LABORATOIRE

INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 6070

HASARD: MINING SEQUENTIAL ASSOCIATION RULES FOR ATHEROSCLEROSIS RISK FACTOR ANALYSIS

Laurent Brisson, Nicolas Pasquier, Martine Collard, Céline Hebert

Projet EXECO

Rapport de recherche
ISRN I3S/RR–2004-26–FR

Octobre 2004

RÉSUMÉ :

MOTS CLÉS :
fouille de données, données médicales, motifs séquentiels

ABSTRACT:
We present the HASARD method that is an hybrid approach for extracting adaptative temporal association rules. This method extracts assocation rules between events ccuring in subsequent time-intervals using closed itemsets xtraction and evolutionary techniques. An important feature is its capacity to consider different time-intervals depending on the analysed attribute. This method was applied for the analysis of long term medical observations of atherosclerosis risk factors for cardio-vascular diseases prevention. Experimental results show that it is well-suited for extracting knowledge from temporal data where interesting patterns have different observation period length.

KEY WORDS :
data mining, medical data, sequential patterns

HASARD: Mining Sequential Association Rules for Atherosclerosis Risk Factor Analysis

Laurent Brisson[1], Nicolas Pasquier[1], Céline Hebert[2], Martine Collard[1]

[1] Laboratoire I3S (CNRS UMR-6070), Université de Nice – Sophia-Antipolis, 2000 route des lucioles, Les Algorithmes, 06903 Sophia-Antipolis, France; {brisson,mcollard,pasquier}@i3s.unice.fr

[2] Laboratoire GREYC (CNRS UMR-6072), Université de Caen – Basse-Normandie, UFR Sciences, Campus II - B.P. 5186, 14032 Caen, France; clecharp@etu.info.unicaen.fr

Abstract. We present the HASARD method that is an hybrid approach for extracting adaptative temporal association rules. This method extracts assocation rules between events occuring in subsequent time-intervals using closed itemsets extraction and evolutionary techniques. An important feature is its capacity to consider different time-intervals depending on the analysed attribute. This method was applied for the analysis of long term medical observations of atherosclerosis risk factors for cardio-vascular diseases prevention. Experimental results show that it is well-suited for extracting knowledge from temporal data where interesting patterns have different observation period length.

1 Introduction

In this paper, we consider *sequential association rules* which express casualty relationships between sets of events occuring in subsequent time-intervals. We developed a specific approach, called HASARD for Hybrid Adaptative Sequential Association Rules Discovery, for finding such patterns of interest. Our methodology is applied to an health care problem, but it is also suited to a broad collection of data mining problems where data are temporal observations on individuals, in customer behaviour or credit risk prediction for instance. Many studies have focused on the efficient mining of sequential patterns or patterns in time-related data. Most of them are based on extensions of the APRIORI algorithm [AMS+96] proposed for extracting association rules. The HASARD approach presented in this paper combines techniques for searching closed itemsets, as defined in the CLOSE algorithm [PTB+04], and an heuristic approach relying on a genetic algorithm.

We used the HASARD approach for the analysis of long term observation data in the STULONG dataset. This dataset was constituted in the framework of a longitudinal study of atherosclerosis risk factors to evaluate the impact of non-pharmacological prescriptions on these risks. It contains data collected between 1975 and 2001 on a population of 1 417 men born between 1926 and 1937 in Czechoslovakia. First, an entry examination was performed and data concerning social characteristics, diet, tobacco and alcohol consumption, physical activities, personal anamnesis and, physical and biochemical examinations were collected. During this examination, patients were classified into three groups according to principal atherosclerosis risk factors (RFs): Arterial hypertension, hypercholesterolemia, hypertriglyceridemia, overweight, smoking and positive family case history. These three groups are the following:

– Normal group (NG): No risk factor and no cardio-vascular disease.
– Risk group (RG): Some risk factors and no cardio-vascular disease.
– Pathological group (PG): Cardio-vascular disease diagnosed or cause of death.

During the twenty one years following the patients' entry, control examinations were performed to record changes in diet, smoking habits, physical activities and responsability level in job, sport practice in leisure time and, physical and biochemical measures. For each control, the patient id, the date and the control number were recorded. In 2001, 389 patients were deceased and, the cause and date of death were recorded. For a group of 403 patients that answered to a postal questionnaire, detailled data – similar to those gathered during controls – were collected.
The STULONG dataset was collected at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomeèková, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. 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 (head. Prof. RNDr. J. Zvárová, DrSc). At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107.

