A Hybrid Classification Method for Database Contents Analysis

Jean-Charles Lamirel, Yannick Toussaint, Shadi Al Shehabi

A Hybrid Classification Method for Database Contents Analysis

Jean-Charles Lamirel and **Yannick Toussaint** and **Shadi Al Shehabi**
LORIA
BP 239
54506 Vandœuvre-lès-Nancy cedex
France
email : <firstname.lastname>@loria.fr

Abstract

The hybridisation of different classification and mining techniques coming from different areas such as the numeric and the symbolic worlds can produce a significant enhancement of the overall classification and retrieval performance in a Data Mining or Information Retrieval context.

This paper introduces an experimental methodology to match an explicative structure issued from a symbolic classification to a numerical classification. The classification models used in the experiment are a boolean lattice on the symbolic side and a Kohonen Self Organising Map model (SOM) on the numerical side.

INTRODUCTION

The aim of this article is to show that the association of various techniques of classification and data mining – although belonging to very separate fields like the symbolic or the numeric ones – may well take benefit from their mutual advantages.

Early work takled with some similar fundamental problems. Carpineto, (Carpineto & Romano 2000), thus compares the classification based on a lattice with the BMR method (Best Ranking Method) to rank the answers in a documentary system. Nevertheless, the main objective of this paper is to go one step further by associating with a numerical classification an explanatory structure. The numerical model of classification can be regarded as our base of classification. The explanatory structure of the numerical classification is established from a symbolic classification and consists of:

- a set of generic properties associated to a group of numerical classes;
- generic-specific relations between groups of numerical classes based on the symbolic properties characterising these groups;
- a set of association rules covering the properties of the individuals and emphasising dependences between them.

This article concentrates on the first two points, the use of the association rules opening numerous prospects and discussions. The classification models used in the experiment are a boolean lattice on the symbolic side and a Kohonen SOM on the numerical side.

Because these two models owns to different paradigms, we adopt an experimental approach for their association. We test some heuristics that we borrow from either paradigm or the other and define a method to evaluate them.

The next section underlines how Kohonen SOM and boolean lattices are complementary. Then, the following section proposes some definitions before describing very shortly the numerical and the symbolic models that are used. The section contains a discussion on the major problems which one encounters in the association between both approaches. The section initially describes our experimental data set which is constituted of textual data. It then presents the different heuristics that are tested and the method of evaluation which is choosen. Lastly, the conclusion summarizes our most promising choices in this association of numerical and symbolic methods.

THE COMPLEMENTARITY OF THE APPROACHES

Each of the two approaches has its own strengths and weaknesses.

The main advantages of the Kohonen SOM model, as compared to other classification models, are its natural robustness and its very good illustrative power. Indeed, it has been successfully applied for several classification tasks (Kohonen 2001).

A topographical classification of the Kohonen type lead nevertheless to major problems of in-deep interpretation mainly because the profiles of the obtained classes represent very complex combination of weighed properties. The principal characteristics of these classes are therefore difficult to emphasise for the user and may induce shortcuts in interpretation or, moreover, misinterpretation. It is indeed what was observed by Lin (Lin, D.Soergel, & Marchionini 1991) when elementary class naming methods based on the weights of class were used for labelling maps generated from textual data. It therefore raises, more generally, the problem of labelling the classes for giving an overview of their contents.

Another problem comes from the learning mechanism of the Kohonen map which is of competitive “winner take

most" type. This may lead to the lost of information issued from marginal data. However, this second problem could be partly solved by an appropriate weighting scheme on the original data, as the one we propose in section .

The lattice advantages are based on the possibilities:

- to update in an incremental way the lattice when a new document or a new property is introduced into the base;
- to extract association rules;
- to take a model of knowledge and to be be able to characterise an individual by multivalued relations instead of single attributs;
- and finally, to take a topdown approach of the set of classes since they are treated on a hierarchical basis.

The generally large number of formal concepts of a lattice, its hierarchical structure and the absence of topography rends difficult the visualisation of a lattice and decreases the overall legibility of the structure. The lattice calculation although represents an expensive process in terms of time and complexity.

As the Kohonen map represents a reliable visualisation support for data analysis, one way to cope with its main defects is to provide him with a sound explicative structure. Our idea is therefore to project the numerical classes (Kohonen) on the symbolic classes (lattice) to benefit from the formal properties of the lattice as elements of explanation for the map. To exploit the synergy between topographical classification and the lattice will therefore make it possible to the user to reach in a formal way the association rules and the correlations which could not be directly highligted in the Kohonen model.

