15

A New Multi-Viewpoint and Multi-Level Clustering Paradigm for Efficient Data Mining Tasks

Jean-Charles LAMIREL,
LORIA,Campus Scientifique, France

1. Introduction

Data mining or knowledge discovery in database (KDD) refers to the non-trivial process of discovering interesting, implicit, and previously unknown knowledge from databases. Such a task implies to be able to perform analyses both on high-dimensional input data and large dataset. The most popular models used in KDD are the symbolic models. Unfortunately, these models suffer of very serious limitations. Rule generation is a highly time-consuming process that generates a huge number of rules, including a large ratio of redundant rules. Hence, this prohibits any kind of rule computation and selection as soon as data are numerous and they are represented by very high-dimensional description space. This latter situation is very often encountered with documentary data. To cope with these problems, preliminary KDD trials using numerical models have been made. An algorithm for knowledge extraction from self-organizing network has been proposed in [8]. This approach is based on a supervised generalized relevance learning vector quantization (GRLVQ) which is used for extracting decision trees. The different paths of the generated trees are then used for denoting rules. Nevertheless, the main defect of this method is to necessitate training data. On our own side, we have proposed a hybrid classification method for mapping an explicative structure issued from a symbolic classification into an unsupervised numerical self-organizing map (SOM) [15]. SOM map and Galois lattice are generated on the same data. The cosine projection is then used for associating lattice concepts to the SOM classes. Concepts properties act as explanation for the SOM classes. Furthermore, lattice pruning combined with migration of the associated SOM classes towards the top of the pruned lattice is used to generate explanation of increasing scope on the SOM map. Association rules can also be produced in such a way. Although it establishes interesting links between numerical and symbolic worlds this approach necessitates the time-consuming computation of a whole Galois lattice. In a parallel way, in order to enhance both the quality and the granularity of the data analysis and to reduce the noise which is inevitably generated in an overall classification approach, we have introduced the multi-viewpoint analysis and multi-level clustering approach based on a significant extension of the SOM model, named MultiSOM [19][25]. The viewpoint building principle consists in separating the description of the data into several sub-descriptions corresponding different property subsets or even different data subsets. In MultiSOM each viewpoint is represented by a single SOM map.

The conservation of an overall view of the analysis is achieved through the use of a communication mechanism between the maps, which is itself based on Bayesian inference [1]. The advantage of the multi-viewpoint analysis provided by MultiSOM as compared to the global analysis provided by SOM [11][12] has been clearly demonstrated for precise mining tasks like patent analysis [19]. Another important mechanism provided by the MultiSOM model is its on-line generalization mechanism that can be used to tune the level of precision of the analysis. Furthermore, using free topology clustering methods like the method of the Neural Gas family [23] or those of the K-means family [22] as a new basis, we have proposed in [2] to extent the MultiSOM model into a more general multi-viewpoint and multi-level clustering paradigm, named Multi-Viewpoints Data Analysis (MVDA). The main advantage of the new MDVA paradigm it that it can imbed various clustering methods which might well prove more efficient than SOM model for classification tasks where explicit visualization of the clustering results is not required. Hence, thanks to the loss of topographic constraints as compared to SOM, the free topology clustering methods, like K-means, Neural Gas or its extensions, like Growing Neural Gas (GNG) [7], Incremental Growing Neural Gas (IGNG) [26], or Improved Incremental Growing Neural Gas (I_2GNG) [9], tends to better represent the structure of the data, yielding generally better clustering results [2].
In this chapter we will propose a new approach for knowledge extraction that consists in using the MVDA paradigm as a front-end for efficiently extracting association rules in the context large datasets constituted by high-dimensional data. In our approach we exploit both the generalization and the intercommunication mechanisms of our new paradigm. We also make use of our original recall and precision measures that derive from the Galois lattice theory and from Information Retrieval (IR) domains. The first introduces the notion of association rules. The second section presents the MVDA model. The third section gives an overview of the specific clustering quality criteria that are used in our approach. The fourth section presents the rule extraction principles based both on the MVDA model and on the formerly presented quality criteria. The experiment that is presented on the last section shows how our method can be used both to control the rules inflation that is inherent to symbolic methods and for extracting the most significant rules.

2. The symbolic model and association rules extraction

The symbolic approach to Database Contents Analysis is mostly based on the Galois lattice model [30]. A Galois lattice, *L(D,P)*, is a conceptual hierarchy built on a set of data *D* which are described by a set of properties *P* also called the intention (Intent) of the concept of the lattice. A class of the hierarchy, also called a "formal concept", is defined as a pair *C=(d,p)* where *d* denotes the extension (Extent) of the concept, i.e. a subset of *D*, and *p* denotes the intention of the concept, i.e. a subset of *P*. The lattice structure implies that it exists a partial order on a lattice such that:

$$\forall C_1, C_2 \in L, C_1 \Leftrightarrow \text{Extent}(C_1) \subseteq \text{Extent}(C_2) \Leftrightarrow \text{Intent}(C_1) \supseteq \text{Intent}(C_2)$$

Association rules are one of the basic types of knowledge that can be extracted from large databases. Given a database, the problem of mining association rules consists in generating all association rules that have some user-specified minimum support and confidence. An association rule is an expression $A \rightarrow B$ where *A* and *B* are conjunctions of properties. It means that if an individual data possesses all the properties of *A* then he necessarily

possesses all the properties of B as regard to the studied dataset[1]. The support $supp(A \cup B)$ of the rule is equivalent to the number of individuals of the verifying both properties A and B, and the confidence $conf(A \cup B)$ is given by: $conf(A \cup B) = supp(A \cup B) / supp(A)$. An approach proposed by [28] shows that a subset of association rules can be obtained by following the direct links of heritage between the concepts in the Galois lattice. Even if no satisfactory solution regarding the rule computation time have been found, an attempt to solve the rule selection problem by combining rules evaluation measures is described in [3].

3. The MVDA model

In [13][14], Lamirel and al. firstly introduced the dynamic cooperation between clustering models in the context of information retrieval. This new approach has been originally used for analyzing the relevance of user's queries regarding the documentary database contents. It represents a major amelioration of the basic clustering approach. From a practical point of view, the *MultiView Data Analysis paradigm* (MVDA), introduces the use of viewpoints associated with the one of unsupervised Bayesian reasoning in the clustering process. Its main advantage is to be a generic paradigm that can be applied to any clustering method and that permits to enhance the quality and the granularity of data analysis while suppressing the noise that is inherent to a global approach.

The principle of the MVDA paradigm is thus to be constituted by several clustering models which have been generated from the same data or even from data that share the same overall description space. Each model is issued from a specific viewpoint and can be generated by any clustering method. The relation between the models is established through the use of an inter-models communication mechanism relying itself on unsupervised Bayesian reasoning.

The inter-models communication mechanism enables to highlight semantic relationships between different topics (i.e. clusters) belonging to different viewpoints related to the same data or even to the same data description space. In the MDVA context, this communication is based on the use of the information that can be shared by the different clustering models, like data associated to clusters or labels associated to their descriptions (see Fig. 1).

