Trabalho apresentado no XXXVIII CNMAC, Campinas - SP, 2018.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Understanding a multivariate semi-metric in the search strategies for attributes subset selection

Gustavo Sosa-Cabrera[1]
Polytechnic School, National University of Asunción , P.O. Box 2111 SL, San Lorenzo, Paraguay
Miguel García-Torres
Computer Science, Universidad Pablo de Olavide, ES-41013, Seville, Spain
Santiago Gómez-Guerrero
Polytechnic School, National University of Asunción , P.O. Box 2111 SL, San Lorenzo, Paraguay
Christian E. Schaerer
CIMA, Centro de Investigación en Matemática, Dr. César López Moreira 693 - P.O.Box: 1766, Asunción, Paraguay
Federico Divina
Computer Science, Universidad Pablo de Olavide, ES-41013, Seville, Spain

1 Introduction

In classification tasks, a feature (i.e. independent variable) is considered relevant, irrelevant or redundant according to the information contained about the class (i.e. dependent variable). Feature selection consists of finding the minimal set of relevant features such that the classification error is optimized. A feature selection method has three components: evaluation criterion definition (e.g. feature relevance), evaluation criterion estimation, and search strategies for feature subset generation. Symmetrical Uncertainty (SU) is a mutual information-based semi-metric measure that has been widely used to identify relevant features, as well as detecting dependencies between two features. The main limitation of SU consists in taking into account only pairwise interactions and so it might lead to failure in the detection of redundancy when dealing with more than two features (e.g. Exclusive-OR function or any real-world dataset where two or more features are needed to determine the class). Multivariate Symmetrical Uncertainty (MSU) is formulated as a generalization of the SU aimed to quantify the redundancy (or dependency) among more than two features [1]. We use the following definition of MSU $\in [0, 1]$ which is a semi-metric in the space of categorical random variables $X_1, X_2, ..., X_n$

$$MSU(X_{1:n}) := \frac{n}{n-1}\left[\frac{C(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right], \tag{1}$$

[1]gdsosa@pol.una.py

where $H(X_i)$ is the information entropy of variable X_i and $C(X_{1:n})$ is the total correlation of the variable set $\{X_1, X_2, ..., X_n\}$.

A search strategy is needed to direct the feature subset selection process as it explores the space of all possible combinations of features. Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are algorithms considered computationally efficient that add or remove features sequentially in one direction respectively.

2 Methods

Despite extensive research in the field of attribute selection, possible conditioning factors between the search strategy and the evaluation measure for the resulting subset have not been fully understood. In addition, previous studies have shown that under certain circumstances the SFS and SBS search strategies present different qualities [3]. Therefore, there is a need to understand possible factors among them.

In this work, we perform several experiments to study the effect of feature group conformation strategies (such as SFS & SBS) while using MSU as a reliable measure of the association of the group with the class. For such purpose we generate several synthetic datasets, with the goal of assessing the strength of candidate groups under different combinations of three factors: informativeness, cardinality and sample size.

3 Conclusion

The contributions of this paper are summarized in the following items: (1) MSU, based on information theory concepts, is studied under the n-bit parity problem and the checkerboard pattern that were mentioned in [2] to show that a variable which is useless by itself can be useful together with others. (2) A more thorough understanding of the performance of the MSU in relation to the number of features that are selected in the sequential forward search (SFS) and in the sequential backward selection (SBS) strategies. (3) A procedure is derived to calculate a threshold for the number of attributes of the subset resulting from the search strategy and which guarantees a controlled bias in the MSU as an evaluation measure.

References

[1] R. Arias-Michel, M. García-Torres, F. Divina and C.E. Schaerer. Feature Selection Using Approximate Multivariate Markov Blankets. In *International Conference on Hybrid Artificial Intelligence Systems*, Springer, Cham, pages 114-125, 2016.

[2] I. Guyon and A. Elisseeff. *An introduction to variable and feature selection.* Journal of machine learning research, 3(Mar), pages 1157-1182, 2003.

[3] D. Huang, T. W. Chow, E. W. Ma and J. Li. *Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis.* IEEE Transactions on Circuits and Systems I: Regular Papers, 52(9), pages 1909-1918, 2005.