In the Proceedings of the PKDD'04 Discovery Challenge on Atherosclerosis Data, co-located with the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04), pages 32-43, Pisa, Italy, September 2004.

Discovery of Overlapping Clusters to detect Atherosclerosis Risk Factors

Nicolas Durand[1], Guillaume Cleuziou[2], and Arnaud Soulet[1]

[1] GREYC, CNRS - UMR 6072, Université de Caen
Boulevard Maréchal Juin
14032 Caen Cédex, France
{ndurand,asoulet}@info.unicaen.fr

[2] LIFO, CNRS - FRE 2490, Université d'Orléans
Rue Léonard de Vinci
45067 Orléans Cédex 2, France
cleuziou@lifo.univ-orleans.fr

Abstract. This work presents a data mining effort to discover pure or almost-pure clusters with respect to atherosclerosis risk factors, from a medical database used by the STULONG project. One originality of this work is to produce overlapping clusters with two recent algorithms: Eclat and PoBOC. Such clusters, described by social characteristics and physical and biochemical examinations on patients, allow to characterize patients affected by disease due to atherosclerosis, and may lead to relevant factors. We compare the two algorithms, and we observe if the results point out the role of some examinations.

Keywords: atherosclerosis, risk factors, clustering, overlapping clusters, class characterization, frequent closed itemsets, pole-based clustering.

1 Introduction

The STULONG project, started in the 1970s, addresses the twenty-year long longitudinal study of the risk factors of the atherosclerosis in a population of 1417 men in the former Czechoslovakia. The main goal of this study is to identify atherosclerosis risk factors and to follow the development of these risk factors and their impacts.

The study was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences. The data resource is on the web pages http://euromise.vse.cz/stulong-en/. At present time the data analysis is supported by the grant of the Czech Ministry of Education.

The use of relevant and efficient methods to explore such large datasets 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. In this paper, we focus on two methods (Ecclat [5] and PoBOC [3]) to discover meaningful clusters (see Section 2). Our aim is to detect associations of examinations expressing risk factors.

These two clustering methods present the originality to discover overlapping clusters i.e. a set of groups where a patient can be present in several groups. This task is also called "soft-clustering". The overlapping is very useful in some applications like web mining or medicine. For instance, we would like to retrieve a user from several kinds of queries corresponding to several centres of interest. This is the same remark in medicine, we can have several points of views to class the patients. Furthermore, we produce the description of each cluster, i.e. a set of examinations. So, we can easily interpret the results. In the KDD vocabulary, such a cluster corresponds to a bi-set [2]: a set of elements and a set of properties describing them. The overlapping and the cluster descriptions are major advantages to a meaningful clustering, especially for the class characterization.

Patients Id.	Items
t_1	$A\ B\ C$
t_2	$A\ B\ C$
t_3	$A\ B\ C$
t_4	$D\ E$
t_5	$D\ E\ \ H$
t_6	$A\ \ D\ E\ F\ G\ H$
t_7	$A\ \ F\ G\ \ I$
t_8	$H\ I$

Table 1. Example of transactional database

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 (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. Table 1 presents an example of transactional database. There are 8 patients (denoted $t_1 \ldots t_8$) and 9 items denoted $A \ldots I$. For example, A denotes an item which is linked to the level of reached education (e.g. secondary school), B means that the level of total cholesterol is greater or equal than 5.2 mmol/l, etc. A bi-set is composed of an itemset and a set of transactions (called *tidset*).

The rest of this paper is organized as follows: Section 2 summarizes the two clustering methods (Ecclat and PoBOC) which produce overlapping clusters of patients. We describe the data preparation stage in Section 3. Results and discussion are respectively presented in Section 4 and Section 5.

2 Discovery of overlapping clusters

2.1 Context and related work

The general meaning of clustering is to decompose or partition a set of objects into groups so that the objects in one group are similar to each other and are as different as possible from the objects of the other groups. We call "hard-clustering" this process. The methods developed in the literature can be identified in three mean types [1]: those based on an attempt to find the optimal partition into a specified number of clusters (for instance, the standard *K-Means* method), those based on a hierarchical attempt to discover cluster structure (like the centroid-based agglomerative hierarchical clustering), and those based on a probabilistic model for the underlying clusters (there is an assumed probability model for each component cluster).

