Combining symbolic and numeric techniques for DL contents classification and analysis

Jean-Charles Lamirel, Yannick Toussaint

Combining symbolic and numeric techniques for DL contents classification and analysis

Jean-Charles LAMIREL, Yannick TOUSSAINT

Abstract

The goal of this article is to prove that the mixture of different classification and mining techniques coming from so different areas such as the numeric and the symbolic worlds can combine their mutual advantages in order to produce a significant enhancement of the overall classification and retrieval performance in a Data Mining or Information Retrieval context.

1 Introduction

The goal of this article is to prove that the mixture of different classification and mining techniques coming from so different areas such as the numeric and the symbolic worlds can combine their mutual advantages in order to produce a significant enhancement of the overall classification and retrieval performance in a Data Mining or Information Retrieval context. In the following sections, we have experimented different heuristics to combine numerical analysis of a data set with a symbolic one. We will first propose some definitions before describing the numerical and symbolic model. Then we describe three different heuristics to combine them. We end with the definition of criteria to evaluate the quality of the different results.

2 The Theoritical Bases

2.1 The numerical model

The numerical model used in our experiment is the MicroNOMAD multiSOM [Lamirel *et al.*, 2000] model, which represents itself a significant extension of the classical Kohonen SOM topographic map model [Kohonen, 1984]. Conversely to the original Kohonen model, the MicroNOMAD multiSOM model is abble to manage the communication between several topographic maps allowing thus the user to exploit dynamic exchanges between multiple viewpoints (i.e. classifications). In the following section we will nevertheless not focus on the MicroNOMAD multiSOM specific capabilities but rather on its direct exploitation of the classical SOM principles.

The SOM approach considers that a data classification can be viewed as a non linear mapping on a 2D neuron grid in which neurons establish predefined neighborhood relation. After the classification process, each neuron of the map will then play the role of a data class representative. The main advantages of SOM are its natural robustness and its very good illustrative power.

A topographic map is initially built up by unsupervised competitive learning carried out on the whole database. This learning takes place through the profile vectors extracted from the individual descriptions. For each neuron of a map M, the basic competitive learning function has the following global form:

$$W_{t+1}^{n_k} = W_t^{n_k} + \alpha(t)k(t)(W_t^{n_k} - P_t^n)$$

where W_t^n is the external weights profile vector (i.e. the class profile vector) of the neuron n at time t, P_t^n is the vector constituted by the weighted properties of the individual i chosen as learning sample at time t, $W_t^{n_k}$ is the profile of the neuron k at time t, $\alpha(t)$ a time decreasing function, $k(t)$ a neighborhood adaptation function.

The topological properties associated with the Kohonen maps make it then possible to project the original individuals onto a map so that their proximity on the map matches as closely as possible their proximity in their original description space. One individual i is then associated to a neuron n of a map M which verifies:

$$\|W_n - P_i\| = min_{n_k \in M}\|W_{n_k} - P_i\|^2$$

Once associated to a neuron, an individual could be considered as a member of the class described by this neuron.

2.2 The symbolic model

The symbolic approach to Database Content Analysis is based upon a Galois lattice. An element in the lattice will be call **formal concept** to distinguish it from the notion of Kohonen class previously introduced.The analysis which results from the overall symbolic process is composed of the two following pieces of information:

- a set of formal concepts structured in a hierarchy (the Galois lattice)
- a set of association rules

We will focus, in this article, on the hierarchical structure of the lattice.

Definition of a lattice 1

Definition 1 (Context) *Let $\mathbb{I}$ being a set of individuals, $\mathbb{P}$ a set of properties. A context is a triple ($\mathbb{I}$, $\mathbb{P}$, $\mathtt{R}$) where $\mathtt{R} \subset \mathbb{I} \times \mathbb{P}$. For any individual $\mathtt{i} \in \mathbb{I}$ and any property $\mathtt{p} \in \mathbb{P}$, $\mathtt{iRp}$ iff the individual $\mathtt{i}$ owns the property $\mathtt{p}$.*

