# Search for Factors Estimating the Stage of Liver Fibrosis Based on the Discovery of Meaningful Clusters

Bruno Crémilleux[1] and Nicolas Durand[2]

[1] GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
F-14032 Caen Cédex France
Bruno.Cremilleux@info.unicaen.fr

[2]France Telecom R&D
42 rue des coutures
F-14066 Caen Cédex 4, France
nicola.durand@francetelecom.com

**Abstract.** This work presents a data mining effort to discover pure or almost-pure 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 use a method, suitable for categorical data, which is able to produce a set of clusters composed of patients with a minimum overlapping or a slight overlapping to catch all the similarities between patients. Results point out the role of some examinations.

**Keywords:** hepatitis, factors of stage of liver fibrosis, cluster, approximate clustering, frequent closed itemsets.

## 1 Introduction

Nowadays we have important medical data stored during the patients' diseases like, for instance, hepatitis data collected at Chiba University Hospital (Japan). 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 we are facing today 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 groups are dissimilar, or to find groups of similar medical features. From KDD techniques point of view, in this paper, we focus on a method to discover meaningful clusters from large data sets, especially with categorical features. We will briefly see in Section 2 that this approach is quite different from usual clustering techniques.

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 2002 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 clusters (see Section 2) in domains like hepatitis.

In this paper, we propose the use of an efficient method called Ecclat [4] (for Extraction of Clusters from Concepts LATtice) to produce a set of clusters composed of patients with a minimum overlapping ("approximate clustering") or a slight overlapping to catch all the similarities between patients. The behavior of the method is parametrized by the user. We will see in Section 2 that such clusters are a subset of concepts from the frequent closed itemsets lattice and are selected according to an evaluation measure. Patients data (and, more generally, examples at hand) are complex and include categorical attributes. In these experiments, clusters gather patients and are described by examinations (and their results) performed on patients (blood test and urinalysis). Biopsy features are not used to build clusters. Selected clusters are then ranked on the biopsy report about progress of fibrosis in order to discover pure (or almost-pure) clusters with respect to the stage of fibrosis. Combinations of examinations characterizing such clusters may be good factors for estimating the stage of fibrosis. Let us note that we have already applied a quite similar approach for searching prognostic factors in patients with supradiaphragmatic early stage Hodgkin's disease. Results brought out some parameters for which classical statistic methods confirmed that they were interesting [5].

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

## 2 Discovery of meaningful clusters

### 2.1 Context and related work

Let us recall that the general meaning of clustering is decomposing or partitioning examples into groups so that the examples in one group are similar to each other and are as different as possible from the examples in other groups. The main methods [10] are those based on an attempt to find the optimal partition into a specified number of clusters (e. g., the standard $K$-means method) and those based on a hierarchical attempt to discover cluster structure (e.g., the centroid-based agglomerative hierarchical clustering).

Usual criterion functions yield satisfactory results for numeric attributes but are not appropriate when examples include categorical attributes. It is not easy

to define fair distances between categorical attributes [3][12]. In medicine, it is common to have also categorical data and some methods have been developed to handle such data [7, 9]. Recent works in Knowledge Discovery in Databases (KDD) revisit this question. More precisely, two main families of clustering methods based on association rules exist [1]. A family of methods [15] consists in grouping examples into clusters in order to minimize an intra-cluster and an inter-cluster costs, but it is not easy to derive a characterization of each cluster. [8] presents a method of association rules hypergraph k-partitioning. A clustering of attributes (i.e. features) is obtained, but transactions are not straightforwardly ranked into clusters.

To cope with these problems, the next section briefly shows the contribution of conceptual classification methods [13] and KDD's results (with frequent closed itemsets) to return a first selection of clusters. Then, in Section 2.3 we briefly presents a method [4] based on a cluster evaluation measure to build the set of the most interesting clusters gathering similar examples.

