Emerging Overlapping Clusters for Characterizing the Stage of Liver Fibrosis

Nicolas Durand and Arnaud Soulet

GREYC - CNRS UMR 6072
Université de Caen Basse-Normandie
F-14032 Caen Cédex - France
{ndurand,asoulet}@info.unicaen.fr

Abstract. This paper presents a data mining effort to discover emerging overlapping clusters with respect to the stage of fibrosis from a medical database collected at the Chiba University Hospital, Japan. Such clusters, described by examinations on patients, may lead to relevant factors for estimating the stage of fibrosis. We used and combined two recent methods: a soft-clustering approach (ECCLAT) and a method to soundly and completely mine patterns under a constraint specified by the user (MUSIC) in order to characterize classes. This new system produces a set of emerging overlapping clusters of patients, based on closed emerging patterns, which summarizes the data set by taking into account the different classes. Results point out the role of some medical examinations.

1 Introduction

The medical data stored during the patients' diseases like, for instance hepatitis data collected at Chiba University Hospital (Japan), represent an important source of knowledge. The use of relevant and efficient methods to explore such large data sets is not easy. Statistics are often used to validate suspected models and today we are facing to a new challenge: how may new models be discovered? By extracting from large amounts of data non trivial "nuggets" of information, Knowledge Discovery in Databases (KDD) is a semi-automatic way which may help the user for this work. We are interested in discovering the structure and relationships within data. For instance, in medicine, it is interesting to find clusters (i.e. groups) of patients having similar characteristics (or close to each other) while patients in different clusters are dissimilar. The general meaning of clustering is partitioning the set of examples (i.e., patients) in clusters [1]. More precisely, it is called hard clustering. Another way is *soft-clustering*. It allows to produce clusters where examples may be present in several clusters: we get overlapping clusters [2, 5]. We claim that the overlapping is an important point in order to capture different aspects of the data. Most of the works address hard clustering and there are few methods in soft-clustering. Moreover, it is even fewer to obtain an explanation of the clusters from the attributes describing the examples. In this paper, we focus on a method to characterize large data sets by discovering overlapping clusters, especially with categorical attributes. This method produces lists of attributes (here, medical examination results) to

describe each discovered cluster of patients. Furthermore, an overlapping among the attributes is allowed. This is precious when an attribute takes part in the characterization of several stages.

Hepatitis B and C are virus infections that affect the liver of the patient. These infections are important because they have a potential risk of developing liver cirrhosis or hepatocarcinoma. Indicators of such diseases is fibrosis of hepatocyte. For instance, liver cirrhosis is characterized as the terminal stage of liver fibrosis. The detailed mechanism of disease progression is unknown yet. The contribution of this paper to the ECML/PKDD 2005 discovery challenge is to better estimate the stage of liver fibrosis from laboratory examinations, this stage is at present determined by biopsy. The idea is to substitute laboratory examinations for biopsy because biopsy is invasive to patients. We would like also to show the potential impact of the discovery of emerging overlapping clusters (see Section 3) in domains like hepatitis.

In this paper, we propose a new approach to produce a set of characterizing clusters, named *emerging overlapping clusters*, composed of patients with a slight overlapping. This approach mixes global and local patterns. Its key idea is to combine *emerging patterns* [4] and *soft-clustering*. We use a soft-clustering method to build a global model from emerging patterns which describe local contrasts between two (or more) classes. This process is different from classical clustering methods, and relies on two efficient methods called MUSIC [9] (for Mining with User-SpecifIed Constraint) and ECCLAT [5] (for Extraction of Clusters from Concepts LATtice). In the experiments of these discovery challenge data, clusters gather patients and are described by examinations (and their results) performed on patients. The combinations of examinations describing such clusters may be good factors for estimating the stage of fibrosis (in order to avoid bias, the biopsy features are not used to build clusters).

