Advances and Applications in Statistical Sciences
Proceedings of The IV Meeting on Dynamics of Social and Economic Systems
Volume 2, Issue 2, 2010, Pages 303-314

ISSUES IN RULE BASED KNOWLEDGE DISCOVERING PROCESS

CLAUDIO RANCAN, PATRICIA PESADO and RAMÓN GARCÍA-MARTÍNEZ

Computer Science PhD Program
Computer Science School
Facultad de Informática, UNLP - CIC Bs As
Instituto de Investigación en Informática LIDI
La Plata National University, Argentina
E-mail: crancan@itba.edu.ar

Intelligent Systems Laboratory
School of Engineering
University of Buenos Aires, Argentina
E-mail: ppesado@lidi.info.unlp.edu.ar

Software Engineering Area
Information Systems Degree Program
Lanus National University, Argentina
E-mail: rgarciamar@fi.uba.ar

Abstract

The improvement of a knowledge base with discovered knowledge pieces (rules) in automatic way can lead to a degradation of the original knowledge base. It is an open issue to establish which the quality of the knowledge discovery process is. This paper introduces a framework for knowledge discovery and expert systems integration experimentation, the experiment protocol for studying rules discovering process quality is described and the preliminary experimental results are shown.

2010 Mathematics Subject Classification: 68T05, 68Q32, 68T35, 97C30.

Keywords: knowledge discovery, top-down induction of decision trees, self organized maps, improving knowledge based systems.

Received April 30, 2009

1. Introduction

The knowledge base of an expert system encapsulates in some representation formalism (rules, frames, semantic nets among other), the domain knowledge that should be used by the system to solve a certain problem [2, 3]. The interaction between knowledge based systems and discovery systems [7, 9] has antecedents in the paradigm of integrated architectures of planning and learning based on theories construction [4] and hybrid architectures of learning [5, 6]. The improvement of a knowledge base with discovered knowledge pieces in automatic way can lead to a degradation of the original Knowledge Base, so it is an open issue which are the curves of degradation of the quality process of knowledge discovery identifying border conditions (at least in a theoretical way). In this context, this paper introduces a framework for knowledge discovery and expert systems integration (Section 2), the experiment protocol for studying rules discovering process quality is presented (Section 3), the experimental results are shown (Section 4), and finally some conclusions are drawn (Section 5).

2. The Framework

In [11, 12] is presented a framework that shows one way of how KBS can be integrated to knowledge discovery processes based on machine learning oriented to improve "on-line" the quality of the knowledge base used for the decision making expert system (See Figure 1). The framework uses the following knowledge and data bases: [a] *Knowledge Base*, this base contains the problem domain knowledge deduced by the knowledge engineer, which contributes the knowledge pieces (rules) applicable to the resolution of the problem outlined by the user of the system, [b] *Concepts Dictionary*, this base stores the registration of all the concepts used in the different knowledge pieces (rules) that integrate the Knowledge Base, for each concept it keeps registration of the corresponding attributes and the possible values of each attribute, [c] *Examples Base*, this base keeps examples of elements that belong to different classes, the attributes of these examples should keep correlativity or should be coordinated with the attributes of the concepts described in the Concepts Dictionary, [d] *Records Base*, this base keeps homogeneous records of information which is associated to some process of

knowledge discovery (I/E clustering), [e] *Clustered Records Base*, this base keeps homogeneous records of information which are clustered in classes without labeling (clusters) as a result of applying the clustering process to the Records Base, [f] *Clustering/Classification Rules Base*: this base keeps knowledge pieces (rules) discovered automatically as a result of applying the induction process to the Clustered Records Base and the Examples Base, [g] *Discovered Rules Base*: this base keeps knowledge pieces (rules) related to the problem domain as result of applying the labeling conceptual process to the discovered knowledge pieces (rules) that are stored in the Clustering/ Classification Rules Base, [h] *Updated Knowledge Base*, this base encapsulates the knowledge that becomes from the integration of the problem domain knowledge pieces (rules) educed by the knowledge engineer and the knowledge pieces (rules) discovered automatically as a result of the application of the processes of clustering/induction to the Records Base or induction to the Examples Base.

