Combining Evidence with Logic and Preferences to Learn Relations from Structured Few Sparse Textual Data

Nadia Zerida and Khaldoun Zreik
Paragraph Laboratory EA 384
Paris 8 University
2, rue de la liberté 93000
Saint Denis - France
firstname.lastname@univ-paris8.fr

In the literature, it is commonly believed that learning from few data problem can be resolved by using classifiers that consider interclass relationships. In this work, we will adopt this point of view in learning from few sparse textual data, essentially, by considering the sparseness of the latter as a good support for inducing theories about generalization. Therefore, we opt for an inductive approach based on combining: evidence-based analysis of patterns, logic and preferences. More precisely, we are interested in supervised learning of biomedical articles by exploiting a multi-scale hybrid description and constrained pattern-based data mining techniques. Unlike existing works, we will highlight the relevance of the absence/weakness of patterns and we will associate to their absence a semantic value compared to their presence. The main characteristic of our approach is that of considering local and global contexts, which connect textual data by introducing regret ratio measures and generalized exclusive patterns in order to avoid a crisp effect between the absence and presence of patterns. Experimental results show the effectiveness of our approach.

Structural Constraints, Analysis of Textual Patterns, Learning from Few Examples, Interclass Relationships, Sparse Textual Data.

1. INTRODUCTION

Learning from few sparse textual data represents a new important need of emergent applications on the web, especially in scientific monitoring domain and many of fields in security. On the other hand, the exploitation of patterns covering few examples, i.e. small disjuncts, has attracted interest of several researches [1, 4, 10]. However, most of the latter were interested in studying the effect on the quality of classification results. By exploring the state of the art, the original article dealing with the problem of learning from small disjuncts [10] provides a comprehensive explanation of why and how small are error prune.

Furthermore, there are few attempts to integrate these patterns to build classifiers [2]; in which the authors integrate only the absent patterns, i.e. negative patterns. This is likely due to: (i) the search space which is much huger by considering these patterns than only frequent ones; (ii) the negative effect of these patterns on inductive learning results as presented in the literature [4].

In this paper, we show how to circumvent this algorithmic difficulty by mining emerging patterns [6]. Those patterns have a frequency whose strongly varies between the class values and can be mined by powerful data mining techniques. We show how the absence or the weak frequency of emerging patterns are highly interesting for classification, thanks to the particularity of multi-scale textual description which provides a minimum of noise and a maximum of preserving consistency. Furthermore, unlike the approaches which consider multi-classes classification problem as a generalization of independent binary classifications; where one class is labelled as a positive class and other classes are grouped in one negative class. We keep

details about the frequency of patterns in all classes and we consider it primordial for our evidence-based analysis., wich compromise the absence/presence of patterns in all classes.

In other words, our analysis is primarily based on facts, wich are represented by the patterns. Therefore, these facts become an evidence if they are relevant to the assumptions, either positively or negatively. Consequently, by referring to Cluxton's principle [5], the parameters of evidence become clearly relevance and plausibility. The strength of this analysis lies in its flexibility; it allows us to switch easily between quantitative to qualitative domain of patterns, by preserving a semantic coherence of the induced theories.

The main stages of induction protocol are those of: (i) designing a hybrid multi-scale description of biomedical articles; (ii) characterizing classes by using an adequate data mining technique; (iii) evidence-based analyzing of emerging patterns strongly connected to multi-scale textual description; (iv) inducing an exclusion-inclusion based classification. In addition, we consider that::

1 The semantic of an article is given by patterns found in top levels. The sparseness of these patterns reflects the force of the auto discrimination of an article.
2 The absence of a pattern is a pattern in itself (default logic for non-monotonic logic by Reiter); it reflects the strength of self-discrimination of a pattern according to other ones.
3 The patterns are semantically related even if they are exclusive with each other because they share the same context. To better understand, it suffices to analyze the reading process of an article; readers generally start by looking at the plan of the article by excluding non-interesting information. They repeated the process on all scales until the inclusion of interesting information.
4 Finally, less informative or redundant patterns are used to improve precision of the latter.

1.1 Contributions

Our contributions in this paper are as follows. We define a multi-scale hybrid description, which combined with linguistic knowledge, are pertinent to characterize related classes of biomedical articles, respectively: reviews, clinical and research. Then, we propose a new method of classification founded on the absence of patterns. The most originality of this work is to associate multi-scale description and machine learning to constrained (local) patterns based techniques. On one hand, assumptions giving priority to evidence related to the structure of documents are considered. On the other hand, assumptions related to classification task are emerging, such as the use of partial or total absence of patterns under certain constraints, which can be useful to build new analogies for text classification.

The robustness of our approach is to combine information from different sources using the experience learned from the past. Its principal strength lies in supporting changes caused by parameters as: contradictions, which generate noise in the coherence of decisions taken by the classifier, affirmations; in order to enhance consistency of decision-making, and the proposal; to be able to be adapted with new knowledge. In addition, it allows us to generate a minimum of global patterns with a minimum of constraints.