### 2.2 Frequent closed itemsets

In the following, we use the most common terms in KDD: *transaction* (instead of *example*) and *item* a pair attribute / value (e. g., `stage of fibrosis = F3`). For a transaction (e.g. a patient), an item has a binary value: `present` (i.e. the patient has the characteristic depicted by the item) or `not`. An *itemset* is a set of items. A transaction $t$ *supports* an itemset $X$ if and only if $X \subseteq t$. $\mathbf{r}$ is a multi-set of transactions (e.g., Table 1 is constituted of 8 transactions, each one identified by its Id, and there are 9 items denoted $A \ldots I$). $|\mathbf{r}|$ (where as usual $|\ldots|$ denotes the cardinality of a set) is the number of transactions.

| Id | Items |
|---|---|
| 1 | $A\ B\ C$ |
| 2 | $A\ B\ C$ |
| 3 | $A\ B\ C$ |
| 4 | $D\ E$ |
| 5 | $D\ E\ \ H$ |
| 6 | $A\ \ D\ E\ F\ G\ H$ |
| 7 | $A\ \ F\ G\ \ I$ |
| 8 | $H\ I$ |

**Table 1.** A transactional database

Even if they do not produce directly a clustering, conceptual classification methods [13] bring out relevant clusters of examples described by categorical attributes. Indeed, these methods create a hierarchy of *concepts*, generally represented by a lattice [6]. Every concept can be seen as a cluster which is a couple $(T, I)$ composed of a set of transactions and an itemset. A key point is that $T$ is the largest set of transactions described by the items found in $I$, and symmetrically, $I$ is the largest set of common items of the transactions supporting $I$. Let us illustrate this idea with the example provided in Table 1. For instance, the

couple composed of transactions 1, 2 and 3 on one side and the items $ABC$[1] on the other side is a concept of the lattice whereas there is no couple composed of transactions 1, 2 (since transaction 3 shares the same items as transactions 1 and 2).

The idea of maximally extending the sets is on the core to highlight meaningful clusters. Indeed, for a group of transactions, we prefer to simply produce the single itemset which is composed of the maximal number of items shared by the group. The key step is to capture the maximum amount of similarity among the data.

This notion is linked to that of a *closed itemset* in the KDD's framework. A closed itemset is a maximal set of items (with respect to set inclusion) shared by a set of transactions. In Table 1, $ABC$ is a closed itemset whereas $AB$ is not a closed itemset since we can add item $C$ to all the transactions supporting $AB$.

Let us note $\mathcal{F}(X)$ the *frequency* of $X$ that is the number of transactions which support $X$. An itemset $X$ is *frequent* if its frequency is at least the frequency threshold $minfr$ fixed by the user. Note that we use here an absolute frequency (a number of examples $\leq |\mathbf{r}|$) instead of the relative frequency $\mathcal{F}(X)/|\mathbf{r}|$ in $[0, 1]$. The frequency is fundamental to extract reliable clusters. It allows to take into account the "importance" (in term of "weight") of a candidate cluster and forget clusters which do not rely on sound relationships within data. A cluster with too few transactions would not be kept by a user. In Table 1, with $minfr = 2$, $DE$ is frequent (its frequency is 3) and $DEF$ is not frequent (its frequency is 1 since only transaction 6 contains $DEF$). From large databases, there are efficient algorithms [2][14] to compute frequent closed itemsets. We have developed such a software, which, moreover, gives the associated transactions.

These two points (the capture of the maximum amount of similarity - i.e. closed itemsets - and the notion of frequency) are the minimal properties required for candidate clusters. After this first selection, we use a cluster evaluation measure (that we present below) to select a set of meaningful clusters.

### 2.3 ECCLAT: Extraction of Clusters from Concepts LATtice

#### Cluster evaluation measure

For the following, $X$ denotes a frequent closed itemset, $T$ the set of transactions associated to $X$ (i.e. transactions supporting $X$) and $L$ the set of the frequent closed itemsets.

A relevant cluster has to be as homogeneous as possible and should gather "enough" transactions. Translated into the usual clustering framework, it means that we have to maximize an intra-cluster similarity (called here *homogeneity*) and minimize an inter-clusters similarity. We use a *concentration* measure to limit the overlapping of transactions between clusters (a relevant cluster should concentrate some transactions).

For homogeneity, we want to favor clusters having many items shared by many transactions. Homogeneity of a cluster $X$ is computed from its size (i.e.

