An approach to the classification and Knowledge Extraction via microarrays by using Formal Concepts Analysis

Haifa Ben Saber, Farah Harrathi, Mohamed Mohsen Gammoudi

An approach to the classification and Knowledge Extraction via microarrays by using Formal Concepts Analysis

[1]H. Ben Saber, [2]F.Harrathi, and [3]M. M. Gammoudi

Abstract—the classification and extraction of knowledge resulting from microarrays are a growing field which gave rise to several research tasks, however these works have not been able yet to be freed with the problems dependant between relations of all genes implied in breast cancer's disease. Moreover, the suggested methods generated some limiting dependency on the format of the treated data and extraction knowledge method. In this article, we present a method of classification and extraction of knowledge to handle data resulting from the microarrays. This method makes it initially possible to classify genes or the patients who have similar profiles of expression, and then to interpret the groups obtained. It is about a coupling between an algorithm of classification and an algorithm of extraction's rules associations. This work was validated by an expert. The obtained results seem to be interesting and promising owing to the fact that extracted knowledge can be used in the clinical medium to discriminate a disease or the state of a patient.

Index Terms— microarrays, classification, association

I. Introduction

The bioinformatics is a domain in full rise ; it results from he need to apply algorithms of biological data to analyze hem[7]. Thus, the principal bioinformatics task is to manage, exploit and analyze the production of voluminous genomics data [13]. In our work, we are interested in the data analysis of genes or patients. The data used to carry out our analysis result from the microarrays: a technology simultaneously to analyze the activity level and the interactions of thousands of genes [15]. The exploitation of the voluminous data by means of the computer programs is essential to extract interesting knowledge.

II. Problems

The traditional data-mining is used to extract data from the significant clustering of the attributes (columns) or objects (lines) or bi-clustering. This task gave place to many works [5, 3, 12, 1, 16, 10, 14, 6]. Nevertheless, this work is not adapted to data of certain contexts to know disproportionately amongst attributes in front of the objects number where we find few columns only comparing lines, which is atypical in the field of data-mining [4].

H.Ben Saber is with the Informatics Department, Hight School of Sciences and Technologie, University of Tunis, 5 Av Taha Hussein, Tunis 1089, Tunisie

F.Harrathi is with the Informatics Department,ISSAIT, University 7th November

M.M.Gammoudi with the Informatics Department ISSAIT, University 7th November

III. Related Work

Multiple methods of classification were proposed to analyze the data resulting from microarrays presenting some biological limits. Indeed, it proves that the discovery of the relations between all genes is partial [15], moreover the associations between genes is not solved. From data processing point of view, for:

- The statistical methods, although they are rather effective, certain works [1, 2] show that the tests on a reduced number of samples do not give satisfactory results for the analysis of differentially expressed genes.
- The supervised classification methods, the method of K neared neighbors is not directly applicable to more than two classes[10]. Moreover, the centroid classification method is sensitive to the aberrant values and disturbed from the microarrays [16, 10]. Then for the linear discriminating analysis method, it is not very effective and sensitive corresponding to data to treat [6].
- The non supervised classification methods, the hierarchical classification method involves the main issue considering the interpretation of a graph as a delicate task [7]. Moreover, the choice of a distance from similarity is not permitted [14]. To avoid the limits of classification methods, we propose a classification and extraction of knowledge's method using Formal Concepts Analysis.

IV. Mathematical foundation used for the classification and the extraction of knowledge

A. Binary relation

A binary relation is a subset of the Cartesian product $E \chi F$. An element of a relation R is noted (x, y) , x is said an argument of R, and y is an image of x by R. We note $x R y$ the fact that an element x of E is related to an element y of F [8].