Why the Galois lattice would be an explanatory structure? First, the definition of a function of projection of a numerical class on a symbolic class makes it possible to associate with a numerical class having a very large number of weighed properties, a symbolic class whose properties, in less number, better represent the principal elements of the numerical class. Moreover, if several numerical classes are projected on a same symbolic class, then the properties of this symbolic class will correspond to the properties shared by this group of numerical classes. To go one step further, the hierarchical structure of the lattice can play the role of a hierarchy of abstraction: the regrouping of numerical classes in increasingly larger sets will be associated with formal concepts increasingly larger relatively to the relation of order partial of the lattice. Lastly – but we will not detail it here – the association rules extracted from this structure are associated with a symbolic class. They therefore bring elements of additional explanation and validation to the numerical classes which were associated with this symbolic class.

THEORETICAL BASES

The basic terminology

Definition 1 (Individual, property) *An entry in the database will be called an* individual. *Each individual in the database has a set of* properties. *Let x beiing an individual and p a property. In the symbolical model, all the couples (x, p) can be described by a boolean matrix $\mathcal{M}$. Hence, if p is a property of x, the value of (x, p) in the matrix $\mathcal{M}$ will be "true". In the numerical model, the properties of individuals are usually weighted. Hence, the matrix $\mathcal{M}$ is a matrix of weights where the value of (x, p) in the matrix $\mathcal{M}$ represents the weight of the property p for x. The overall weight vector of an individual x is then represented by the line x of the matrix $\mathcal{M}$.*

Definition 2 (Extension, intension) *The* intension *of an individual is equivalent to the set of the properties which are associated to an individual. The* intension *of a class is the set of properties which characterises the class. The* extension *of the class is the set of the individuals belonging to the class.*

The numerical model

The SOM model approach consists in considering the classification of the individuals as a non-linear process of projection on a two-dimensional grid of neurons in which the neurons maintain predefined neighbourhood relations (Kohonen 2001). At the end of the classification process, each neuron of the map plays the role of a representative of a class of individuals.

A topographic map is initialised through one unsupervised competitive learning conduced on the whole database of the individuals, thanks to their associated weight vectors (see def. 1).

The topographical properties of the Kohonen map model make possible the reprojection of the individuals on a map, in such a way that their proximity on this map illustrates as much as possible their proximity in their original description space. Once associated with a neuron, an individual can be regarded as a representative of the class of individuals materialised by this neuron. After the whole projection phase some of the neurons of the map will have several associated individuals. In contrast, other neurons of the map will not have any. The latter will then materialise said "intermediate classes" whose role is that to maintain topographical continuity, rather than that to be the representatives of precise thematic tendencies associated with the data of the map.

After the preliminary learning phase, each map is organised so as be made "legible" to the user. This organisation is mainly based on the analysis of the dominant components of the profiles of the neurons of the map. A first phase of this analysis consists to define labels of classes which optimally represent the contents of the latter when they are presented to the user. One second phase of the analysis consists to divide the map into coherent logical areas. The two phases of analysis which are mentionned above are more precisely described in (Lin, D.Soergel, & Marchionini 1991) or, in an alternative manner, in (Lamirel & Crehange 1994) (see figure. 1).

The symbolic model

The symbolic approach for the analysis of the contents of the database is based on the Galois lattice model (Wille 1982). We present it hereafter in a rather informal way. From now, an element of a Galois lattice is called a "formal concept" to distinguish it from an element of the Kohonen map which is

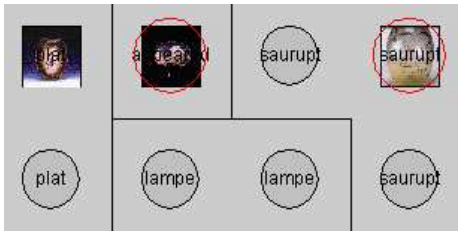

Figure 1: Partial view of a topographic Kohonen map: icons show the most representive member of the classes, labels represent class names, circles represent intermediate classes, lines represent logical area borders

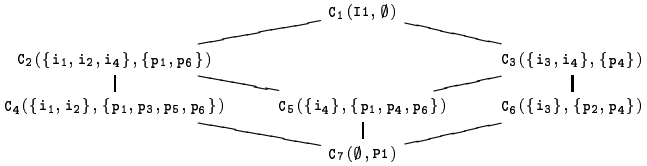

Figure 2: Graph representing the lattice

called a “class”. The analysis of the database which results from the use of the symbolic model is thus composed of a set of formal concepts structured as a hierarchy (i.e. the Galois lattice) and a set of association rules.

