Mining the Strongest Emerging Patterns Characterizing Patients Affected by Diseases due to Atherosclerosis

Bruno Crémilleux, Arnaud Soulet, and François Rioult

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
F-14032 Caen Cédex France
{Forename.Surname}@info.unicaen.fr

Abstract. This work presents a data mining effort to discover emerging patterns from an important medical database concerning a longitudinal study of the risk factors on atherosclerosis (STULONG project). These emerging patterns characterize patients affected by diseases due to atherosclerosis. Such emerging patterns may lead to improve the definition of risk groups of patients. Searching emerging patterns is a difficult task and the main originality of the approach is to use recent results in condensed representations of patterns in order to mine efficiently the strongest emerging patterns (i.e. with the best growth rate).

Keywords: atherosclerosis, risk factors, emerging patterns, class characterization, frequent closed patterns.

1 Introduction

The STULONG data address the twenty-year long longitudinal study of the risk factors of the atherosclerosis in a population of 1417 men in the former Czechoslovakia (this project started in the 1970s). The main goal of this study is to identify atherosclerosis risk factors (and their combinations) and follow the development of these risk factors and their impacts. The prevention and detection of atherosclerosis and cardiovascular diseases are undoubtedly a very important task for public health.

Data have been collected at the Institute of Clinical and Experimental Medicine (IKEM) in Praha (i.e. Prague) and the Medicine Faculty at Charles University in Plzen (i.e. Pilsen). Most of the data were transferred into electronic form by EuroMISE (European Centre for Medical Informatics, Statistics and Epidemiology) with the support of an European project. These data were kindly provided for the PKDD'03 discovery challenge. Four tables are included (see Section 3.1 for details on the data set) and let us note that the first two tables (named `Entry` and `Control`) were already available for the PKDD'02 discovery challenge [1].

Even if there are a lot of works in Knowledge Discovery in Databases (KDD), it is still a challenge to design processes in order to extract useful information for real data and the use of relevant and efficient methods to explore such large data sets is not easy. From KDD techniques point of view, in this paper, we focus on the discovery of a kind of patterns called emerging patterns (EPs). An EP captures significant changes and differences between two (or more) datasets. We will see in Section 2.3 that we propose a new approach which takes advantages on recent results in condensed representations of patterns and more precisely on closed frequent patterns. This new approach enables to mine efficiently the strongest emerging patterns (i.e. with the best growth rate). We call "strong emerging pattern" (SEPs) such EPs. The contribution of this paper to the PKDD'03 discovery challenge dealing with the data mining tasks, is to suggest SEPs characterizing patients according to whether they will be affected or not by a disease due to atherosclerosis. These SEP are combinations on features available when patients are entered in the study (i.e. during the initial examination). We know the future of the patients thanks to the study protocol which includes a long-term observation methodology (during 20 years) for patients belonging to the risk group and few patients from the normal group. We hope that the discovered SEP might help to better understand the risk factors and then improve the definition of the risk groups. We think that a fruitful data mining task needs interaction and collaborative work between, on the one hand the experts and providers of data, and on the other hand, the "data miners". The feedback and discussions with the medical experts during the workshop are important to keep this work so as to improve it.

The remainder of this paper is organized as follows: Section 2 outlines the notion of EP and presents our method to extract EPs. Section 3 gives our work for the data preparation stage. We show in Section 4 results and discussions on the medical data.

2 Discovery of strong emerging patterns

2.1 Emerging pattern: context and related work

An emerging pattern (EP) is defined as a pattern whose frequency increases significantly from one data set to another [7]. More precisely, an EP is a pattern whose *growth rate* (the ratio of the two frequencies computed on the two data sets) is larger than a given threshold. An EP can capture useful contrasts between data classes and EPs were proposed for classification [8, 9].