The inter-models communication is established by standard Bayesian inference network propagation algorithm which is used to compute the posterior probabilities of target model's nodes (i.e. clusters) T_k which inherited of the activity (evidence Q) transmitted by their associated data or descriptor nodes. This computation can be carried out efficiently because of the specific Bayesian inference network topology that is associated to the set of models by the MVDA paradigm [1]. Hence, it is possible to compute the probability $P(ac_m t \mid t_k, Q)$ for an activity of modality act_m on the model node t_k which is inherited from activities generated on the source model. This computation is achieved as follows:

$$P(act_m \mid t_k, Q) = \frac{\sum_{d \in act_m, t_k} Sim(d, s_d)}{\sum_{d \in t_k} Sim(d, s_d)} \quad (1)$$

such that s_d is the source node to which the data d has been associated, $Sim(d, s_d)$ is the cosine correlation measure between the description vector of the data d and the one of its source

[1] An association rule cannot be considered as a logical implication, because his validity directly depends on the dataset from which it is extracted.

node s_d and $d \in act_m, t_k$, if it has been activated with the positive or negative modality act_m from the source model.
The nodes of the target model getting the highest probabilities can be considered as the ones who include the topics sharing the strongest relationships with the topics belonging to the activated nodes of the source model.
One of the richness of this paradigm is that there are very various ways to define viewpoints. One possible way consists in separating the description space of the data into different subspaces corresponding to different criteria of analysis. As an example, an image can be simultaneously described using 3 different viewpoints represented by: (1) a keyword vector; (2) colour histogram vector; (3) a feature vector. A multi-view analysis that is performed on such data can thus highlight the most interesting correspondences between the domains of colours, shapes and image topics while letting the opportunity to figure out specific relationships existing inside each specific domain.

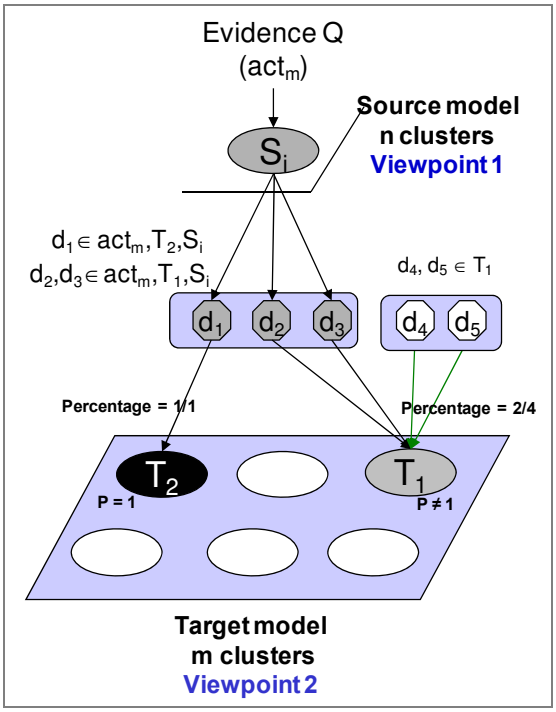


Fig. 1. The MVDA inter-models communication principle.

The relation between maps is established through the use of two main mechanisms: the inter-maps communication mechanism presented formerly and the generalization mechanism that we present hereafter.
The main roles of the MVDA generalization mechanism are both to evaluate the coherency of the topics that have been computed on an original clustering model and to summarize the contents of this later into more generic topics with the advantage of avoiding applying any further learning process. Our generalization mechanism [2] creates its specific link structure

in which each node of a given level is linked to its 2-nearest neighbours (see Fig. 2). For each new level node n the following description vector computation applies:

$$W_n^{M+1} = \frac{1}{3}\left(W_n^M + \sum_{n_k \in V_n^M} W_{n_k}^M \right) \quad (2)$$

where V_n^M represents the 2-nearest neighbour nodes of the node n on the level M associated to the cluster n of the new generated level $M+1$.

After description vectors computation, the repeated nodes of the new level (i.e. the nodes of the new level that share the same description vector) are summarized into a single node. Our generalization mechanism can be considered as an implicit and distributed form of a hierarchical clustering method based on neighbourhood reciprocity [21]. Existing clustering algorithms, such as growing hierarchical self-organizing map (GHSOM) [24], represents a dynamically growing architecture which evolves into a descending hierarchical structure. Nevertheless, the weak point of such methods is to isolate lower level models without regards to their potential links with the other levels. As opposed, our generalization method has the advantage of preserving the original neighbourhood structure between nodes on the new generated levels. Moreover, we have shown in [2] that this method produces more homogeneous results than the classical training approach which should be repeated at each level, while significantly reducing time consumption. Lastly, the inter-model communication mechanism presented in the former section can be used on a given viewpoint between a clustering model and its generalizations as soon as they share the same projected data.

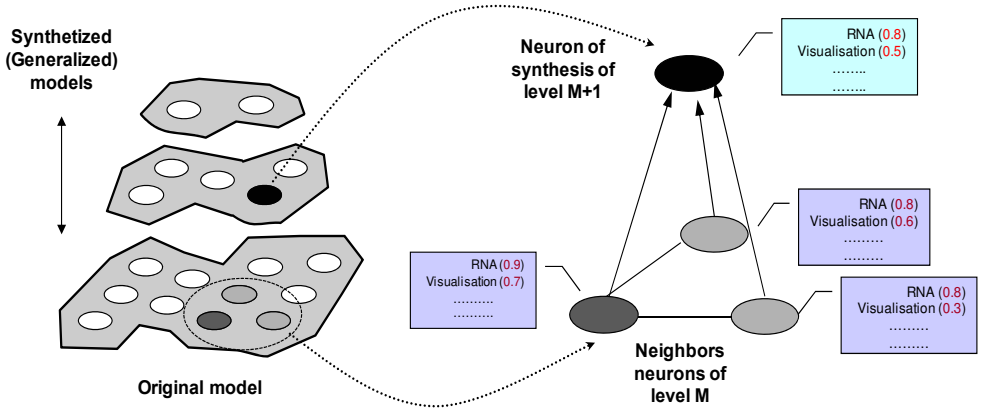


Fig. 2. MVDA generalization mechanism applied on a Neural Gas clustering model (2D representation of gas is used only for the sake of clarity of the figure)

The MVDA paradigm has been chosen as one of the two reference approaches of the IST-EISCTES European project [6]. Its most recent version has opened new perspectives for automatic link analysis in webometrics by allowing to automatically combining referencing and textual information [18]. In section 4, we show this model can be exploited for efficient rule extraction. Our association rule extraction approach makes use of this model in combination with specific clustering quality indexes that we present in the next section.

4. Quality of classification model

When anyone aims at comparing clustering methods, or even evaluating clustering results, he will be faced with the problem of choice of reliable clustering quality indexes. The classical evaluation indexes for the clustering quality are based on the intra-cluster inertia and the inter-cluster inertia [4] [5][21]. Thanks to these two indexes, a clustering result is considered as good if it possesses low intra-cluster inertia as compared to its inter-cluster inertia. However, as it has been shown in [17], the distance based indexes are often strongly biased[2] and highly dependent on the clustering method. They cannot thus be easily used for comparing different methods, or even different clustering results issued from data whose description spaces have different sizes. Moreover, as it has been also shown in [Ka], they are often properly unable to identify an optimal clustering model whenever the dataset is constituted by complex data that must be represented in a both highly multidimensional and sparse description space, as it is often the case with textual data. To cope with such problems, our own approach takes its inspiration both from the behavior of symbolic classifiers and from the evaluation principles used in Information Retrieval. Our Recall/Precision and F-measures indexes exploit the properties of the data associated to each cluster after the clustering process without prior consideration of clusters profiles [17]. Their main advantage is thus to be independent of the clustering methods and of their operating mode.

In IR, the **Recall** ***R*** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of relevant documents which should have been found in the documentary database [27]. The **Precision** ***P*** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of documents returned for the said query. **Recall** and **Precision** generally behave in an antagonist way: as **Recall** increases, **Precision** decreases, and conversely. The ***F*** function has thus been proposed in order to highlight the best compromise between these two values [29].