The rest of paper is organized as follows. Section 2 outlines the structural constraints of multi-scale data description and classes. Section 3 proposes an evidence based analysis to post process emerging patterns. Section 4 will show how combining evidence and lexmin-ordering allow us to exploit the absence of patterns preserving local and global context by introducing the notion of regret ratio measures. The detailed process of classification will be given in Section 5. Experimental results will be provided in Section 6. Finally, Section 7 closes the paper and will be projected in future works.

2. STRUCTURAL CONSTRAINTS

2.1 Relational constraints on classes

2.1.1 Context

As many researchers, we strongly believe to necessity of context in order to control data structure which makes it possible to adapt a given task to a particular situation. Therefore, the context is considered at two levels: first, by choosing specific topics as "Brain and Glioblastoma" and "Glioblastoma and prostate"; second, by using inheritance notion; which will be given in multi-scale description sub-section.

2.1.2 Interclass relationships
As mentioned in introduction, in this work we handle three classes of documents; reviews, clinical and research articles.

- **Review article:** is a scientific article in which are analyzed, evaluated, confronted then synthesized information previously published in literature.
- **Clinical article:** reports one or a series of original clinical cases, where observations are limited to significant facts and demonstrations.
- **Research article:** is a personal work done by authors over the state of current knowledge. It forms a systematic review of its purpose, which is to inform readers about a precise topic.

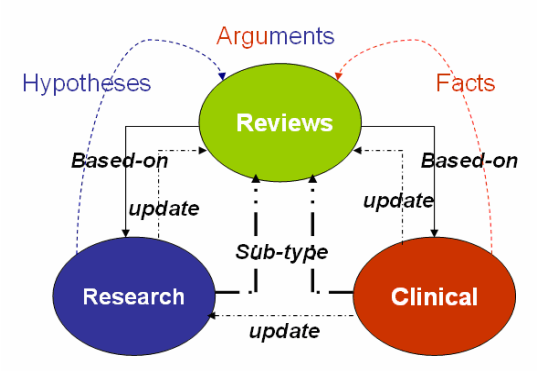

FIGURE 1: Interclass relationships

As shown in Figure 1, there are three types of interclass relationships:

1 "Hierarchical" relationship, i.e. a document is a sub-type or super-type of another one (research and clinical articles are a sub-types of reviews).
2 "Update" relationship, i.e. a document update another one (clinical article updates a research paper).
3 "Based-on" relationship, i.e. a document is based on the work of another one (reviews are based on clinic and research articles).

Furthermore, the transferred information between different classes can be perceived as a set of elements of evidence, whereas, clinical articles provide facts, research articles provide hypotheses and reviews provide arguments.

2.1.3 Overlapping of classes
The classes do not overlap and each document belongs to exactly one class.

2.2 Relational constraints on classes

2.2.1 Multi-scale Document Description
Figure 1 provides an example of global and local regularities wich can be found in the article, body and section scales.

In this example, we remark a progress of information flows across different textual units and logical organizational structure of the article. Then, we note that the 1st sentence, of the 1st paragraph, of the 1st section, i.e. introduction, is very short (it contains 6 words) and it starts by chronic, followed directly by disease, which is the topic of the article. We note also that, the 1st sentence of the last paragraph of the 1st section entitled introduction, begins with the personal pronoun We followed by the verb report, and after a systematic review, which induces the type of article.

In addition, when we analyze the 4th paragraph of the 1st section. The assertion: *There is a little evidence suggests that the author will give his opinion when he is simply a report of results found by [7]*. On the one hand, the presence of *There is* introduced the sub-topic on *QOL measures*. On the other hand, the absence of a personal pronoun *we* affirmed that it is about a report of work and not a position on what is referred, the reference comes at the end of the 1st comma unit to affirm that it is a report.

The following patterns formulate the type of the article, the general topic, one sub-topic and the population:

FIGURE 2: Multi-scale description of articles. The text is marked by sections, paragraphs, sentences and comma units. The sections between the introduction and conclusion have been removed in order to indicate most interesting elements in this example.

Description 1:

((Level=Body)∧ ((Section:position=1)∧ (Section:Title=Introduction))∧ (Paragraph:position=last) ∧ ((Sentence: position=1)∧ (Sentence:word:position=1)=We)∧ (Sentence:voice=active)∧ (Sentence:word:verb=report)))

Description 2:

((Level=Body)∧ ((Section:position=1)∧ (Section:Title=Introduction))∧ ((Sentence:position=1)∧ ((Sentence:mot:position=1)=adjectif)∧ (Sentence:length= 6)∧ (Sentence:voice=active))

Description 3:

((Level=Body)∧ ((Section:position=1)∧ (Section:Title=Introduction))∧ (Scale=paragraph)∧ ((Sen tence:position=last)∧ (Sentence:length=short)∧ (Sentence:BeginWith(adverb))=Consequently) ∧ ((Sentence:Contains(advreb)=about): followedBy (of):followedBy(with)))

Description 4:

((Section:position=1∧ Section:Title=Introduction)∧ (Sentence:position=1∧ Sentence:length=6∧ Sentence:voice=active∧ (Sentence:contains(adv)∧ Sentence:contains(of)∧ Sentence:contains(PluralForm)))

The given example shows that descriptors are organized according to a certain hierarchy that represents the logical and cognitive model of the article. Thus, the words will not have the same role, nor the same importance, according to their place in different textual units. First, the semantic of article is given by the top level (at the content table level, i.e. plan of article) and it is catched early in the article, the more precise information is found in the lowest levels. Therefore, the lower levels provide more facts, and the semantic projection is given by the highest levels.