Let $\mathbb{I}$ beiing a set of individuals, and $\mathbb{P}$ beiing a set of properties. A formal concept is a couple $(\mathtt{I}, \mathtt{P})$ where $\mathtt{I}$ represents a set of individuals and $\mathtt{P}$ a set of properties such that $\mathtt{I}$ includes all the individuals with all the properties in $\mathtt{P}$, and conversely, $\mathtt{P}$ contains all the properties shared by the individuals of $\mathtt{I}$.

Introduction of a partial order: A relation of partial order on the set F_C of the formal concepts $\mathbb{P}, \mathbb{R})$ can be defined by:
$(I_1, P_1) \preccurlyeq (I_2, P_2) \leftrightarrow I_1 \subseteq I_2$ or, in an equivalent way
$(I_1, P_1) \preccurlyeq (I_2, P_2) \leftrightarrow P_1 \supseteq P_2$.
$\preccurlyeq$ is a partial order.

Property 1 (Lattice) $(F_C, \preccurlyeq)$ *define a lattice. For each pair of formal concept, there exist both a minimal and a maximal element thanks to the relation* $\preccurlyeq$.

The following example illustrates the notion of lattice and the figure 2 gives a view on such a type of structure. Let $\mathcal{M}$ be the following matrix (where $\times$ stands for the boolean value “true”):

	p_1	p_2	p_3	p_4	p_5	p_6
i_1	×		×		×	×
i_2	×		×		×	×
i_3		×		×		
i_4	×			×		×

Different algorithms for building lattices can be found in (Bordat 1986),(Ganter, Stahl, & Wille 1986). Incremental approaches are proposed in (Carpineto & Romano 1993). The lattice building complexity is also adressed in (Carpineto & Romano 1993).

Extraction of association rules: We will not detail this section because we do not exploit these rules for the moment. It is however important to stress that it is possible to extract from a lattice association rules on properties of type A - > B where A and B are conjunctions of properties. It means that if an individual possesses all the properties of A then he has necessarily the properties of B. The algorithms of extraction of these rules are, in particular, described in (Duquenne 1996),(Godin & Missaoui 1994).

THE EXPERIMENT

The general outline of the experiment consists in building the topography of Kohonen and the lattice on the same initial set of data. From a purely formal point of view, the comparison between the two structures is not possible as one uses a distance and the other a partial order (see section). This leads us to define one or more heuritics to project each Kohonen class onto one or more formal concept of the lattice.

The evaluation of this methodology can be summed up by the following questions: on what level[1] are the classes of the Kohonen map projected onto the lattice? Are Kohonen classes that are very close on the map projected on brother classes in the lattice? What are the variations observed in these according to the type of projection used? Might we observe a certain correlation between the relation of generalisation in the lattice and the concept of connexity in the topography of Kohonen, which would then make it possible to define explanatory areas of increasing size and level of generality on the map?

The data set

We led our experimentation on the iconographic database related to the artistic period "Art Nouveau" managed by the Web server BIBAN(Lamirel, Ducloy, & Oster 2000). The iconographic database contains approximately 200 images of various artistic works out of the period of "Art Nouveau". Various fields are covered such as architecture, painting, and scuplture. The images are associated with a description of the bibliographical type with a title, a set of keywords, and authors. Each image constitutes an individual and each keyword is regarded as a property of this individual.

Before using a numerical model, a common operation consists in weigthing the properties of the original data. Weigthing mainly aims at reducing the influence on the classification of the properties which are the most widespread among the data. In our SOM model, we successfully tested a weighting scheme which we have called IDF-Norm. IDF-Norm uses the classical IDF (Inverse Document Frequency) (Jones 2000) weighting, issued from information retrieval, as the principal method of weighting. According to this weighting scheme, the higher frequency has a property $\mathtt{p}$ in the whole set of individuals, the lower $\mathtt{IDF}(\mathtt{p})$ is. Additionally to IDF, the weights of the properties are normalised by individual, so that the amount of the weights is equal to 1. The weight of the same property $\mathtt{p}$ therefore becomes variable from one individual to another. This additional method has the advantage of strengthening the weight of the properties of the individuals who possess only few properties.