The framework uses the following processes: [a] *Cluster*: this process is based in the use of self organized maps (SOM) to generate groups of records that are in the Records Base, these groups are stored in the Clustered Records Base; [b] *Inducer*: this process is based in the use of induction algorithms to generate clustering rules beginning from the records groups that are in the Clustered Records Base and Classification Rules beginning from the records that are in the Examples Base, [c] *Conceptual Labeler*: this process is based on the use of the Concepts Dictionary and the Clustering/ Classification Rules Base to generate the Discovered Rules Base, this process transforms the knowledge pieces obtained into pieces of coordinated knowledge with the Knowledge Base, [d] *Knowledge Integrator*, this process generates the Updated Knowledge Base from the Discovered Rules Base and the Knowledge Base, solving all the integration problems between them, [e] *Inference Engine*, it is the process that automates the reasoning to solve the problem outlined by the user, beginning from the pieces of knowledge available in the Updated Knowledge Base or Knowledge Base.

The dynamic of the framework is: Knowledge Base encapsulates the necessary pieces of knowledge (rules) for the resolution of domain problems. This interaction with the inference engine constitutes the Knowledge Based System (Expert System).

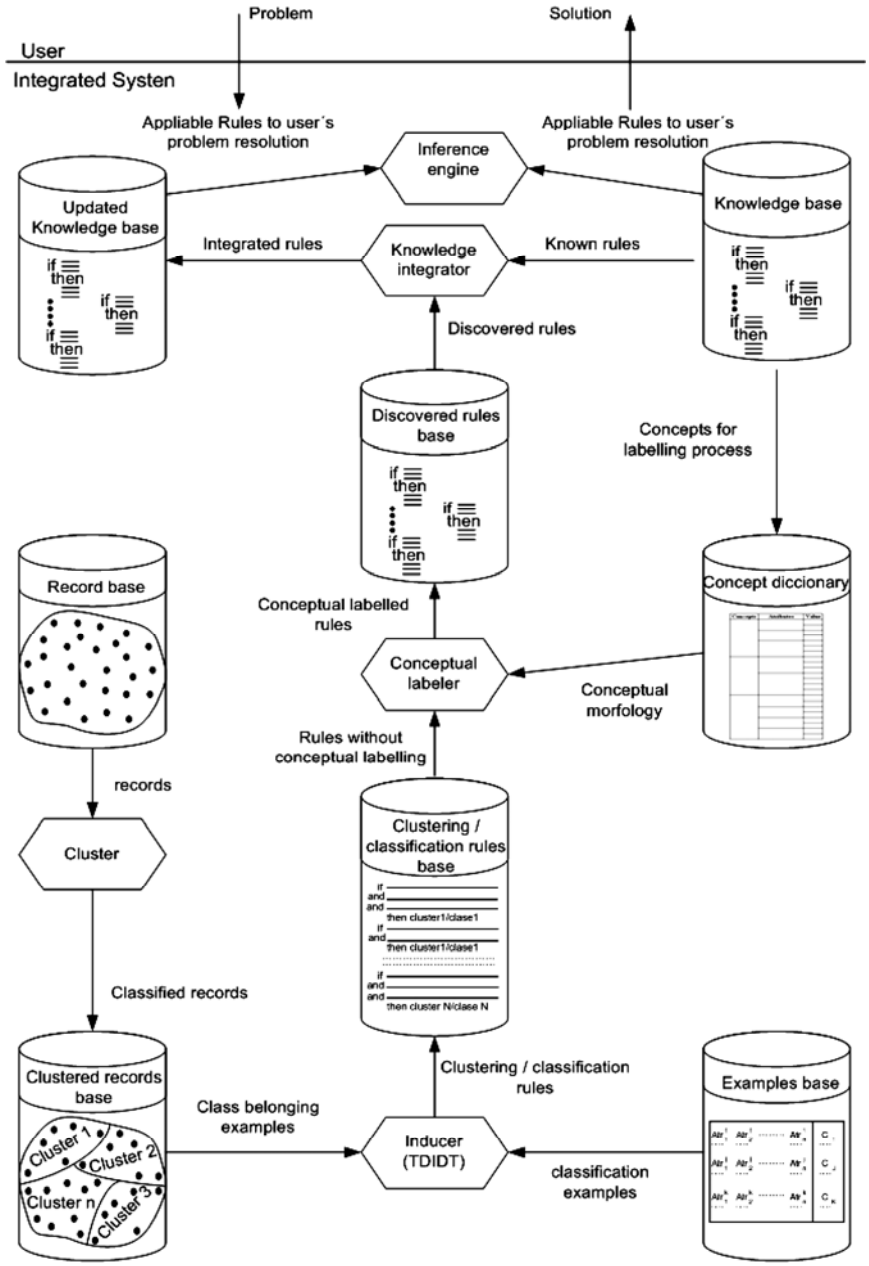

Figure 1. Interaction among different components.

Beginning from the concepts/attributes/values that are present in the different pieces of knowledge inside the Knowledge Base, the Concepts Dictionary is built. When a situation of knowledge discovery takes place

because the Inducer generated a Clustering/Classification Rules Base, or because this has became from an Examples Base or a Clustered Records Base resultanting of applying the Cluster to a Records Base, the pieces of knowledge (rules) that are in the Clustering/Classification Rules Base can present the characteristic of not being coordinated with the available pieces of knowledge in the Knowledge Base. In this context the Conceptual Labeler transforms the knowledge pieces of the Clustering/Classification Rules Base into coordinated knowledge pieces with those rules corresponding to the Knowledge Base generating the Discovered Rules Base. The Knowledge Integrator takes the Discovered Rules Base and (solving the emergent integration problems) integrates it into the Knowledge Base, generating the Updated Knowledge Base, that becomes the new Knowledge Base and the cycle is restarted.

3. The Experiment

The experiments purpose is to explore the quality of the rule discovering process used in the framework in domains where:

- classes have associated different amounts of classification rules and
- the amount of attributes per classification rule can vary and
- in domains where amount of classes can vary and each class has associated classification rules in which the amount of attributes per each one can vary.

A three step experiment has been carry out.

The step 1 consists in experiment preparation. This step involves:

- [a] domain generation based on generation of classes and generation of classification rules for each class and
- [b] examples generation for each classification rule.

The output of this step is a classification rules set and a domain records set.

The step 2 consists in experiment execution. This step involves:

- [a] to apply the cluster process to domain records (examples) set to obtain the domain clusters set and

- [b] to apply the inducer process to the domain clusters set to obtain the discovered rules set.

The step 3 consists on the comparison of the classification rule set from step 1 with the discovered rules set from step 2 the percentage of matching rules defines the experiment success. The variables used in the experiment are shown in Table 1.

4. The Results

The results are grouped into two different approaches: one is domain oriented, where it is studied how variations on independent variables values associated with domain characteristics influence the percentage of well discovered rules (Figures 2 to 4); and in the other hand, the examples oriented approach that focuses on how variations on independent variables values associated with original rule's characteristics influence the percentage of well discovered rules (Figures 5 to 9).

The more classes the domain has more associated rules, the lower is the performance of the proposed method (Figures 2 and 4). The more attributes the examples are composed, the lower is the performance of the proposed method (Figures 3 and 4).

Table 1. Variables used in the experiments.

Variable	Type	Variable's description
attPosibleValues	independent	Amount of possible values an attribute can take
attributesNumber	independent	Amount of attributes in each classification rule
attributesNumber	independent	Amount of attributes in each examples
attUsedInRule	independent	Specificity of the covering of each rule over its examples
classAttPossibleValues Percentage	independent	Concentration of the rules that determine a class
classPosibleValues	independent	Amount of domain different classes
instancesByRule	independent	Amount of examples that support each rule
rulesCorrectlyCovered	dependent	Percentage of matching rules among classification rules set and discovered rules set (well discovered rules)
rulesPerClass	independent	Amount of classification rules for determining each domain class

The more possible values each attribute can take, the higher is the performance of the proposed method (Figure 3). The higher amount of rules that determine the sense of belonging to each class, the lower is the performance of the proposed method (Figure 2). The higher amount of rules that determine the sense of belonging to each class, the lower is the performance of the proposed method, showing an asymptotic behavior towards a minimum when the amount of rules that determine the sense of belonging to each class is high (Figure 2). From certain amount of possible values each attribute can take, raising this amount does not seem to improve the performance of the proposed method (Figure 3). From certain amount of classes that rules the domain, raising this amount does not seem to be worse the performance of the proposed method (Figure 4).