Objectives. One of the main objectives of this study was to evaluate the impact of behavioural changes – starting or stopping a diet or sport practice for instance – on the RF and the development of cardio-vascular diseases (CVDs). The initial analytical questions in STULONG evolved after the first experimentations and discussions with medical experts, mainly because of the evolutions of medical knowledge since the beginning of the study and missing data (e.g. uriacid measure is given for less than 10 % of controls). During this work, we have developped a new method that, we believe, will help to answer the following questions defined throughout discussions with medical experts and related to the long-term observations:

– Are there differences between men of the normal, risk and pathological groups from the viewpoint of the impact of behavioural changes on RF and CVD developement ?
– What characterizes men who developed a CVD and those who stayed healthy on the global population and the risk group ?
– Are the education level and the responsability in job good criteria for segmenting patients with perilous or safe behaviours and high or low RF ?

HASARD is an association rules extraction approach incorporating temporal relationships. It extracts *sequential association rules* which, we believe, are well fitted to analyse casualty relationships between behavioural changes and, RF evolutions and CVD development. Association rule extraction, first introduced in [AIS93], aims at discovering casualty relationships between sets of attribute values, called *itemsets*, in large datasets. An example association rule, fitting in the context of market basket analysis is: *buy(cereal)* $\wedge$ *buy(sugar)* $\rightarrow$ *buy(milk)*, *support* = 20 %, *confidence* = 75 %. This rule states that customers who buy cereal and sugar also tend to by milk. The support measure indicates that 20 % of all customers bought both three items and the confidence measure shows that 75 % of customers who bought cereal and sugar also bought milk. Informaly, the support represents the range of the rule and the confidence indicates the precision of the rule. In order to extract only statisticaly significant association rules, only those with support and confidence at least equal to some user defined *minsupport* and *minconfidence* thresholds are generated.

Organization. In section 2, we present preparation and transformation methods applied to the dataset for extracting long-term observation related patterns. Sequential association rules and the techniques we used for their extraction are defined in section 3. In section 4, we show experimental results and section 5 concludes the paper.

2 Data preparation

Data for the long-term observation were collected during the twenty one years controls after the patients' entry. We used both entry and control data to generate multiple datasets that are adapted to the kind of knowledge we were interested in. Attributes of interest were selected and prepared according to discussions with medical experts, for determining threshold values of physical and biological measures for instance.

2.1 Sequential rules and search strategy

Since our main objective concerns the effect of behaviour on risk factor development, we decided to look for sequential rules involving casualty relationships between patient behavioral changes and risk factor changes on subsequent time intervals.
Sequential rules with the form $X \rightarrow Y$ we search for involve both itemsets and time-itemsets. We call *time_item* an attribute value occuring in a particular temporal window. We call *time_itemset* a set of time_items. Our sequential rules have the following structure:

$$IDE\_itemset \ \wedge \ BEH\_time\_itemset \ \rightarrow \ RF\_time\_itemset$$

where the components are:

- *IDE_itemset*: an itemset of static identifiaction attributes,
- *BEH_time_itemset*: a time_itemset of behavioural attributes,
- *RF_time_item*: a time_item on a risk factor attribute.

An example sequential rule may be:

$$ALCOHOL{=}regularly \wedge BEH\_PHA{=}decreased\_sits \rightarrow RF\_CHOLEST{=}increased.$$

Such a rule must be interpreted as a casualty relationship between changes on risk factors occuring on an *observation temporal window* of O months and induced by static data and by changes on behaviour on a previous *action temporal window* of A months. We also defined a *latency temporal window* L between the *action period* and the *observation period* which allows a waiting time to observe the impact of some behavioural changes.
The example rule should be interpreted as follows:
if *the patient regularly consumed alcohol when he entered the study*
and *his physical activity after job decreased over a* A *month period*
then *his cholesterol rate increased at a control which occured* L *months after and over the subsequent* O *month observation time.*
An important element for the method flexibility is that temporal window sizes O, A and L are defined as parameters of sequential rules. The strategy we applied for extracting these rules was first running wide data transformations to tailor data to the specific task. This first step consisted in applying corrections, replacing missing values, creating new attributes and a new table for saving behavioural and risk factor changes and flattening data on changes.