Definition 2 (Concept) *A formal concept of the context ($\mathbb{I}$, $\mathbb{P}$, $\mathtt{R}$) is defined to be a pair $(\mathtt{I}, \mathtt{P})$ where $\mathtt{I} \subseteq \mathbb{I}, \mathtt{P} \subseteq \mathbb{P}$ and $\mathtt{I} = \{i \in \mathbb{I} \mid (\forall p \in \mathbb{P})\mathtt{iRp}\}$, $\mathtt{P} = \{p \in \mathbb{P} \mid (\forall i \in \mathbb{I})\mathtt{iRp}\}$. That means that $\mathtt{I}$ is the set of individuals that owns all the properties in $\mathtt{P}$ and P is the set of properties common to all the individuals in $\mathtt{I}$. If the pair $(\mathtt{I}, \mathtt{P})$ is complete (there do not exist any other $\mathtt{i} \in \mathbb{I} - \mathtt{I}$ which owns all the properties of $\mathtt{P}$ and there do not exist any other $\mathtt{p} \in \mathbb{P} - \mathtt{P}$ which is owned by all $\mathtt{i} \in \mathtt{I}$), then it represents an admisible formal concept.*

Definition 3 (Partial Order) *We define an ordering relation on the set F_C of formal concepts of the context ($\mathbb{I}$, $\mathbb{P}$, $\mathtt{R}$) by:*
$(I_1, P_1) \leq (I_2, P_2) \leftrightarrow I_1 \subseteq I_2$ *or (equivalently)*
$(I_1, P_1) \leq (I_2, P_2) \leftrightarrow P_1 \supseteq P_2$.
$\leq$ *is a partial order.*

Definition 4 (Lattice) $(F_C, \leq)$ *defines a lattice. That means that each pair of formal concepts has a unique meet and join.*

The most popular Galois lattice construction algorithm was proposed by [Ganter *et al.*, 1986]. This algorithm constructs first the entire set of formal concepts which are then organised with the $\preccurlyeq$ relation.

3 The complementarity of the two approaches

Each approach has its own strength and weakness that we underline in this section. The results of topographical classification methods such as MicroNOMAD results lead to interpretation problems due to the fact that the profile of the obtained classes are mostly complex weighted combination indexes extracted from the documents. The main characteristics of the classes are therefore difficult to highlight to the user and could cause shortcomings or mistakes in the interpretation of the database content. This was previously observed in the first labelisation methods proposed for Kohonen topographies and meets also the more general problem of labeling classes whatever the method is.

The usually high number of classes of a lattice, its hierarchical structure and the lack of any topological structure makes of course harder its visualization and decrease the global readability of the analysis. Even if building a lattice is a time-consuming process with a high complexity, their interests in information retrieval and in text mining activities relies on the possibility:

- to update incrementally the lattice when a new document or a new property is introduced,
- to extract association rules,
- to take into account background knowledge and to represent an individual by a set of relations instead of valued properties,
- and the main point, to adopt a top-down analysis of the set of classes as they are structured in a hierarchy.

Exploiting a synergy between topographical classification and lattice seems then to be very promising way for enhancing the capabilities of data mining and knowledge discovery applications. However, a very first step for this consists in finding reliable connections between the two approaches.

4 The experiment

The core experiment consists in building both the Kohonen topography and the lattice using the same initial data set.It has been carried out on the iconographic database of the "Art Nouveau" period managed by the BIBAN server. BIBAN is a research prototype for iconographic Digital Libraries. It is an application of a generic XML workbench DILIB [Lamirel *et al.*, 2000] and has been designed for investigating digital libraries containing images and heterogeneous documents in a multilingual context. This database contains approximately 300 images related to the various artistic works of the Art Nouveau School. It covers several domains, such as architecture, painting and sculpture. The images are associated to a bibliographic description containing a title, keywords and author information. Each image constitutes an individual and each of its keywords is considered as a property.

4.1 The methodology

To study the complementarity of the numerical and symbolic models, we will adopt a three-level approach:

1. **projection:** we first project each Kohonen class onto one or more formal concept of the lattice. "Projection" designates either the process of calculation or, for a given Kohonen class, the formal concept(s) it has been projected on. We aims at defining criteria to evaluate the quality of the pairs (kc, fc) where kc is a Kohonen class and fc its projection.

2. **grouping:** grouping is not a process but uses the structure provided by the projection process. Instead of evaluating the pairs (kc, fc) we will study the pairs $(\bigcup_{i=1}^{n} \texttt{kc}_\texttt{i}), fc)$ where $\texttt{kc}_\texttt{i}$ are the different classes projected onto $\texttt{fc}$.

3. **agglomerating:** Definition 3 show that for two formal concepts, $(I_1, P_1) \leq (I_2, P_2) \leftrightarrow I_1 \subseteq I_2$. Following this definition, if $projection(kc) = fc_1$ and if $fc_1 \leq fc_2$ then, we would like to verify whether the Kohonen classes which are projected onto fc_1 could be propagated to fc_2. The agglomeration algorithm associates to the formal concept fc_2 the set of Kohonen classes associated to his sons by projection or by agglomeration.