Patient Id.	Items	Class
P_1	$A\ B\ C$	$\mathcal{D}_1$
P_2	$A\ B\ C$	
P_3	$A\ B\ C$	
P_4	$D\ E$	
P_5	$D\ E\ H$	$\mathcal{D}_2$
P_6	$A\ D\ E\ F\ G\ H$	
P_7	$A\ F\ G\ I$	
P_8	$H\ I$	

Table 1. A transactional database $\mathcal{D}$

In the following discussion, we use the most common terms in KDD: each data record is called a *transaction* and is described by *items*. For a transaction (i.e. a patient), an item has a binary value: present (i.e. the patient has the characteristic depicted by the item) or not. A *pattern* is a set of items (also called itemset). Table 1 presents an example of transactional database. There are 8 patients (denoted $P_1 \dots P_8$) and 9 items denoted $A \dots I$. For example, A denotes an item which is linked to result of the level of lymphocytes. There are two classes associated to the two sub-databases $\mathcal{D}_1$ and $\mathcal{D}_2$. The class $\mathcal{D}_1$ corresponds, for example, to the first stage of the liver fibrosis. A bi-set [8] is

composed of a pattern and a set of transactions (for instance: $(ABC; P_1, P_2, P_3)$). In this paper, we focus on a particular kind of bi-sets stemming from the closed patterns (see Section 2.2) to characterize a phenomenon.

Section 2 presents the essential material which is required to understand the work done in this discovery challenge. Section 3 details our method to produce emerging overlapping clusters of patients. Section 4 gives our work for the data preparation stage. Results (including out-hospital and in-hospital examinations) and discussion are presented in Section 5.

2 Closed emerging patterns and soft-clustering

2.1 Using MUSIC to mine closed emerging patterns

Initially introduced in [4], emerging patterns (EPs) are patterns whose frequency strongly varies between two data sets (i.e. two classes). EPs characterize the classes in a quantitative and qualitative way. Thanks to their capacity to emphasize the distinctions between classes, EPs allow to build classifiers or to propose a help for diagnosis. From an applicative point of view, we can quote various works on the characterization of biochemical properties or medical data [6]. The concept of emerging patterns is related to the notion of frequency. The frequency of a pattern in a data set $\mathcal{D}$, denoted by $\mathcal{F}(X, \mathcal{D})$, is the number of transactions which contain X in $\mathcal{D}$. The *growth rate* of a pattern X from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$ is defined as:

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

Definition 1 (emerging pattern or EP). *Given a threshold* $mingr > 1$*, a pattern* X *is an* emerging pattern *from* $\mathcal{D}\backslash\mathcal{D}_i$ *to* $\mathcal{D}_i$ *if* $GR_i(X) \geq mingr$*.*

The quantitative evaluation of the contrast between classes brought by a pattern is measured by its growth rate. The more $GR_i(X)$ is high, the more X characterizes the data set $\mathcal{D}_i$. Let us give some examples from Table 1. With $mingr = 2$, AB and ABC are EPs from $\mathcal{D}_2$ to $\mathcal{D}_1$. Indeed, $GR_1(AB) = \infty$ and $GR_1(ABC) = \infty$. With this threshold, the pattern A is not an EP for $\mathcal{D}_1$ because $GR_1(A) = 3/2 < 2$.

Now, we introduce the concept of closed pattern. X is a closed pattern if its frequency only decreases when any item is added. For instance, in the Table 1, AB is not closed because we can add the item C without changing its frequency. So, ABC is a closed pattern corresponding to P_1, P_2 and P_3. ABC is the closure of AB. Given an EP, we consider a significant property to quantify the interest of its closure:

Property 1 *The growth rate of a pattern is equal to the growth rate of its closure.*

The proof of this property is in [10]. This property means that a pattern has the same growth rate than its closure (it shows that the closed patterns with their growth rates are a condensed representation of the whole set of EPs with their growth rates [10]). Thereafter, an emerging pattern which is also a closed pattern, is named a *closed emerging pattern* (or a CEP). Even if the closed emerging patterns have been introduced in [10], to the best of our knowledge, this is the first proposition to use CEPs. In our running example, AB and ABC have the same growth rate for $\mathcal{D}_1$, but only ABC is a closed pattern.