B. *Formal context*

Let (E, F, R) be a triplet where E is a set of objects, F is a set of attributes and R is a binary relation between E and F called relation of incidence of C where $R \subseteq E \times F$. A couple $(x, y) \in R$ means that $x \in E$ has the attribute where $y \in F$. In a context's extraction, the objects are indicated by numbers and the attributes by chars. We define two functions f and g, $f(x_i) = \{y \in F / \forall x \in X_i, (x, y) \in R\}$ and $g(y_i) = \{x \in E / \forall y \in Y_i, (x, y) \in R\}$ which form the Galois Connexion between the parts $P(E)$ and $P(F)$ [8].

C. *Formal concept*

Let R be a binary relation. Let A, A', B, B' be some subsets such as $A \subseteq E$, $A' \subseteq E$ and $B \subseteq F$, $B' \subseteq F$. We say that the paire (A, B) is a concept if and only if there exists (A', B') such as: $A \chi B \subseteq A' \chi B' \subseteq R \Rightarrow A = A'$ and $B = B'$. For a context (E, F, R) the paire (A, B) such as $A \subseteq E$, $B \subseteq F$ is a concept if $f(A) = f(B)$ and $g(B) = A$; with f and g are the Galois Connexion. A is called the extension; B is called the intension [8].

D. *Galois lattice*

Let R be a binary relation, T is the whole of extracted concepts from this relation. $(T, \leq\leq)$ is a Galois lattice. In other words, the Galois lattice T for a binary relation R is the whole of all extracted concepts from this relation of partial order $\leq\leq$ [8].

E. *Order relation*

Let $RE_1 = (A_1, B_1)$ and $RE_2 = (A_2, B_2)$ be two concepts of R, $(A_1, B_1) \leq\leq (A_2, B_2) \Leftrightarrow A_1 \subset A_2$ and $B_1 \subset B_2$. $\leq\leq$ is a partial relation[8].

F. *Graph of concepts*

Let R be a binary relation definite on the sets E and F. let G a set of optimal concepts obtained from the conceptual decomposition of R. $(G, \leq\leq)$ forms a graph of concepts. The obtained concepts are a subset of covered Galois Lattice. If G would be the set of concepts, $(G, \leq\leq)$ form a galois lattice [8].

G. *Definitions related to associations rules*

An association rule allows correlations between the elements of a formal context which are related. The rule $X_i \Rightarrow X_j$ occurs in the set of transactions with a support S and a confidence C such as: The support is the percentage of transaction which contains $X_i \cup X_j$, called P, where : $Support(X_i \Rightarrow X_j) = P(X_i \cup X_j) = S$ and confidence is the percentage of transaction containing X_i which also contains X_j, called P, where : $Confidence(X_i \Rightarrow X_j) = P(X_j / X_i) = C$. The rules having the form $X \Rightarrow X_j$ are known as relevant if Support $(X_i \Rightarrow X_j) \succ MinSupp$ and Confiance $(X_i \Rightarrow X_j) \succ MinConf$ such as $MinConf$ and $MinSupp$ are minimal thresholds. Association rules are said to be valid if the measurement of confidence is higher or equal to a minimal threshold, which is called MinConf. Two types of association rules are related to the measurement of confidence. Thus, a rule is called exact if $Confidence(X_i \Rightarrow X_j) = 1$, and a rule is known as approximate if $Confiance(X_i \Rightarrow X_j) \prec 1$: [11, 9]

V. CONTRIBUTION

In this section we present some suggestions to explain the process of coupling algorithms "DCRB - INC" and "AprioriT ID" method.

A. *Structure of the method*

Our method is composed of four stages (Fig 1), namely the matrix transposition, the discretization of the selected data, the extraction of concepts from the binary relation and the generation of association rules.