It is given by:

$$F = \frac{2(R * P)}{R + P} \tag{3}$$

Based on the same principles, the *Recall* and *Precision* quality indexes which we introduce hereafter evaluate the quality of a clustering method in an unsupervised way[3] by measuring the relevance of the clusters content in terms of shared properties. In our further descriptions, a cluster's content is supposed to be represented by the data associated with this latter after the clustering process and the descriptors (i.e. the properties) of the data are supposed to be weighted by values within the range [0,1].

Let us consider a set of clusters *C* resulting from a clustering method applied on a set of data *D*, the local *Recall* (Rec) and *Precision* (Prec) indexes for a given property *p* of the cluster *c* can be expressed as:

$$\underset{c}{\mathrm{Rec}}(p) = \frac{\left|c_p^*\right|}{\left|D_p^*\right|}, \quad \underset{c}{\mathrm{Prec}}(p) = \frac{\left|c_p^*\right|}{\left|c\right|} \tag{4}$$

[2] A bias can occur when the intrinsic dimensions of the obtained clusters (number of non-zero components in the reference vectors describing the clusters) are not of the same order of magnitude than the intrinsic dimensions of the data profiles (see [17] for more details).

[3] Conversely to classical **Recall** and **Precision** indexes that are supervised.

where the notation X_p^* represents the restriction of the set X to the set members having the property p.
Fig. 3 illustrates the basic principle of the new unsupervised **Recall** and **Precision** indexes that have been formerly presented.

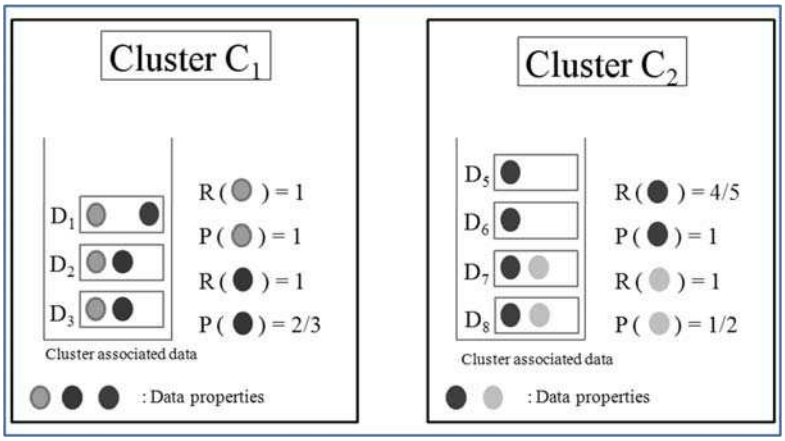


Fig. 3. Principle of Recall(R)-Precision(P) quality indexes
(in this example, for the sake of simplicity data are considered to have Boolean properties).

Then, for estimating the overall clustering quality, the averaged **Recall** (R) and **Precision** (P) indexes can be expressed as:

$$R = \frac{1}{\left|\overline{C}\right|}\sum_{c\in\overline{C}}\frac{1}{S_c}\sum_{p\in S_c}\underset{c}{\text{Rec}}(p)\,,\quad P = \frac{1}{\left|\overline{C}\right|}\sum_{c\in\overline{C}}\frac{1}{S_c}\sum_{p\in S_c}\underset{c}{\text{Prec}}(p) \tag{5}$$

where S_c is the set of properties which are peculiar to the cluster c that is described as:

$$S_c = \left\{ p\in d,\, d\in c \;\middle|\; \overline{W}_c^p = \underset{c'\in C}{Max}\left(\overline{W}_{c'}^p\right) \right\} \tag{6}$$

where $\overline{C}$ represents the peculiar set of clusters extracted from the clusters of C, which verifies:

$\overline{C}$ verifies $$\overline{C} = \left\{\; c\in C \,\middle|\, S_c \neq \varnothing \;\right\} \tag{7}$$

and:
$$\overline{W}_c^p = \frac{\sum_{d\in c} W_d^p}{\sum_{c'\in C}\sum_{d\in c'} W_d^p}$$

where W_x^p represents the weight of the property p for element x.
Similarly to IR, the **F-measure** (described by Eq. 3) could be used to combine averaged **Recall** and **Precision** results. Moreover, we demonstrate in Annex A that if both values of

averaged **Recall** and **Precision** reach the unity value, the peculiar set of clusters C represents a Galois lattice. Therefore, the combination of this two measures enables to evaluate to what extent a numerical clustering model can be assimilated to a Galois lattice natural classifier. The stability of our **Quality** criteria has also been demonstrated in [20].
Macro-Recall and *Macro-Precision* indexes defined by (Eq. 5) can be considered as cluster-oriented measures because they provide average values of *Recall* and *Precision* for each cluster. They have opposite behaviors according to the number of clusters. Thus, these indexes permit to estimate in a global way an optimal number of clusters for a given method and a given dataset. The best data partition, or clustering result, is in this case the one which minimizes the difference between their values (see Fig. 4).

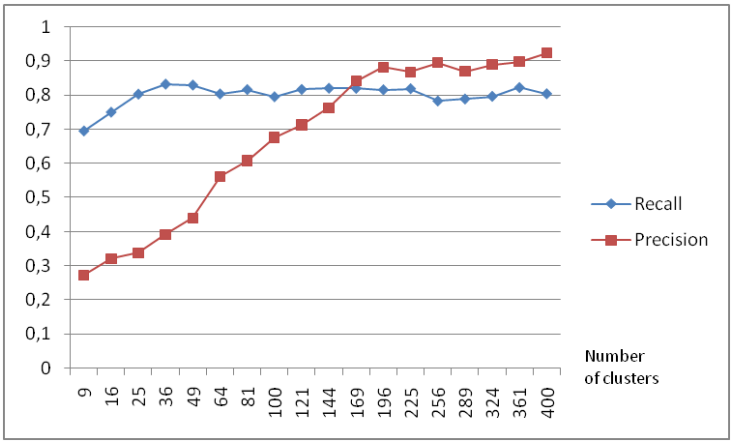


Fig. 4. Evolution of quality criteria for optimal clustering model detection. The optimal model is found at the break-even point between the Recall and Precision quality indexes, letting varying the number of clusters. Here the optimal clustering model is obtained at 169 clusters.

5. Rules extraction from a MultiGAS model

An elaborated unsupervised clustering model, like a MultiGAS model, which represents itself an extension of the Neural Gas model relying on the MDVA paradigm, is a natural candidate to cope with the related problems of rule inflation, rule selection and computation time that are inherent to symbolic methods. Hence, its synthesis capabilities that can be used both for reducing the number of rules and for extracting the most significant ones. In the knowledge extraction task, the generalization mechanism can be specifically used for controlling the number of extracted association rules. The intercommunication mechanism will be useful for highlighting association rules figuring out relationships between topics belonging to different viewpoints.