Let us give an example. Table 1 provides an excerpt of a database (noted $\mathcal{D}$) including 8 patients (noted $P_1 \ldots P_8$) and 7 items denoted C_1, C_2, $A \ldots E$. 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. Items C_1 and C_2 have a specific role: they indicate the data set in which each patient is located. Let us suppose that there are two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$. For instance, $\mathcal{D}_1$ gathers the patients not affected by a disease due to atherosclerosis and $\mathcal{D}_2$ the others. C_1 (resp. C_2) is the item used to indicate that a patient belongs to $\mathcal{D}_1$ (resp. $\mathcal{D}_2$). In other words, C_1 and C_2 can be seen as the class values. For the sake of clarity, this example contains only two data sets but, as we will see below, the search of EP can be straightforwardly generalized in presence of more than two classes. Table 1 is used as the running example throughout the paper.

$\mathcal{D}$

Patient	Items					
P_1	C_1	A	B	C	D	
P_2	C_1	A	B	C	D	
P_3	C_1	A	B	C		
P_4	C_1	A			D	E
P_5	C_2	A	B	C		
P_6	C_2		B	C	D	E
P_7	C_2		B	C		E
P_8	C_2		B			E

Table 1. Excerpt of a transactional database

In the following, each data record (i.e. a patient) is called a *transaction* (this term comes from the data mining context) or an *example*. A transaction is described by *items* (i.e. features). A *pattern* is a set of items (or an itemset). A transaction t *supports* (or contains) a pattern X if and only if $X \subseteq t$. $|\mathcal{D}|$ (where as usual $|\ldots|$ denotes the cardinality of a set) is the number of transactions of $\mathcal{D}$.

To highlight EPs, it is necessary to introduce the notion of frequency. The *frequency* of a pattern X is the number of transactions which support X. X is *frequent* if its frequency is at least the frequency threshold γ fixed by the user. We note $\mathcal{F}(X, \mathcal{D})$ the frequency of X with respect to the data set $\mathcal{D}$. Unless otherwise indicated, we use in this paper an absolute frequency of X in $\mathcal{D}$ (i.e. a number of examples $\leq |\mathcal{D}|$) instead of the relative frequency $\mathcal{F}(X, \mathcal{D})/|\mathcal{D}|$ in $[0, 1]$. For instance, in Table 1, $\mathcal{F}(ABC, \mathcal{D}) = 4$ (the transactions P_1, P_2, P_3 and P_5 support ABC[1]).

Let us recall that $\mathcal{D}_1$ and $\mathcal{D}_2$ are two data sets. The growth rate of a pattern X from $\mathcal{D}_2$ to $\mathcal{D}_1$ denoted as $GR_1(X)$ is defined as:

$$\begin{cases} 0, & \text{if } \mathcal{F}(X, \mathcal{D}_1) = 0 \text{ and } \mathcal{F}(X, \mathcal{D}_2) = 0 \\ \infty, & \text{if } \mathcal{F}(X, \mathcal{D}_1) \neq 0 \text{ and } \mathcal{F}(X, \mathcal{D}_2) = 0 \\ \frac{\mathcal{F}(X, \mathcal{D}_1)}{\mathcal{F}(X, \mathcal{D}_2)}, & \text{otherwise} \end{cases}$$

Given $\rho > 1$ as a growth rate threshold, a pattern X is said to be an ρ-emerging pattern (or simply EP) from $\mathcal{D}_2$ to $\mathcal{D}_1$ if $GR_1(X) \geq \rho$. For instance, in Table 1, with $\rho = 3$, A, ABC and $ABCD$ are EP from $\mathcal{D}_2$ to $\mathcal{D}_1$. Indeed, $GR_1(A) = 4/1 = 4$, $GR_1(ABC) = 3/1 = 3$ and $GR_1(ABCD) = 2/0 = \infty$. Conversely, BCD is not an EP : $GR_1(BCD) = 2/1 = 2$.