Rule quality is computed according statistical criteria like traditional association rules. Statitical measures are presented in sections 3.1 and 3.3. HASARD flexibility is also provided by the evolutionary algorithm involved for searching for rules. We defined a Genetic Algorithm (GA) for this task. GAs [GR87] are well suited to large combinatorial search spaces. GAs are adaptive procedures that evolve a population of structures in order to find the best individual. The evolution is performed by specific genetic operators like mutation and crossover. They have a long history of being exploited for rule manipulation [Freitas02,SAL03]. As it was suggested, they offer techniques such as niching which allow not only to find the best rule, but a selection of good rules.

2.2 Patient classes

In order to observe potential differences between classes of patients, we distinguished the following groups:

- NG, PG and RG assigned to patients in the STULONG study.
- CVD and NCVD which respectively represent patients who had and did not have a cardio cascular disease during the STULONG study.
- Classes of patients based on their education level and job responsability criteria.

Classes CVD and NCVD were obtained by splitting the EXPERIMENT table using attributes CONTROL.HODNx, with $x \neq 0$ and $x \neq 15$, that indicates if a cardio-vascular disease was diagnosed and DEATH.PRICUMR that indicates if his death was due to a cardio-vascular disease.
Classes of patients based on their education level and job responsabilities criteria were obtained as follows. In a first attempt, we extracted all closed patterns containing at least the social factors but those gathering a relevant number of patients (at least 200) did not reveal a significant medical interpretation. This is due to the fact that some attribute values cover a large number of patients (e.g., 1023 patients among the 1199 are married). After talking with a physician, his main opinion was that "education level" (`VZDELANI`) and "responsibilities in job" (`ZODPOV`) are most likely to influence atherosclerosis. So we decided to start building the clusters from these two attributes. We investigated the closed patterns containing the items coming from these attributes. We got the following 11 closed patterns (or potential clusters):

1. basic school and others (for responsibilities in job)
2. primary school and managerial worker
3. primary school and partly independent worker
4. primary school and others
5. secondary school and managerial worker
6. secondary school and partly independent worker
7. secondary school and others
8. university and managerial worker
9. university and partly independent worker
10. university and others
11. university and pensioner (not because of ICHS)

As values `secondary school` and `university` are very close, we merge closed patterns number 5 and 8 to produce the first cluster. We perform a similar process with closed patterns number 6 and 9, 7 and 10, 1 and 4. The fifth cluster contains the 151 remaining transactions (closed patterns 2, 3 and 11). Finally, we obtain the non-overlapping clusters described in table 1.

	Social description		Number of patients		
Cluster	Secondary school	Responsibility in a job	`healthy`	`atherosclerosis`	Total
1	yes	managerial worker	150	60	210
2	yes	partly independent worker	227	82	309
3	yes	others	127	59	186
4	no	others	221	122	343
5	-	-	94	57	151
		Total of patients	819	380	1199

Table 1. Description of clusters.

2.3 Attributes of change and new tables

A preliminary step consisted in correcting some errors or contradictions and replacing missing values. We replaced both null values and values explicitly by the same value. We built new tables CHANGES and EXPERIMENT more fitted to the task from the initial tables ENTRY, CONTROL. The DEATH table was used in order to split patient set into two classes CVD and NCVD. Discussions with medical experts allowed us to identify some guidelines for building tables and to understand which initial variables we had to keep an which ones we have to build. First, we kept existing identification variables (IDE variables) about patients. These variables were named according to the expression IDE_*variable_name*. They are listed below:

- the age of the patient, his education level,
- the initial group of the patient when he came into the study,
- the alcohol consumption at the beginning of the study since STULONG data do not provide this information for each control.

For informations varying from one control to another, we built variables related to behavioural changes on one hand and variables related to risk factors changes on the other hand. Behavioural change variable (BEH change variable) were named according the expression BEH_*variable_name* and risk factor change variable (RF change variable) were named according to the expression RF_*variable_name*. These attributes are listed in table 2.

Variables for behavioural changes	Variables for risk factors
Consumption of cigarettes a day	Cholesterol level
Physical activity after job	HDL cholesterol level
Different kinds of diet	LDL cholesterol level
Physical activity in job	Triglycerides level
Medecine for cholesterol	Overweight or obesity
Medecine for blood pressure	Bloodpressure measures
	Glycemia level

Table 2. Attributes of change.