Among these methods, there are some fuzzy-clustering methods (like *Fuzzy c-Medoids* [7]) which use a fuzzy membership function computing a membership value of an element for each class. Let us note that it is necessary to perform a post-processing task in order to exploit the real organisation of the groups. These approaches compute fuzzy clusters without explanations on the gathering. Soft-clustering algorithms are not very abundant. We can mention for example, some methods based on pyramids [4]. Algorithms providing an explanation on the gathering are also uncommon. We can cite *Bi-Clust* [11] which produces bi-sets representing a bi-partition (on the transactions and on the items), and *COBWEB* [6] using probabilistic distributions as descriptions.

2.2 Ecclat

Ecclat (Extraction of Clusters from Concept LATtice) [5] produces bi-sets from large categorical datasets. These bi-sets represent a set of overlapping clusters described by itemsets. 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. Let us note that the number of clusters is not set in advance.

Ecclat discovers the frequent closed itemsets [10] (seen as potential clusters), evaluates them and selects some. An itemset X is frequent if the number of transactions which contain X is at least the frequency threshold (called $minfr$) set by the user. X is a closed itemset if its frequency only decreases when any item is added. A closed itemset checks an important property for clustering: it gathers a maximal set of items shared by a maximal number 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 meaningful clusters selection. $minfr$ corresponds to the minimum number of transactions in a cluster.

Ecclat 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 favoured (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 *concentration*.

ECCLAT uses the *interestingness* to select clusters and to produce a clustering with a slight overlapping between clusters (which is called "*approximate clustering*") or a set of clusters with overlapping. This functionality 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 algorithm of selection, 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". When the M value increases, the number of clusters decreases. With M near to $minfr$, a pseudo-partition is found.

2.3 POBOC

POBOC (Pole-Based Overlapping Clustering) [3] takes a similarity matrix as input and produces a set of overlapping clusters. Like for ECCLAT, the final number of clusters is unknown *a priori*.

From the similarity matrix $\mathcal{S}$ over a set of transactions $\mathcal{T}$, the algorithm proceeds as follows:

(1) construction of a similarity graph $G_{\mathcal{S}}(\mathcal{T}, \mathcal{V})$ with $\mathcal{V}$ the set of edges,
(2) extraction of complete sub-graphs from $G_{\mathcal{S}}(\mathcal{T}, \mathcal{V})$, the "poles",
(3) multi-assignment of the transactions to the poles.

The similarity measure we use in our study [9] consists in defining a new description language (new items), derived from the initial set of items. Each new item is obtained by random conjunctions of initial items. This measure allows to deal with categorical and/or quantitative variables and is thus adapted to our problem.

In the first step, the similarity graph $G_{\mathcal{S}}(\mathcal{T}, \mathcal{V})$ is based on the set of transactions as vertices. $\mathcal{V}$ is the set of edges so that there is an edge between two transactions t_i and t_j if t_i belongs to the neighborhood of t_j and vice versa. This step allows to take into account the density and to isolate the outliers.

The second step uses two main heuristics for extract the poles from the similarity graph. Poles correspond to homogeneous areas into the similarity graph. A first heuristic consists in finding a "well-defined" vertice t_i in $G_{\mathcal{S}}(\mathcal{T}, \mathcal{V})$ and the second heuristic allows to approximate the maximal clique-graph which contains t_i. This two-stage process is iterated until no "well-defined" vertices can be

found. The number of poles is determined automatically, and corresponds to the number of final clusters.

The last step is the one which assigns each transaction from $\mathcal{T}$ to one or several poles among $\{P_1, \ldots, P_k\}$. A new heuristic is used for this assignment stage, so that a transaction t_i is only assigned to its most similar poles. The similarity between a transaction t_i and a pole P_u is the average similarity between t_i and each transaction from P_u ($sim(t_i, P_u) = \frac{1}{|P_u|} \sum_{t_j \in P_u} s(t_i, t_j)$). Finaly, a clustering with PoBOC results in k overlapping clusters $Y_1, \ldots, Y_k$ so that $Y_u = \{t_i \in \mathcal{T} \mid t_i \textit{ is assigned to } P_u\}$.