[1] Note that we use a string notation (e.g. ABC) to denote a set of items

If there are more than two datasets, the growth rate of a pattern X from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$ is simply:

$$GR_i(X) = \frac{\mathcal{F}(X, \mathcal{D}_i)}{\mathcal{F}(X, \mathcal{D}) - \mathcal{F}(X, \mathcal{D}_i)} \tag{1}$$

If the data sets have unbalanced numbers of transactions, it is necessary to correct $GR_i(X)$ by multiplying it by $\sum_{j \neq i} |\mathcal{D}_j|/|\mathcal{D}_i|$ (otherwise, there is a bias which favors the discovery of EPs to the data sets having the largest numbers of transactions and it is impossible to compare fairly EPs to different data sets). Nevertheless, this coefficient of correction is not required to compare EPs to a same data set (this coefficient is a constant and the order of EPs with respect to the growth rate remains identical). So, for the sake of clarity, we omit this coefficient in this section which provides the background on strong emerging patterns. But, in the data mining tasks of this challenge, we will see in Section 3 that we obtain files with unbalanced numbers of patients and experiments in Section 4 were performed with this correction.

Efficient computation of EPs remains a challenge because the property of anti-monotonic constraint with respect to the specialization relation of patterns does not hold longer for EPs. Let us recall that this property ensures an efficient search for the classical framework of level-wise algorithms [12] thanks to a safe pruning criterion within the search space. This property is a key point in the success of condensed representations [11, 5, 13] and, previously, for the Apriori algorithm [2]. There are usually a lot of candidates when mining EP in large databases or for small frequencies, and naive algorithms are too costly (the number of candidate patterns may be exponential). Dong et al. [7] propose the use of borders (the pair of sets of the minimal and of the maximal patterns in the collections) for concisely representing EPs and improving this search. EPs are discovered by manipulating only these borders using a multi-border differential algorithm. A particular type of EPs is called jumping emerging patterns (JEPs). A JEP corresponds to a pattern whose support increases abruptly from zero in one data set to non-zero in the other data set. Due to this infinite increase in support, JEPs represent knowledge that discriminates between the classes and they can be applied in classification [10]. Bailey et al. [3] propose a new tree-based data structure for storing the raw data. Their mining algorithm operates directly on the data contained in the trees. This approach is 2-10 times faster than the previous ones.

Yet even using borders, the discovery of EPs and JEPs has still exponential complexity and methods for improving efficiency are an open issue. This is crucial especially in large data sets. In the next section, we propose to revisit this task by using the last progress in condensed representations of patterns.

2.2 Condensed representations based on closed patterns

Let us have a look on recent results in Knowledge Discovery in Databases and more precisely in condensed representations of patterns. Briefly speaking, a condensed representation of patterns provides useful syntheses of large data sets, highlighting the correlations embedded in data [5, 13]. There is a twofold advantage of such an approach. First, it allows to improve efficiency of algorithms for usual tasks such as frequent patterns. For instance, the criterion of freeness enables tractable extractions (and to achieve *all* frequent patterns) in contexts where usual APRIORI-like algorithms fail [5, 13]. Second, condensed representations enable multiple uses of frequent patterns [11, 16, 6] (e.g., strong rules, informative rules or rules with minimal body, non-redundant rules, clustering, classification), which is a key point in many practical applications.

Let us focus now on closed patterns. A closed pattern in a database $\mathcal{D}$ is a maximal set of items (with respect to the set inclusion) shared by a set of transactions of $\mathcal{D}$. It relies on lattice theory [4] and Galois connection, which leads learning concepts [15]. In our example (see Table 1), ABC is a closed pattern because it is the maximum set of items shared by transactions containing at least ABC (here, transactions P_1, P_2, P_3 and P_5). On the opposite, AB is not a closed pattern since it is not a maximal group of items common to the data: all patients having the items AB also have the item C. The frequency provides another definition of a closed pattern. X is a closed pattern if and only if the addition of any item to X leads to decrease the frequency of X. Let us give the relationship with the *closure* which is necessary for the rest of the paper.

