Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 109C (2017) 448–457

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Ambient Systems, Networks and Technologies (ANT2017)

A new predictive approach to variables selection through Genetic Algorithm and Fuzzy Adaptive Resonance Theory Using medical diagnosis as a case

Choukri Djellali, Mehdi Adda

Mathematics, Computer Science and Engineering Dep.
University of Quebec At Rimouski
300, Allée des Ursulines, Rimouski , QC G5L 3A1, Canada

Abstract

Variables selection is challenging task due mainly to huge search space. This study addresses the increasingly encountered challenge of variables selection. It addresses the application of machine learning techniques to the problem of variables selection. We detailed the various models of the variables selection and examined the basic steps that are used to select the cost-effective predictors. We also walked through the initial settings and all variables selection stages, including architecture configuration, strategy generation, learning, model induction, and scoring. Results from this study show that the cost and generalization were seen to improve significantly in terms of computing time and recognition accuracy when the proposed system is applied for medical diagnosis. Good comparisons with an experimental study demonstrate the multidisciplinary applications of our approach.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining, Pattern Recognition, Machine Learning, Clustering, Variables Selection, Genetic Algorithm.

1. Introduction

Most machine learning algorithms are sensitive to noise and Bellman's curse of dimensionality. It is challenging task due mainly to huge search space. Irrelevant (and/or redundant) variables increase the search size and decrease the generalization ability. The most common machine learning technique is variables selection (also known as subset selection or features selection), which is used for exploratory subset analysis to find the relevant variables or cost-effective predictors.

* Corresponding author. Tel.: +1-514-987-3000 ; fax: +1-514-987-8477.
E-mail address: Choukri.Djellali@uqar.ca

Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.05.305

Variables selection is defined as follows: is the machine learning task of selecting a subset of relevant features for use in model construction[7].
With the proliferation of Big Data, there has been a lot of interest in recent years to develop effective methods for variables selection. It is one of the most active research and application areas of Data Mining. Applications range from Information Retrieval, image Mining,[2] , Robotic[3], Big Data[4], Network security[5], Cloud computing[6], Bioinformatics[11], Text Mining[12], etc. All these disciplines show the practical importance of variables selection algorithms.
However, the variables selection is a crucial challenge, especially for pattern recognition. The curse of dimensionality associated with the exponential increase of the space size adds unnecessary noise within the decision boundary and learning generates significantly lower performance.
On one hand, the curse of dimensionality is very expensive and many machine learning algorithms provide a significantly lower performance.
On the other hand, the search space increases exponentially with respect to the number of possible combinations. Therefore, exhaustive search strategies are prohibitively time-consuming and computationally expensive even for problem instances of moderate size of the search space.
In order to avoid these limitations, we used a new variables selection scheme based on meta-heuristic search, efficient Wrapper model, and subset validity assessment criterion.
This paper is structured as follows: In Section 2, we present the current state of the art, our research questions and the problematic of variables selection. The conceptual architecture of our variables selection is given in Section 3. We present in Section 4 a short evaluation with a benchmarking model for variables selection. Finally, a conclusion (Section 5) ends the paper with future works.

2. State of the Art, Problem and Research Questions

Variable selection is the process of selecting a subset of the relevant variables (also known as predictors) occurring in the training set and using this subset in model construction.
As shown in Figure 1, there are three main models for variables selection, namely Filter model, Wrapper model, and Embedded model.

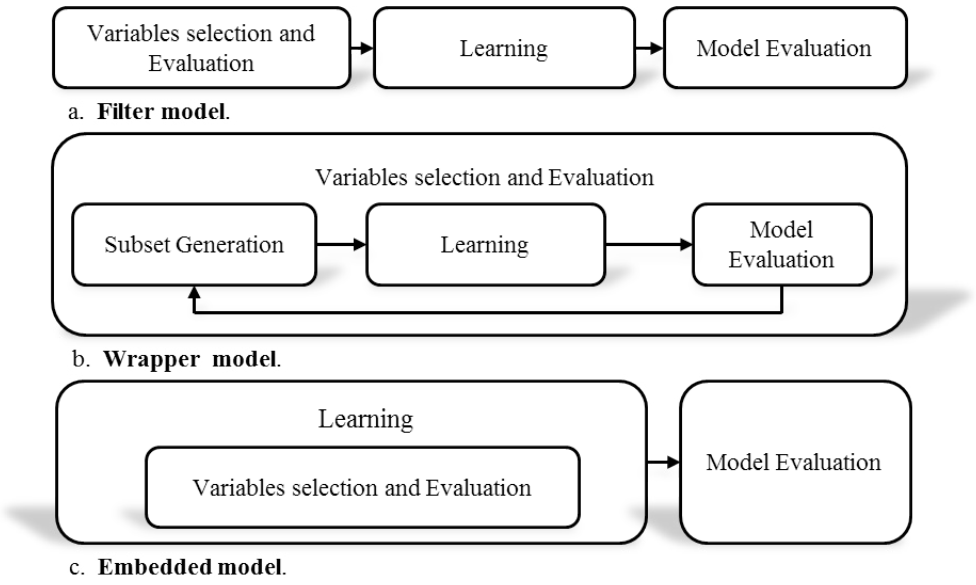

Fig. 1. Filter, Wrapper, and Embedded variables selection model[23].

First, the Filter Model is the simplest method where a statistical measure is applied to score each variable individually. Second, the Wrapper Model is an extension of Filter Model using techniques of Data Mining algorithms to score