. E-mail: st@casre.kiev.ua

Most often, the techniques of Machine learning are used for the decision of problems in Reliability Analysis. In this study, we propose to consider the application of the Reliability Analysis based method application for the decision problem in Machine Learning, in particular, the analysis of the influence of input attributes on the classification result. Some attributes are most important for the classification because they significantly influence the classification result than others. A new method for the determination of the most important attributes is proposed. This method is developed based on the approach of Importance Analysis, which is widely used in Reliability Analysis. The attribute's importance is evaluated by structural importance.

Keywords: Classification, Importance analysis, Multi-state system, Structure function, attributes selection.

1. Introduction

Classification is one of the principal approaches in machine learning. The classification efficiency depends on many factors. One of the factors affecting classification is the quality of the initial data. The initial data used for the classifier induction assumes sufficient training samples, their balances among specified classes, acceptable data dimension, authenticity, and absence of redundant data or noise. The choice of the input attributes that have the maximum impact on the result is one of the tasks of the data pre-processing. This problem is known as feature extraction/selection (Bolon-Canedo et al, 2013). The feature selection/extraction, on the one hand, reduces the dimensions of the studied data sets, on the other hand, the required classification accuracy is of paramount importance and must be maintained despite the reduction of attributes. There are many approaches to feature selection/extraction. As a result of feature extraction, a new set of attributes from the original set is formed (Bruni et al., 2022). The often-used methods for feature extraction are based on techniques of Linear Discriminant Analysis, Principal Component Analysis and Independent Component Analysis. As a result of feature selection, an acceptable set of attributes from the original dataset is selected without any transformation (Naheed et al., 2020). Feature selection methods can be divided into filter, wrapper, and embedded methods. Filters are independent of an algorithm of the classifier induction and choose attributes based on the characteristics of initial data. The wrapper methods depend on the algorithm of classifier induction. A classification algorithm evaluates a subset of selected attributes according to a classification measure. These methods have high performance but require much computation time to execute. Embedded methods perform attribute selection in the process of the classifier induction and depend on the classifier and algorithm of its induction. This study proposes a new method for attribute selection based on the importance analysis of the initial data set of attributes.

2. Method

A new method for analysis of the influence of the classification factors (input attributes) on the result is proposed. It allows for studying the sensitivity of decision making. This method can be thought of as a feature selection method that combines the advantages of filter and wrapper based methods. The proposed method has the universality of filters. Attributes are evaluated and quantified by Structural Importance (SI) (Zaitseva et al., 2023).

The proposed method consists of two steps (Fig. 1). The construction of the mathematical model acceptable for reliability analysis is implemented in the first step. This step is implemented based on the classifier induction. The result of this step is a decision table of the classification, which is interpreted as the structure function of a system (mathematical model of a system for its reliability analysis). The analysis of important components (attributes) is performed in the second step based on reliability engineering based methods. The result of this step is a quantification of all attribute importance for the result of classification.

The proposed method has been used for the induction of a Fuzzy Decision Tree (FDT). The analysis of input attributes' importance based on the proposed method is considered by the example of the analysis of the factor of the timing of tracheostomy in COVID-19 patients (Zaitseva et al., 2023). The attributes with the biggest or non-zero value of SI are selected for the induction of the final classifier (Fig.2). The classification performance in this case study is not changed after eliminating non-important attributes with SI=0. Therefore, if we omit attributes with zero or small values of SI, the classification model can be simplified, the time for training will be shorter, we can avoid large dimensional, and also the data compatibility with the input model can be improved.

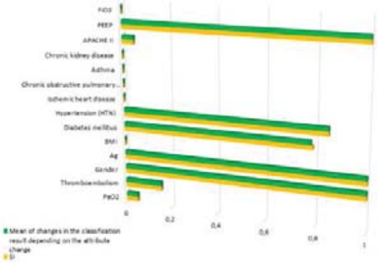

Fig. 2. The evaluation of attributes' importance

3. Conclusion

The advantages of the proposed method are (a) evaluation and selection of attributes depending on the classification result (similar to wrapper methods) and (b) can be used for any classifier without fundamental changes similarly to filters. At the same time, the application of the proposed method for the analysis of big data or data drift will need additional modification. The present version of the method can be effective for minor changes in attribute importance since the uncertainty arising, in this case, can be covered by the use of a fuzzy classifier.

Acknowledgment

This study is supported by the project "New Approaches of Reliability Analysis of non-coherent Systems" (VEGA 1/0165/21) and the project EWALD which has received funding from the European Union's Framework Programme for Research and Innovation Horizon Europe – the Framework Programme for Research and Innovation (2021-2027), Grant Agreement No. ID 101086250.

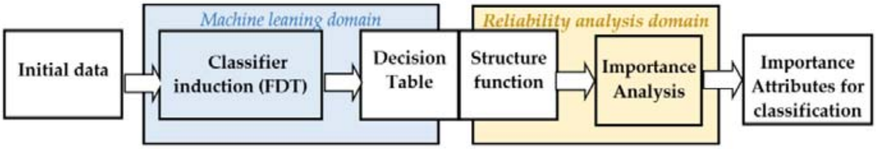
Machine leaning domain
Reliability analysis domain
Initial data
Classifier induction (FDT)
Decision Table
Structure function
Importance Analysis
Importance Attributes for classification

Fig. 1. The method for quantification of input attributes' importance.

References

Bolon-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems 34*, 483–519.

Bruni, V., Cardinali, M.L., Vitulano, D. (2022). A short review on minimum description length: An application to dimension reduction in pca. *Entropy 24*.

Naheed, N., Shaheen, M., Khan, S.A., Alawairdhi, M., Khan, M.A. (2020). Importance of features selection, attributes selection, challenges and future directions for medical imaging data: A review. *Computer Modeling in Engineering & Sciences 125*, 315–344.

Zaitseva, E., Rabcan, J., Levashenko, V., Kvassay, M. (2023). Importance analysis of decision making factors based on fuzzy decision trees. *Applied Soft Computing 134*, 109988