Rather than give an intensional description of the clusters via itemsets, PoBOC induces clusters defined on an extentional way (set of transactions). Thus, in order to compare the two clustering, we propose a step of characterization of the clusters. Given a cluster $Y = \{t_i, \ldots, t_j\}$ (a tidset), we search the set of items $i(Y)$ which belong to all the transactions from Y. Finaly, the tidset is extended to a "closed tidset" $t(i(Y))$ with all the transactions which contain the itemset $i(Y)$.

3 Data preparation

The database contains four tables available on the web (`http://lisp.vse.cz/challenge/ecmlpkdd2004/`). They have been loaded using the relational database management system (`Mysql 4.0.18`). Even if the web site associated to the discovery challenge provides a lot of useful information (e.g. a clear meaning of each attribute, the frequency of each attribute value), one advantage of using a relational database is to achieve easily an overview of the data.

3.1 Overview of tables

The table `Entry` contains 1417 men who have been examined during the entry examination. Each patient is described by 64 attributes. Most of them are qualitative (biochemical examinations mainly gather continuous attributes). We use this table to get the features describing the patients when they are entered in the study (i.e. during the initial examination).

The table `Control` gathers risk factors and clinical demonstration of atherosclerosis during the examinations of the patients followed during 20 years (i.e. patients from `normal group`, `intervened risk group` and `control risk group`). There are 10572 examinations. We use this table to collect patients affected by a disease due to atherosclerosis during the study. This table has 66 attributes.

The table `Death` indicates the 389 patients who died during the study. The causes of death can be different from atherosclerosis. We use this table to pick out the patients who died from atherosclerosis during the study. The attributes of this table are the patient identification number, the date and cause of death.

Finally, the table `Letter` provides additional information (received via a postal questionnaire) about the health status of 403 patients. We do not use this table in this work.

3.2 Aim of experiments and resulting files

Let us recall that we are here interested in characterizing patients (by using overlapping clusters) according to whether they will be affected or not by a disease due to atherosclerosis. This topic corresponds to the analytic questions related to the long-term observation depicted on the web pages of this discovery challenge. We then performed the two following experiments.

In the first experiment (named `AthDeath`), from the features available in the table `Entry`, we would like to distinguish the patients who died from atherosclerosis from the others. We focus on patients from `normal group`, `intervened risk group` and `control risk group` because only patients of the above-mentioned groups are followed during the period. So, thanks to the long-term observation, by using the table `Death`, we know the patients who died. The attribute (`PRICUMAR`) of the table `Death` which provides the cause of death has 8 values. We consider (from a medical point of view) that the values `myocardial infarction`, `coronary heart disease`, `stroke` and `general atherosclerosis` are the causes of death due to atherosclerosis: these four values correspond to 165 patients. When we join them with the groups of patients who are followed during all the study, 124 patients remain. This work is done under the assumption that all patients dying from atherosclerosis are recorded in the table `Death`. We obtain overall 748 patients.

With the second experiment (named `AthRisk`), we would like to distinguish the patients of the `risk group` (merging of the `intervened risk group` and the `control risk group`) from patients of the `normal group` (studied or not). The dataset contains all the patients who are not in the `pathologic group`. Finally, we have 1,303 patients. We kept the 168 patients who do not have a class value. We will see how our algorithms will class them.

We decided *a priori* to keep all attributes of the table `Entry`. Nevertheless, we deleted the attribute `KONSKUP` (the class of patients) because its value may introduce a bias. We also removed attributes relating to the personal anamnesis due to the very low frequencies of values. We replaced the attributes `ROKNAR` (year of birth) and `ROKVSTUP` (year of entry into the study) by the age of the patient when he was entered in the study. For attributes having only two values, only the item corresponding to the value `true` (i.e. presence of the characteristic) has been kept. For the categorical attributes, we created as many items than values. The attribute `STAV` (marital status) gave 4 items : `STAV=1` (married), `STAV=2` (divorced), etc. The attributes `CHLST` (cholesterol) and `TRIGL` (triglycerides) were segmented into binary attributes according to the thresholds given in the web pages. We used the following equivalences: for `CHLST`: 5.2 mmol/l = 200 mg/dL and for `TRIGL`: 2.0 mmol/l = 150 mg/dL. For example, the item `CHLST-` indicates a value lower than the threshold, and `CHLST+` a value greater than the threshold. The other continuous attributes (e.g. `SYST1`: blood pressure systolic I, mm Hg) were cut into qualitative attributes, each of these attributes having 3 values with an even number of patients per value. For example, the item `SYST1<=116` indicates a value lower than 116, `SYST1[117-135]` corresponds to a value between 117 and 135, etc.