The more specific is the covering of each rule over its examples, the higher is the performance of the proposed method (Figures 5, 7, 8). The lesser concentration of the rules that determine the sense of belonging to each class, the higher performance of the proposed method (Figures 6, 7). The higher amount of examples that support each rule, the higher is the performance of the proposed method (Figures 6, 8, 9). The concentration of the rules that determine the sense of belonging to each class seems to not modify the performance of the proposed method being evaluated on domains where the specificity of the covering of each rule about its examples can vary (Figure 5). The more specific is the covering of each rule over its examples, the more pronounced is the growth of the performance of the proposed method as the concentration of the rules that determine the sense of belonging to each class decreases (Figure 7). The more specific is the covering of each rule over its examples, the more pronounced is the growth of the performance of the proposed method as amount of examples that support each rule increases (Figure 8). The concentration of the rules that indicate the sense of belonging to each class seems to not modify the performance of the proposed method being evaluated on domains where the rules are supported by a different amount of examples (Figure 9).

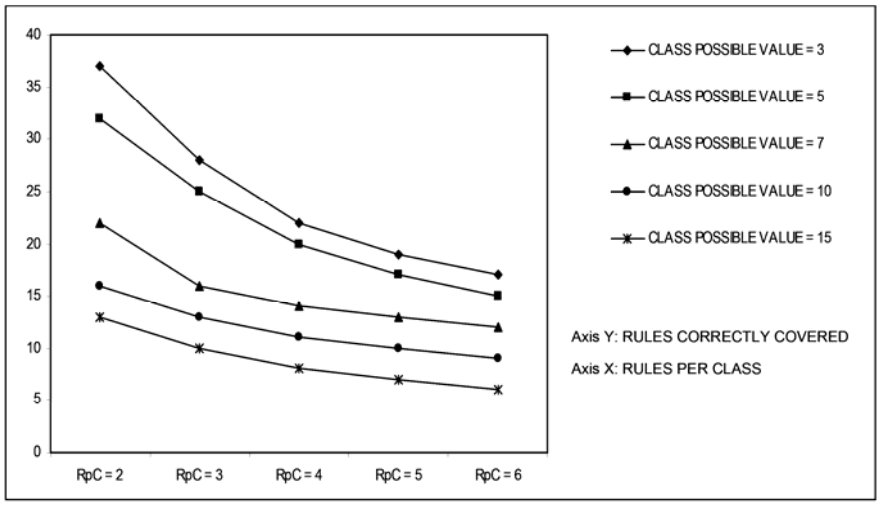

Figure 2. Domains ruled by different amount of classes and the number of rules that determine the sense of belonging to each class can vary.

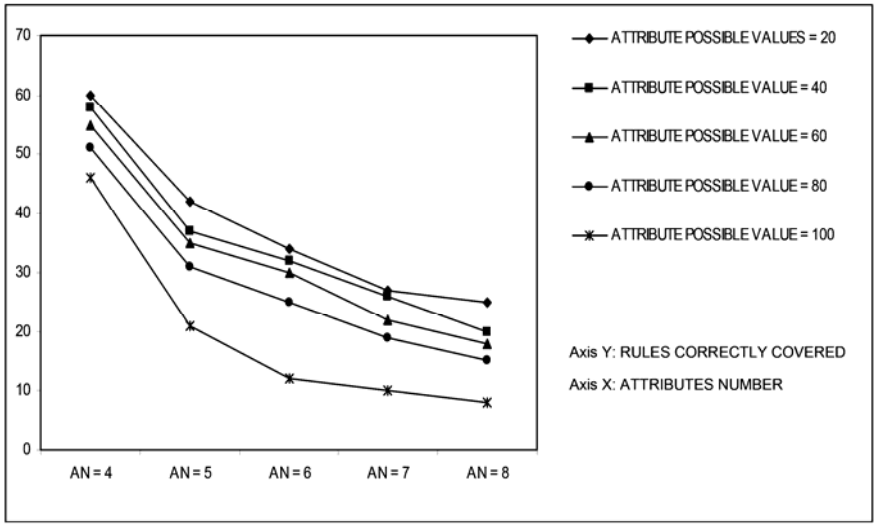

Figure 3. Domains where attributes can take a different amount of possible values and the number of attributes that compose each example can vary.