[1] Note that we use a string notation for sets.

its number of items), $\mathcal{F}(X)$ and a divergence measure. The *divergence* is the number of items not in $X$, for each transaction of $T$.

$$homogeneity(X) = \frac{\mathcal{F}(X) \times |X|}{divergence(X) + (\mathcal{F}(X) \times |X|)}$$

where $divergence(X) = \sum_{t \in T} |f(t) - X|$.

We have $0 \leq homogeneity(X) \leq 1$. If a cluster is pure (i.e. $\forall t \in T \quad f(t) = X$), its divergence is equal to 0, and its homogeneity equals 1. The more a cluster supports transactions with items not belonging to $X$, the more its homogeneity leads to 0. Let us remark that the homogeneity of a cluster $X$ depends only on $X$ and can be computed simultaneously to $X$.

For concentration, we want to favor clusters having transactions appearing the least in the whole set of clusters. Concentration limits the overlapping of transactions between selected clusters. Concentration of a cluster $X$ is defined by taking into account the number of clusters where each transaction appears.

$$concentration(X) = \frac{1}{\mathcal{F}(X)} \times \sum_{t \in T} \frac{1}{\mathcal{F}'(t)}$$

where $\mathcal{F}'(t)$ is the number of clusters where $t$ occurs (i.e. absolute frequency of $t$ in $L$).

We have $0 \leq concentration(X) \leq 1$. If all transactions of $T$ occur only in $X$, then $concentration(X) = 1$. The more frequent the transactions of $T$ in the whole set of clusters, the more $concentration(X)$ leads to 0.

Finally, we define the *score* of a cluster as the average of its homogeneity and concentration. We have $0 \leq score(X) \leq 1$.

$$score(X) = \frac{homogeneity(X) + concentration(X)}{2}$$

Let us give a short example: in Table 1, we have $homogeneity(DE) = (3 \times 2)/((0 + 1 + 4) + (3 \times 2)) = 0.545$: transaction 5 has an item which diverges (i.e. does not belong to the closed $DE$) and four items diverge for transaction 6. $concentration(DE) = 1/3 \times (1/1 + 1/3 + 1/5) = 0.511$. Transaction 4 only supports the closed itemset $DE$ while transaction 5 supports three closed itemsets and transaction 6 supports five closed itemsets. Finally, $score(DE) = (0.545 + 0.511)/2 = 0.528$.

The idea is to select clusters with high scores and the next paragraph presents an algorithm for this task.

**Clusters procedure: selection algorithm**

Ecclat [4] uses the score defined above to select clusters from the frequent closed itemsets lattice. It has the originality to produce a clustering with a minimum overlapping between clusters (that we call "*approximate clustering*") or a set of clusters with a slight overlapping. This functionality depends on the value of a parameter called $M$. $M$ is an integer corresponding to a number of transactions that a new selected cluster must classify. With $M = 1$, we assure

to classify all transactions in at least one cluster (except if $minfr$ is very high). Nevertheless, a slight overlapping between clusters may appear. $M$ should be set near 1 if we are interested in discovering meaningful clusters. The more the value of $M$ increases, the more the overlapping decreases but some transactions may not belong to any cluster. We refer to these unclustered transactions as *trash* (i.e. remaining transactions are grouped in a trash cluster).

The sketch of the algorithm is the following. At first, the score of each cluster of $L$ is computed. The cluster having the highest score is selected. Then as long as there are transactions to classify (i.e. which do not belong to any selected cluster) and clusters remain, we select the cluster having the highest score and containing at least $M$ transactions not yet classified.

The results of the method are linked to the value of $M$. In Table 1, with $M = 1$, all transactions are classified in four clusters ($ABC$, $DE$, $AHG$ and $I$). We get a slight overlapping: transaction 6 is ranked in $DE$ and $AFG$ and transaction 7 belongs to $AFG$ and $I$. Intuitively, observing transactions 6 and 7 (see Table 1), there is no sound reason to classify them in one cluster or the other. Note that item $A$ appears in two clusters. With $M = 2$, we get clusters $ABC$, $DE$ and $I$. There is no overlapping and we obtain a partition. With $M = 3$, only two clusters are selected ($ABC$ and $DE$) and a trash cluster is built with transactions 7 and 8.