Plan of article. This set of descriptors reflects the textual organization of the article. The global unit of the article is preserved in order to present logical structure. The titles of sections constitute the plan descriptors at the article level.

Multi-scale metrics. Another set of descriptors contains the length of the various textual units: the length of the body of text (expressed by the number of sections). Sections (expressed by the number of sub-sections or paragraphs). Sub-sections (expressed by the number of paragraphs or sentences). Paragraphs (expressed by the number of sentences). Sentences (expressed by the number of comma units). We also took as a descriptor the length of the title and sub-titles of the article (expressed by the number of words).

Linguistic descriptors. These set of descriptors are used in~\cite{Lucas}. They have been improved, adapted, and organized in classes in~\cite{Zerida}. This set of descriptors is based primarily on two concepts: inheritance (i.e. a level can inherit information from other levels.) and salience. A descriptor is more significant when it occurs at the first and the last unit included in each level. For example, at paragraph level, coordination's class as moreover, is more significant when it occur at the first sentence.

2.2.2 Data sparseness

The nature of multi-scale descriptors or their association's do that they are never totally absents at once. They vary from the densest to relatively absent, the most frequent are found in the finest level of the document structure. This bias would make it impossible to retrieve less frequent descriptors that characterize documents at a higher level. Therefore, to avoid the dominance of the lowest levels, we need to evaluate a degree of dominance of each textual unit in relation to others.

2.2.3 Independence of levels

Each level of the hierarchy is handled separately in order to better profile the content. We assume that each level is designed in order to achieve a specific sub-goal issued by its adjacent higher level; the success of any level in solving its generalized content is independent of that of any other level. This property allowed us to start from any level to build classifier.

2.2.4 Data progression

The contextual progression is provided by the progress of the global context through the various levels of hierarchy, via the inheritance concept in order to preserve the global coherence.

3. EVIDENCE-BASED POST-PROCESSING OF EPS

The following analysis of Emerging Patterns is purely empirical and without any a priori axiom. The developing reasoning has emerged naturally and intuitively. However, it refers to the structural constraints cited above to check coherence in reasoning. Our evidence parameters are relevance and plausibility, and we constructed consistent structured arguments using a simple reasoning: an argument is the evidence of facts to conclusion via a logical sequence.

Definition 3.1 Emerging patterns are patterns whose frequency strongly varies between the class values. The capture of the contrast between the class i containing the objects Di and the objects belonging to the other classes is measured by the Growth rate.

$$GR_i(X) = \frac{|D|-|D_i|}{|D_i|} \times \frac{F(X,Di)_i}{F(X,D)-F(X,D_i)}$$

We say that X is an emerging pattern from D\Di in Di if GRi(X)≥rho, with rho> 1.

Let a selected subset of patterns as collected in Table 1. In the first column, we have the characterized class versus the other ones. In the second column, we have the value of the emerging pattern, the plan of article. In the third column, we have the Growth Rate value of each pattern. Finally, we have the support of a pattern p_i in the class$_i$, given by $F(p_i, class_i)$.

Class	Emerging Patterns (ρ =2)	GR	F(pi, classi)		
			Clinic	Review	Research
Clinic vs. Reviews& Research	EP1= {Footnotes, Acknowledgement} {Abstract, Introduction, Material&Methods, Results}	2.7451	88.23%	00.00%	100%
Reviews vs. Research & Clinic	EP2= {Conclusion, abstract}	10.4615	05.88%	61.53%	05.88%
Research vs. Reviews & Clinic	EP3= {Discussion, Footnotes} {Abstract, Introduction, Material&Methods, Results}	2.0000	82.35%	00.00%	100%

TABLE 1: Excerpt of patterns

We note that the pattern EP1 is totally absent in reviews articles, and totally present in research articles, but it is in 88.23% of clinical ones. In addition, the pattern EP3 is present in the totality of research articles; also, it is present in 82.35% of clinical articles. On the other hand, this pattern is absent in the totality of reviews. That implies negative characterization of review articles, thereby we formulate:

$$review \Leftrightarrow NOT(Non(review))$$

The second interesting remark is, by combining, i.e. sequencing the three patterns EP1, EP2 and EP3, we could infer the three classes of articles by using a negative characterization, i.e., a pattern type:

$$X \rightarrow \neg class_i$$

We call this type of patterns, exclusive patterns. Thus, we found that the combination (or sequence) of a set of these exclusive patterns provide to conclude on classes. These first analyses conduct us to determine essential points to retain for further work:

1. The Growth Rate only is not able to provide interesting results given by the use of patterns frequencies (10.4615 for EP2 and less than 3 for EP1 and EP3).
2. The affirmation property of the reviews exclusion by the EP3 and proposed by the EP1 give an argumentative aspect to our analysis.
3. The absence of a pattern in one class can be interesting but it is always compared to its presence in the other classes. Consequently, finding a compromise between absence/presence will be useful.