Table 2 presents the characteristics of the obtained datasets. For `AthDeath`, there are 748 patients described by 117 items. There are two class values : `dead` (124 patients are dead from atherosclerosis) and `other`. The classes for `AthRisk` are `risk`, `normal` and `unknown`.

	No. of patients	Distribution according to the classes	No. of items
`AthDeath`	748	124 (dead), 624 (other)	117
`AthRisk`	1303	859 (risk), 276 (normal), 168 (unknown)	108

Table 2. Characteristics of `AthDeath` and `AthRisk`

4 Experimentations

4.1 Protocol

In order to discover combinations of examinations associated to a group, we would like to rank clusters with respect to their "purity" score, measured according to the class. The items corresponding to a pure or almost-pure cluster, present a high interest to characterize the main class of this cluster.

An usual way to measure impurity is to use an entropy function [8]. Let $P = (p_1, \ldots, p_j)$ be the frequency distribution of the classes on a cluster, the entropy of P denoted $\varphi(P)$ is $\varphi(P) = -\sum_{i=1}^{j} p_i \times log\ p_i$. 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. all the transactions belong to the class i).

4.2 Results

Table 3 summarizes the results obtained with ECCLAT. For each datasets and a $minfr$ value, we can observe the number of frequent closed itemsets (fci). We also present the number of clusters, the number of transactions in the trash cluster, the average overlapping between the clusters (number of transactions), and the average number of transactions contained in the clusters, according to a M value (see Section 2.2). We fixed $minfr$ to 5%. For `AthDeath`, this represents a minimum number of 38 patients per cluster. We obtain the smallest number of clusters who class all the patients with M=3. For `AthRisk`, $minfr$ represents 66 patients. With M=5, we reduce the number of clusters by minimizing the trash cluster. Only one patient remains (of the class `unknown`).

Table 4 presents the results obtained with POBOC. For `AthDeath`, 155 clusters are produced. The average size of the clusters is 24.5 before characterization and the average overlapping between two clusters is 1.4. The closing of the tidsets leads to larger clusters (367.7) with more intersections between them (204). With `AthRisk`, we obtained 264 clusters having an average size equal to 32 transactions and an average overlapping about 2. The closed tidsets induced contain 670.2 transactions in average with an overlapping equal to 492. Let us remark that two clusters are not characterized: the itemsets describing them are an empty set (see Section 2.3).

	minfr (%)	No. of fci	M	No. of clusters	Size of the trash cluster	Overlap (avr.)	Size of clusters (avr.)
AthDeath	5	1,412,883	1	418	0	7	41.1
			3	181	0	5	41.5
AthRisk	5	1,263,715	1	626	0	13	73.3
			5	179	1	10	77.7

Table 3. Results with ECCLAT

	No. of clusters	Tidsets before characterization		Closed tidsets	
		Overlap (avr.)	Size of clusters (avr.)	Overlap (avr.)	Size of clusters (avr.)
AthDeath	155	1.4	24.5	204	367.7
AthRisk	264	2	32.0	492	670.2

Table 4. Results with POBOC

We give now some clusters from those which maximize the purity according to the class of the patients. On examples given below, each line corresponds to a cluster: its definition (i.e. a itemset), the number of patients and the frequency distribution of the classes (a ":" is inserted between these three informations).