Definition 1 (closure operator) *Given a pattern X, its* closure *in $\mathcal{D}$, noted $h(X, \mathcal{D})$ is the maximal superset (w.r.t. set inclusion) of X that has the same frequency as X in $\mathcal{D}$. This closure is composed of X and the items $A \notin X$ such that*

$$\mathcal{F}(XA, \mathcal{D}) = \mathcal{F}(X, \mathcal{D}) \tag{2}$$

The closure of X is a closed pattern and $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$. For instance, in Table 1, $h(AB, \mathcal{D}) = ABC$ (i.e. C is always true when AB is true) and $\mathcal{F}(AB, \mathcal{D}) = \mathcal{F}(ABC, \mathcal{D})$. From large databases, there are efficient algorithms [5, 13] to compute all the frequent closed patterns (from a technical point of view, it is easy to produce the closed patterns from the free patterns because each closed pattern is the closure of a free pattern). We have developed such a software called MVMINER, which, moreover, is able to tackle data with missing values [14]. We now show the usefulness of such a condensed representation for emerging patterns.

2.3 From closed patterns to strong emerging patterns

As the collection of all frequent closed patterns enables to compute the frequency of every frequent pattern, we can expect that frequent closed patterns have relevant properties to discover emerging patterns. We start by giving a new characterization of the jumping emerging patterns (i.e. EPs whose growth rate is infinity).

Let us consider a database $\mathcal{D}$ which is split into k data set called $\mathcal{D}_1 \ldots \mathcal{D}_k$. We have $\mathcal{D} = \bigcup_i \mathcal{D}_i$ (i.e. each transaction belongs to one (and only one) data set $\mathcal{D}_i$). Let us call $C_1 \ldots C_k$ the items meaning the membership of $\mathcal{D}_1 \ldots \mathcal{D}_k$ (C_i is present for a transaction t if and only if t belongs to $\mathcal{D}_i$). For the following, X notes a pattern which does not get any item C_i (i.e., $\forall i : 1 \ldots k,\ \{C_i\} \not\subseteq X$).

Property 1 (characterization of JEP by closed patterns)

$$X \text{ is a } JEP \text{ from } \mathcal{D}\backslash\mathcal{D}_i \text{ to } \mathcal{D}_i \iff C_i \in h(X,\mathcal{D})$$

Proof $C_i \in h(X,\mathcal{D}) \iff \mathcal{F}(XC_i,\mathcal{D}) = \mathcal{F}(X,\mathcal{D})$ (Equation 2). By the definition of $\mathcal{D}_i$, $\mathcal{F}(X,\mathcal{D}_i) = \mathcal{F}(XC_i,\mathcal{D})$. Then $\mathcal{F}(X,\mathcal{D}) = \mathcal{F}(X,\mathcal{D}_i)$ and the denominator of $GR_i(X)$ is null (see Equation 1) and X is a JEP.

This characterization is helpful: it shows that it is possible to compute *all* JEPs above the frequency threshold from the frequent closed patterns containing the class value provided by the condensed representation. But, the condensed representation has another usefulness: it is easy to build EPs from frequent closed patterns having a class value (i.e. C_1 or C_2 in our example). We call Strong Emerging Patterns (SEPs) these EPs based on closed patterns because, as we will see, these EPs have the highest growth rate.

Definition 2 (strong emerging pattern) *A strong emerging pattern (SEP) from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$ is an EP based on a closed frequent pattern in $\mathcal{D}_i$.*

The following lemma indicates that it is easy to get the growth rate of a SEP.