The algorithm MUSIC (Mining with a User-SpecifIed Constraint) mines soundly and completely the intervals synthesizing the collection of patterns satisfying a given primitive-based constraint [9]. The set of accepted constraints is very varied and it includes monotonous, convertible and tougher ones. The efficiency of the extraction is ensured by powerful pruning criteria tackling intervals. These pruning criteria are automatically deduced from the constraint given by the user thanks to formal operators.

MUSIC is well adapted to extract closed emerging patterns. Property 1 ensures that the pruning is correct and each right bound of the extracted intervals is a closed pattern whose the growth rate exceeds the threshold $mingr$. Thus, MUSIC enables us to directly mine the closed emerging patterns.

2.2 ECCLAT: a soft-clustering method

ECCLAT (Extraction of Clusters from Concepts LATtice) [5] produces bi-sets from large categorical data sets. These bi-sets represent a set of overlapping clusters described by patterns. The approach used by ECCLAT is quite different from usual clustering techniques. Unlike existing techniques, ECCLAT does not use a global measure of similarity between elements but is based on the discovery and the evaluation of potential clusters coming from the set of frequent closed patterns [7]. Moreover, the number of clusters is not set in advance.

Let us recall that a pattern is frequent if its frequency is at least the frequency threshold (called $minfr$) set by the user. ECCLAT starts from the set of all frequent closed patterns. Indeed, a closed pattern checks an important property for clustering: it gathers a maximal set of items shared by a set of transactions. In other words, this allows to capture the maximum amount of similarity. These two points (the capture of the maximum amount of similarity and the frequency) are the basis of the approach of clusters selection. In practice, $minfr$ can be seen as the minimum number of transactions in a cluster.

ECCLAT evaluates and selects the most interesting clusters by using a cluster evaluation measure. All computations and interpretations are detailed in [5]. The cluster evaluation measure is composed of two criteria: *homogeneity* and *concentration*. With the *homogeneity* value, clusters having many items shared by many transactions are favored (a relevant cluster has to be as homogeneous as possible and should gather "enough" transactions). The *concentration* measure limits an excessive overlapping of transactions between clusters. Finally, the *interestingness* of a cluster is defined as the average of its *homogeneity* and its *concentration*.

ECCLAT uses the *interestingness* to select clusters and to produce a clustering with a slight overlapping between clusters (i.e., a soft-clustering). The overlapping depends on the value of a parameter M corresponding to the minimal number of different transactions between two selected clusters. The algorithm performs as follows. The cluster having the highest *interestingness* is selected. Then as long as there are transactions to classify (i.e. which do not belong to any selected clusters) and some clusters are left, the cluster, having the highest *interestingness* and containing at least M transactions not classified yet, is selected.

The number of clusters is established by the selection process, and is linked to the M value. Let n be the number of transactions, at worst there are $1 + \lfloor \frac{n-minfr}{M} \rfloor$ clusters. In practice, this does not happen. With M=1, the overlapping is "free". The more the value of M increases, the more the overlapping decreases but some transactions may not belong to any cluster (the remaining transactions are grouped in a *trash* cluster).

3 Discovery of emerging overlapping clusters

We have seen in the previous sections the interest of local patterns (i.e., CEPs) to characterize a phenomenon and the use of ECCLAT to provide global views of the data, these views being composed of overlapping clusters.

The frequent closed emerging patterns are very interesting as potential emerging clusters. At first, CEPs characterize classes because their frequency increases significantly from one class to another. Secondly, as they are also closed, CEPs capture the maximum amount of similarity. Besides, Property 1 guarantees that no interesting EP is lost by only mining CEPs. However, the number of CEPs remains huge. Thus, it is necessary to select the most interesting ones in order to build a global model. As ECCLAT picks clusters from the collection of frequent closed patterns, we propose to replace this collection by the CEPs. Thereby, we characterize the data set by discovering emerging overlapping clusters which are frequent closed emerging patterns. Such clusters are called *emerging clusters*. Let us note that an emerging cluster corresponds to a CEP.