BEH and RF change variables were built from attributes of the STULONG tables ENTRY, CONTROL and DEATH. For instance, we show how variables BEH_PHA and RF_CHOLEST are computed. The BEH_PHA variable was deduced from variable CONTROL.AKTPOZAM, which indicates the physical activity after job (see table 3).
RF_CHOLEST was deduced from variable CONTROL.CHLST, which indicates the global cholesterol rate, as shown in table 4.
The new CHANGES table is composed with IDE variables, BEH and RF change variables by using the CONTROL table. This tables contains as many tuples as the CONTROL

Value	CONTROL.AKTPOZAM (N)	CONTROL.AKTPOZAM (N+1)
stay_sits	he mainly sits	he mainly sits
decreased_sits	moderate or great activity	he mainly sits
increased_modest	he mainly sits	moderate activity
stay_modest	moderate activity	moderate activity
decreased_modest	great activity	moderate activity
increased_great	he mainly sits or moderate activity	great activity
stay_great	great activity	great activity

Table 3. Variable *BEH_PHA*.

Value	CONTROL.CHLST (N)	CONTROL.CHLST (N+1)
stay_normal	<6	<6
decreased	≥ 6	<6
increased	<6	≥ 6
stay_high	≥ 6	≥ 6

Table 4. Variable *RF_CHOLEST*.

table. In order to get a temporal description of each patient all along the study, we flattened the CHANGES table. The EXPERIMENT table is the result of the flattening operation; it contains as many tuples as patients in CONTROL. One tuple in EXPERIMENT contains IDE variables of the patient and changes variables for every control the patient passed through.

3 Strategies

3.1 Target model and definitions

We consider the time segment which represents the intervention on a patient from the time he entered the study until the time he left it. Our discussions with medical expert led us to fix a month as the time unit. Let us consider a patient and the time-interval [*In ; Out*] during he was in the STULONG study.
C(T, patient) refers to a control occuring at time T months after *In* for the *patient*. A control *C(T, patient)* is characterized by a BEH_time_itemset and a RF_time_item which represents behavioural and risk factors changes observed for the patient at this control time. A *ControlPeriod CP(T, patient)* is a A-month size time interval [*T ; T+A*] where it occurs one control at least for the *patient*. A *Temporal Configuration TC(T1,T2, patient)* is a time interval [*T1 ; T2*] such as:

- it exists a *ControlPeriod* γ=*CP(T,patient)* with T1 $\in$ [*T ; T+A*],
- T2 $\in$ [*T1+L ; TI+L+O*],
- it occured a control *C(T2, patient)*,
- it did not occur any control in the interval [*T+A ; T2*] for the *patient*.

We say that two Temporal Configurations *TC(T1,T2, patient)* and *TC(T3,T4, patient)* are *compatibles* if $T2 \leq T3$ or $T4 \leq T1$.

Statistical measures. We define the measure *Max_Support* as the whole number of possible compatible *Temporal Configurations* for all records (patients) in the flattened table EXPERIMENT. For a rule antecedent X = *IDE_itemset* $\cup$ *BEH_time_itemset*, we say that a temporal configuration*TC(T1,T2, p) contains* X if:

– IDE_itemset is observed for the patient p,
– BEH_time_itemset is observed over a *ControlPeriod* γ=*CP(T,patient)*.

For a rule consequent Y =*RF_time_ item*, we say that a temporal configuration *TC(T1,T2, p) contains* Y if RF_time_item is observed at control *C(T2, patient)*. For a rule antecedent X, we define the cardinality measure of X, *Card(X)* as the number of possible compatible *TC(T1,T2, p)* which contains X. For a rule consequent Y, we define the cardinality measure of Y, *Card(Y)* as the number of possible as the number of possible compatible *TC(T1,T2, p)* which contains Y.
For a rule $X \rightarrow Y$, we define the cardinality measure of $X \wedge Y$, *Card*($X \cup Y$) as the number of possible *TC(T1,T2, p)* which *contains* X and Y.

3.2 Association rules based approach

Two main approaches for extracting association rules can be distinguished.
In the first approach, all itemsets with *support* $\geq$ *minsupport*, called *frequent itemsets*, are extracted and all association rules with *confidence* $\geq$ *minconfidence* are generated from them. This approach is very efficient when data are weakly correlated, such as market basket data, but performances drastically decrease when data are dense or correlated, such as statistical data for instance. A comprehensive survey of this approach can be found in [AMS+96].
The second approach is based on the extraction of *generators* and *frequent closed itemsets* defined using the Galois closure operator. From these, the *informative basis for association rules* containing non-redundant association rules with minimal antecedent and maximal consequent. This approach both improves the extraction efficiency, by reducing the search-space, and the result relevance, by suppressing redundant rules, in the case of dense or correlated data. A summary of this approach can be found in [PTB+04].