Based on previous observations, we are automatically projected in evidence theory with clear parameters as belief and plausibility measures, and more particularly: necessity and plausibility measures.

Property 3.1 An Event A is necessary therefore certain ($N(A)=1$), if only if its complement is impossible ($\Pi(\neg A) = 0$). Moreover, $N(A) \leq \Pi(A)$ for any event A. In particular, if one has even moderate certainty on the realization of A ($N(A) \neq 0$), then A is quite possible ($\Pi(A) = 1$).

Indeed, we have assigned not null values of believes to the well consistent rules, the ones with each other. Therefore, referring to property 3.1, we deduced that the absence of a pattern p_i in a class c_i makes $\Pi(\neg p_i) = 0$ that makes its necessity $N(p_i) = 1$.

We can also formulate this in terms of confidence of a rule as follows:

$$Confidence(p_i \rightarrow \neg class_i) = \frac{F(p_i \cup class_i)}{F(p_i)} = 1$$

4. COMBINING LOGIC AND LEXMIN-ORDERING

4.1 Exclusive Patterns

Definition 4.1 Exclusive patterns are patterns; which are completely absent in at least one class.

Definition 4.2 The relevance of an exclusive absence of a pattern (in one class) than its presence in (several classes) is formulated by:

$$(\forall p_j, \exists! c_i, F_{ij}=0) \Rightarrow (p_j \rightarrow \neg c_i)$$

Definition 4.3 The relevance of an exclusive rare pattern (in one class) than its absence in (several classes) is formulated by:

$$(\forall p_j, \exists! c_i, F_{ij} \gtrsim 0, \forall k \neq i, F_{ij} = 0) \Rightarrow (p_j \rightarrow c_i)$$

In addition, the relevance of an exclusive frequent pattern (in one class) than its absence (in several classes) is given by:

$$(\forall p_j, \exists! c_i, F_{ij} >> 0, \forall k \neq i, F_{ij} = 0) \Rightarrow (p_j \rightarrow \neg c_i)$$

4.2 Regret ratio to quantify weak patterns

Definition 4.4 Weak patterns are patterns with at least one not null low frequency in all classes (see Table 1).

As different patterns have different levels of efficiency, it is necessary to weight them according to their global performance, expressed by their frequency. We propose a method to provide the best sequence of these patterns based on assessing two measures, called regret ratio from patterns frequency matrix.

Definition 4.5 The regret growth ratio, rgr, quantifies the importance of a pattern in a class according to the others classes (i.e. the global context). It is based on BFD measure: Best Frequency in Data.

$$rgr(i, j) = \frac{F_{ij} - BFD(i)}{BFD(i)}, where \qquad (1)$$

$$BFD(i) = \max_{j \in \{1..k\}} F_{ij}$$

Definition 4.6 The regret frequency ratio, rfr, measures how much a pattern is important in a given class according to the other patterns (i.e. the local context). It is based on BFP measure: Best Frequency of Pattern.

$$rfr(i, j) = \frac{F_{ij} - BFP(j)}{BFP(j)}, where \qquad (2)$$

$$BFP(i) = \max_{i \in \{1..N\}} F_{ij}$$

The regret ratio measures quantifies the loss caused by the absence of a pattern in a class compared to its presence in other classes and the absence/presence compared to other

patterns of the same class. Thus, it provides a semantic value to the local context and global context of patterns.

Definition 4.7 From the measures (1) and (2), we define a new matrix called, Matrix of regrets MR, where for each pattern *i* and a class *j*, we assign the couple (rgr(i,j),rfr(i,j)).

$$MR = ((rgr(i, j), rfr(i, j)))_{1 \le i \le N, 1 \le j \le K}$$

4.3 Lexicographic ordering (π_{lex})

We use the measures given in the matrix MR to compare patterns; afterwards, we apply lexmin-ordering algorithm.

Consequently, we prefer a pattern *j* over pattern *j'* in a *class i*, when the couple of regret scores given in MR of *j* are lexicographically preferred over the ones of *j'*, more formally,

$$(rgr(i, j), rfr(i, j)\ \pi_{lex}\ (rgr(i', j), rfr(i', j))$$

By applying the difference principle of Rawls [9], we ordered weak patterns following a lexicographic order starting from the lowest. This means that a pattern is preferable to another if the situation of lower is better and the situations of the lowest for two patterns are identical. Therefore, we compared the second lowest to decide between them and so on.

This choice comes from the fact that a simple combination with Maxmin/Minmax will not work because the normalization of frequencies by regret ratio will cause the existence of at least one pattern with a null value. Therefore, only the patterns which are equals to '0' will be considered. In order to overcome the disadvantage of the Maxmin/Minmax combination, we adopt the natural extension of the latter: Lexmin/Lexmax [8], which have proved its performance in [3].