On AthDeath, we are able to produce pure clusters with ECCLAT but only for the class other. For instance, the two following clusters are obtained :

```
STAV=1 AKTPOZAM=2 DOPRATRV=5 PIVO10 VINO LIHOV PIVOMN=2 PIVOMN=5      : 39 : dead=0 other=100
LIHMN=8 CAJ=5 BOLDK=1 DUSNOST=1 MOC=1

TELAKTZA=1 AKTPOZAM=2 DOPRATRV=5 DOBAKOUR=10 VINO LIHOV PIVOMN=2      : 39 : dead=0 other=100
PIVOMN=5 LIHMN=8 BOLHR=1 BOLDK=1 DUSNOST=1 MOC=1
```

As we can see, they correspond to consumers of alcohol (beer: PIVO, wine: VINO, liquor: LIHOV) and smokers (DOBAKOUR). We can make similar observations with the clusters obtained with POBOC. They provide other items concerning tobacco (KOURENI, BYVKURAK):

```
VINO PIVO12 KOURENI=6 BOLHR=2 TRIGL-                        : 186 : dead=15.1 other=84.9

ZODPOV=3 BYVKURAK=12 SUBSC[24-50] KOURENI=6 AGE<=48         : 137 : dead=14.6 other=85.4
TRIGL-

PIVO12 AGE<=48 STAV=3 TRIGL-                                : 289 : dead=14.2 other=85.8

BYVKURAK=12 CUKR[3-6] SUBSC[24-50] AGE<=48 BOLDK=1          : 184 : dead=16.9 other=83.1
BOLHR=2 TRIGL-
```

Let us note that initially POBOC produces some pure clusters, but the computing of the closed tidsets extends the clusters and they become almost-pure.

The items which are not related to alcohol and tobacco, are different with POBOC or ECCLAT. For instance, ECCLAT finds some items concerning physical

activity after job (moderate activity, AKTPOZAM=2), in job (mainly sit, TELAKTZA=1), marital status (married, STAV=1) and the duration to the way to work (around 0.5 hours, DOPRATRV=5). PoBOC finds some items related to marital status (single, STAV=3), the food habits (sugar, CUKR[3-6]), physical examination (skinfold above musculus subscapularis, SUBSC[24-50] mm). We remark that PoBOC brings out low values of blood pressures (systolic and diastolic):

```
SYST2<=118 DIAST2<=75 SYST1<=120 BOLHR=2   : 175 : dead=24.6 other=75.4

DIAST1<=75 DIAST2<=75 TRIGL-               : 144 : dead=24.3 other=75.7

SYST2<=118 SYST1<=120 SUBSC[24-50] TRIGL-  : 184 : dead=23.9 other=76.1
```

PoBOC and Ecclat have some difficulties to discover clusters for the class dead, because of the disproportions in the initial data. For example, we present some clusters produced by Ecclat, having a maximum number of patients of this class :

```
KOURENI=4 DOBAKOUR=10 BOLHR=1 SYST1[136-225]              : 39 : dead=33.3 other=66.7
SYST2[139-215] DIAST2[86-140] CHLST+

PIVOMN=4 LIHMN=7 BOLHR=1 BOLDK=1 SYST1[136-225]           : 45 : dead=33.3 other=66.7
DIAST1[89-145] DIAST2[86-140]

STAV=1 AKTPOZAM=2 DOBAKOUR=10 PIVOMN=4 BOLHR=1 BOLDK=1    : 39 : dead=33.3 other=66.7
SYST1[136-225] SYST2[139-215]

STAV=1 DOBAKOUR=10 BOLHR=1 BOLDK=1 SYST1[136-225]         : 42 : dead=33.3 other=66.7
DIAST1[89-145] SYST2[139-215] DIAST2[86-140] CHLST+ MOC=1
```

We remark that the items SYST1[136-225], SYST2[139-215], DIAST1[89-145], DIAST2[86-140] and CHLST+ are often present. We go back over this remark with AthRisk. The clusters produced by PoBOC do not present these items on this dataset.

In order to discover more clusters of this class with Ecclat, a lower value for $minfr$ should be necessary. But, with $minfr$=3%, more than 5 millions of frequent closed itemsets are obtained, and we come up against the computing time.