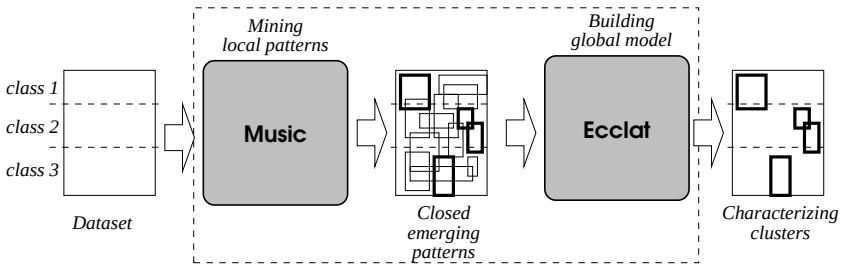

Fig. 1. The process of discovery of emerging overlapping clusters

Figure 1 depicts the whole process in order to discover emerging clusters. At first, MUSIC mines all the frequent CEPs (according to the thresholds $minfr$ and $mingr$) starting from the data set. The output corresponds to the complete collection of frequent CEPs. It is important to say that the property of homogeneity (see Section 2.2) is straightforwardly computed by MUSIC thanks to the generic power of the primitives which are on the core of MUSIC. Secondly, starting from these potential emerging clusters, ECCLAT selects the most significant ones (according to the *interestingness* and M) in order to produce the global model.

Figure 1 also illustrates several points of our method:

- *high selection*: ECCLAT only selects few emerging clusters. It reduces significantly the number of CEPs obtained by MUSIC (see Table 4 in Section 5).
- *almost-pure emerging clusters*: each selected emerging cluster mainly describes one class.
- *soft-clustering*: the intersection of clusters (see at the right of Figure 1, the second row symbolizes the class 2) shows clearly the overlapping. Indeed, a transaction may belong to several clusters. This allows to capture different aspects of the data.

According to us, this discovery process of emerging overlapping clusters has a twofold advantage of benefiting from particular local phenomena and the global aspect of the data set. It builds a global model which concisely emphasizes the main characteristics for each class. It can also be seen as a global and coherent selecting method of interesting local information provided by CEPs.

4 Data preparation

Seven tables have been downloaded from the Web[1]. We give below the main transformations that we performed.

4.1 Overview of tables

Table `patient` contains 771 tuples. There is no missing value. Most of the patients are males (70,69%).

Table `biopsy` contains 694 tuples. There are only 49 missing values for the `Activity` attribute. Let us note that this table has been very improved compared to the ECML/PKDD 2002 Discovery Challenge.

Table `out-hospital_examinations` reports on results of out-hospital examinations for patients. 31,040 tuples were downloaded (the guide to the hepatitis data set indicates 30,243). There are many missing values for `Evaluation`, `Eval_SubCode`, `Condition`, `Comment1` and `Comment2` attributes. There are also 7,314 (23.56%) missing values for the `Exam_Result` attribute, 16,051 (51.71%) for the `Unit` attribute and 19,309 (62.21%) for the `Qualitative_Interpretation` attribute. There are 844 distinct values for the `Name` attribute, ten of them occur more than 500 times. The `Qualitative_Interpretation` attribute has 41

[1] `http://lisp.vse.cz/challenge/ecmlpkdd2005/`

distinct values and, without the help of a physician, we are not able to interpret some of them (e.g. `******`, `10*2`, `< (+)`).

The `in-hospital_examinations` is a large table (1,565,876 tuples). It stores results of in-hospital examinations for patients. The attribute `Name` has 230 distinct values. This table gives the numeric value of an examination result and we will use the table `measurements_in-hospital` (see Section 4.2) to get the lower and upper bounds of these examinations. Let us note that 212 of the 459 examinations have a lower bound and a upper bound equal to 0. These examinations are not here taken in account.

The tables `interferon_therapy` and `hematological_analysis` are not used in this work.

4.2 Resulting files

To discover the relationships between the stage of liver fibrosis and laboratory examinations, transactions are built as follows: each transaction gathers a biopsy and examinations of the patient associated to this biopsy (in fact, we will see below that the obtained files have a single biopsy per patient). The idea is to discover clusters described by examinations (and their results) and which are pure or almost-pure with regard to the stage of the liver fibrosis. Biopsy features are not used during the discovery stage so that the combination of examinations given by a cluster may be a good indicator for estimating the stage of fibrosis.