[1]The level of a formal concept in a lattice is defined as the length of the longest path between it and the top of the lattice.

The methodology

To study the complementarity of the symbolic and numerical approach, we have adopted a methodology in four stages that we detail in the following subsections: the projection, the grouping, the agglomeration, and finally, the pruning.

Definition of the projection

Each Kohonen class is projected onto one or more formal concepts. The formal concept is considered as an explicative structure for the Kohonen class (see section). The definition of the heuristics of projection of the Kohonen classes on the formal concepts and the quality of this projection are, of course, of primary importance for our future explanatory goals. We tested 3 different heuristics:

Subsumption: Subsumption is the first heuristics which comes to the mind as it is completely in the logic of construction of the lattices. Nevertheless, strong differences of cardinality between the property sets of Kohonen classes and those of the formal concepts result from the different principles of the two classification methods. Moreover, in most cases, several properties of the Kohonen classes are associated with a very low weight. Thereby, we chose to use a threshold under which, a property of a Kohonen class would not be taken into account for the calculation of subsumption[2]. We tested 4 various values of threshold: 0.0 – who conserves all the properties of the Kohonen class–, 0.1 who conserves the properties whose weight was greater than 0,1, then than 0,2 and 0,3.

The choice of the subsumption as the projection strategy has for consequence that a Kohonen class is generally projected on several formal concepts of the lattice.

The definition of a distance: One of the main weaknesses of the subsumption heuristics is to lead to complex explanatory structures because Kohonen classes are often projected onto several formal concepts. To cope with this problem, we search for another heuristics liable to produce a more simple structure: a Kohonen class is projected onto the formal concept which is the nearest relatively to a distance. Among possible distances, cosine was the most appropriate because it is a norm independent distance[3]. Even with the use of cosine, we have nevertheless to ensure that the distance between a formal concept and a Kohonen class owning an equivalent set of properties is minimal. For this, we consider that a formal concept – intially described by a boolean vector – can be rewritten as a weighted vector such that for any property $\mathtt{p}$ if the initial boolean value is "false", then the weight of $\mathtt{p}$ is $\mathtt{0}$ else, the weight of $\mathtt{p}$ is $\mathtt{IDF(p)}$.

As for any given property $\mathtt{p}$, the value $\mathtt{IDF(p)}$ is independent of both the individuals and the formal concepts, it is obvious that the above described weighted method do not alter the initial lattice structure[4]. Cosine distance can thus be expressed as: $\frac{T.K}{\|T\|*\|K\|}$, where K and T represent respectively the vector of weights of the Kohonen class and those of the formal concept.

Combination of subsumption and distance: This method seeks among the formal concepts subsuming a Kohonen class the one which is the closest. As for cosine, the interest is to reduce the number of formal concepts who receive a projection of a Kohonen class.

Grouping

The grouping does not change the repartition of the Kohonen classes over the formal concepts: instead of evaluating the quality of the couples $(\mathtt{kc}, \mathtt{fc})$, we study the pairs $((\bigcup_{i=1}^{n} \mathtt{kc_i}), \mathtt{fc})$ where $\mathtt{kc_i}$ are the various Kohonen classes projected on the same $\mathtt{fc}$. The case when several Kohonen classes have been associated to a formal concept defines a first level of area definition.

Agglomeration

The agglomeration process be viewed as a hierarchical classification process over an initial Kohonen map. It is used to conceive explanatory areas of increasing size on the map. The semantics of the agglomeration is given by the lattice. As a matter of fact, the $\preccurlyeq$ relation ensures that if $(\mathtt{I_1}, \mathtt{P_1}) \preccurlyeq (\mathtt{I_2}, \mathtt{P_2})$ then $\mathtt{I_1} \subseteq \mathtt{I_2}$. The greater is the formal concept (wrt. $\preccurlyeq$), the larger si the set of individuals which share the properties of this formal concept.

This principle has been used to set up the following agglomeration algorithm:

Let `Agg` be an array such that `Agg(fc)` is the set of Kohonen classes associated to the formal concept `fc`. For each `fc` of the lattice `Agg(fc)` is initialised with the set of Kohonen classes projected onto `fc`.
Let `Bottom` be the smaller formal concept of the lattice (wrt. $\preccurlyeq$).