5.1 Rules extraction by the generalization mechanism

We will rely on our local cluster quality criteria (see Eq. 4) for extracting rules from the classes of the original gas and its generalizations. For a given class c, the general form of the extraction algorithm (**A1**) follows:

$\forall p_1, p_2 \in P_c^*$

1. ***If*** (Rec(p_1) = Rec(p_2) = Prec(p_1) = Prec(p_2) = 1) ***Then***: $p_1 \leftrightarrow p_2$ (equivalence rule)
2. ***ElseIf*** (Rec(p_1) = Rec(p_2) = Prec(p_2) = 1) ***Then***: $p_1 \rightarrow p_2$
3. ***ElseIf*** (Rec(p_1) = Rec(p_2) = 1) ***Then***
   If (Extent(p_1) $\subset$ Extent(p_2)) ***Then***: $p_1 \rightarrow p_2$
   If (Extent(p_2) $\subset$ Extent(p_1)) ***Then***: $p_2 \rightarrow p_1$
   If (Extent(p_1) $\equiv$ Extent(p_2)) ***Then***: $p_1 \leftrightarrow p_2$

$\forall p_1 \in P_c^*, \forall p_2 \in P_c - P_c^*$

4. ***If*** (Rec(p_1) = 1) ***If*** (Extent(p_1) $\subset$ Extent(p_2)) ***Then***: $p_1 \rightarrow p_2$ (*)

where Prec and Rec respectively represent the local *Precision* and *Recall* measures, Extent(p) represents the extension of the property p (i.e. the list of data to which the property p is associated), and P_c^* represent the set of peculiar properties of the class c.
The optional step 4) (*) can be used for extracting **extended rules**. For **extended rules**, the constraint of peculiarity is not applied to the most general property. Hence, the extension of this latter property can include data being outside of the scope of the current class c.

5.2 Rules extraction by the inter-gas communication mechanism

A complementary extraction strategy consists in making use of the extraction algorithm in combination with the principle of communication between viewpoints for extracting rules. The general form of the extraction algorithm (**A2**) between two viewpoints v_1 and v_2 will be:

$\forall p_1 \in P_c^*, \forall p_2 \in P_{c`}^*$ and $c \in v_1$, $c` \in v_2$

1. ***If*** (Rec(p_1) = Rec(p_2) = Prec(p_1) = Prec(p_2) = 1) ***Then*** *Test_Rule_Type*;
2. ***ElseIf*** (Rec(p_1) = Rec(p_2) = Prec(p_2) = 1) ***Then*** *Test_Rule_Type*;
3. ***ElseIf*** (Rec(p_1) = Rec(p_2) = Prec(p_1) = 1) ***Then*** *Test_Rule_Type*;
4. ***ElseIf*** (Rec(p_1) = Rec(p_2) = 1) ***Then*** *Test_Rule_Type*;

where *Test_Rule_Type* procedure is expressed as:

1. ***If*** (Extent$_{v1}$(p_1) $\subset$ Extent$_{v2}$(p_2)) ***Then***: $p_1 \rightarrow p_2$
2. ***If*** (Extent$_{v2}$(p_2) $\subset$ Extent$_{v1}$(p_1)) ***Then***: $p_2 \rightarrow p_1$
3. ***If*** (Extent$_{v1}$(p_1) $\equiv$ Extent$_{v2}$(p_2)) ***Then***: $p_1 \leftrightarrow p_2$

Extended rules will be obtained as:

a. $\forall p_1 \in P_c^*, \forall p_2 \in P_{c`}^*$: Substituting respectively Rec(p_2) and Prec(p_2) by the *viewpoint-based measures* Rec$_{v1}$(p_2) and Prec$_{v1}$(p_2), related to the source viewpoint, in the previous algorithm.
b. $\forall p_1 \in P_c, \forall p_2 \in P_{c`}^*$: Substituting respectively Rec(p_1) and Prec(p_1) by the *viewpoint-based measures* Rec$_{v2}$(p_1) and Prec$_{v2}$(p_1), related to the destination viewpoint, in the previous algorithm.

6. Experimental results

Our test database is a database of 1000 patents that has been used in some of our preceding experiments [19]. For the viewpoint-oriented approach the structure of the patents has been parsed in order to extract four different subfields corresponding to four different viewpoints: **Use**, **Advantages**, **Patentees** and **Titles**. As it is full text, the content of the textual fields of the patents associated with the different viewpoints is parsed by a standard lexicographic analyzer

in order to extract viewpoint specific indexes. The obtained indexes are then normalized by an expert of the patent domain. Table 1 summarizes the results of the indexing phase.
Each of our experiments is initiated with optimal gases generated by means of an optimization algorithm based on our quality criteria [17] (see also Fig. 4). In a first step, original optimal gases are generated for all the viewpoints. In a second step, generalized gases are generated for each viewpoint by applying successive steps of generalization on the original optimal gases. The results of these two steps are summarized in Table 2.
Our first experiment consists in extracting rules from each single viewpoint. Both the original gases and their generalizations are used for extracting the rules. The algorithm is first used without its optional step, and a second time including this step (for more details, see Algorithm A1). The overall results of rule extraction are presented in Table 3.
Some examples of extracted rules related to each viewpoint are given hereafter:

- *Refrigerator oil* → *Gear oil* (supp = 7, conf = 100%) (Use)
- *Wide viscosity range* → *Thermal and oxidative stability* (supp = 3, conf = 100%) (Advantages)
- *Surfactant system* → *Calcium* (supp = 7, conf = 100%) (Title)
- *Infineum* → *Hexxon* (supp = 10, conf = 100%) (Patentees)

	Use	**Advantages**	**Patentees**	**Titles**
Number of indexed documents	745	624	1000	1000
Number of rough indexes generated	252	231	73	605
Number of final indexes (after normalization)	234	207	32	589

Table 1. **Summary of the patent indexation process.** Some remarks must be made concerning the results shown in this table. (1). The index count of the **Titles** field is significantly higher than the other ones. Indeed, the information contained in the patent titles is both of higher sparseness and of higher diversity than in the **Use** and **Advantages** fields. (2) The number of final patentees has been significantly reduced by grouping the small companies in the same general index. (3) Only 62% of the patents have aN **Advantages** field, and 75% an **Use** field. Consequently, some of the patents will not be indexed for these viewpoints.

	Original level: Optimal	**Generalized levels**												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Use	100	79	62	50	40	31	26	16	11	-	-	-	-	-
Advantages	121	100	83	75	64	53	44	34	28	23	18	13	-	-
Patentees	49	31	24	19	16	12								
Titles	144	114	111	105	95	83	71	59	46	35	27	22	18	14

Table 2. **Summary of the gas generation process.** For all viewpoints, the generalization limit has been fixed to levels that have more than 10 neurons. Hence, for a given viewpoint, the number of generalization levels depends both on the initial count of neurons of its associated optimal gas and on the homogeneity of the data distribution relatively to this viewpoint.

For evaluating the complexity of our algorithm based on a numerical approach, as compared to a symbolic approach, we use following efficiency factor (EF) computation:

$$EF = (RC * MLC) / (MRC * LC) \quad (9)$$

where RC=rules count, MRC=maximum rules count (symbolic), LC=loops count, MLC=maximum loop count (symbolic).

A global summary of the results is given in Table 3. Said table includes a comparison of our extraction algorithm with a standard symbolic rule extraction method as regards to the amount of extracted rules. In single viewpoint experiment, when our extraction algorithm is used with its optional step, it is able to extract a significant ratio of the rules that can be extracted by a classical symbolic model basically using a combinatory approach. In some case, such as the **Patentees** viewpoint, all the rules of 100% confidence can be extracted from a single level of the gas (see Fig. 9). Alternatively, as in the case of the **Use** viewpoint, the combination of gas levels of the same viewpoint can be used for extracting all the rules of 100% confidence (see Table 3 and Fig. 6). The worse extraction performance is obtained with the **Titles** viewpoint. This relatively low performance (58% of rules of 100% confidence extracted using all the gas levels) can be explained both by the higher sparseness and by the higher diversity of the data related to this viewpoint. Nevertheless, it is compensated by the much better extraction efficiency, as compared to the symbolic model. Moreover, in the case of this viewpoint, the extracted rules have an average support which is higher than the average support of the overall rule set (see Table 3 and Fig. 10).