Lemma 1 (SEP and growth rate) *Let X be a SEP from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$, then $GR_i(X)$ can be directly obtained from the frequencies provided by the condensed representation of frequent closed patterns of $\mathcal{D}$.*

Proof Assuming that X is a closed frequent pattern in $\mathcal{D}_i$ (i.e. a SEP). To compute $GR_i(X)$ (see Equation 1), we need to compute $\mathcal{F}(X,\mathcal{D}_i)$ and $\mathcal{F}(X,\mathcal{D})$. By the definition of $\mathcal{D}_i$, we have $\mathcal{F}(X,\mathcal{D}_i) = \mathcal{F}(XC_i,\mathcal{D})$. As X is a frequent closed pattern in $\mathcal{D}_i$, XC_i is a closed frequent pattern in $\mathcal{D}$ and its frequency in $\mathcal{D}$ is known from the condensed representation, so $\mathcal{F}(X,\mathcal{D}_i)$ is known. Let us consider now $\mathcal{F}(X,\mathcal{D})$. As $X \subset XC_i$, then $\mathcal{F}(X,\mathcal{D}) \geq \mathcal{F}(XC_i,\mathcal{D}) \geq \gamma$. It means that X is frequent, consequently its frequency can be inferred from the condensed representation [5]. Nevertheless, in our situation, $\mathcal{F}(X,\mathcal{D})$ can be known without any supplementary effort. Indeed, if X is a closed frequent pattern in $\mathcal{D}$, its frequency is immediately known from the condensed representation and $GR_i(X)$ is obtained. Otherwise, it means that X is not closed. As XC_i is the smallest *closed* superset including X, then we have $h(X,\mathcal{D}) = XC_i$ or $C_i \in h(X,\mathcal{D})$. We recognize Property 1: X is a JEP and we know its frequency. In both cases (X closed or not), Lemma 1 is proved.

Now, we are proving an important property showing that the SEPs are a cover of all EPs in the sense that any EP from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$ based on a pattern X is included in the EP based on the pattern $h(X,\mathcal{D}_i)$ with $GR_i(X) \leq GR_i(h(X,\mathcal{D}_i))$. In other words, any EP which is not extracted is covered by a *stronger* EP (i.e. a SEP) with respect to the growth rate.

Property 2 (SEPs as a cover of EPs) *Let X be a pattern. Then the SEP $h(X,\mathcal{D}_i)$ provides a upper bound for $GR_i(X)$. We have $GR_i(X) \leq GR_i(h(X,\mathcal{D}_i))$.*

Let us note that EPs based on X and $h(X)$ have the same frequencies (i.e. the same quality with respect to this criterion).

Proof Thanks to the property of closure, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(h(X, \mathcal{D}_i), \mathcal{D}_i)$. We can then write $GR_i(h(X, \mathcal{D}_i)) = \frac{\mathcal{F}(X,\mathcal{D}_i)}{\mathcal{F}(h(X,\mathcal{D}_i),\mathcal{D})-\mathcal{F}(X,\mathcal{D}_i)}$. The extensivity of the closure operator allows to write $X \subseteq h(X, \mathcal{D}_i)$ then $\mathcal{F}(X, \mathcal{D}) \geq \mathcal{F}(h(X, \mathcal{D}_i), \mathcal{D})$ because of the anti-monotonicity of frequency. We then have:

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

and the claim holds.

Let us come back to our running example. BC is not a closed pattern in $\mathcal{D}_1$. $h(BC, \mathcal{D}_1) = ABC$. We have $GR_1(BC) = 3/3 = 1$ and $GR_1(ABC) = 3/1 = 3$, ABC is a stronger EP than BC to characterize D_1. Remark that $ABCD$ is a closed pattern in $\mathcal{D}_1$ and it is a JEP.

We have seen that there are efficient algorithms to extract frequent closed patterns and we claim that SEPs are an efficient approach to extract all the strongest EPs.

Finally, let us specify that we have proposed a method (incorporated in our prototype MVMINER) to build condensed representations in presence of missing values [14]. This method provides an adequate condensed representation of free and closed patterns in presence of missing values. As the achieved files contain some missing values (see Section 3.2), we use MVMINER in the experiments.