   Agglomeration is supposed to answer the question "how could we group Kohonen classes into areas, and areas into bigger ones?" We have experimented two different strategies for the agglomeration process.

The definition of the heuristics to project a Kohonen class onto a formal concept and the quality of this projection is of course crucial for further observations. We tested three of them :

The subsumption : The subsumption is the first heuristics coming to mind to projet a Kohonen class onto a formal concept in the lattice. It is in complete accordance with the lattice building principle. The major problem comes from the fact that a Kohonen class is characterized by a number of properties – sometimes with a very low weight – much higher than the formal concepts in the lattice. Carpineto [Carpineto and Romano, 2000] kept the k first properties, k being the average of the properties in the numeric method. We adopted the notion of treshold and tested four different ones: 0.0 – keeping all the properties of the Kohonen class, 0.1 keeping properties which value is under 0.1, then 0.2 and 0.3. Using subsumption, a Kohonen class is generally projected onto more than one formal concept.

The definition of a distance : In order to prune the structure obtained by the subsumption projection and to obtain an easier-to-read hierarchical structure with less formal concepts, another heuristics for the projection is based upon the definition of a distance between a Kohonen class and a formal concept. The Kohonen class is then projected onto only one formal concept, the closest one.As the weights of the profile of the individuals have been normed in the Kohonen model, the cosine distance C has been choosen for the comparison :

$$C = \frac{T.K}{\|T\| * \|K\|}$$

where K and T represent respectively the weighted vector of the Kohonen class and of the formal concept. N is the length of these vectors.

Combining subsumption and distance : This method searches for all the subsumants to the Kohonen class and then choose the closest one following the distance.

Heuristics for agglomeration : Starting from the Kohonen map, groups define a first level of areas. Then, these areas are agglomerated into larger ones until all the map is grouped into a single vaste area. The hierarchical structure of the lattice seems a good structure to build these areas. The first heuristics for agglomeration is the most immediate. Each formal concept fc_i groups one or more Kohonen classes. Agglomeration propagates this set of classes to the immediate fathers of fc_i. And so on, until being at the top of the lattice which is though associated to the whole set of Kohonen classes. However, because of multiple inheritance in the lattice, this heuristics "activates" a great number of nodes. Indeed, an expert of the domain will have a hard time in analysing this structure.The second heuristics uses the same principle except that we only keep one father among all the fathers in the lattice. We keep the closest one to fc_i using the cosine distance. That is the heuristics we will use in the rest of the paper.This heuristic leads to a tree structure. In this case, a only partial inheritance from top to leaves of the tree is preserved as the partial order is not preserved.

4.2 Comparative analysis of the results of the different heuristics and distances

We present in this section a comparative analysis of the experiments depending on the heuristics for the projection and for the agglomeration. We choose two criteria to evaluate the final structure obtained. The first uses the well known recall-precision measure issued from Information Retrieval, completed by elementary measures for evaluating the quality of the projections. The second one, more subjective, is concerned by how Kohonen classes are grouped together and agglomerated. It relies on the mesure of the connexity of Kohonen classes (how closely related are these classes). Both aims at defining which final hierarchical structure is better-suited for a user to analyse the content of the data set. We have explored six types of projection : subsumption with a treshold at 0.0, 0.1, 0.2 and 0.3, cosine and, finally, subsumption at 0.0 combined with cosine.

The initial set of data is composed of 162 individuals characterized by 191 properties. The lattice built upon these data has 307 formal concepts; the Kohonen map has 100 classes. The lattice has 11 level including top (level 0) to bottom (level 10). The level for a formal concept is defined as being the shorter way from the top to itself.

4.2.1 Recall and precision

In Information Retrieval, recall (R) is the proportion of relevant materials retrieved and precision (P) is the proportion of retrieved materials which are relevant. We will use these two indicators to evaluate each projection of a Kohonen class onto a formal concept as well as to evaluate the groups and agglomerations of Kohonen classes onto a formal concept. R and P are calculated upon extensions: P = $\frac{|K \cap T|}{|T|}$, R = $\frac{|K \cap T|}{|K|}$, where K and T represent (resp.) the extension of the Kohonen class and of the formal concept.

Two other kinds of measure can be used with benefit to measure the quality of the projection.The average projection level (APL) in the lattice give an information that is complementary to the precision for measuring the accuracy of the projection and also its generalization sharpness. The discriminating power of a projection can be directly estimated by comparing the number of Kohonen classes to be projected with the number of formal concept involved in the projection (InvP).