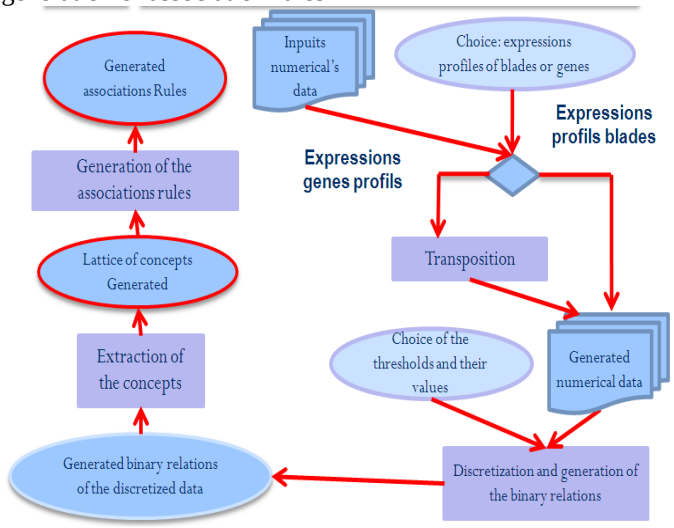

Fig.1. General principle of EC-ACF@DNA-Array

Through this method, we initially compute many lines and columns, we use the transposed matrix in the objective to use a reduced number of columns compared to the number of lines [2]. We generate in the second time associations between genes or patients by using an algorithm of conceptual graph generation covering the associated binary relation. After that, we apply an algorithm generating the association's rules to these concepts.

a- Data organization

We present an example of microarrays extraction context. G is the set of L genes (itemsets) presenting columns of the matrix called concept intension. C is the set of N conditions (objects) presenting line set called the concept extension. R is

a binary relation between C and G. This step consists in extracting from knowledge of transposing matrix instead of making the extraction starting from original data to control proportionately between lines and columns.

b- Data selection

The selection includes the discretization which consists in transforming the numerical data in a binary matrix. The form of discretization is established after having selected the data input, by using intervals of discretization published by the expert. As the end result depends on the choice of the discretization threshold, we left this task to the expert specialized in the domain. To classify genes, we suppose that each column represents a gene and that each line gives the results of gene level experiment measures .The genes which are over-expressed; in other words those which have important biological activities. For our study, the genes which are over expressed in a biological situation are coded by the Boolean value 1, and the others, are coded by the Boolean value 0. The obtained matrix form a formal context of the genes G set and the conditions C set.

c- Graph of concept extraction

We generate according to algorithm DCRB- INC, the graph of concepts which represents in our case sub lattice of concepts

d- Generation of the association rules

In the context of the microarrays extractions, we use the rules of association to identify groups of co-expressed genes or classifications of blades. Let $G = \{1,.....,L\}$ be the set of genes (itemsets) and $L = \{1,.....,N\}$ a set of conditions (objects). For our study case, a reason is an itemset presenting a subset of G. A reason $I \subseteq G$ is included in an object $J \in C$, if I and J are in relation by binary relation R. Let an itemset $I \subseteq G$, the support of itemset I in the formal context $C = (C, G, R)$ be defined as follows : $Supp(I) = Card(I) / Card(C)$.

B. Classification and generation of the association rules prototype

This section presents the software components developed through a process of knowledge extraction applied to a case study which is Breast cancer. Through this prototype, we show the contributions of this practice on data containing a large number of attributes, like the data of genome. With this intention, we treat a data file rather large compound of 77 blades of 9216 genes of breast cancer [15]. To meet this need, we implemented a prototype, called EC -ACF@DNArray.

a- Management of data

We use a form to be filled with the names of the input files (data file, file of the gene names, and files of the blade names), the values of the thresholds to separate the levels from kind of genes and the choice of the expression profile from genes or of the blades. Indeed, to manage the diagonal data of our case study, these data are standardized and cleaned for the study of the case of expression gene in two different cultures. We point out that the first culture (of reference) is marked with the green fluorochrome (CY_3) and that of the second culture (of test) is with the red fluorograme (CY_5): The level of gene expression is determined by the $\log_2$ of the luminous intensities between red fluorescence (R) and green fluorescence (V) report's $M = \log_2(R/V)$. Thus, we start from a collection of L genes in N experimental different measurement forming a matrix from profile of gene expression. We determine the level of gene expression change between two samples CY_3 and CY_5. For example, a spot representing a gene in sample CY_5 must be controlled twice as much as same gene in sample CY_3, what amounts saying that the $\log_2$ of the report is equal to 1 or to -1. Therefore, if $M_{ij} \geq 1$, gene i is called over regular under the condition j associated with culture CY_5. If $-1 \leq M_{ij} \leq +1$, gene i is called unregular under the condition j associated with culture CY_5. If $M_{ij} \leq +1$, gene i is known as lower regular under the condition j associated with culture CY_5.

b- Discretization

This stage consists in generating of the context binary contexts structure, by carrying out the choice of the thresholds according to the biological instructions. We present in the following figure a formal context (Fig 2) to treat patients.