Frequents closed itemsets and generators. Frequent closed itemsets and generators are defined according to the closure operator ϕ of the Galois connection. This operator associates with an itemset l its closure $\phi(l)$ that is the maximal set of items common to all objects containing l. That is, the closure of l is the intersection of all objects containing l. The minimal closed itemset containing an itemset l is its closure $\phi(l)$ and we say that an itemset l is a *closed itemset* if $\phi(l) = l$. The generators of a closed itemset c are the minimal[3] itemsets which closure is c. Generators are the minimal itemsets we can consider for discovering frequent closed itemsets, by computing their closures. Since the support of a frequent itemset is equal to its closure support and since maximal frequent itemsets are maximal frequent closed itemsets, the frequent closed itemsets constitute a minimal non-redundant generating set for all frequent itemsets and thus, for all association rules.
Consider the dataset $\mathcal{D}$, constituted of six objects identified by their *OID* and five items, represented in figure 1(a). The eight generators and five frequent closed itemsets, with their supports, in $\mathcal{D}$ for *minsupport* = 2/6 are given in figure 1(b).
The itemset {A} is the generator of the frequent closed itemsets {AC}: the intersection of all objects containing {A}, that are objects 1, 3 and 4, gives {AC} and no subset of {A} has {AC} as closure. {A} and {AC} both have a support of $\frac{\|\{1,3,4\}\|}{\|\mathcal{D}\|} = 3/6$. The frequent closed itemset {BCE} has two generators: {BC} and {CE}. {BE} is not a generator of {BCE} since it is a frequent closed itemset, {B} and {E} are generators of {BE} and {C} is itself its own generator.

[3] With respect to the inclusion relation.

OID	Items			
1	A	C	D	
2	B	C	E	
3	A	B	C	E
4	B	E		
5	A	B	C	E
6	B	C	E	

(a) Dataset $\mathcal{D}$.

Generator	Frequent closed itemset	Support
{A}	{AC}	3/6
{B}	{BE}	5/6
{C}	{C}	5/6
{E}	{BE}	5/6
{AB}	{ABCE}	2/6
{AE}	{ABCE}	2/6
{BC}	{BCE}	4/6
{CE}	{BCE}	4/6

(b) Extraction result.

Fig. 1. Generators and frequent closed itemsets.

3.3 Evolutionnary approach

Genetic Algorithms are robust, flexible algorithms which tend to cope well with attribute interaction in atherosclerosis data. Furthermore, the comprehensibility of the discovered knowledge is important and GAs allow us to extract comprehensible rules evolving populations of prediction patterns. Our GA implementation uses EO, a templates-based, ANSI-C++ compliant evolutionary computation library.

Genome. The first issue in designing a GA is how to encode each individual in the population. To represent variable length rule we use a fixed-length genome which contains a gene for each IDE_attribute and BEH_attribute and another gene for one of the RF_attributes. Each gene contains three elements: attribute name, attribute value and an activation flag indicating whether or not an item in the rule is associated to the gene.

Generation of initial population. The method used to generate the initial population is based on CLOSE algorithm results. CLOSE generate a list with generators and their frequent closed itemset associated, that allows us to initialize the population in three steps:

- rule antecedents are created from generators. Generators containing only RFs are skipped,
- rule consequents are created with the first RF found in the frequent closed itemset or the generator. If no RF is found, we generate randomly one.
- for each atribute not represented in generators the value of the corresponding gene is randomly defined and the activation flag is set to false.

Genetic operators. We use the tournament selection with size of 2. The selection is deterministic, starting from the best ones down to the worse ones. If the total number to select is less than the size of the source populations, the best individuals are selected once. If more individuals are needed after reaching the bottom of the population, then the selection starts again at top. It the total number required is N times that of the source size, all individuals are selected exactly N times. For replacement we use the most straightforward method, called generational replacement where all offspring replace all parents; however weak elitism is used.
New patterns are generated by combining existing patterns using a crossover operator or by modifying existing patterns via a mutation operator. Crossover is a recombination operator that swaps genetic material between two individuals. We used a one point crossover method. Three mutation operators were used: first one simple changes the

attribute of a gene with a random attribute, the second generates a random transition value (domain of this value is a parameter of the GA) which is randomly added or subtracted to the current gene value and the last inverts the current value of the activation flag. All of the mutation rates can be define independently.