Even if no rule selection is performed when the extraction algorithm is used with its optional step, the main advantage of this version of the algorithm, as compared to a classical symbolic method, is the computation time. As a matter of fact, the computation time is significantly reduced, since our algorithm is class-based. Moreover, generally speaking, the lower the generalization level, the more specialized will be the classes, and hence, the lower will be the combinatory effect during computation (see Fig. 6,7,8, 9 and 10). Another interesting result is the behaviour of our extraction algorithm when it is used without its optional step. Fig. 7 shows that, in this case, a rule selection process that depends on the generalization level is performed: the higher the generalization level, the more rules will be extracted. We have already done some extension of our algorithm in order to search for partial rules. Complementary results showed us that, even if this extension is used, no partial rules will be extracted in the low level of generalization when no optional step is used. This tends to prove that the standard version of our algorithm is able to naturally perform rule selection.

Our second experiment consists in extracting rules using the intercommunication mechanism between the Use and the Advantage viewpoints. The communication is achieved between the original gas of each viewpoint, and furthermore, between the same levels of generalization of each viewpoint (see Fig. 5). For each single communication step the extraction algorithm is applied is a bidirectional way.

Some examples of extracted rules are given hereafter:

- *Natural oil* (Advantages) → *Catapult oil* (Use) (supp = 8, conf = 100%)
- *Natural oil* (Advantages) → *Drilling fluid* (Use) (supp = 8, conf = 100%)

The results of our multi-viewpoint experiment are similar to the ones of our single viewpoint experiment (see Table 3). A rule selection process is performed when the standard version of our algorithm is used. The maximum extraction performance is obtained when *viewpoint-based Recall* and *viewpoint-based Precision* viewpoint are used (see Algorithm A2).

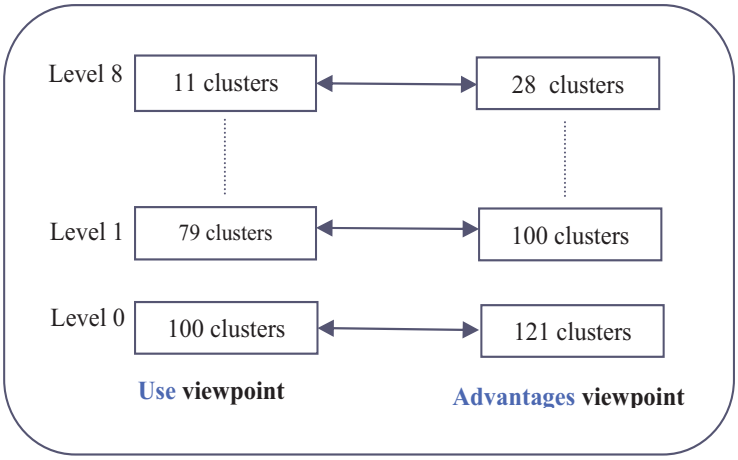


Fig. 5. **Overview of the communication process.** For the generalization process, only 9 levels have been used for both viewpoints, starting at level 0 from the optimal gases.

		Patentees	Titles	Use	Advantages	Use ↔ Advantages
Symbolic model	**Total rule count**	12	2326	536	404	649
	Average confidence	100%	100%	100%	100%	100%
	Average support	3.583	1.049	1.139	1.042	1.081
	Global rule count	26	4912	2238	1436	2822
	Average confidence	53%	59%	59%	44%	45%
MultiGAS model	**Peculiar rule count**	12	422	251	287	250
	Average confidence	100%	100%	100%	100%	100%
	Extended rule count	12	1358	536	319	642
	Average confidence	100%	100%	100%	100%	100%
	% of symbolic total	100%	58%	100%	79%	99%
	Average support	3.583	1.081	1.139	1.050	1.073

Table 3. **Summary of results.** The table presents a basic comparison between the standard symbolic rule extraction method and the MultiGAS-based rule extraction method. The global rule count defined for the symbolic model includes the count of partial rules (confidence<100%) and the count of total rules (confidence=100%). In our experiments, the rules generated by the MultiGAS model are only total rules. The peculiar rule count is the count of rules obtained with the standard versions of the extraction algorithms. The extended rule count is the count of rules obtained with the extended versions of the extraction algorithms including their optional steps.

7. Conclusion

In this paper we have proposed a new approach for knowledge extraction based on a MultiGAS model, which represents itself an extension of the Neural Gas model relying on the MDVA paradigm. Our approach makes use of original measures of unsupervised Recall and Precision for extracting rules from gases. Thanks to the MultiGAS model, our experiments have been conduced on single viewpoint classifications as well as between multiple viewpoints classifications on the same data. They take benefit of the generalization and the inter-gas communication mechanisms that are embedded in the MVDA paradigm. Even if complementary experiments must be done, our first results are very promising. They tend to prove that a clustering model, as soon as it is elaborated enough, represents a natural candidate to cope with the related problems of rule inflation, rule selection and computation time that are inherent to symbolic models. One of our perspectives is to more deeply develop our model in order to extract rules with larger context like the ones that can be obtained by the use of closed set in symbolic approaches. Another interesting perspective would be to adapt measures issued from information theory, like IDF or entropy, for ranking the rules. Furthermore, we plan to test our model on a reference dataset on genome. Indeed, these dataset has been already used for experiments of rule extraction and selection with symbolic methods. Lastly, our extraction approach can be applied in a straightforward way to a MultiSOM model, or even to a single SOM model, when overall visualization of the analysis results is required and less accuracy is needed.

8. References

[1] Al-Shehabi S., and Lamirel J.-C. (2004). Inference Bayesian Network for Multi-topographic neural network communication: a case study in documentary data. Proceedings of ICTTA, Damas, Syria, April 2004.

[2] Al Shehabi S., and Lamirel J.C. (2005). Multi-Topographic Neural Network Communication and Generalization for Multi-Viewpoint Analysis. International Joint Conference on Neural Networks - IJCNN'05, Montréal, Québec, Canada, August 2005.

[3] Bayardo Jr. R. J., and Agrawal R. (1999). Mining the most interesting rules. In Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, p.145-154, August 1999.

[4] Calinski T. and Harabasz J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3 (1974), 1–27.

[5] Davies D., and Bouldin W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. 1 (1979) 224–227.

[6] François C., Hoffmann M., Lamirel J.-C., and Polanco X. (2003). Artificial Neural Network mapping experiments. EICSTES (IST-1999-20350) Final Report (WP 9.4), 86 p., September 2003.

[7] Frizke B. (1995). A growing neural gas network lerans topologies. Tesauro G., Touretzky D. S., leen T. K., Eds., Advances in neural Information processing Systems 7, pp 625-632, MIT Press, Cambridge MA.

[8] Hammer B., Rechtien A., Strickert M., and Villmann T. (2002). Rule extraction from self-organizing networks. ICANN, Springer, p. 877-882.

[9] Hamza H., Belaïd Y., Belaîd. A, and Chaudhuri B. B. (2008). Incremental classification of invoice documents. 19th International Conference on Pattern Recognition - ICPR 2008.

[10] Kassab R., and Lamirel J.-C., (2008). Feature Based Cluster Validation for High Dimensional Data. IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008.

[11] Kaski S., Honkela T., Lagus K., and Kohonen, T. (1998). WEBSOM-self organizing maps of document collections, Neurocomputing, vol. 21, pp. 101-117.

[12] Kohonen T. (2001). Self-Organizing Maps. 3rd ed. Springer Verlag, Berlin.