## 3 Data preparation

The six tables available on the web (`http://lisp.vse.cz/challenge/ecmlpkdd2002/`) have been loaded using the relational database management system (`Mysql 3.23.49`). We noticed that some attributes were missing for table `biopsy` (i.e., delimiters between attributes are missing, which is different from missing values). For example, the patient of `MID`#727 and `Exam_Date` 1990-12-28 has 6 attributes whereas 8 are expected. Nevertheless, the type of attributes seem to indicate that, when two attributes are missing, it is a matter of the last two. We use this table with this assumption.

One advantage of using a relational database is to provide an overview of the data as we will see below. Such a preliminary inspection based on SQL queries helps for the understanding of the required data transformations, i.e., to prepare the data for the data mining task. Let us note that this idea can be carried on with contingency tables which show the number of tuples (i.e. transactions or examples) for each value combination of two or more variables that constitute the table (application in medical data on patients suffering from collagen diseases is given in [16]). We give below the main transformations that we performed.

### 3.1 Overview of tables

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

The table `biopsy` contains 960 tuples. There are two missing values for the `Exam_Date` attribute, four for the `Facility` attribute, 833 for the `Fibrosis` attribute and 841 for the `Activity` attribute. This large number of missing values

on Fibrosis unfortunately leads to a loss of data to study the relationships between the stage of liver fibrosis and laboratory examinations. 136 patients had more than one biopsy. Two tuples have identical values, except for Fibrosis and Activity attributes which are unknown for one tuple (the patient having ID#251 and the Exam_Date value of 1999-10-13). Surprisingly, there is one hepatitis A (the patient having ID#103).

The table out-hospital_examinations reports on results of out-hospital examinations for patients. 31,040 tuples were downloaded (the guide to the hepatitis dataset indicates 30,243). There are many missing values for Condition, Comment1, Comment2, Evaluation and Eval_SubCode attributes. There are also 7,314 (23.56%) missing values for the Exam_Result attribute, 16,051 (51.71%) for the Unit attribute and 19,315 (62.23%) 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 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 (1,565,877 tuples). It stores results of in-hospital examinations for patients. There are two missing values for the Exam_No attribute, two for the Exam_Name attribute and 204,424 (13.05%) for the Exam_Result attribute. 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 3.2) to get the lower and upper bounds of these examinations. We remark that the table measurements_in -hospital contains more potential examinations than those really performed in-hospital.

### 3.2 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. For the sake of clarity, we deleted three patients (ID#179, ID#597 and ID#930) for which two or three biopsy have been done (two biopsy for a same

patient can have different Fibrosis values). 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 also removed the patient of ID#923 because its Fibrosis value (F3-4) is ambiguous. We get a temporary table called biopat composed of 119 patients. For the Fibrosis attribute, 1 patient has the value F0, 49 have the value F1, 32 have the value F2, 24 have the value F3 and 13 have the value F4.

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 1,122 examinations with values + or - for Qualitative_Interpretation and dealing with the patients of biopat. Nevertheless, these examinations correspond only to 59 distinct patients (in other words, some patients of biopat have not examinations in out-hospital_examinations with an understandable value for Qualitative_Interpretation). For bioexain, there are 243,653 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 118 distinct patients.

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. We kept only features on examinations to better highlight the role of examinations. 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 upper bounds, more than the upper bound. For instance, for the examination GLU, these three values will be denoted respectively GLU-, GLU= and GLU+. An item is a pair examination / result (e.g., HBE-AB-) 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 | No. of distinct examinations | No. of items |
|---|---|---|---|---|
| bioexaout | 59 | 1122 | 11 | 17 |
| bioexain | 118 | 243,653 | 172 | 308 |

**Table 2.** Characteristics of bioexaout and bioexain

## 4 Results and discussion

The results reported below are achieved by a prototype running on a PC with 900 MB of memory and a 1,7 GHz P4 processor. We start by giving an overview of the obtained clusters.