As out-hospital and in-hospital examinations are not straightly comparable, we construct two data files: file called `bioexaout` for examinations out-hospital and `bioexain` for examinations in-hospital. The tables `out-hospital_examinations` and `in-hospital_examinations` show that a same examination can be performed several times on a same biopsy, sometimes with different results. In this case, we decide to keep only the examination (with its result) which is the closest of the date of the biopsy.

The process to obtain `bioexaout` and `bioexain` is the following. First, we joined the tables `biopsy` and `patient` for biopsy where the `Fibrosis` attribute is known. In this way, an element of a cluster can as well be seen as a patient or a biopsy and clusters might be easier to interpret. We get a temporary table called `biopat` composed of 599 patients.

Secondly, we joined `biopat` with `out-hospital_examinations` to produce `bioexaout` and with `in-hospital_examinations` to produce `bioexain`. During the join, we computed the number of days between the date of the biopsy and the date of the examination. For `bioexaout`, we kept only examinations for which the `Qualitative_Interpretation` attribute is known and can straightly be recoded in `+` or `-` values. More precisely, we grouped values `1+`, `2+`, `3+`, `4+`, `(+)` and `+` in a single value denoted `+` and we gathered values `(-)` and `-` in the value coded `-`, other values are ignored. We obtain 9,956 examinations with values `+` or `-` for `Qualitative_Interpretation` and dealing with the patients of `biopat`. Nevertheless, these examinations correspond only to 342 distinct patients (in other words, some patients of `biopat` have no examinations in `out-hospital_examinations` with an understandable value for `Qualitative_`

Interpretation). For bioexain, there are 1,499,280 examinations without missing values for the Exam_Result attribute and for which the qualitative interpretation can be inferred from measurements_in-hospital. These examinations concern 499 distinct patients. For the Fibrosis attribute, 13 patient has the value F0, 216 have the value F1, 109 have the value F2, 78 have the value F3 and 83 have the value F4.

Final files are obtained by gathering for each patient all his examinations. Let us recall that a patient occurs once and corresponds to a biopsy and in case of several occurrences of an examination for a patient, we keep only the examination which is the closest of the date of the biopsy. Table 2 summarizes the characteristics of bioexaout and bioexain. One examination can lead to two qualitative results on bioexaout: for instance, we will denote the two results for the examination HBE-AB by using HBE-AB+ (positive) and HBE-AB- (negative). On bioexain, three qualitative results can appear for an examination: less than the lower bound, between the lower and the upper bound (normal values), and more than the upper bound. We only focussed on abnormal values because most of the pairs examination / results concern normal values [3]. This can involve noise and hide abnormal situations in the results. For instance, for the examination GLU, two values will be denoted GLU- and GLU+, respectively for less and greater than normal values. An item is a pair examination / result (e.g. GLU-) and the number of items indicated in Table 2 is the number of pairs observed in a file.

	No. of patients	No. of performed examinations (final)	No. of items
bioexaout	342	1,400	42
bioexain	499	14,226	168

Table 2. Characteristics of bioexaout and bioexain

5 Results and discussion

Let us remind that starting from the two prepared data sets, we search for emerging overlapping clusters in order to propose a model to estimate the stage of liver fibrosis (see Section 3). CEPs are extracted from each class of the liver fibrosis.

5.1 Quantitative results

In the following experiments, Tables 3 and 4 give an overview of the results from a quantitative point of view, respectively for bioexaout and bioexain. They provide quantitative results on the union of CEPs extracted from each class. They indicate for a minimal frequency value ($minfr$) and a minimal growth rate value ($mingr$), the number of all the frequent CEPs (seen as candidate clusters). For a M value (the minimal number of distinct patients between clusters), we can observe the number of obtained clusters, the average overlap, and the size of the trash cluster.