[13] Lamirel J.-C., and Créhange M. (1994). Application of a symbolico-connectionist approach for the design of a highly interactive documentary database interrogation system with on-line learning capabilities. Proceedings ACM-CIKM 94, Gaitherburg, Maryland, USA, November 94.

[14] Lamirel J.C. (1995). Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif. PhD Thesis, Université of Nancy 1, Henri Poincaré, Nancy, France.

[15] Lamirel J.C., Toussaint Y., and Al Shehabi S. (2003). A Hybrid Classification Method for Database Contents Analysis, FLAIRS 03 Conference, p. 286-292.

[16] Lamirel J.C., Al Shehabi S., Hoffmann M., and Francois C. (2003). Intelligent patent analysis through the use of a neural network : experiment of multi-viewpoint analysis with the MultiSOM model. Proceedings of ACL, Sapporo, Japan, p. 7-23.

[17] Lamirel J.C., Al Shehabi S., Francois C., and Hoffmann M. (2004). New classification quality estimators for analysis of documentary information: application to web mapping. Scientometrics, Vol. 60, No. 3, p. 445-462.

[18] Lamirel J.C., Al Shehabi S., François C., and Polanco X. (2004). Using a compound approach based on elaborated neural network for Webometrics: an example issued from the EICSTES Project. Scientometrics, Vol. 61, No. 3, p. 427-441.

[19] Lamirel J.-C., and Al Shehabi S. (2006). MultiSOM: a multiview neural model for accurately analyzing and mining complex data. Proceedings of the 4th International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV), London, UK, July 2006.

[20] Lamirel J.C., Ghribi M., and Cuxac P. (2010). Unsupervised Recall and Precision measures: a step towards new efficient clustering quality indexes. 19th International Conference on Computational Statistics (COMPSTAT'2010), Paris, France, August 2010.

[21] Lebart, L. , Morineau A., and Fénelon J. P. (1982). Traitement des données statistiques. Dunod, Paris, France.

[22] McQueen J.B. (1967). Some methods of classification and analysis of multivariate observations. L. Le Cam and J. Neyman (Eds.), Proc. 5th Berkeley Symposium in Mathematics, Statistics and Probability, Vol. 1: 281-297, Univ. of California, Berkeley, USA.

[23] Martinetz T., and Schulten K. (1991). A "neural-gas" network learns topologies. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, Artificial neural networks, North-Holland, Amsterdam, p. 397-402.

[24] Ontrup J. and Ritter H. (2005). A hierarchically growing hyperbolic self-organizing map for rapid structuring of large data sets. In Proceedings of 5th Workshop On Self-Organizing Maps - WSOM 05, Paris 1 Panthéon-Sorbonne University.

[25] Polanco X., Lamirel, J.C., and François C. (2001). Using Artificial Neural Networks for Mapping of Science and technology: A Multi self-organizing maps Approach. Scientometrics, Vol. 51, N° 1, p. 267-292.

[26] Prudent Y., Ennaji A. (2005). An Incremental Growing Neural Gas learns Topology. ESANN2005, 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, 27-29April 205, published in Neural Networks, 2005. IJCNN apos;05. Proceedings. 2005 IEEE International Joint Conference , vol. 2, no. 31 pp 1211 - 1216, July-4 Aug. 2005.

[27] Salton G. (1971). The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall Inc., Englewood Cliffs, New Jersey.

[28] Simon A., and Napoli A. (1999). Building Viewpoints in an Object-based Representation System for Knowledge Discovery in Databases. Proceedings of IRI'99, Atlanta, Geogia, S. Rubin editor, The International Society for Computers and Their Applications, ISCA, p. 104-108, 1999.

[29] Van Rijsbergen C. J. (1975). Information Retrieval. Butterworths, London, England, 1975.

[30] Wille R. (1982). Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In Ordered Sets, (I. Rival, ed.), D. Reidel: 1982, p. 445-470.

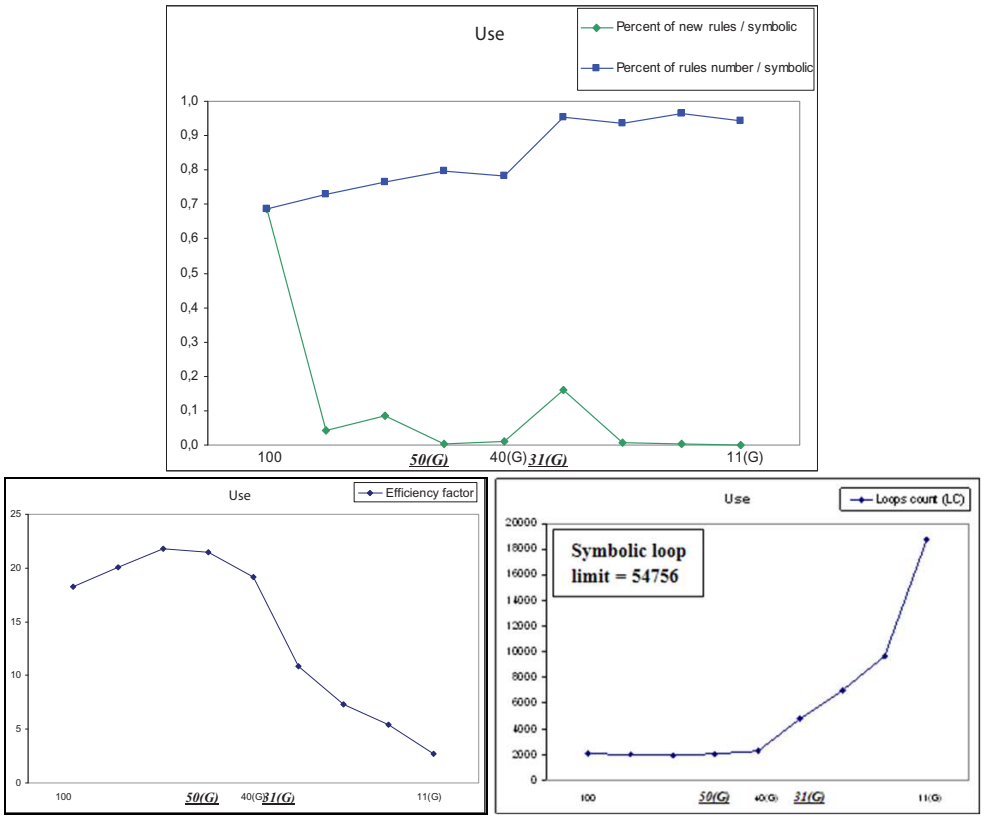


Fig. 6. **Rule extraction results for the Use viewpoint (extraction algorithm A1 with optional step).** The 50(G) level can be considered as an optimal level since it provides the best compromise between the percentage of extracted rules (80% of all the rules) and the computation complexity (2033 loops). However, if the percentage of extracted rules is considered as prior to the computation complexity, the 31(G) level (95% of all the rules, 4794 loops) should be considered as a more optimal level.

New rules: rules that are found at a given level but not in the preceding ones.
Symbolic loop limit: number of loops used by a symbolic approach for extracting the rules.
x(G): represents a level of generalization of x neurons).