On AthRisk, we are able to produce almost-pure clusters for the class risk. For instance, we have the following clusters with Ecclat:

```
STAV=1 AKTPOZAM=2 DOPRATRV=5 BOLDK=1 DUSNOST=1   : 69 : normal=4.35 risk=89.85 unknown=5.8
SYST1[136-225] DIAST1[89-145] SYST2[139-215]
DIAST2[86-140] CHLST+ MOC=1

STAV=1 CAJ=4 DUSNOST=1 SYST1[136-225]            : 69 : normal=2.90 risk=89.85 unknown=7.25
DIAST1[89-145] SYST2[139-215] DIAST2[86-140]
MOC=1

STAV=1 DOPRATRV=5 DOBAKOUR=10 BOLDK=1 DUSNOST=1  : 67 : normal=7.45 risk=89.55 unknown=3
SYST1[136-225] DIAST1[89-145] SYST2[139-215]
DIAST2[86-140] CHLST+ MOC=1

STAV=1 DOBAKOUR=10 BOLHR=1 BOLDK=1 DUSNOST=1     : 73 : normal=8.25 risk=89.0 unknown=2.75
SYST1[136-225] DIAST1[89-145] SYST2[139-215]
DIAST2[86-140] CHLST+ MOC=1
```

We note the same remark than we had with `AthDeath` and the class `dead`, but with more precise observations. Indeed, the association of all the items `SYST1[136-225]`, `SYST2[139-215]`, `DIAST1[89-145]` and `DIAST2[86-140]` is present. Sometimes `CHLST+` also appears.

Some of these items seem to be very important for atherosclerosis, because PoBOC also produced clusters containing them:

```
SYST1[136-225] SYST2[139-215]      : 246 : normal=30.1 risk=56.1 unknown=13.8
CUKR[3-6] BOLDK=2 BOLHR=2 TRIGL-

DIAST1[89-145] AGE[44-47] BOLDK=2  : 397 : normal=29.2 risk=58.2 unknown=12.6
BOLHR=2 TRIGL-

SYST1<=116 SYST2<=118 DIAST2<=75   : 249 : normal=8 risk=92
```

Nevertheless, we observe that some patients from the class `risk` have low values of blood pressures (systolic I, systolic II and diastolic II).

The class `normal` is characterized by few clusters. This is not very important, because we want to detect risk factors. We present some clusters of the class `normal` with Ecclat:

```
STAV=1 KOURENI=1 BOLHR=1 DUSNOST=1     : 77 : normal=58.4 risk=26.0 unknown=15.6
SYST1[117-135] SYST2[119-138] MOC=1

AKTPOZAM=2 BOLDK=1 DUSNOST=1           : 86 : normal=51.2 risk=39.5 unknown=9.3
DIAST2[76-85] CHLST- MOC=1
```

Most of the items concern normal values of examinations. The items `STAV<=1` and `MOC<=1` are present in most of `normal` clusters and `risk` clusters. `STAV<=1` is present in 73.2% of the clusters, and `MOC<=1` in 91.6%, so their role does not seem to be significant.

We also observe normal values or mean values of examinations with the clusters obtained by PoBOC having the most of `normal` patients:

```
VAHA[75-84] DIAST1[76-88] SYST2[119-138]    : 140 : normal=35.7 risk=64.3
TRIGL-

BOLDK=2 SYST1[117-135] TRIGL-               : 254 : normal=34.7 risk=65.3

SYST1[117-135] SYST2[119-138] PIVOMN=3      : 122 : normal=27.0 risk=64.8 unknown=8.2
PIVO10 AGE[44-47] TRIGL-
```

Let us note that 50.4% of the clusters contain the item `TRIGL-`. The triglycerides examination is normal for a lot of patients from the both classes: `risk` and `normal`. So, it seems difficult to use this examination in pronostic profiles (with this database).

On `AthRisk`, we also observe the position of the 168 patients with no class label and try to detect patients who present a risk. A set of almost-pure clusters[1] is selected from each of the two clustering proposed by PoBOC and Ecclat.

[1] 80% of the patients belong to the same class.