Example 4.1 From Table 2, classically, by using contingency table, we note that p_7 contributes more in the discrimination of the review class, p_9 contribute more in the discrimination of both research and clinic classes, and the pattern p_8 contributes less in this discrimination.

	P7	P8	P9
Clinic	90.12	51.34	10.25
Review	05.33	05.03	40.33
Research	05.33	45.18	50.12

TABLE 2: Example of weak patterns

We conclude that p_9 and p_8 are not informative for the discrimination. However, if we analyze this example basing on exclusion principle, we conclude that p_8 and p_9 participate respectively in the exclusion of reviews and clinic classes. By calculating rgr rfr matrix, we obtain the sequence (p_9, p_8, p_7), and we deduce that:

$$p_9 \rightarrow \neg review,\ p_8 \rightarrow \neg clinic,\ p_7 \rightarrow research$$

5. GENERALIZATION BY TOTAL AND PARTIAL EXCLUSION

5.1 Generalization principle

Let *C* a set of classes and "$\models$" a semantic consequence relation. We define total exclusion and partial exclusion as follows:

Definition 5.1 Total exclusion is a sequence of local decisions that excludes all classes at once, in order to include a single class.

$$\neg c_1 \wedge \neg c_2 \wedge \neg c_3 \ldots \wedge \neg c_{n-1} \models c_n$$

Definition 5.2 Partial exclusion is a sequence of local decisions that excludes a subset of classes.

$$\neg c_1 \wedge \neg c_2 \wedge \neg c_3 .. \wedge \neg c_{i-1} \models \neg c_i \vee \neg c_{i+1} \vee c_{i+2} \vee c_{i+3} .. \vee c_n$$

The left part of "$\models$" the two generalized rules defined above represents the premise of the rule, and the right part represents the conclusion of the rule. The strength of this type of rule is that the premise is connected with the conclusion by using a *disjunction*. Therefore, the strength of these rules lies in the fact to preserve consistency by generating disjunctive rules in extensible way. The latter is provided by the most interesting properties of the semantic relation "$\models$", which are:

Property 5.1 (Extension property)
If we extend a class C_i by new exclusive patterns, the classes which are related semantically whith this latter remain in relation of the extended class.

Property 5.2 (Semantic aspect)
If each pattern of a class C_i is a pattern of a class C_j ; which is a semantic consequence of C_i, then any pattern of superclass C_k of C_i is also a pattern of C_i, i.e. if $C_i \models C_j$ and $C_i \subseteq C_k$, then $C_k \models C_j$.

5.2 Qualitative argued measure to preserve coherence

In this stage of analysis, the question that we must to ask is how to combine extracted exclusions? For example, if a pattern p_1 propose to exclude a class c_1 and a pattern p_2 excluded the same class c_1, it is quite natural to think to associate a weight to this decision because it represents an affirmation of the decision taken by the pattern p_1. The same reasoning is valid in the sense that a pattern p_3 contradicted the proposed exclusion of p_1 and asserted by p_2, so it is also quite natural to think to quantify this contradiction.

Therefore, in order to preserve coherence of exclusions, we formulate the combination of generalized rules by using preferences, which can illuminate decision making by these rules to obtain classification by regrouping all or a part of generalized rules in equivalence classes. These ones will be ordered with a complete or partial way, according to preferences.

Let P be a set of patterns, for each pattern p_i, we define a preference function $P(p_1, p_2)$ to assign a degree of preference of a pattern p_1 to a pattern for the exclusion criterion of classes.

Affirmation. In general, an affirmation represents a strict preference; it corresponds to the existence of clear and positive reasons that justify a significant preference in favour of p_1 or p_2. It is formalized as follows:

$$P(p_1, p_2) = 1, if : P(p_1) - P(p_2) >>> 0$$

P is asymmetric and irreflexive, s.t:

- Irreflexive: $\forall p \in P, Non[pPp]$
- Asysmmetric: $\forall p_1, p_2 \in P, p_1 P p_2 \Rightarrow Non[p_2 P p_1]$

Infirmation. An information is represented by a low preference and it corresponds to the existence of clear and reasons that invalidate a strict preference (i.e. affirmation) in favour of p_1 or p_2. However, these reasons are insufficient to infer either a strict preference in favour of one pattern or indifference against p_1 and p_2. It is given by:

$$Q(p_1, p_2) \approx 0, if : Q(p_1) - Q(p_2) > 0$$

Q is asymmetric relation (irreflexive):

- Irreflexive : $\forall p \in P, Non[pQp]$
- Asymmetric : $\forall p_1, p_2 \in P, p_1 Q p_2 \Rightarrow Non[p_2 Q p_1]$

In our case, we simply onsider that an infirmation is a negative affirmation, more specifically, a strict and negative affirmation, it is given by:

$$Q(p_1, p_2) = -P(p_1, p_2) = -1$$

Proposal. A proposal is a situation of incomparability, if it is not an affirmation and nor contradiction, it is a proposal. It corresponds to the absence of clear and positive reasons to justify one of the two previous situations. It is normalized as follows:

$$R(p_1, p_2) = 0, if : R(p_1) - R(p_2) < 0$$

R is also an asymmetric relation (irreflexive):

- Irreflexive : $\forall p_1, p_2 \in P, Non[pRp]$
- Asymmetric : $\forall p_1, p_2 \in P, p_1 R p_2 \Rightarrow p_2 R p_1$

Since the three relations *(P,Q,R)* define a set of exclusif patterns, we say they contain a relational system of preferences of an expert *Z* on *P* if they are:

1 consistent with the definitions and properties above.
2 exhaustive: for any pair of patterns, at least one is verified.
3 mutually exclusive for any pair of patterns, two separate relations are never verified.