About `bioexaout`, Table 3 shows the number of frequent closed itemsets, the number of selected clusters, the number of transactions in the trash cluster (0 means there is no trash cluster) and the sizes of clusters according to $minfr$ and $M$ (see Section 2.3). Ratio is the percentage of selected clusters among the frequent closed itemsets. MIS means MInimal Size of clusters (i.e. the size of the smallest cluster), MAS: MAximal Size of clusters (i.e. the size of the largest cluster) and AVS: AVerage Size of clusters. The size of a cluster is its number of items. Table 4 presents these results on `bioexain`.

| $minfr$ | No. of frequent closed itemsets | M | No. of selected clusters | Ratio % | No. of transactions in trash | MIS | MAS | AVS |
|---|---|---|---|---|---|---|---|---|
| 3 | 93 | 1 | 23 | 24.73 | 0 | 1 | 6 | 3.52 |
| | | 3 | 14 | 15.05 | 2 | 1 | 6 | 3.14 |
| 5 | 65 | 1 | 21 | 32.31 | 1 | 1 | 5 | 3.09 |
| | | 5 | 9 | 13.85 | 2 | 1 | 5 | 2.66 |
| 7 | 44 | 1 | 15 | 34.09 | 1 | 1 | 5 | 2.6 |
| | | 7 | 6 | 13.64 | 9 | 1 | 5 | 2.33 |

**Table 3.** Clusters on `bioexaout`

| $minfr$ | No. of frequent closed itemsets | M | No. of selected clusters | Ratio % | No. of transactions in trash | MIS | MAS | AVS |
|---|---|---|---|---|---|---|---|---|
| 53 | 462,041 | 1 | 34 | 0.007 | 0 | 1 | 20 | 9.55 |
| | | 10 | 7 | 0.001 | 0 | 1 | 20 | 8.35 |
| 59 | 217,952 | 1 | 28 | 0.013 | 0 | 11 | 19 | 16.92 |
| | | 10 | 7 | 0.003 | 4 | 14 | 19 | 14.57 |
| 65 | 99,968 | 1 | 23 | 0.023 | 0 | 10 | 17 | 15.74 |
| | | 10 | 6 | 0.006 | 2 | 7 | 17 | 12 |
| 71 | 43,396 | 1 | 19 | 0.044 | 0 | 9 | 16 | 14.52 |
| | | 10 | 5 | 0.011 | 9 | 13 | 16 | 12.2 |

**Table 4.** Clusters on `bioexain`

Especially on large data sets (like `bioexain`), the number of selected clusters is much lower than the number of frequent closed itemsets. When $minfr$ decreases, for an identical value of $M$, we see that the number of selected clusters does not increase as much as the number of frequent closed itemsets (in other words, ratio decreases). With $M = 1$, generally all transactions are classified. Let us recall (Section 2.3) that to discover all meaningful clusters it is suitable

to set $M$ to 1. When the value of $M$ increases, we note experimentally a great decrease of the number of selected clusters.

In order to discover combinations of examinations associated to a stage of fibrosis, we would like to rank clusters on their "purity" score, measured according to the stage of fibrosis. An usual way to measure impurity is to use an entropy function [11]. Let $P = (p_1, \ldots, p_5)$ be the frequency distribution of the stage of fibrosis on a cluster (5 stages exist), the entropy of $P$ denoted $\varphi(P)$ is $\varphi(P) = \sum_{i=1}^{5} p_i \times log\ p_i$. We know that the lower $\varphi(P)$ is, the purer the cluster is. $\varphi(P) = 0$ if and only if $\exists\ i$ with $p_i = 1$ (i.e. there is a single value for the stage of fibrosis).

