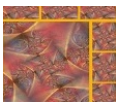
Garcia-Martinez, R., Lopez Nocera, M. 2012
Behavioral Variability of Datamining Algorithms Based on Domain Complexity
Proceedings of VI Meeting on Dynamics of Social and Economic Systems. Pages 37-38
Ushuaia, October 1-4, Argentina



VI Dyses Meeting

Behavioral Variability of Datamining Algorithms Based on Domain Complexity

Ramon Garcia-Martinez, Marcelo Lopez Nocera

Information Systems Research Group. Department of Technological and Productive Development. Lanus National University. Buenos Aires, Argentine

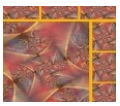
rgarcia@unla.edu.ar

1. INTRODUCTION

Britos [1] has defined five information mining processes: discovery of rules of behavior, groups discovery, discovery of significant attributes, discovery of group membership rules and weighting of rules of behavior or group membership. The information mining processes use different data mining algorithms in order to obtain patterns of knowledge from the existing examples (instances) with in the problem domain [2]. One of the implicit assumptions these algorithms work with is that once the algorithms for the information mining process are set, the complexity of the domain on which information is working, has no bearing on the quality of the patterns obtained. However, there are preliminary research results [8] which show that the complexity of the domains in terms of pieces of knowledge that describe them and that information mining processes are seeking to discover, emerges as a not worthless component when analysis of the quality of results to be obtained is being performed. In this context, an experiment for the case of the information mining process of discovery of group membership rules demonstrates that the quality of the patterns which are obtained differs depending on the complexity of the domains to which it is applied and of the algorithms used in the process.

2. CHARACTERIZATION OF DOMAIN COMPLEXITY

Based on the classification attributes given and the domains classification protocol proposed in [3], domains can be classified according to their complexity in the following types: [a] Simple Complexity domains, those domains where increasing the number of



samples by rule improves rules coverage (independently of the other dimensions used); [b] Medium Complexity Domains: those domains that are explained with examples with few attributes and a few classes, or few attributes and many classes or few classes and few rules per class; [c] Oscillating domains: those domains are explained with examples where the number of attributes per example, or the number of examples supported by a rule, or common values of attributes in a set of examples covered by the same rule can vary; [d] Complex domains: those domains that are explained with examples with few attributes and many possible values per attribute, or with many attributes and a few possible values per attribute, or with many attributes and many possible values per attribute; [e] Hyper complex domains: those domains that are explained with examples where the number of possible values that can take the attributes, the number of attributes that cover examples, the amount of rules covering examples, or the number of classes can vary.

3. RESULTS

As a result of the analysis of experimental results the following experimental propositions emerged: [i] For medium complexity domains, we verified experimentally that the combination which offers better coverage of rules is the pair <SOM, AQ15>, with a value close to 86 percent on average; [i] For hyper complex type domains, we verified experimentally that the combination which offers better coverage of rules is also <SOM, AQ15>, but with a lower yield, nearly 65 percent on average; [iii] For domains of type complex, we verified experimentally that the combination which offers better coverage of rules is <SOM, AQ15>, with 49 percent as average value; [iv] For oscillating type domains, it was verified experimentally that the combination which offers better coverage of rules is <SOM, M5>, with 60 percent on average; [v] It is observed and it is experimentally determined that the underlying domain characteristics influence the obtained experimental result. Regarding the performance in terms of coverage of rules of clustering and induction algorithms combined, results have been obtained which distort the assumption that the performance of the algorithms used for clustering and induction is independent of the complexity of the domain.

AKNOWLEDGEMENTS
The research reported in this article has been partially financed by projects UNLa 33A105 and UNLa 33B102 of the National University of Lanus.

REFERENCES
[1] Britos, P. (2008). Procesos de explotación de información basados sobre sistemas inteligentes. Thesis submitted for obtaining the degree of Doctor of Computer Science, School of Computing, National University of La Plata, Argentina; **[2]** Britos, P., García-Martínez, R. (2009). Propuesta de Procesos de Explotación de Información. Proceedings XV Congreso Argentino de Ciencias de la Computación. Workshop de Base de Datos y Minería de Datos. Pages 1041-1050. ISBN 978-897-24068-4-1; **[3]** Lopez-Nocera, M., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. (2011). Un Protocolo de Caracterización Empírica de Dominios para Uso en Explotación de Información. Proceedings XVII Congreso Argentino de Ciencias de la Computación. Pages 1047-1055. ISBN 978-950-34-0756-1.