Condition 3 states that if *H1* and *H2* are two distincts relations among the three :

- $p_1 H1 p_2 \wedge p_2 H1 p_1$ is excluded, except in exceptional cases that it's not restrictive to exclude.
- $p_1 H1 p_2 \wedge p_1 H2 p_2$ is excluded, It is restrictive because $p_1 P p_2 \vee p_1 P p_2$ characterize consistent situations.

5.3 Scores of generalized rules

Algorithm 1 calculates the score associated with each exclusive generalized rule for each class. It takes as input the exclusive patterns and preferences on the patterns in different levels, it updates the list of scores of different patterns, either by incrementing affirmations, decreasing contradictions or adding proposals.

Algorithm 1: upDateScoresList (i,scoresList)

Input: Emerging Patterns {Eps}, Levels{Li}, ϕ on levels
Output: upDatedscoresList
1: scoresList = [proposal, affirmation, contradiction]
2: Assign(pj ,scoresList)
3: **for each** class i **do**
4: **for each** level Li **do**
5: **for each** pattern j **do**
6: **if** proposition **then**
7: count=1
8: **endif**
9: **if** affirmation OR contradiction **then**
10: count++
11: **endif**
12: addCount(scoresList,count)
13: **return** upDatedscoresList
14: **endfor**
15: **endfor**
16: **endfor**

5.4 Classify a new document

Classifying a new document *d** by using the exclusion-inclusion based classifier is given by a simplified version in the Algorithm 2.

Algorithm 2: Exclusion-inclusion based classifier

Input: new document d*, multiscale descriptors *D*
Output: classify a new document d*
1: Preprocess(*d**,*D*)
2: scoresList = [proposal, assertion, opposition]
3: **for all** *class* i **do**
4: **if** IsExlusif(pattern) **then**
5: Apply ExclusiveAbsentPatterns(*d**,i)
6: **else**
7: Apply *rgr_rfr*_DecisionSequences(*d**,,i)
8: **endif**
9: upDateScoresList(i,scoresList)
10: **return** Decision = ((*d**,i),scoresList)
11: **endfor**

As showed in Figure 3, the exclusion-inclusion based classifier proceeds in two stages. First, it tries to predict the class of the document by applying exclusion because it is the least expensive in terms of computing. Second, if it is not possible to include a global decision using only exclusive patterns, it applies the sequence of patterns calculated from of regret measures. Therefore, a pattern may, affirm or oppose a decision given earlier by previous patterns. The first step exploits the qualitative argued measure of exclusive patterns and the second one exploits the sequence returned by applying Lexmin ordering on weak patterns.

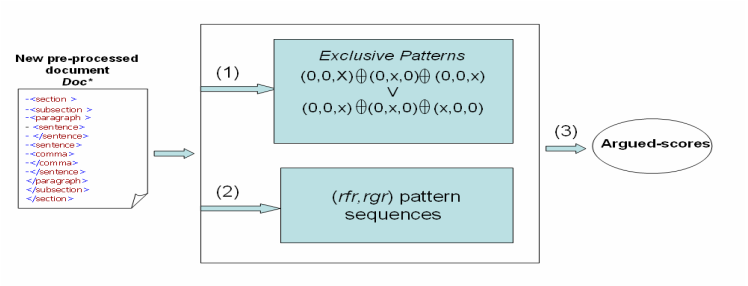

FIGURE 3: Global decision for a new document

The utility of argued score is to assess and improve the quality of the current generalization in order to decide whether it would be necessary to continue in the same level or it is necessary to change the level and use patterns that are more specific. Therefore, argued score takes into account the structure of a pattern, its evolution through time and strength of supporting evidence, so it clearly identifies low, forgotten and negative evidence. Consequently, for each class, we will prepare a list of patterns ordered by their level of preference, and we calculate the weight of a decision that is associated with a new document d* for a class ci by using the following formula:

$$score(d^*, c_i) = \sum Affirmations - \sum \Pr oposls - \sum Infirmations$$

6. EXPERIMENTAL RESULTS

6.1 Pre-processing step

The segmentation into different textual units (body, parts, sections, etc.) is an important step of our approach, because we based all multi-scale linguistic and structural descriptors on it. We pre-processed documents into manageable representations by using XML and Xpath technologies. Table 3 contains statistics on data.