Table 5 reports that 22 patients appear in at least one selected cluster, on the two clustering process. Hence, one can consider these patients should be subject to medical controls. Furthermore, 6 patients (resp. 7 patients) appear into at least 5 clusters with PoBOC (resp. Ecclat). These patients are different with the two clustering processes and most of them do not belong to the 22 previous detected patients. The two clustering techniques then lead to complementary results which tend to finaly select 32 patients with suspicious profile with respect to atherosclerosis.

	PoBOC	Ecclat
No. of selected clusters	40/264	31/179
No. of patients in at least 1 cluster	56/168	69/168
No. of patients in common	22/168	
No. of patients in at least 5 cluster	6/56	7/69

Table 5. Analysis of the patients with unknwon class label

5 Discussion

From a technical point of view, Ecclat presents the advantage of directly producing bi-sets, but has some difficulties to identify clusters corresponding to minority classes. PoBOC find more clusters of the different classes but requires an post-processing step to complete the bi-sets.

We observed that the items related to blood pressures (systolic and diastolic) seem to have an important role. Ecclat finds the highest values for patients dead of diseases due to atherosclerosis (`AthDeath`). PoBOC discovers low values for patients of the other class. We can say that the results are complementary. For `AthRisk`, Ecclat and PoBOC bring out high values of blood pressures for risk patients. We noted that the other items present in the clusters produced by our algorithms are not the same. They find some physical examinations, items related to consumption of alcohol and tobacco, and physical activity. But PoBOC finds other items concerning sugar, age and weight. The tea, the coffee and the reached education are not present among the characteristics from the discovered clusters.

Our methods brought out that some attributes are not significant. With Ecclat, we remarked the marital status and the urine examinations. For PoBOC, the triglycerides examination is debatable. Indeed, the item corresponding to a normal value is present in half of the clusters (of any classes). We can wonder if this attribute is useful to find risk factors (based on this database).

First evaluations to detect the groups of the patients without label, show that 32 patients present some risks with respect to atherosclerosis. PoBOC and Ecclat agree to class 22 patients of them in the `risk group`.

6 Conclusion

Using two recent algorithms to discover meaningful clusters, we have searched atherosclerosis risk factors. ECCLAT and PoBOC did not produce the same clusters but the first interpretations lead to complementary conclusions. The important role of blood pressures are noted by the two algorithms.

In perspective, we will perform some experimentations with a more precise characterization for PoBOC and a lower threshold for ECCLAT to better capture the minority classes. In that way, we could identify some interesting roles of other items. We will also realize other evaluations to class the patients without label.

In order to take into acount the evolution of the patients, we note that it could be interesting to use overlapping clusters in the discovery of emerging patterns [12].

References

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

[2] J. Besson, C. Robardet, J-F. Boulicaut, and S. Rome. Constraint-based Concept Mining and its Application to Microarray Data Analysis. *Intelligent Data Analysis journal, IOS Press*, April 2004.

[3] G. Cleuziou, L. Martin, and C. Vrain. PoBOC: an Overlapping Clustering Algorithm, Application to Rule-Based Classification and Textual Data. In *16th European Conf. on Artificial Intelligence ECAI (To appear)*, Valencia, Spain, 2004.

[4] E. Diday. Une représentation visuelle des classes empiétantes : Les pyramides. In *Rairo : Analyse des Données (vol. 52)*, pages 475–526, 1986.

[5] N. Durand and B. Crémilleux. ECCLAT: a New Approach of Clusters Discovery in Categorical Data. In *the 22nd Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence (ES'02)*, pages 177–190, Cambridge, UK, December 2002.

[6] D. H. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, 2:139–172, 1987.

[7] R. Krishnapuram. Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining. *IEEE Transactions on Fuzzy Systems*, 9(4):596–607, 2001.

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

[9] L. Martin and F. Moal. A Language-based Similarity Measure. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01), LNAI 2167*, pages 336–347, Freiburg, Germany, September 2001.

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

[11] C. Robardet and F. Feschet. Efficient Local Search in Conceptual Clustering. In *Proceedings of the Int. Conf. Discovery science (DS'01), LNCS 2226*, pages 323–335, Washington, USA, November 2001.

[12] A. Soulet, B. Crémilleux, and R. Rioult. Condensed Representation of Emerging Patterns. In *the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*, pages 127–132, Sydney, Australia, May 2004.