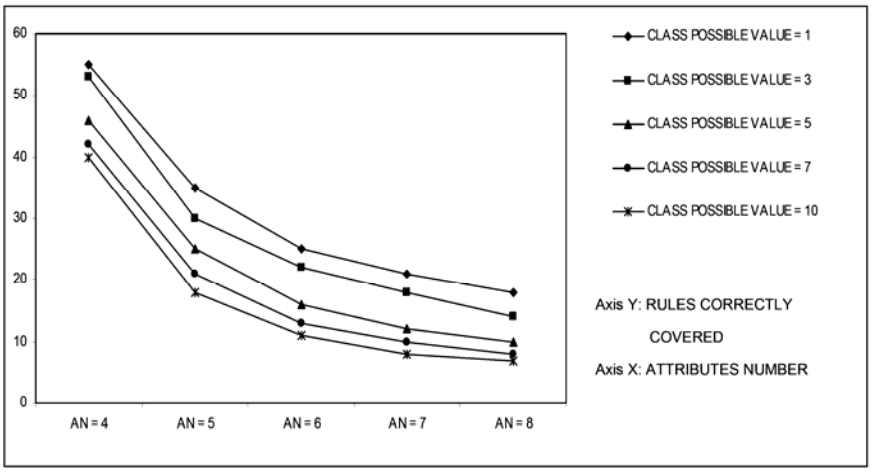

Figure 4. Domains ruled by different amount of classes and the number of attributes that compose each example can vary.

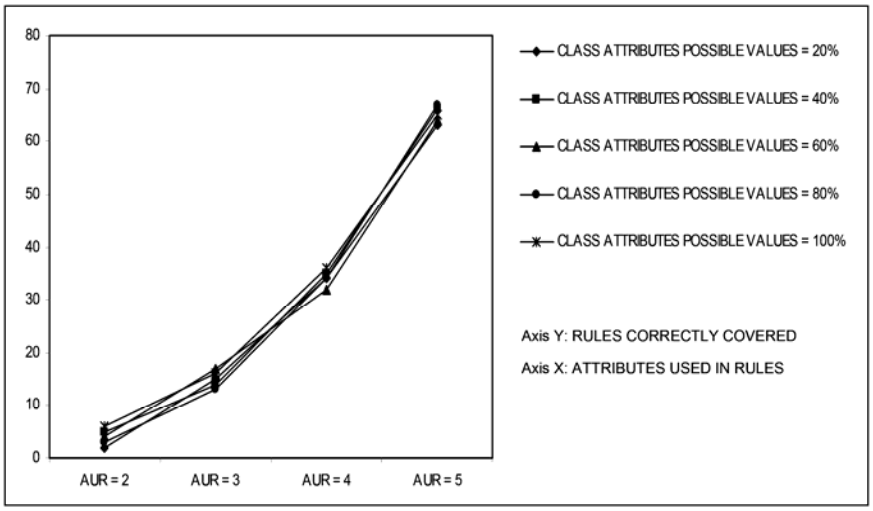

Figure 5. Domains with different concentrations of the rules that determine each class and the specificity of the covering of each rule about its examples can vary.