Level	# Transactions				# Items
	Clinic	Reviews	Research	Total	
Article level : -Body -Plan -Article Title -Section Title	37	33	37	107	48
Sections	216	148	229	593	114
Sub-Sections	300	164	340	804	141
Paragraphs	1.346	1.050	1.214	3.610	219
Sentences	3.404	3.032	3.960	10.396	168
Comma-units	7.044	10.972	9.053	27.069	201

TABLE 3: Statistics on data sets

The constrained based mining technique that we use is method that we use is a set-based method, where each pattern is represented by a set of boolean attributes described in attribute-value formalism and stored in a table that contains items and transactions. Items are categorical and take values in a finite and discrete set. However, multi-scale descriptors represent two types of attributes: (i) symbolic descriptors: for the plan and linguistic descriptors; (ii) digital descriptors: for metric ones. Consequently, for the first type, the transformation of textual data into transactional one is almost done automatically. In contrast, for the second type, a grouping of these values is necessary to obtain boolean data. In order to establish good choice, we adopted a discretization approach, which is based on a priori knowledge of the expert in the field namely, the linguist. This has helped us to define intervals to be considered relevant. Also, we sought to minimize the presence of too small or too large intervals, which may influence the results of extracted patterns. Finally, thanks to property inheritance, the attributes are discretized independently from each other.

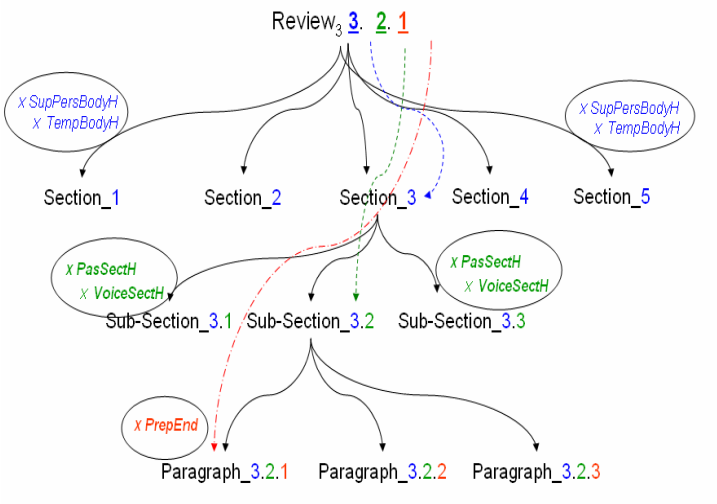

FIGURE 4: The 1st paragraph of 2nd Sub-Section of the 3th Section of the review article number 3 is described by: the presence of a preposition at the end of paragraph, the inherited voice and Past from Section level, and the inherited temporal and superpersonnal from body level.

On the other hand, salience and inheritances constraints are managed intrinsically in textual pre-processing step. It consists in annotating each level by their specific descriptors by respecting the property of salience and inherited information from the highest levels. The example in Figure 4 illustrates the semantic of the following transaction:

Textual unit	Items
Review3 3.2.1	PrepEnd VoiceSectH PastSectH SupPersBodyH TempBodyH

Finally, different transactional tables are generated for each level and several data sets are built.

6.2 Classification step

In order to highlight our classification method based on combining evidence, logic and preferences, we compared it with ten other classifiers including SVM, ID3, Meta Logit Boost,

Naive Bayes, Meta RandomComittee, Bayes Waode, Meta Logit Boost, Misc OSDL Boos, NNge, Rules Jrip and Rules Ridor.

The performance of each classifier is evaluated by calculating standards metrics including micro and macro-average, precision, recall and F-measure values.

- Micro average precision is calculated by dividing the number of correctly classified articles to the number of all articles, which were classified. Micro average recall is the number of the correctly classified articles divided by the total number of articles. Micro F-measure is calculated from micro precision and micro recall. However, in our case, the three measures are equals and we consider micro measures like the percent of correctly classified instances.

- Macro average measures are calculated as average of corresponding measures for each class. Then, the macro precision is the average of precisions of all classes, macro average recall is the average of all classes and macro average F-measure is average of F-measures.

As shown in Table 4, a series of experiments is performed with different classifiers. We note that the worst obtained results in terms of macro and micro average measures are those of Misc ODSL Boost classifier with Macro average (precision = 05.41%. recall = 40.00%. F-measure = 32.26%) and Micro average =34.58%. and NNge classifier with Macro average (precision = 27.03%. recall = 40.00%. F-measure = 32.26%) and Micro average measure =46.73%. In addition, with ID3, SimpleLogistic, NaiveBayesSimple, Rules Jrip and Rules Ridor classifiers. We observe an increase of the Macro average precision; which attains 48.65% for Id3 and SimpleLogistic, and 64.86% for Rules Jrip and Rules Ridor. However, in terms of recall, these latter provide a comparable score to NNge and Misc ODSL Boost.

The best competitive results are obtained by Meta Logit Boost, SVM, Meta RandomComittee, Bayes WAODE and exclusion-inclusion classifier. Although Meta logit Boost attains the maximal macro average precision value 78.68%, followed by Meta RandomComittee and Bayes WAODE with 75.68%. They remain less efficient in terms of recall compared with SVM and our exclusion-inclusion based classifier. In contrast, if we consider only Macro average F-measure, we note that SVM classifier attains the top value with 77.14%, followed by our approach with 71.16%.