Fitness function. A crucial issue in the design of a GA is the choice of the fitness function. In a first approach we only consider support to select the most frequent rules, confidence to consider reliable rules and lift to ensure a high level of dependence between antecedent and consequent part of a rule. To evaluate the quality of a $X \rightarrow Y$ rule, our GA applies the fitness function on the individual associated to the rule.

$$Fitness(rule) = Support(rule) * Confidence(rule) * Lift(rule) \quad (1)$$

$$Support(rule) = \frac{Card(X \cup Y)}{Max\_Support} \quad (2) \qquad Confidence(rule) = \frac{Card(X \cup Y)}{Card(X)} \quad (3)$$

$$Lift(rule) = \frac{Card(X \cup Y)}{Card(X) * Card(Y)} \quad (4)$$

4 Experimental results

4.1 Patients classes comparison

Patient groups. The experience consisted in extracting rules on PG and testing them on NG and RG. One may observe that the best rules found on PG are not valid on NG. Most of them have a good support but a weak fitness on RG. Thus these results show quite different relationships between the patient behaviour and their risk factors among initial groups.

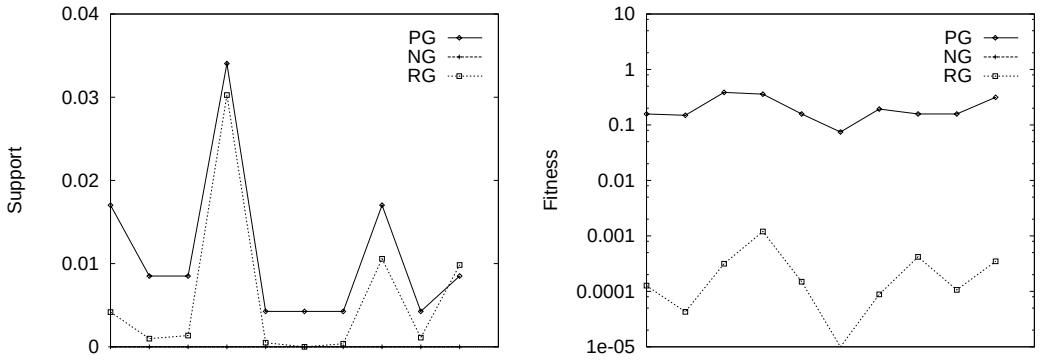


Fig. 2. Best rules on PG versus NG and RG.

Cardiovascular disease. The experience consisted in extracting rules on CVD and testing them on NCVD. One may observe that the best rules found on CVD have similar support and fitness on NCVD. These results seem to show that relationships between behavioural changes and risk factors are not really different between CVD and NCVD patients.

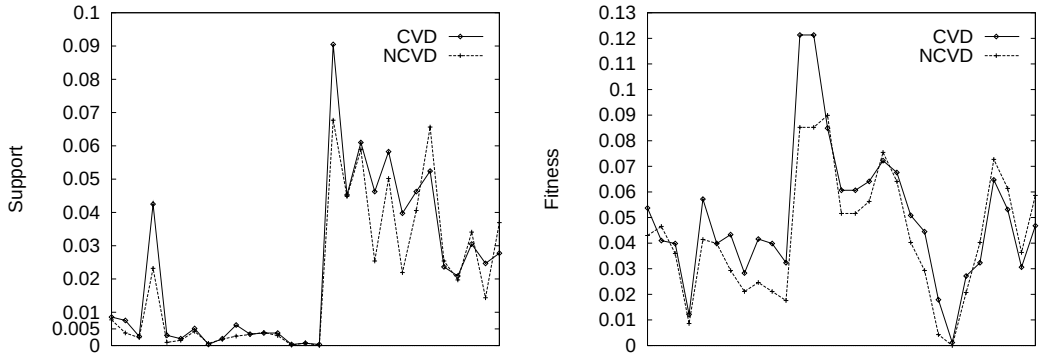


Fig. 3. Best rules on CVD versus NCVD.

Social classes. The experience consisted in extracting rules on Cluster1 and testing them on Cluster3 and Cluster4. One may observe that best rules found on Cluster1 stay good on Cluster4. Cluster3 does not give too much different results. Thus it seems that social factors like education level and job responsability do not allow to distinguish different behaviour related to cardiovascular risks. distinguish different behaviour against cardiovascular.