We give now some clusters among them maximizing the purity according to the stage of fibrosis. On examples given below, each line corresponds to a cluster: its definition (i.e. a set of pairs examination / result), its frequency and the frequency distribution of the stage of fibrosis (a colon is inserted between these elements). On `bioexaout`, with $minfr = 3$, we are able to produce pure clusters but with few examples. For instance, the following cluster gathers 4 patients:

```
HBC-AB+,HBS-AG+,HBE-AG-,HBE-AB+ : 4 : F0=0.0 F1=0.0 F2=0.0 F3=100.0 F4=0.0
```

Some clusters seem more associated with severe stages of fibrosis. For instance:

```
HBS-AB+,HBE-AG-,HBE-AB+ : 5 : F0=0.0 F1=0.0 F2=0.0 F3=25.0 F4=75.0
HBC-AB+,HBE-AG-,HBE-AB+ : 8 : F0=0.0 F1=14.29 F2=0.0 F3=57.15 F4=28.57
HBE-AB+ : 13 :  F0=0.0 F1=16.67 F2=16.67 F3=41.67 F4=25.0
```

The clusters depicted above seem to indicate that `HBE-AB+` may have a great role with regard to severe stages. Let us note that these three clusters are described only by five items: `HBC-AB+, HBE-AB+, HBE-AG-, HBS-AB+` and `HBS-AG+` Some other clusters conclude rather on mild stages:

```
HBE-AB-,HCV5'NCRRT-PCR+,HBE-AG- : 7 : F0=0.0 F1=66.67 F2=33.33 F3=0.0 F4=0.0
HBE-AB-,HBE-AG+ : 13 : F0=0.0 F1=58.33 F2=41.67 F3=0.0 F4=0.0
HBE-AB-,HCV5'NCRRT-PCR+,HBE-AG-,HBS-AG- : 6 : F0=0.0 F1=80.0 F2=20.0 F3=0.0
                                               F4=0.0
HBE-AG+ : 16 :   F0=0.0 F1=53.33 F2=40.0 F3=6.67 F4=0.0
```

It should be useful to perform statistical analysis in order to confirm (or not) these observations. Due to the space limitation,[2] we do not give here all the selected clusters (Tables 3 and 4 indicate their main characteristics).

As `bioexain` has a rather large number of items and data are highly correlated, the computation of clusters require a value for $minfr$ not too low. That leads to produce clusters quite general having a mixture of degrees of fibrosis. Clusters are described with more items than on `bioexaout` (see Tables 3 and 4 for more details). Here are some examples:

```
ALB=,CL=,K=,oudan+,G-GTP= : 54 : F0=0.0 F1=56.60 F2=30.19 F3=7.55 F4=5.66
ALB=,CL=,LAP=,nyuubi=,oudan+,U-BIL=,U-GLU=,U-KET=,U-PH+,U-PRO=,U-RBC=,
```

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

```
        U-SG+,G-GTP= : 53 : F0=0.0 F1=55.77 F2=30.77 F3=9.61 F4=3.85
ALB=,CRP=,K=,oudan+,U-BIL=,U-GLU=,U-PH+,U-PRO=,U-RBC=,U-SG+,F-A2.GL= :
        53 : F0=1.92 F1=55.77 F2=26.92 F3=15.38 F4=0.0
CHE=,CL=,D-BIL=,HCV-AB=,I-BIL=,K=,LAP=,oudan+,T-BIL=,UA=,UN=,NA= : 56 :
        F0=1.79 F1=41.07 F2=35.71 F3=14.29 F4=7.14
D-BIL=,nyuubi=,oudan+,T-BIL=,U-BIL=,U-GLU=,U-KET=,U-PH+,U-PRO=,U-RBC=,
        U-SG+,G-GTP= : 56 : F0=0.0 F1=45.45 F2=40.0 F3=12.73 F4=1.82
```

Let us note that most of the selected pairs examination / result concern the normal values of examinations. In a cluster, there are always few examinations with a non-normal value. On the experimentation with $minfr = 53$ and $M = 1$ (34 selected clusters), only three non-normal values arise: `oudan+`, `U-PH+` and `U-SG+`. We do not know if these examinations have a special role or if they are included in a cluster because they are common. It is likely that it should be useful to mine `bioexain` with a lower value for $minfr$.

## 5 Conclusion

Using a new method, suitable for categorical data, to discover meaningful clusters, we have searched factors for estimating the stage of liver fibrosis from hepatitis data. These factors are combinations of examinations performed on patients. The number of selected clusters (between 5 and 34, according to $minfr$, $M$ and the data set) is much lower than the number of potential clusters given by conceptual classification methods or frequent closed itemsets.