The $minfr$ value represents the minimum number of patients in each cluster. The higher $minfr$ is, the more the result is reliable. Nevertheless, $minfr$ has to correspond to the stage distribution, in order to characterize each class ($minfr$ is a relative threshold depending on the data set). To keep a relevant characterization, we fix the $mingr$ value to 3. For example, on `bioexaout` (see Table 3), for $minfr = 2\%$ (7 patients) and $mingr = 3$, we obtain 41 frequent CEPs. With $M = 1$, we get 14 clusters. There are in average 1.11 common patients between each pair of clusters. 244 patients do not belong to a cluster. We note that only a small part of patients belongs to a cluster. The best rate is 36.25% (with $minfr$=0%, $mingr$=3 and M=1). Nevertheless, this low $minfr$ value does not have sense (the obtained clusters are composed of one patient). Considering the weak number of CEPs for each $minfr$ value, we did not continue experiments with `bioexaout`. We concentrated our mining effort on `bioexain`.

$minfr$	$mingr$	No. of CEPs	M	No. of clusters	Avr. overlap	# trash
3 (10)	3	24	1	10	2	265
2 (7)	3	41	1	14	1.11	244
0 (1)	3	180	1	52	0.22	218

Table 3. Quantitative results on `bioexaout`

$minfr$	$mingr$	No. of CEPs	M	No. of clusters	Avr. overlap	# trash
8 (40)	3	106,237	1	269	12.24	17
			20	15	9.43	148
	3.5	57,707	1	273	13.91	37
			20	13	11.15	196
	4	30,481	1	234	15.27	64
			20	11	8.14	214
	5	8,119	1	174	17.86	154
			20	6	12	340
6 (30)	3	287,122	1	310	10.38	13
			20	16	8.29	141
	3.5	176,136	1	314	7.24	20
			20	13	4.64	208
	4	106,024	1	298	8.5	32
			20	13	6.96	211
	5	41,303	1	273	10.15	83
			20	9	5.3	294

Table 4. Quantitative results on `bioexain`

For `bioexain` (see Table 4), we fix $minfr$ to 8% (40 patients) and to 6% (30 patients). We use different $mingr$ values. If $mingr$ is high, we observe a high diminution of the number of frequent CEPs. Nevertheless, the higher $mingr$ is, the stronger the characterization of CEPs is. We choose a tradeoff: $mingr = 3.5$.

With $minfr = 8\%$ and $mingr = 3.5$, we obtain 57,707 closed emerging patterns. With $M = 1$, the number of clusters is equal to 273 (the lowest value). In order to reduce this number, we increase M to 20. Finally, 13 clusters are selected. The average overlap is equal to 11.15 patients. Let us remark that 196 patients do not belong to a cluster. The next section analyzes more precisely this set of clusters.

5.2 Interpretation of the results

Due to the space limitation[2], we only detail the 13 clusters previously discovered on `bioexain` with $minfr = 8\%$ and $mingr = 3.5$ (see Table 5). In this table, the clusters are sorted according to the interestingness measure of ECCLAT. For each cluster, we give the CEPs describing this cluster, the number of corresponding patients ($\mathcal{F}$), the growth rate (GR) of the stage associated to its cluster (we will describe below how the stage attribution is done). Let us note that 30 items (17.85%) are used to describe all the clusters.