3 Data preparation

The four tables available on the web (`http://lisp.vse.cz/challenge/ecmlpkdd2003/`) have been loaded using the relational database management system (`Mysql 3.23.49`). Even if the web site associated to this 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. Furthermore, a database management system is also useful to perform elementary transformations of data (e.g., computing the age of patients) and to run joins without writing scripts with a programming language.

3.1 Overview of provided 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 (physical examination and biochemical examination 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 studied 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 following 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 SEPs) 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. For that, we performed the two following experiments. As we need for these experiments information (death or disease due to atherosclerosis) arising during the 20-year observation, we focus on patients from `normal studied group`, `intervened risk group` and `control risk group` because only patients of the above-mentioned groups are followed during the period. We get then 899 patients.

In the first experiment (named `Experiment 1`), from the features available in the table `Entry`, we would like to distinguish the patients who died from atherosclerosis from the others. 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 did a similar work in the second experiment (named `Experiment 2`), except that we aim to distinguish patients who came down with the observed cardiovascular diseases from those who stayed healthy (but these patients may suffer from another disease). Following the recommendation given on the web pages (question 5 of the analytic questions), we consider that a patient was affected by a cardiovascular disease when he had an illness based on attributes `HODN1`, `HODN2`, `HODN3`, `HODN11`, `HODN12`, `HODN13`, `HODN14`, `HODN21`, `HODN23` (group of attributes stemmed from the "A_2 `questionnaire`" of the table `Control`). As with `Experiment 1`, we suppose that all patients who suffer from atherosclerosis are recorded in the data base. Finally, there are 281 patients who came down with the observed cardiovascular diseases and belonging to one of the groups of followed patients.

We decided a priori to keep all attributes of the table `Entry`. Nevertheless, we deleted the attribute `KONSUP` (studied group of patients) because its value `normal` `studied group` may introduce a bias and the attributes relating to risk factors (information represented by these attributes is already taken into account by other attributes). 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 its value `true` (i.e. presence of the characteristic) has been kept. The attributes `CHLST` (cholesterol) and `TRIGL` (triglycerides) were segmented in 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. The other continuous attributes (e.g., `VYSKA` (height)) were cut into qualitative attributes, each of these attributes having 3 values with an even number of patients per value. Finally, we get a total of 119 items, each patient being described by at most 37 items.

The first part of Table 2 indicates the characteristics of the obtained data sets. We call `atherosclerosis` the name of the data set of the patients who died from atherosclerosis (`Experiment 1`) or who came down with the observed cardiovascular diseases (`Experiment 2`). `healthy` means the data sets containing the other patients. In both data sets, the attribute `TRIGL` (triglycerides) has 19% of missing values, the attribute `DOPRATRV` (how long to get to work) 7% and the attribute `DOPRAVA` (used means of transport for getting to work) 6%. There are also few missing values on 21 other attributes.

4 Results and discussion

For both experiments, we fixed to 15 the frequency threshold to search SEPs. For each data set, Table 2 gives the threshold of minimum relative frequency (γ) to get a SEP and the numbers of SEPs (with at most 8 items) according to their growth rates (noted GR).

	`Experiment 1`		`Experiment 2`	
	`atherosclerosis`	`healthy`	`atherosclerosis`	`healthy`
No. of patients	124	624	281	618
γ (%)	12.1%	2.4%	5.3%	2.4%
$GR \in [1..2[$	32606	2,278,346	510,901	2,845,756
$GR \in [2..5[$	6254	1,229,359	69609	605,312
$GR \in [5..\infty[$	47	94,921	1038	61168
JEP	132	387,203	2690	16916

Table 2. Numbers of patients, minimum frequencies and numbers of SEPs according to their growth rates

Table 2 summarizes results. It shows that, for the two experiments, the number of SEPs to `healthy` is higher than `atherosclerosis`. This may be explained by the relative frequencies which are lower to `healthy` than `atherosclerosis` or maybe by the fact there are more medical factors to highlight `healthy` than `atherosclerosis`. Even if SEPs are a cover of EPs (see Section 2.3), the total numbers of SEPs remain high and a further work is to extract only SEP with a growth rate specified by the user.