On out-hospital examination data, this work suggests an interesting role of some examinations (e. g., `HBC-AB+`, `HBE-AB+` and `HBE-AG-` seem more associated with severe stages). On in-hospital examination data, except three non-normal examinations values (`oudan+`, `U-PH+` and `U-SG+`), the other selected examinations have a normal value, which might make the medical interpretation of the associated clusters difficult. It should be useful to have more biopsy with a known value for the stage of fibrosis (due to missing values on this attribute, we had to remove in this work many biopsy of the Table `biopsy`). More biopsy data would solely lead to a very slight increase of the computation cost because the algorithmic cost of such methods is chiefly due to the number of items.

## References


[1] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of ACM SIGMOD 93*, pages 207–216. ACM Press, 1993.

[2] J. F. Boulicaut and A. Bykowski. Frequent Closures as a Concise Representation for Binary Data Mining. In *Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 00*, volume 1805 of *Lecture notes in artificial intelligence*, pages 62–73, Kyoto, Japan, 2000. Springer-Verlag.

[3] G. Das and H. Mannila. Context-based Similarity Measures for Categorical Databases. In *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 00*, volume 1910 of *Lecture notes in artificial intelligence*, pages 201–210, Lyon, F, 2000. Springer-Verlag.

[4] N. Durand and B. Crémilleux. Extraction of a Subset of Concepts from the Frequent Closed Itemset Lattice: a New Approach of Meaningful Clusters Discovery. In *Proceedings of the Advances in Formal Concept Analysis in Knowledge Discovery in Databases workshop, co-located with ECAI 02*, Lyon, France, July 2002.

[5] N. Durand, B. Crémilleux, and M. Henry-Amar. Discovering Associations in Clinical Data: Application to Search for Prognostic Factors in Hodgkin's Disease. In S. Quaglini, P. Barahona, and S. Andreassen, editors, *Proceedings of the 8th Conference on Artificial Intelligence in Medecine in Europe (AIME 2001), Lecture Notes in Artificial Intelligence*, volume 2101, pages 50–54, Cascais, Portugal, July 2001. Springer-Verlag.

[6] R. Godin, R. Missaoui, and H. Alaoui. Incremental Concept Formation Algorithms based on Galois (concept) Lattices. *Computational Intelligence*, 11(2):246–267, 1995.

[7] S. Guha, R. Rastogi, and K. Shim. ROCK: A Robust Clustering Algorithm for Categorical Attributes. In *the 15th International Conference on IEEE Data Engineering (ICDE'99)*, pages 512–521, 1999.

[8] E. H. Han, G. Karypis, V. Kumar, and B. Mobasher. Hypergraph Based Clustering in High-Dimensional Data Sets : a Summary of Results. *Bulletin of the Technical Commitee on Data Engineering*, 21(1), 1998.

[9] Z. Huang. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Kownledge Discovery*, 2(3):283–304, 1999.

[10] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[11] S. Kullback. *Information theory and statistics*. Chapman and Hall, New York - Dover, 1967.

[12] W. Z. Liu and A. P. White. Metrics for Nearest Neighbour Discrimination with Categorical Attributes. In *proceedings of the Seventh International Annual International Conference of the British Computer Society Specialist Group on Expert Systems (ES 97)*, Cambridge (UK), 1997.

[13] R. S. Michalski and R. E. Stepp. Learning from Observation: Conceptual Clustering. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach*, volume 1, pages 331–363. Morgan Kauffmann, 1983.

[14] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing Iceberg Concept Lattices with TITANIC. *Journal on Knowledge and Data Engineering*, 2002.

[15] K. Wang, X. Chu, and B. Liu. Clustering Transactions Using Large Items. In *ACM Conference on Information and Knowledge Management (CIKM)*, USA, 1999.

[16] J. Zytkow and S. Gupta. Mining medical data using sql queries and contingency tables. In *proceedings of the PKDD 01 Discovery Challenge on Thrombosis Data co-located with the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD 01*, pages 61–73, Freiburg, Germany, September 2001.