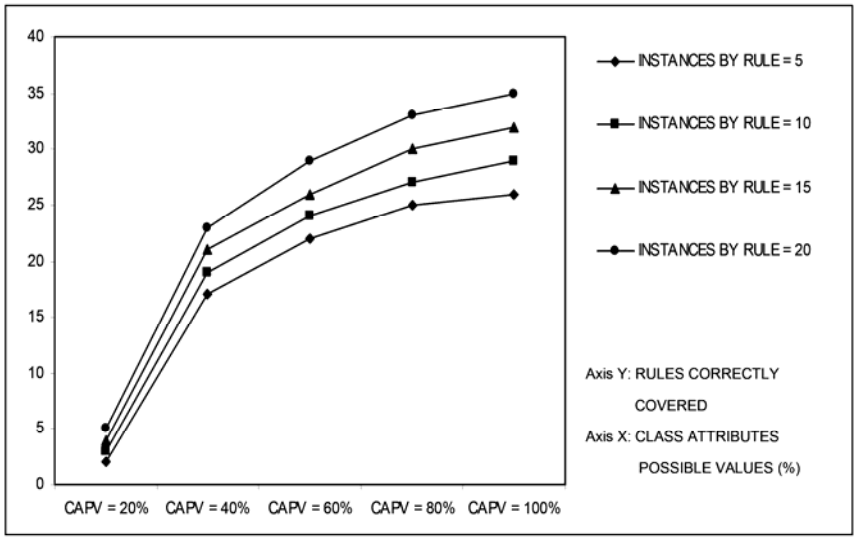

Figure 6. Domains where rules are supported by different amount of examples and the concentration of the rules that determine each class can vary.

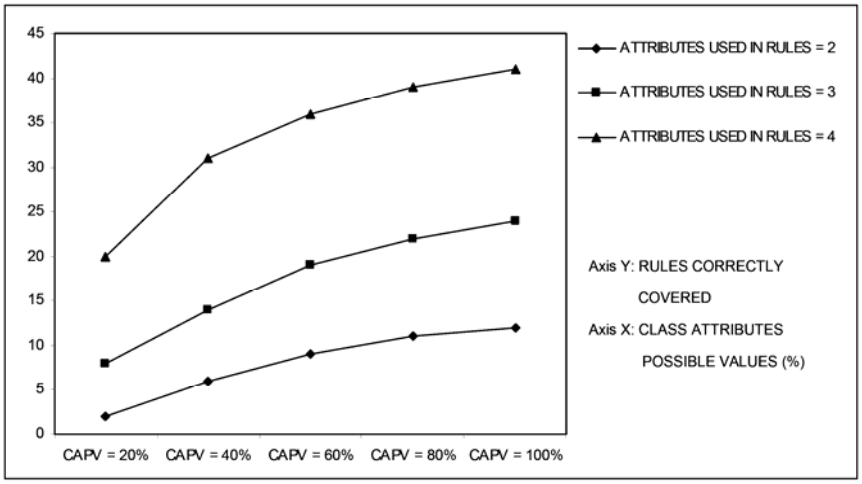

Figure 7. Domains with different specificity of the covering of each rule over its examples and the concentration of the rules that determine each class can vary.

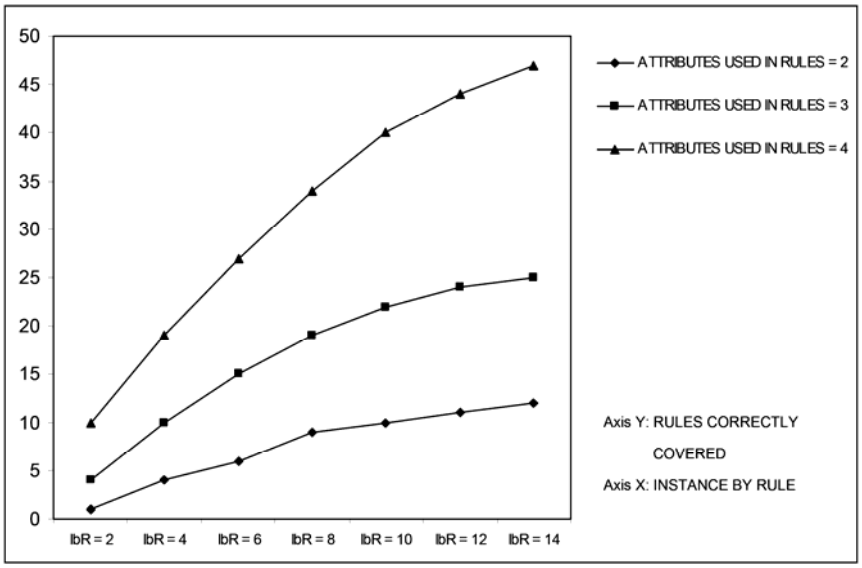

Figure 8. Domains with different specificity of the covering of each rule over its examples and the amount of examples that support each rule can vary.