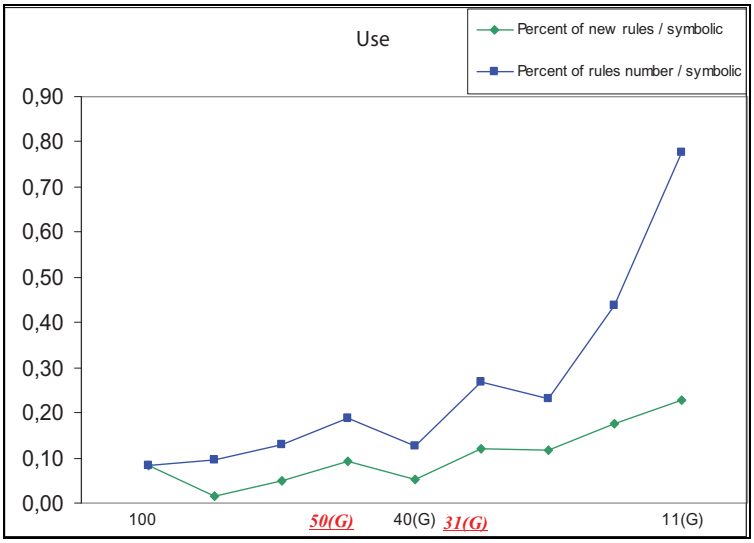


Fig. 7. **Rule extraction results for the Use viewpoint (extraction algorithm A1 without optional step).** Only peculiar rules are extracted with this version of the algorithm. In this case, a rule selection process that depends on the level of generalization is performed: the higher the level of generalization, the lower will be the selection. The good performances of the respective 50(G) and 31(G) levels are also highlighted with this version of the algorithm (see fig. 3 for a comparison with the other version of the algorithm).

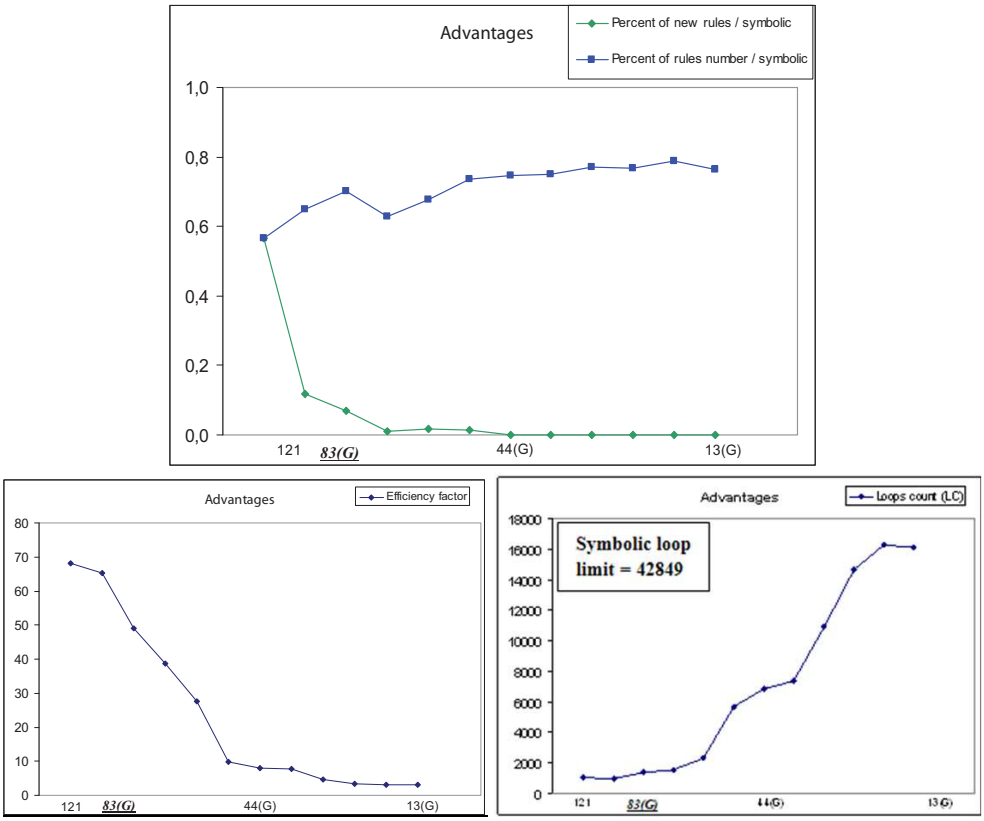


Fig. 8. **Rule extraction results for the Advantages viewpoint.** The 83(G) level can be considered as the level representing the optimal compromise: 70% of all the rules are extracted using 1402 loops.

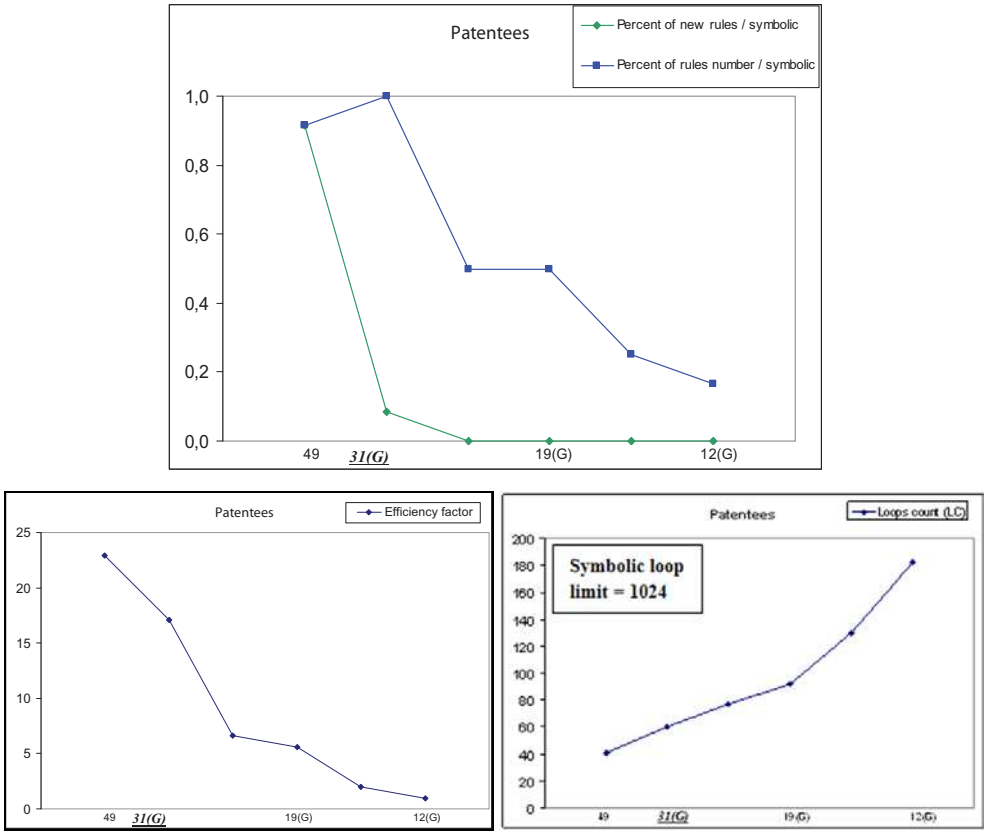


Fig. 9. **Rule extraction results for the Patentees viewpoint.** The 49(G) level can be considered as an optimal level since it provides the best compromise between the percentage of extracted rules (92% of all the rules) and the computation complexity (41 loops). Nevertheless, if the percentage of extracted rules is considered as prior to the computation complexity, the 31(G) should be considered as a more optimal level: all the rules are extracted using 61 loops. The decrease of the rules extraction performance for high generalization levels is due all together to the specific indexation characteristics of this viewpoint (the average number of indexes per document is near to 1), to the initial gathering effect of the winner-takes-most NG learning strategy, to the further effect of the generalization process and to the rule extraction strategy based on the distribution of single properties. Hence, when generalization is performed, the documents described by combination of indexes that have been initially gathered into the same classes by the initial learning strategy will be spread into specific classes. In such a way, the distribution of single properties into the classes will become more heterogeneous for intermediary generalization levels than for the initial level.