$\mathtt{CurrentFC\_Set} = \mathtt{Bottom}$ *#the current set of formal concepts*
While ($\mathtt{CurrentFC\_Set} <> \emptyset$) do
  foreach formal concept $\mathtt{fc_i}$ of $\mathtt{CurrentFC\_Set}$ do
    $\mathtt{FatherSet} = Select\_Fathers(\mathtt{fc_i})$
    foreach father $\mathtt{fc_j} \in \mathtt{FatherSet}$
      $\mathtt{Agg}(\mathtt{fc_j}) = \mathtt{Agg}(\mathtt{fc_j}) \cup \mathtt{Agg}(\mathtt{fc_i})$
    done
  $\mathtt{CurrentFC\_Set}$ = Union of all the fathers of each element of $\mathtt{CurrentFC\_Set}$
  done

The result of agglomeration strongly depends on the $Select\_Fathers$ procedure. We tested two heuristics, the first one being the most immediate. In this case, $Select\_Fathers(\mathtt{fc_i})$ produces a set constituted by the entire list of $\mathtt{fc_i}$ fathers. This heuristics preserve the multiple

[2] Carpineto (Carpineto & Romano 2000) – whose purpose was very different – proposed to keep only the $\mathtt{k}$ first properties, $\mathtt{k}$ being the average of the properties associated to the individuals in the numerical approach

[3] Distances which are dependent of the norm such as, for example, Euclidean or Inclusion distances, cannot properly deal with the difference in the cardinality of the property sets.

[4] Applying a full IDF-Norm weigthing on the lattice, similarly to SOM, would not preserve its structure

heritage of the lattice but induces a very complex final explicative structure.

The second heuristics uses the same algorithm but $Select\_Fathers(\mathtt{fc_i})$ produces a singleton which is the nearest father within the meaning of the cosine distance retained. This is this heuritics that will be reported hereafter.

Pruning of the herarchical structure

The lattice can be simplified in order to visualise only the formal concepts with which a Kohonen class was associated. According to the choosen agglomeration heuristics, one finally obtains a tree structure with multiple heritage for which the relation of partial order is different from that of the initial lattice.

Comparative analysis of the heuritics

We firstly base our evaluation on criteria which could reflect, in an overall way, the soundness of our projection and agglomeration heuristics. We thus focused on the use of – precision and recall – which are well-known measures in the world of information retrieval (Salton 1983). The role is these criteria is thus to measure the similarities, in terms of repartition of individuals, bewteen the formal concepts and their associated Kohonen classes. We secondly used complementary criteria whose role is to evaluate the quality of the generated structure, when it is viewed as an explanatory structure for an end-user.

Overall evaluation of heuristics : Recall and precision In the information retrieval domain, the recall (R) is the proportion of correct results found relatively to the whole set of correct answers and precision (P) is the proportion among the provided answers, of correct answers. At the projection level, these two values can be interpreted in the following way: the more the individuals of a Kohonen class are included in the extension of the formal concept which is associated to it, the higher is the recall. The more the individuals of a formal concept are included in the Kohonen class, the better the precision will be[5]. Let $\mathtt{k}$ be a Kohonen class, $\mathtt{c}$ its associated formal concept, and $\mathtt{K}$ and $\mathtt{T}$ their respective extensions. Thus, $\mathtt{P}$ and $\mathtt{R}$ are expressed by: $\mathtt{P(k,c)} = \frac{|\mathtt{K} \cap \mathtt{T}|}{|\mathtt{T}|}$, $\mathtt{R(k,c)} = \frac{|\mathtt{K} \cap \mathtt{T}|}{|\mathtt{K}|}$.

At the grouping level, recall of a group (GR) and the precision of a group (GP) are defined as: $\mathtt{GP(c)} = \frac{|\bigcup_{i=1}^{n}\{\mathtt{K_i}\} \cap T|}{|T|}$, $\mathtt{GR(c)} = \frac{|\bigcup_{i=1}^{n}\{\mathtt{K_i}\} \cap T|}{|\bigcup_{i=1}^{n}\{\mathtt{K_i}\}|}$ where $\mathtt{k_i}$ (for $\mathtt{i} = 1, \mathtt{n}$) are the $\mathtt{n}$ Kohonen classes projected on the formal concept $\mathtt{c}$ of the lattice.