cluster	CEP	$\mathcal{F}$	GR	stage
1	ALB- GOT+ GPT+ ZTT+ ALP+ F-ALB- G-GTP+ TTT+ CHE- D-BIL+ G.GL+ I-BIL+ LDH+ T-BIL+	40	6.12	F4
2	ALB- GOT+ GPT+ ZTT+ ALP+ CRE- F-ALB- G-GTP+ TTT+ G.GL+ LDH+ T-BA+	43	3.61	F4
3	GOT+ GPT+ ZTT+ F-ALB- G-GTP+ LAP+ TTT+ AMY+ CHE- D-BIL+ G.GL+ I-BIL+ T-BIL+ IG-G+ PT+	40	4.53	F4
4	GOT+ GPT+ ZTT+ ALP+ F-A2.GL+ F-ALB- TTT+ D-BIL+ G.GL+ I-BIL+ T-BIL+	46	3.85	F4
5	ALB- GOT+ GPT+ ZTT+ F-ALB- F-B.GL+ TTT+ CHE- D-BIL+ G.GL+ T-BIL+ PT+	53	3.84	F4
6	ALB- GOT+ GPT+ ZTT+ ALP+ F-ALB- G-GTP+ TTT+ G.GL+ PT+	77	3.56 2.02	F4 F3
7	GOT+ GPT+ LDH- TP- F-ALB- LAP+ TTT+ CHE- D-BIL+ G.GL+ T-BIL+	52	3.67	F4
8	GOT+ GPT+ ZTT+ G-GTP+ TTT+ D-BIL+ I-BIL+ LDH+ T-BIL+	70	3.54	F4
9	ALP- F-A/G+ GOT+ GPT+ LDH- TTT+ G.GL+ HBD-	42	3.62	F2
10	ALB- GOT+ GPT+ ZTT+ F-ALB- CHE- G.GL+ F-A/G-	89	3.56 2.47	F4 F3
11	ALB- F-A1.GL- F-A/G+ GOT+ GPT+ TG+	48	3.93	F1
12	F-A/G+ GOT+ GPT+ I-BIL+ T-BIL+ F-CHO-	48	5.34	F0
13	F-A1.GL- GOT+ GPT+ ALB+	42	3.93	F0

Table 5. Results on `bioexain` ($minfr$=8%, $mingr$=3.5, M=20)

In order to discover examinations associated to a stage of fibrosis, it is necessary to label the obtained clusters with the stages. This is done thanks to the growth rates. Table 6 presents the growth rates of each stage for the 13 clusters. The numbering of the clusters is the same as the one in Table 5. Let us provide an example with the cluster number 1. For this cluster, the highest growth rate corresponds to the `F4` stage. Moreover, 6.12 is rather a high value for a growth rate. So, we consider that the cluster 1 characterizes `F4`. We perform a similar process with the other clusters.

[2] More results are available for readers, just contact the authors.

Let us remark that the growth rate is linked to the purity (see Section 2.1). If the cluster is pure (only one stage F_j), then the growth rate for F_j (GR_j) is infinite and the growth rate for the other stages is equal to 0.

cluster	F0	F1	F2	F3	F4
1	0	0.11	0.52	1.8	**6.12**
2	0	0.4	0.83	1.05	**3.61**
3	0	0.23	0.52	1.8	**4.53**
4	1.69	0.41	0.65	0.81	**3.85**
5	0	0.13	1.06	1.75	**3.84**
6	0	0.29	0.54	***2.02***	**3.56**
7	0	0.2	0.97	1.62	**3.67**
8	0.54	0.42	0.53	1.35	**3.54**
9	0	0.41	**3.62**	0.73	0.83
10	0	0.2	0.56	***2.47***	**3.56**
11	0.79	**3.93**	0.72	0.36	0
12	**5.34**	1.31	0.52	0.49	1
13	**3.93**	1.19	0.85	0.57	0.83

Table 6. Growth rates (for each stage) on the results of `bioexain`

We remark that most of clusters characterize `F4` (9 clusters). For `F3`, we choose to characterize it with the clusters 6 and 10, for `F2` with the cluster 9, for `F1` with the cluster 11 and for `F0` with the clusters 12 and 13. Let us note that the clusters 1, 3, 11, 12 and 13 have good growth rate values.

By observing Table 5, we can express hypotheses about the links between some examinations and the stages of liver fibrosis.

First, we note that some items are present in many clusters and for different stages. We can think that `GOT+` (13 clusters), `GPT+` (13 clusters), `TTT+` (9 clusters), `G.GL+` (9 clusters), `T-BIL+` (7 clusters) and `I-BIL+` (5 clusters) are not significant to estimate the stage of liver fibrosis.