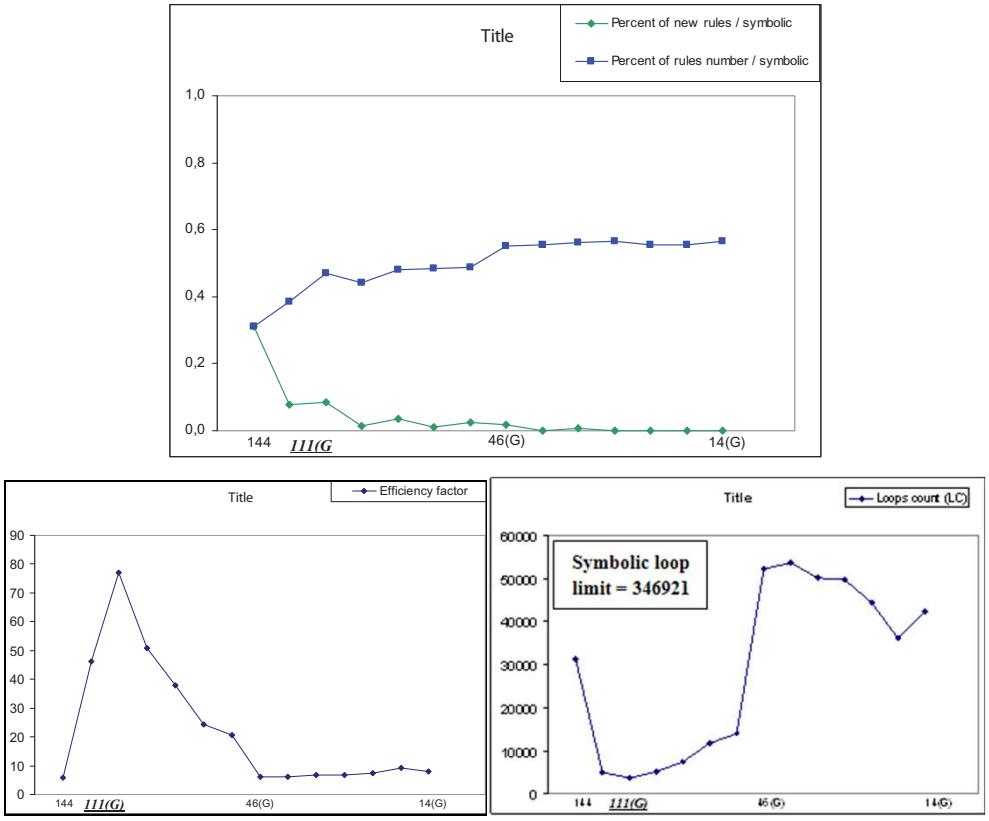


Fig. 10. **Rule extraction results for the Titles viewpoint.** The 111(G) level can be considered as an optimal level. Hence, even if the percentage of extracted rules is relatively low (48% of all the rules), this level provides a very good extraction efficiency (Efficiency factor=76 and 3619 loops) as compared to the symbolic model. The high rule computation complexity of the first gas level (optimal level), as compared to the next generalization levels, is due both to the sparseness of indexation of the **Titles** viewpoint and to the gathering effect of the winner–takes-most NG learning strategy (see also fig. 6 concerning that point). Hence, when data distribution is sparse, this learning strategy tends to gather the data into a few number of classes, letting some other classes empty. This effect will increase significantly the complexity of rule computation on the original level since non empty classes will have relatively heterogeneous descriptions. It should be noted that this complexity will be reduced at intermediary steps of generalization since unrelated data will be spread into different classes by the generalization process.

Annex 1: Galois lattice equivalence

Let D being a given dataset and P being the set of properties associated to the data of D,
Let $f : P \rightarrow D$ being a function associating to a set of properties p, the set of data which possess at least all the properties of p
Let $g : D \rightarrow P$ being a function associating to a set of data d, the set of properties which are at least common to all the data of d

Let $\bar{C}$ be a set of peculiar clusters issued from D with:
(see Section 3 for peculiar cluster definition)

$\forall c \in \bar{C}, D_c = \{d_1, d_2, \ldots, d_n\}$ being the set of data associated to the cluster c, $\forall d_i \in D_c, p_i$ being the set of peculiar properties of the data d_i and $P_C = p_1 \cup p_2 \cup \ldots p_n$ being the whole set of peculiar properties of the cluster c

$\bar{C}$ represents a Galois lattice if it verifies the condition:

$\forall c \in \bar{C}, h(P_c) = P_c \text{ with } h(P_c) = g \text{O} f(P_c)$
(Galois lattice definition given in [Lamirel and Toussaint, 2000])

Let c being a cluster of $\bar{C}$:

$$\begin{aligned} f(P_c) &= f(p_1 \cup p_2 \cup \ldots p_n) \\ &= f(p_1) \cap f(p_2) \cap \ldots f(p_n) \text{ by construction of } f \\ &= \{d_1, d_2, \ldots, d_n\} \\ &= D \end{aligned}$$

A value of **Precision** of **1** implies that all each peculiar property p_i of a cluster c is possessed at least by all the data of D_c , *thus:*

$$\forall p_i, f(p_i) \supseteq D_c \tag{1}$$

A value of **Recall** of **1** implies that all the data possessing a peculiar property p_i of a cluster c are included in this cluster, *thus:*

$$\forall p_i, D_c \supseteq f(p_i) \tag{2}$$

Hence, by (1) and (2),

$$f(p_1) = f(p_2) = \ldots f(p_n) = D_c$$

and subsequently:

$$D = D_c \tag{3}$$

$$\begin{aligned} g(D_c) &= g(d_1 \cup d_2 \cup \ldots d_n) \\ &= g(d_1) \cap g(d_2) \cap \ldots g(d_n) \text{ by construction of } g \\ &= \{p_1, p_2, \ldots, p_n\} \\ &= P \end{aligned}$$

A value of **Precision** of **1** implies that each data d_i of a cluster c possesses at least all the properties of P_c, *thus:*

$$\forall d_i, g(d_i) \supseteq P_c \tag{1}$$

A value of **Recall** of **1** implies that each property p_i of a cluster c is belonged exclusively by the data of this cluster*, thus:*

$$\forall d_i, P_c \supseteq g(d_i) \tag{2}$$

Hence, by (1) and (2),

$$g(d_1) = g(d_2) = \ldots g(d_n) = P_c$$

and subsequently:

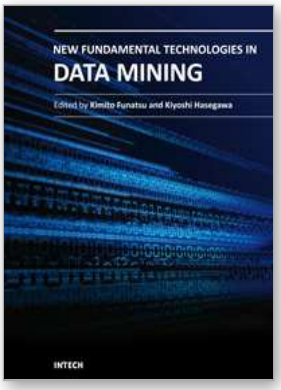
$$P = P_c \tag{4}$$

Finally, $g \mathrm{O} f(P_c) = g(f(P_c))$

$= g(D_c)$ from (3)

$= P_c$ from (4)

Conclusion: Joint unity values of **Recall** and **Precision** implies that the set of peculiar clusters of a numerical classification could be assimilated to a Galois lattice.



New Fundamental Technologies in Data Mining

Edited by Prof. Kimito Funatsu

ISBN 978-953-307-547-1
Hard cover, 584 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by "Data Mining" address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821