The recall and precision for the agglomeration (resp. AR and AP) are calculated on the same basis as GR and GP but for each formal concept involved in the process of agglomeration.

[5] In our approach, precision is a more important criteria than recall. As a matter of fact, a high precision garanties a minimum of dissimilarities between a Kohonen and its associated formal concept.

Heuristics	S0.0	S0.1	S0.2	S0.3	Cos.	S-Cos
Recall(AR)	0.05	0.28	0.87	0.99	0.52	0.25
Precision(AP)	0.26	0.59	0.35	0.17	0.81	0.86
Group Recall(AGR)	0.06	0.28	0.75	0.99	0.49	0.35
Group Prec.(AGP)	0.73	0.63	0.70	0.61	0.82	0.94
Agg Recall(AAR)	0.15	0.40	0.78	0.99	0.61	0.54
Agg Prec.(AAP)	0.88	0.72	0.84	0.69	0.83	0.88
# of formal cpt(InvP)	180	108	20	9	44	41
Proj.Level(APL)	3.83	3.26	1.5	1.33	4.20	4.51
step of agg. 1	**126**/138	**76**/80	**14**/14	**6**/6	**36**/36	**38**/39
step of agg. 2	**33**/41	**22**/27	**5**/5	**2**/2	**11**/12	**8**/11
step of agg. 3	**17**/21	**7**/10	**1**/1	**1**/1	**5**/7	**2**/4
step of agg. 4	**9**/11	**4**/5	-	-	**3**/3	**3**/3
step of agg. 5	**4**/6	**1**/1	-	-	**1**/1	**1**/1
step of agg. 6	**1**/1	**1**/1	-	-	-	-
step of agg. 7	**1**/1	**1**/1	-	-	-	-
step of agg. 8	**1**/1	**1**/1	-	-	-	-

Table 1: Comparative analysis of the heuristics. In the last part of the table, the boldface values represent the number of formal concepts associated to connexe Kohonen classes.

Explanatory quality A first set of two criteria may be use to compare the heuristics by evaluating the quality of the initial projection:

- the average level of projection (APL). The higher it is, the most explicative will be the formal concept as its intension is larger.
- the number of formal concepts involved in a projection (InvP). The higher it is, the more discriminative is the projection[6].

The hierarchical structure obtained after pruning can finally be evaluated by a last set of criteria: the number of levels, the number of formal concepts per level, and the balance of the hierarchy. It also includes the connexity criterion: at which point the Kohonen classes which are grouped or agglomerated on the same formal concept are close or related on the Kohonen map? The idea is that if the agglomeration makes it possible to group Kohonen classes which are topographically close, then the resulting hierarchical structure will be more easier to comment by the expert. Additionally, if connexity is checked at the time of the agglomeration, the properties of the formal concepts involved in the process of agglomeration could be used directly as an explanation of the Kohonen map.

Results The initial set of data is composed of 162 individuals characterised per 191 properties. The Kohonen map has 100 classes. The lattice built on this data makes up of 307 formal concepts. The lattice has 11 levels. Only 48 Kohonen classes have a non empty extension and so could be used for the evaluation. Moreover, for the sake of statistical validity, only 29 classes having an extension higher than 3 individuals can be used for the evaluation in R and P.

The top part of table 1 contains the average values for precision and recall. They allow us to focus on the most interesting heuristics as regard to values R and P which are both high and stable among projection, grouping and agglomera-

[6] InP is meaningful only when a Kohonen class is projected on a single formal concept (cosine and subsumption-cosine projections)

tion. R and P stability is suppose to maintain the homogeneity of the overall explanatory structure for the Kohonen map. Cosine and subsumption 0.1 are top-ranked here.

The recall and precision can however be regarded as imperfect measures to evaluate the intrinsic quality of the projection. Indeed, they do not make it possible to emphasise which properties of a Kohonen class are not taken into account in its associated formal concepts: are these properties important or not? It seems obvious that the risk of suppressing important properties during projection is minimised by the cosine method in comparison with pure subsumption. In fact, cosine presents the interest to take directly into account the weight of the properties in the process of projection.