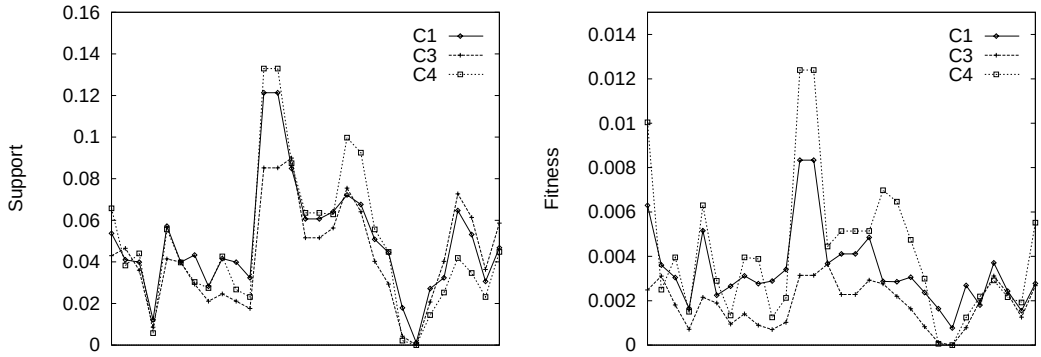
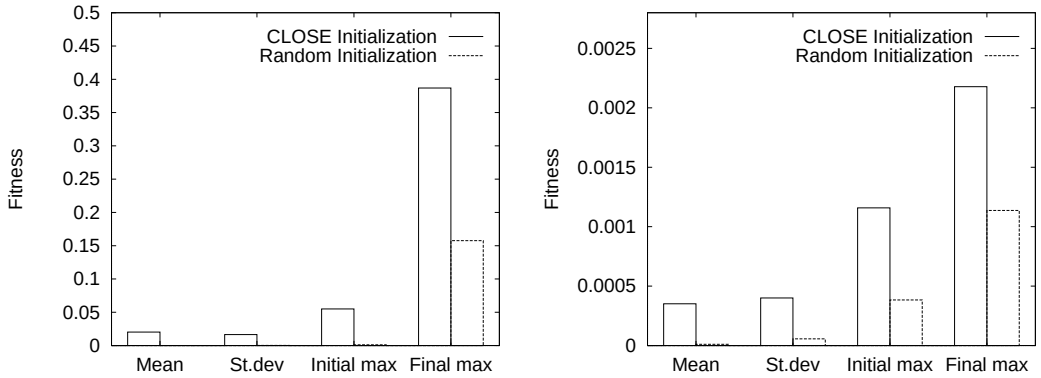


Fig. 4. Best rules on Cluster1 versus Cluster3 and Cluster4.

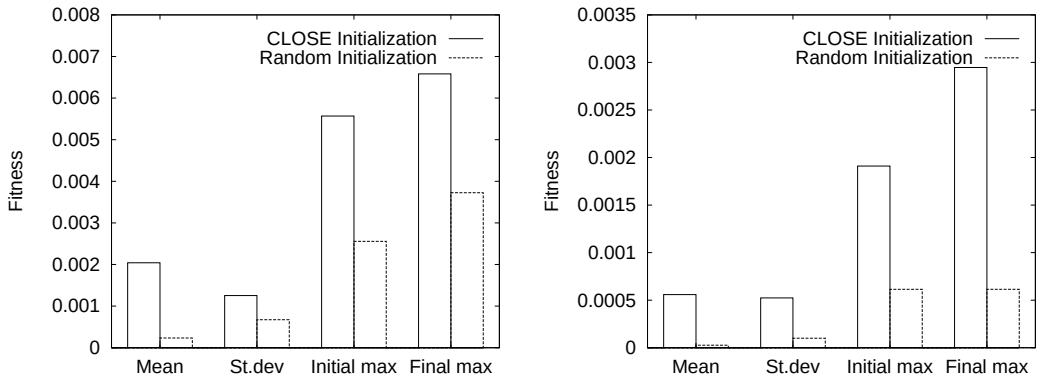
4.2 Initialisation methods

In order to evaluate performances of our initialization method we compare it with a random initialization method. Results on PG and RG groups and on CVD and NCVD classes are shown in figure 5 and 6 respectively. The first three columns show statistics about initial populations and the last column show the fitness of best individuals after a GA run. We can observe that mean fitness of populations generated using CLOSE is 8.75 to 400 times better than those of randomly generated population. It is not surprising since CLOSE optimize rules support and confidence, two mains criteria of our fitness function. However, it is interesting to note that after a GA run the fitness of the best offspring of populations generated using CLOSE is 1.55 to 4.79 times better than the best offspring of randomly generated population. Furthermore, an analyze of rules show that GA doesn't converge toward local optima given by CLOSE. Then we succeed in improving GA performances conserving diversity in solutions proposed.



(a) PG (b) RG

Fig. 5. Comparing initialization methods on patient groups.