Table 3 (resp. 4) depicts some SEPs from the `Experiment 1` (resp. `Experiment 2`) with the best growth rates and frequencies[2]. The items of an EP are separated by ';'. The first parts of these tables provide EPs to `atherosclerosis` and the second parts to `healthy`. The frequency (noted $\mathcal{F}$) of a SEP to a data set $\mathcal{D}_i$ is its relative frequency with respect to $\mathcal{D}_i$ (for instance, a frequency of 11.3% of a SEP to `atherosclerosis` means that 11.3% of patients of the data set `atherosclerosis` support this SEP).

atherosclerosis		
items of SEPs	GR	$\mathcal{F}$ (%)
the way to work takes around 1 hour ; smoker of 21 and more cigarettes per day ; smoking during 21 and more years ; do not drink liquors	6.71	11.3
weight $\leq$ 74 kg ; blood pressure II diastolic $>$ 92 mm Hg ; normal urine	6.29	11.3
height $\leq$ 1.72 m ; blood pressure II diastolic $>$ 92 mm Hg	3.91	16.9
blood pressure II diastolic $>$ 92 mm Hg	1.72	32.3
age of entry in the study $\in$ [43,47] ; moderate activity after his job ; level of total cholesterol $\geq$ 200 mg/dL	∞	18.5
healthy		
items of SEPs	GR	$\mathcal{F}$ (%)
partly independent worker ; blood pressure II systolic $\leq$ 118 mm Hg	11.7	9.46
reached education: university ; level of total cholesterol $<$ 200 mg/dL	8.35	6.7
age of entry in the study $\in$ [44,47] ; level of total cholesterol $<$ 200 mg/dL	8.15	6.6
age of entry in the study $\leq$ 43 years	2.21	30.3
reached education: university ; blood pressure II diastolic $\leq$ 78 mm Hg	∞	8.7
age of entry in the study $\leq$ 43 years ; mainly standing at work	∞	5.0
non-smoker ; blood pressure I systolic $\leq$ 120 mm Hg	∞	4.5

Table 3. SEPs on `Experiment 1`

In `Experiment 1`, a lot of SEPs to `atherosclerosis` have the item "smoking during 21 and more years" (for instance, 67 JEPs among the 132 include this item). The blood pressure seems to have an important role. All the JEPs to `healthy` have at least two items. In `Experiment 2`, there are 11 JEPs to `atherosclerosis` having 4 items (there are no JEP with less items) and 7 EPs composed of 2 items and having a growth rate greater than 2. Contrary to `Experiment 1`, there is no SEP having a single item with a growth rate greater than 1.7. To `atherosclerosis`, `height` seems to have a role. To `healthy`, there is a single JEP having 2 items (see Table 4).

We can think that most associations highlighting by SEPs are expected (and already known) by physicians. Nevertheless, according to us, an interesting result brought by SEPs is to quantify such associations (how much increases the

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

atherosclerosis		
items of SEPs	GR	$\mathcal{F}$ (%)
1 or 2 cups of coffee per day ; height $<$ 1.72 m ; blood pressure I diastolic $\in$ [75,92] ; skinfold above musculus triceps $>$ 11	7.48	6.0
more than 6 sugar lumps per day ; skinfold above musculus triceps $>$ 11	2.30	8.2
height $\leq$ 1.72 m ; blood pressure I systolic $>$ 135	2.00	14.6
drinking of alcohol: occasionally ; drinking of vine ; up to half a litre of vine per day ; level of triglycerides $>$ 150 mg/dL	∞	14.2
reached education: secondary school ; drinking of vine ; up to half a litre of vine per day ; blood pressure II diastolic $\in$ [78,92]	∞	12.5