Looking in the middle part of the table, we will prefer heuristics for which the InvP value is in the middle range. In fact a too high value of Invp corresponds to a complex explanatory structure, and, conversly, a too low value corresponds to a weak explanatory one. That also means that the projection is not capable to reflect the smoothness of the description of Kohonen classes: either the Kohonen classes share the same properties but with different weights, or the subsuming formal concepts share only few properties with the Kohonen classes.

Moreover, a high APL value means that the formal concepts which are implied in the projection have a large description relatively to their intension. The top-ranked methods concerning these two last criteria are subsumption 0.1, cosine and subsumption-cosine.

Finally, the last part of the table brings us to the following conclusions:

- In terms of number of formal concepts and by looking at the final structure of the hierarchy, subsumption 0,2, sosine and subsumption-cosine are structures which can be interpreted by an expert. Subsumption 0,0 and 0,1 are not well-balanced hierarchies. Subsumption 0,3 is too simple.
- At first sight, subsumption 0,2 seems better than Cosine or than subsumption-cosine insofar as the Kohonen classes grouped or agglomerated on the same formal concept are connexe.
- However, the hierarchy corresponding to Cosine is better balanced than subsumption 0,2 and the three formal concepts which do not correspond to related Kohonen classes are distributed among two levels of the lattice.

CONCLUSION

We have presented an experimental methodology to match an explicative structure issued from a symbolic classification to a numerical classification. This association gives access to an analysis which one could not have by using each method separately. Our experiment highlighted that cosine distance seems to be the most interesting heuristics to project Kohonen classes on formal concepts of a lattice as the process of agglomeration leads to a well-balanced explanatory hierarchy. Our criterion of connexity could nevertheless be refined in order to have a better estimation of the accuracy of the heuristics we proposed. Indeed, thanks to this criterion, a formal concept is considered related if each class of Kohonen grouped or agglomerated on this formal concept if it is adjoining to at least another Kohonen class also associated with this formal concept. It is a very strong condition which could be modulated.

To be able to test our approach on different sets of data, we developed a tool which enables us to carry out all these analyses. Experimenting on another set of data will enable us to see in which way the choice of heuristics could depend on the data. In our future work, we will analyze the interest to exploit, in addition to this stage, the association rules extracted from the lattice. We will particularly link the exploitation of these rules with the capacities of communication between points of view proposed by the MicroNOMAD multiSOM model (Lamirel & Crehange 1994) – our own multimap extension of the classical SOM model – since this model makes it possible to find in an unsupervised way thematic correlations between the data.

References

Bordat, J. 1986. Calcul pratique du treillis de galois d'une correspondance. *Maths. Sci. Hum.* 24ième année(96):31–47.

Carpineto, C., and Romano, G. 1993. Galois: An order-theoretic approach to conceptual clustering. In *Proceedings of the 10th International Conference on Machine Learning (ICML'93)*. Morgan Kaufmann.

Carpineto, C., and Romano, G. 2000. Order-theoritical ranking. *Journal of the American Society For Information Science* 51(7):587–601.

Duquenne, V. 1996. On lattices approximations: Syntactic aspects. *Social Networks* 18:189–199.

Ganter, B.; Stahl, J.; and Wille, R. 1986. Conceptual measurement and many-valued contexts. In Gaul, W., and Schader, M., eds., *Classification as a Tool of Research*.

Godin, R., and Missaoui, R. 1994. An incremental concept formation approach for learning from databases. *Theoretical Computer Science* 133(2):387–419.

Jones, K. S. 2000. Search terms relevance weighting. *Journal of Documentation* 35(1).

Kohonen, T. 2001. *Self-Organization and Associative Memory*. Springer Verlag, 3 edition.

Lamirel, J., and Crehange, M. 1994. Application of a symbolico-connectionnist approach for the design of a higly interactive documentary database interrogation system with on-line learning capabilities. In *Proceedings of the ACM/CIKM*.

Lamirel, J.; Ducloy, J.; and Oster, G. 2000. Adaptive browsing for information discovery in an iconographic context. In *Proceedings of RIAO*.

Lin, X.; D.Soergel; and Marchionini, G. 1991. A self organizing semantic map from information retrieval. In *Proceedings of 4th International SIGFIR Conference on R&D in Information retrieval*, 262–269.

Salton, G. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1 edition.

Wille, R. 1982. Restructuring lattice theory : ,an approach based on hierarchies of concepts. In *Ordered sets*. Boston: O. Rival.