Second, we observe that the item `ZTT+` appears in 8 clusters characterizing `F3` and `F4`, whose 6 clusters only for `F4`. It is the same case for `F-ALB-`. `PT+` is present in 3 clusters (`F3` and `F4`), whose 2 only for `F4`. `ALP+` appears in 4 clusters (3 for `F4`). `LDH+` is only present for `F4` (3 clusters). Let us note that all these items are exclusively present in clusters characterizing `F3` and `F4`. Moreover, `ZTT+` is associated with `F-ALB-` in 7 clusters. We also remark that `PT+` and `ALP+` are always present with `ZTT+` and `F-ALB-`. So, we can think that these results of examinations are linked to the most severe stages of liver fibrosis. These items or their associations {`ZTT+`, `F-ALB-`, `ALP+`} and {`ZTT+`, `F-ALB-`, `PT+`} could be used to estimate the stages `F3` and `F4`. Let us note that it is difficult to characterize only `F3`. Maybe, `F3` is very linked to `F4`, and these stages might be merged.

Concerning the other stages, we note that `HBD-` is only present for `F2` (cluster 9). The item `F-A1.GL-` seems to be associated to the first stages (`F0` and `F1`). The items `F-CHO-`, `ALB+` and `TG+` only occur respectively for `F0` and `F1`. Nevertheless, they appear in one cluster each. We can remark that the severe stages are easier to characterize than the first stages of the liver fibrosis.

Furthermore, the item `F-A/G-` (cluster 10) seems to be associated to severe stages (`F3`, `F4`), while `F-A/G+` (clusters 9, 11, 12) seems to correspond to the

other stages (F0, F1, F2). This could be an interesting way to explore. Let us remark that we do not have the description of this examination.

6 Conclusion

We have presented a new method to characterize classes by discovering emerging overlapping clusters which are based on closed emerging patterns. This method combines the advantages of a local characterization of classes brought by the emerging patterns with a global model of data performed by a soft-clustering approach. We use it to search for factors estimating the stage of the liver fibrosis from hepatitis data. These factors are combinations of examinations measured on patients.

We have focussed essentially on in-hospital examination data. This work suggests an interesting role of some examinations. ZTT+, F-ALB-, ALP+, and PT+ seem to be associated to severe stages of the liver fibrosis. We have noticed that the first stages are more difficult to characterize than the severe stages. Nevertheless, F-A/G seems to have a lower value than the lower bound for the severe stages, and a higher value than the upper bound for the first stages. It may be interesting to perform more investigations on this examination. Finally, we have remarked some examinations which seem to be insignificant regarding to the stage of liver fibrosis.

References

[1] P. Berkhin. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[2] G. Cleuziou, L. Martin, and C. Vrain. PoBOC: an Overlapping Clustering Algorithm, Application to Rule-Based Classification and Textual Data. In *ECAI'04*, pages 440–444, Valencia, Spain, August 2004.

[3] B. Crémilleux and N. Durand. Search for Factors Estimating the Stage of Liver Fibrosis Based on the Discovery of Meaningful Clusters. In *the PKDD 2002 Discovery Challenge on Hepatitis Data*, Helsinki, Finland, August 2002.

[4] G. Dong and J. Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *ACM SIGKDD'99*, pages 43–52, San Diego, USA, August 1999.

[5] N. Durand and B. Crémilleux. ECCLAT: a New Approach of Clusters Discovery in Categorical Data. In *ES'02*, pages 177–190, Cambridge, UK, December 2002.

[6] J. Li and L. Wong. Emerging patterns and gene expression data. In *Genome Informatics 12*, pages 3–13, 2001.

[7] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient Mining of Association Rules Using Closed Itemset Lattices. *Information Systems*, 24(1):25–46, 1999.

[8] C. Robardet. *Contribution à la classification non supervisée : proposition d'une méthode de bi-partitionnement.* PhD thesis, Université de Lyon 1, 2002.

[9] A. Soulet and B. Crémilleux. An Efficient Framework for Mining Flexible Constraints. In *PAKDD'05*, pages 661–670, Hanoi, Vietnam, May 2005.

[10] A. Soulet, B. Crémilleux, and F. Rioult. *Knowledge Discovery in Inductive Databases: KDID 2004*, volume 3377 of *LNCS*, chapter Condensed representation of EPs and patterns quantified by frequency-based measures, pages 173–190. Springer, 2005.