Fig.2. Generated binary context

It is the same to give of a matrix which refers to the genes and which represents the transposed matrix.

c- Classification method of the microarrays data

The EC-ACF@DNArray proposal is effective research tools genes and blood groups. During the development of our prototype DCRB-INC, we generated for a context (of 125 Objects and 25 attributes) 100 concepts (Fig 3) and 246 associations' rules. In our case study, a bi-cluster is a node in

the Galois lattice. The gene group having the same profile of expression will be represented in the same one under tree of the lattice. It defines a group of co-regulated genes (resp. for patients) (Fig 4).

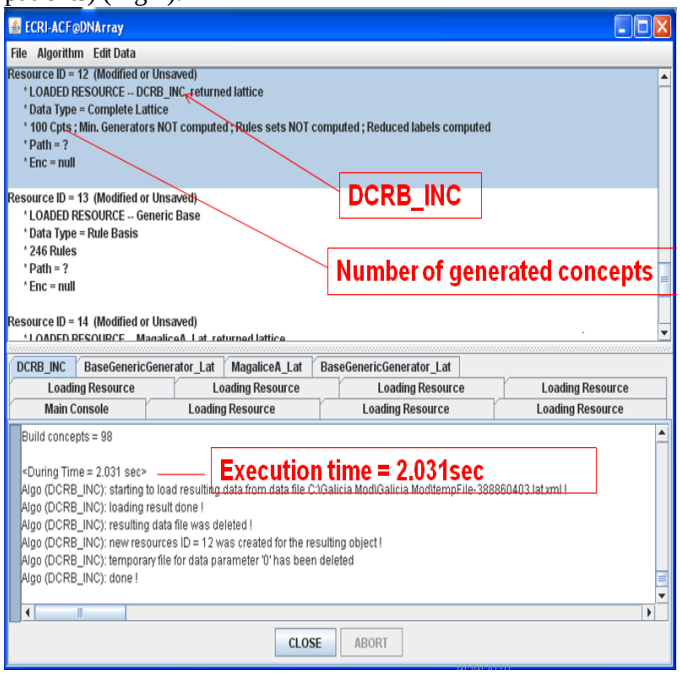

Fig.3. Resulting binary context

According to the obtained results, we confirm that we could minimize the concept number presented by the nodes compared to the other methods. At this level, we could reach our primary purpose which consists in generating the gene classes or the patient classes to have a certain gene characterization or patients. This characterization facilitates to the specialised in this field the making of certain decisions, such as the setting of the patients under the same treatment, or the application of certain analyzes for co-expressed gene units. To give more significations to the generated classes, we propose to release the whole association rules. It is the extraction of relevant knowledge starting from the microarrays.

d-The results visualization by generating the association rules

In our approach, we work on scanned data, quantified, and deferred on a matrix level gene differentially expressed. With this intention, we carry out gene regroupings and/or samples according to their expression profile. In figure (Fig 5), we present the association's rules generated by the blades profiles where each gene plays the part of an object and each sample represents an attribute of gene.