(d) CVD (e) NCVD

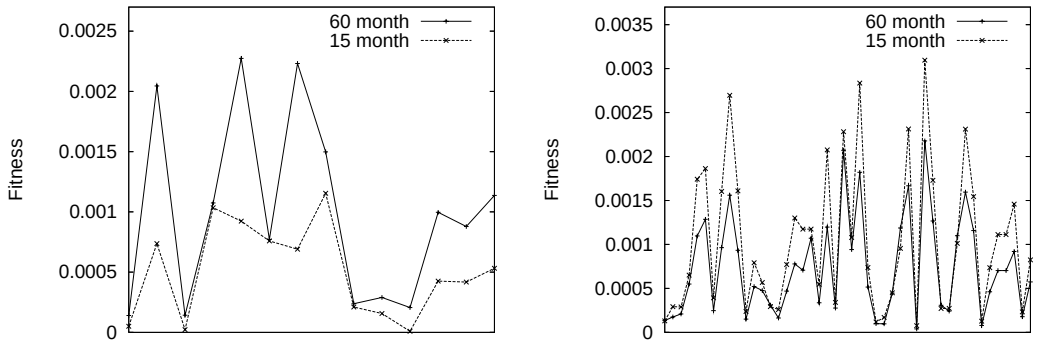
Fig. 6. Comparing initialization methods on patient clusters.

4.3 Observation time windows variations

Time windows A, L and O, used to extract antecedents and consequents of rules, define *time_itemsets*. An important aspect of the HASARD method is its capacity to make vary these time windows. In the STULONG analysis, the observation time of RF evolutions depends on the considered risk. For instance, according to physicians' knowledge, the effects of a diet on the weight are perceptible after a few months whereas the effects on the cholesterol measures are most often perceptible after a longer period.
We evaluated the effect of RFs observation time window variations on the fitness of rules for two RFs: RF_CHOLEST and RF_BLOODPRESS. The results are shown in figure 7. For RF_CHOLEST, rules generated for a 60 months window have a much better fitness than those generated for a 15 months window. For RF_BLOODPRESS, the situation is the opposite: a 15 months window gives better fitness than a 60 months window. This shows that effects of non pharmacological prescriptions on hypercholesterolemy must be observed on much longer time period than effects on arterial hypertension.

5 Conclusion

In this paper, we have presented an innovative method for extracting adaptative sequential rules. We have applied the method on the atherosclerosis STULONG dataset. While



(a) RF_CHOLEST (b) RF_BLOODPRESS

Fig. 7. Effects of observation time window length variations on rules fitness.

previous works on this dataset essentially focused on static information from the ENTRY table, we have investigated the analysis of temporal data in the CONTROL table. Our approach is based on two main points:

- a set of data transformations which is suited to various temporal data,
- an hybrid strategy which combines advantages from quite different techniques: an exhaustive search for frequent closed itemsets which are used as the initial population of an evolutionnary algorithm.

Experimental results allowed to point out different tendencies among patient groups and confirmed prior medical knowledge. In order to answer the analytical questions, a complete analysis of the sequential rules with the assistance of medical experts will be required.
In the future, we plan to apply the HASAR approach to other temporal datasets and to extend it by integrating background knowledge, such as medical ontologies, in the search process.

References

[AIS93] Agrawal R., Imielinski T. and Swami A. Mining Association Rules Between Sets of Items in Large Databases. *Proceedings of the SIGMOD conference*, pp 207–216, May 1993.

[AMS+96] Agrawal R., Mannila H., Srikant R., Toivonen H. and Verkamo A.I. Fast Discovery of Association Rules. *Advances in Knowledge Discovery and data mining*, pp 307–328, Fayyad U.M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R. editors, AAAI Press, 1996.

[PTB+04] Pasquier N., Taouil R., Bastide Y., Stumme G. and Lakhal L. Generating a Condensed Representation for Association Rules. *Journal of Intelligent Information Systems*, Kerschberg L., Ras Z. and Zemankova M. editors, Kluwer Academic Publishers, to appear.

[Freitas02] Freitas A.A. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer-Verlag, Berlin, 2002.

[GR87] Goldberg D. E., Richardson J. Genetic Algorithms with Sharing for Multimodal Function Optimization. Proc. Int. Conf. on Genetic Algorithms (ICGA-87), 1987, 41-49.

[SAL03] Sebag M., Azé J., Lucas N. ROC-based Evolutionary Learning: Application to Medical Data Mining. Artificial Evolution'03, p. 384-397. Springer Verlag LNCS 2936.