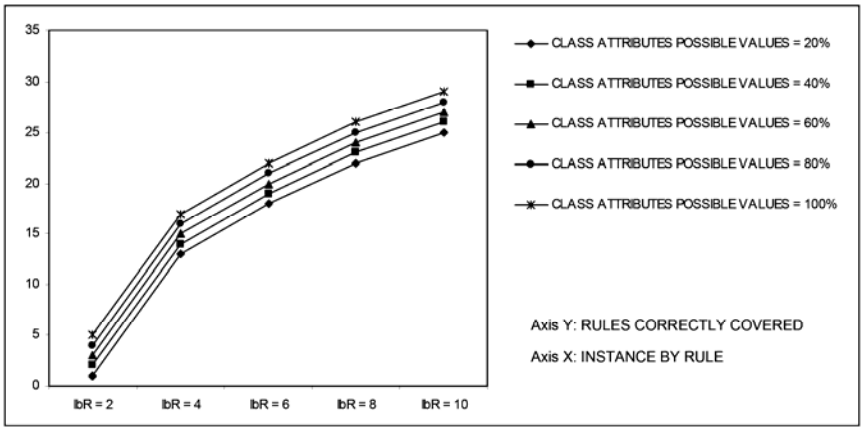

Figure 9. Domains with different concentration of the rules that determine each class and the amount of examples that support each rule can vary.

5. Conclusions

The automatic discovery of useful knowledge pieces is a topic of growing interest in the expert systems engineering community [1, 8 and 10]. Our work differs from those mentioned before in the proposal of a combined mechanism for obtaining rules, using self-organized maps based clustering and induction algorithms.

As our research also focuses on the identification of the necessary processes to allow the autonomous assimilation of the knowledge pieces generated for the expert system, it is necessary to explore which is the quality of the rules discovering process.

References

[1] H. Cao, F. Recknagel, G. Joo and D. Kim, Discovery of predictive rule sets for chlorophylla dynamics in the Nakdong River (Korea) by means of the hybrid evolutionary algorithm HEA, Ecological Informatics 1(1) (2006), 43-53.

[2] J. Debenham, Knowledge Systems Design, Prentice Hall, 1990.

[3] J. Debenham, Knowledge Engineering: Unifying Knowledge Base and Database Design, Springer-Verlag, 1998.

[4] R. García Martínez and D. Borrajo, Planning, learning and executing in autonomous systems, Lecture Notes in Artificial Intelligence, Springer -Verlag 1348 (1997), 208-210.

[5] R. García Martínez and D. Borrajo Millán, An integrated approach of learning, planning and executing, Journal of Intelligent and Robotic Systems 29(1) (2000), 47-78.

[6] R. García Martínez, D. Borrajo, P. Britos and P. Maceri, Learning by knowledge sharing in autonomous intelligent systems, Lecture Notes in Artificial Intelligence, Springer-Verlag 4140 (2006), 128-137.

[7] R. Grossman, S. Kasif, R. Moore, D. Rocke and J. Ullman, Data Mining Research: Opportunities and Challenges, A Report of three NSF Workshops on Mining Large, Massive and Distributed Data, Chicago, 1999.

[8] F. Hoffmann, B. Baesens, C. Mues and J. Vanthienen, Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms, European J. Oper. Res. 2006, in press.

[9] R. Michalski, I. Bratko and M. Kubat, Eds., Machine Learning and Data Mining, Methods and Applications, John Wiley & Sons Ltd., West Sussex, England, 1998.

[10] V. Podgorelec, P. Kokol, M. Stiglic, M. Heričko and I. Rozman, Knowledge discovery with classification rules in a cardiovascular dataset, Comp. Meth. Prog. Biomed. 80 (2005), S39-S49.

[11] C. Rancan, A. Kogan, P. Pesado and R. García-Martínez, Knowledge discovery for knowledge based systems, Some Experimental Results, Res. Comp. Sci. J. 27 (2007a), 3-13.

[12] C. Rancán, P. Pesado and R. García-Martínez, Toward integration of knowledge based systems and knowledge discovery systems, J. Comp. Sci. Tech. 7(1) (2007b), 91-97.