We think that the results obtained by ID3 can be improved by considering patterns appearing in very few texts to build the tree. For SVM, although it is commonly known that its strength lies in its independence from the number of objects to classify. Thus it is not sensitive to data density and can deal with sparse data. The only explanation that we can give to this result is :

- either the classifier has been influenced by the variation in the number of positive and negative examples,
- or the order in which classes were obtained influenced the results.

Classifier	Macro average			Micro average
	Precision	Recall	F-measure	
Meta RandomComittee	75.68%	54.90%	63.64%	58.88%
ID3	48.65%	69.23%	57.14%	57.01%
SimpleLogistic	48.65%	72.00%	58.06%	61.68%
NaiveBayesSimple	56.13%	57.17%	56.64%	56.07%
Bayes WAODE	75.68%	63.64%	69.14%	65.42%
Meta Logit Boost	78.38%	64.44%	70.73%	63.55%
Misc OSDL Boos	05.41%	40.00%	9.52%	34.58%
NNge	27.03%	40.00%	32.26%	46.73%
Rules Jrip	64.86%	42.86%	51.61%	47.66%
Rules Ridor	64.86%	40.68%	50.00%	41.12%
Exclusion-Inclusion	72.36%	70.00%	71.16%	56.07%
SVM	72.97%	81.82%	77.14%	64.49%

TABLE 4: Performances on different classifiers

Rank	*Patterns*
1	{Article_Title_length ∈ [35,195]}
2	{Body_length < 6 }
3	{Sections_length ∈]5,10] }
4	{Footnotes, Aknowledgement} {Abstract, Introduction, Material&Methods, Results}
5	{conclusion, abstract}
6	{Discussion, Footnotes} {Abstract, Introduction, Material&Methods, Results}
7	Body_level : {TEMP_Start, SUPPERS_End}
8	Body_level : {MOD_End, SUPPERS_End}
9	Body_level : {SUPPERS_Start, SUPPERS_End}
10	Section_level: {NEG_Start,DET_End}{DET_Start,ANAPH_End, SUPPERS_Start}

TABLE 5: Top 10 used patterns

7. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a new hybrid method for classification based on combining different scales of evidence. This method has proven efficiency of proposed combination, but can be improved to be more generic for an adaptation into other domain applications. It is also clearly for us that our method can be easily transposed in several domains application, namely: genomic, named entity extraction, multilingualism, images and video analysis, etc. This is the next step of our research.

8. KNOWLEDGEMENTS

We want to thank those who have known to give us valuable advice to achieve this work, in particular: Dr. Steven Simske, a Distinguished Technologist at HP Labs and Prof. Patrice Enjalbert.

REFERENCES.

[1] K. M. Ali, and M.J. Pazzani, " Reducing the small disjuncts problem by learning probabilistic concept descriptions", Computational Learning Theory and Natural Learning Systems, MIT Press , Cambridge, Massachusetts, 1995, pp. 183-199.
[2] M-L. Antonie, and O. R. Zaiane, "An Associative Classifier based on Positive and Negative Rules", 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-04), Paris, France, 2004, pp .64-69.
[3] N. Ben-Amor, S. Benferhat and Z. Elouedi, "Qualitative classification and evaluation in possibilistic decision trees", IEEE, volume 2, ,2004, pp. 653- 657.
[4]D.R. Carvalho, and A.A. Freitas," Evaluating Six Candidate Solutions for the Small-Disjunct Problem and Choosing the Best Solution via Meta-Learning", Artificial Intelligence Review, Kluwer Academic Publishers, USA, 2005, pp. 61-98.
[5] D. Cluxton, S.G. Eick, and J. Yun, "Hypothesis visualization", In Proceedings of the IEEE Symposium on Information Visualization, IEEE, Washington, DC, USA : IEEE Computer Society., pp. 215.4.
[6] G. Dong, and J. Li "Efficient mining of emerging patterns: discovering trends and differences", proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD'99), ACM Press, , San Diego, CA, 1999, pp. 43-52.
[7] N. Lucas, B. Crémilleux, and L. Turmel, "Signalling well-written academic articles in an English corpus by text-mining techniques", UCREL technical papers, 16 (Special issue Proceedings Corpus Linguistics 2003), p.465-474.
[8] H. Moulin, Axioms of Cooperative Decision Making, Cambridge University Press, 1991.
[9] J. Rawls, The theory of Justice, Cambridge: Harvard University Press, 1971.
[10] G.M. Weiss, "Learning with Rare Cases and Small Disjuncts", In Proceeding of Twelfth International Conference on Machine Learning, Morgan Kaufmann, California, USA, 1995, pp. 558-565.
[11] N. Zerida, N. Lucas, B. Crémilleux, "Combining linguistic and structural descriptors for mining biomedical literature", ACM Symposium on Document Engineering, ACM, Amsterdam, The Netherlands, 2006, p.62-64.