Abstract

A Fast Multivariate Symmetrical Uncertainty Based Heuristic for High Dimensional Feature Selection [†]

Miguel Garcia Torres [1], Federico Divina [1], Francisco A. Gómez Vela [1] and José Luis Vázquez Noguera [2,3]

[1] Universidad Pablo de Olavide, Sevilla, Spain
[2] Universidad Nacional de Asunción, San Lorenzo, Paraguay
[3] Universidad Americana, Asunción, Paraguay

[†] Presented at the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May 2021; Available online: https://sciforum.net/conference/Entropy2021/.

Published: 5 May 2021

In classification tasks the increase in the number of dimensions of a data makes the learning process harder. In this context feature selection usually allows to induce simpler classifier models while keeping the accuracy. However, some factors, such as the presence of irrelevant and redundant features, make the feature selection process challenging. Symmetrical Uncertainty (SU) is an entropy-based measure widely used to identify subsets of useful features for the classification task. However, SU is a bivariate measure and, so, it ignores possible dependencies among more than two features. In order to overcome this issue, SU has been extended to the multivariate case. This extension, called Multivariate Symmetrical Uncertainty (MSU), is time-consuming and may become impracticable when evaluating larger subsets of features during the search. In this work we propose a MSU based Feature Selection (MSUFS) heuristic to address feature selection on high-dimensional data. In order to design MSUFS, the concept of Approximate Markov Blanket is redefined to take into account the MSU measure. The performance of MSUFS is tested on high-dimensional datasets from different domains and its results where compared with popular and competitive techniques. Results show that MSUFS is capable of identifying possible correlations and interaction among features and, therefore, it achieves competitive results. Finally, the proposed strategy is also applied to a case study regarding melanoma skin cancer.

Acknowledgments: This work is partially supported by the research project PINV15-0257 from CONACyT-Paraguay. Authors are also thankful to the Andalusian Scientific Computer Science Centre (CICA) for allowing us to use their computing infrastructures.