The projection of Kohonen class onto formal concept

Subsumption at 0.0: The projection produces 360 couples (Kohonen class, formal concept). The number of formal concept receiving at least the projection of a Kohonen class is 180. The average projection level (APL) onto the lattice is 3.83. It's the lowest value comparing to the other subsumption experiments. This can be explained as all the properties of the Kohonen classes are taken into account during the subsumption. The average number of Kohonen class projected per formal concept is 7.41. That means that this projection is not able to reflect the

	Subsp0.0	Subsp0.1	Subsp0.2	Subsp0.3	Cos.	Subsp-Cos
Nb of fc involved in a projection (InvP)	180	108	20	9	44	41
Nb of selected pairs (kc,fc) (Nbp)	360	75	20	31	29	29
Average projection level (APL)	3.83	3.26	1.5	1.33	4.20	4.51
Average Recall (AR)	0.05	0.28	0.87	0.99	0.52	0.25
Average Precision (AP)	0.26	0.59	0.35	0.17	0.81	0.86
Average Group Recall (AGR)	0.06	0.28	0.75	0.99	0.49	0.35
Average Group Precision (AGP)	0.73	0.63	0.70	0.61	0.82	0.94
Average Agglomeration Recall (AAR)	0.15	0.40	0.78	0.99	0.61	0.54
Average Agglomeration Precision (AAP)	0.88	0.72	0.84	0.69	0.83	0.88

Table 1: Indicators for projection

sharpness of the description of the Kohonen classes: either the Kohonen classes share the same properties with a different weight, either the formal concept subsumant share only a few properties with the Kohonen classes. The recall average is very low (0.05) and the precision average is low (0.26) too.

Subsumption from 0.1 to 0.3: At 0.1, the Kohonen class are projeted at a higher level, due to the higher treshold. This treshold is the best one for the subsumption approach. AR is low and AP is acceptable. At 0.2, and moreover at 0.3, we observe a degradation of most of the indicators. AR is getting higher (for 0.2 and then 0.3) while AP is decreasing. Formal concepts onto which a Kohonen class is projected include the individuals of the Kohonen class but they are much more wider.

Cosine: We observe that the Kohonen classes are projected onto a low level in the lattice. This indicates that the formal concepts have a description in intension relatively important (The average is 4.72 properties per formal concept). However, AR is good and AP very good. The average projection level (APL) is higher than for all the pure subsumption experiments. Moreover, the number of different formal concept (InvP) which are involved in the projection (44) is very near from the inital number of Kohonen projected classes (48). This indicates a very good discriminating power of the cosine projection.

Subsumption and Cosine: We observe that the Kohonen classes are projected onto the lowest level in the lattice compared to previous approaches. These values are close to these of cosine. Formal concepts (involved in a projection) have a description in intension better than for cosine (5.19 properties per formal concept). However, AR is worst than for cosine and AP is a little better. We will therefore prefer the use of the cosine distance for the projection.

Recall and Precision on the extension of the groups of Kohonen classes

Let K_i for i=1,n be the n Kohonen classes projected over one formal concept of the lattice C. Group Recall (GR for group recall) and Group precision (GP) are defined as follows:

$$GP = \frac{|\bigcup_{i=1}^{n}\{K_i\} \cap T|}{|T|}, GR = \frac{|\bigcup_{i=1}^{n}\{K_i\} \cap T|}{|\bigcup_{i=1}^{n}\{K_i\}|}$$

In case of subsumption, whatever is the treshold, the averages of GR (AGR) is globally stable compared to AR and the average of GP(AGP) is higher than AP. Best values are obtained for 0.2 (AGR=0.75 ,AGP=0.70) and for 0.3. (AGR=0.99 ,AGP=0.61). This is normal as we grouped the extensions of Kohonen classes. On the opposite, AGR and AGP for cosine and subsumption-cosine are stable compared to AR and AP.

Recall and Precision for agglomerations

Agglomeration recall AggR and precision AggP are calculated upon the same basis as GR and GP but for each formal concept which is associated to Kohonen classes when coming back up to the top of the hierarchy. The tendency is the same as for group recall/precision.