healthy		
items of SEPs	GR	$\mathcal{F}$ (%)
single ; do not drink coffee	8.64	3.1
lower limbs pain is non-ischaemic ; blood pressure I diastolic $\leq$ 75 mm Hg	8.64	3.1
mainly walks at work ; drink daily more than 1 litre of beer	5.12	7.3
partly independent worker ; blood pressure I systolic $\leq$ 120 mm Hg ; blood pressure I diastolic $\leq$ 75 mm Hg	5	7.1
drinking of 10° beer ; daily consumption of 2 at 6 sugar lumps ; blood pressure I diastolic $\leq$ 75 mm Hg ; normal urine	∞	7.3
blood pressure II systolic $\in$ [118,138] ; blood pressure II diastolic $>$ 92 mm Hg	∞	3.8

Table 4. SEPs on `Experiment 2`

risk of atherosclerosis with respect to precise features?). Let us note that some surprising SEPs may be discovered like "mainly walks at work and drink daily more than 1 litre of beer" (see Table 4). Is it a bias of this data set or should we advise employees to walk at work to drink beer?

5 Conclusion

We have defined the strongest emerging patterns as the EPs having the best growth rate (they are, a priori, the most interesting EPs for real-world data mining tasks). We have proposed an efficient method to extract the strongest emerging patterns based on the recent results in condensed representations of patterns.

Dealing with the data mining questions of this discovery challenge, we have suggested SEPs characterizing patients with respect to atherosclerosis. More precisely, from the features available when patients are entered in the study, we mined at first SEPs to distinguish the patients who died from atherosclerosis with regards to the other patients, and then, patients who came down with the observed cardiovascular diseases compared to those who stayed healthy. We extracted more SEPs to `healthy` than `atherosclerosis`. Experiments highlighted SEPs with a quite high growth rate and frequency especially in the `Experiment 2` (i.e. patients affected by the observed cardiovascular diseases). We hope that this work may be useful about the study aims of this discovery challenge and we believe that the feedback and discussions with the medical experts during this workshop are of great importance to keep this work so as to improve it.

References

[1] http://lisp.vse.cz/challenge/ecmlpkdd2002/.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. *Fast Discovery of Association Rules*, chapter 12, pages 307–328. AAAI/MIT Press, Menlo Park, CA, 1996.

[3] J. Bailey, T. Manoukian, and K. Ramamohanarao. Fast algorithms for mining emerging patterns. In *proceedings of the Sixth European Conference on Principles Data Mining and Knowledge Discovery, PKDD'02*, volume 2431 of *Lecture notes in artificial intelligence*, pages 39–50, Helsinki, Finland, 2002. Springer.

[4] G. Birkhoff. Lattices theory. *American Mathematical Society, vol. 25*, 1967.

[5] J. F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal*, 7(1):5–22, 2003. Kluwer Academics Publishers.

[6] B Crémilleux and J. F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. pages 33–46, December 2002.

[7] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD'99*, pages 43–52, San Diego, CA, 1999. ACM Press.

[8] G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In *proceedings of the Second International Conference on Discovery Science, DS'99*, pages 30–42, Tokyo, Japan, 1999. Springer-Verlag.

[9] J. Li, G. Dong, and K. Ramamohanarao. Instance-based classification by emerging patterns. 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 191–200, Lyon, F, 2000. Springer-Verlag.

[10] J. Li, G. Dong, and K. Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowledge and Information Systems*, 3(2):131–145, 2001.

[11] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. In *proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD 96*, pages 189–194, Portland, Oregon, 1996.

[12] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.

[13] 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.

[14] F. Rioult and B. Crémilleux. Condensed representations in presence of missing values. In *proceedings of the 5th International Conference on Intelligent Data Analysis (IDA'03)*, Lecture notes in Computer Science, Berlin, Germany, to appear (August 03) 2003. Springer.

[15] R. Wille. Concept lattices and conceptual knowledge systems. *Computer mathematic applied*, 23(6-9):493–515, 1992.

[16] M. Zaki. Generating non-redundant association rules. In *proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, ACM SIKKDD'00*, pages 34–43, 2000.