Recall and Precision: conclusions Cosine and Subsumption-cosine are stable on projection, grouping and agglomerating (up in the lattice) with good values for R and P. Subsumption 0.2 and 0.3 has lower values for projection but higher for groups or agglomeration. However, recall and precision measures should be considered as unperfect measures for estimating the intrisic quality of a projection method. Indeed, they do not permit to explicitly highlight which is the nature of the properties of the Kohonen classes that are dropped out by the projection process, either important ones or marginal ones. It appears clearly that the risk of eliminating important properties during the projection process is minimized by the cosine method as compared to the pure subsumption methods. The former one is indeed the only one that takes directly into account the property weights in the projection process.

4.2.2 Connexity and agglomerations

The hierarchical structure obtained by agglomeration can be evaluated through its global characterics like its number of nodes, the number of different levels, the balance between the levels. Another important criteria is the connexity : how closely related are the Kohonen classes which are grouped or agglomerated onto the same formal concept. The idea is that if agglomeration enables us to relate Kohonen classes which are topologically close, then we suppose that it will be easier for an expert to comment the hierarchical structure. In the same way, if the connexity is verified through all the agglomeration process, the property of the formal concepts belonging to the

agglomeration could then be automatically used to generate explanation of different order of magnitude on the Kohonen maps. The analysis the table 2 leads to the following conclusions:

- In term of number of nodes and looking at the graph of the final hierarchy, subsumption at 0.2, cosine and Subsumption-cosine are structures that could be interpreted by an expert. Subsumption at 0.0 and at 0.1 are too complex. Subsumption at 0.3 is too simple.
- At a first glance, subsumption at 0.2 looks better than cosine or subsumption-cosine as all the Kohonen classes grouped or agglomerated onto a formal concept are closely related. Subsumption-cosine has lower performance than cosine.
- However, the hierarchy corresponding to cosine is well balanced and the three formal concepts which do not correspond to closely related Kohonen classes are distributed on two level.

5 Conclusion

Cosine seems to be the best criteria to project the Kohonen classes onto formal concepts of the lattice. The agglomeration process coming back up to the top of the lattice enable us to simplify the hierarchy and to obtain a well-balanced hierarchical structure.

We plan to go on further experiments to see how data dependent this approach can be. However, we have shown that numerical methods can be positively enhanced by symbolic methods.We also plan to involve a user in order to validate the idea that the hierarchical structure is useful for analysing the data and to validate if the criteria which leads us to choose cosine hierarchy among all the others is good. Finally, the notion of connexity could also be refined: in table 2, a formal concept fc_1 is considered as "connexe" if any the Kohonen class grouped or agglomerated onto this formal concept is closely related to at least one other Kohonen class associated to fc_1. This is a very strong condition that could be modulated.

Combining both numerical and symbolic approaches can be sucessfully applied to Digital Librairies. It provides the user with an intelligent way of analyzing, of visualizing the data set including some generalisation operations. It also brings some new elements that each method could not separately provide. The first results of our experiment tends to prove that our techniques can be used with strong benefits to highlight properties from the most specific to the most general ones which can be derived from larger area of the Kohonen map. This can be performed considering the conservation of the topographic coherence of the merged classes when coming back up to the top of the lattice.

References

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

[Ganter *et al.*, 1986] B. Ganter, J. Stahl, and R. Wille. Conceptual measurement and many-valued contexts. In W. Gaul and M. Schader, editors, *Classifi cation as a Tool of Research*, pages 169–176, North-Holland, Amsterdam, 1986.

[Kohonen, 1984] T. Kohonen. *Self-Organization and Associative Memory*. Springer Verlag, 2 edition, 1984.

[Lamirel *et al.*, 2000] J.C. Lamirel, J. Ducloy, and G. Oster. Adaptive browsing for information discovery in an iconographic context. In *Proceedings of RIAO*, 2000.

heuristics	Subsp 0.0	Subsp 0.1	Subsp 0.2	Subsp 0.3	Cos	Subsp-Cos
level 0	**126**/138	**76**/80	**14**/14	**6**/6	**36**/36	**38**/39
level 1	**33**/41	**22**/27	**5**/5	**2**/2	**11**/12	**8**/11
level 2	**17**/21	**7**/10	**1**/1	**1**/1	**5**/7	**2**/4
level 3	**9**/11	**4**/5	-	-	**3**/3	**3**/3
level 4	**4**/6	**1**/1	-	-	**1**/1	**1**/1
level 5	**1**/1	**1**/1	-	-	-	-
level 6	**0**/1	**1**/1	-	-	-	-
level 7	**1**/1	**1**/1	-	-	-	-

Table 2: Connexity among the different step of agglomeration. The bold values represent the number of connexe classes for each projection method and for each level of agglomeration