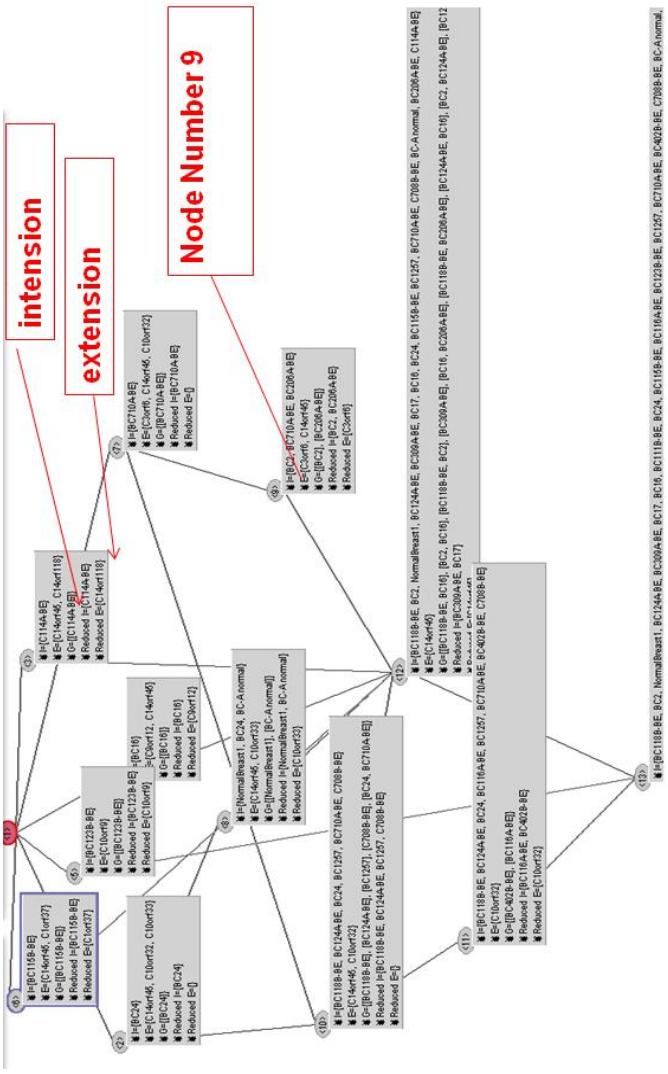

Fig.4. Hasse Diagram generated to classify blades

.

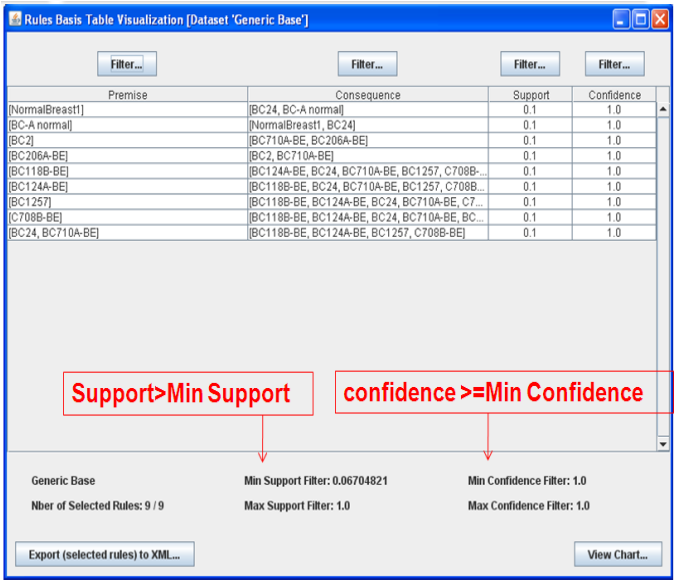

Fig.5. Associations rules of the classified blades

We can also express the results obtained in the form of histogram (Fig 6). Moreover, the association rules generated by the gene profiles detect the whole of Co-on-expressed genes. For our case study, we notice that genes BRCA1, BRCA2, BRCATA, BRCA3, BWSCR1A, TP53 BRIP1, RB1CC1, RAD51, CHEK2, BARD1 are implied in the disease of breast cancer. Moreover, genes PPM1D, PIK3CA, AKT1, PALB2, CASP8, TGFB1, NQO1, HMMR, PTEN, STK11, ATM, NCOA3, ZNF217 are also implied in the disease but with a low support of confidence.

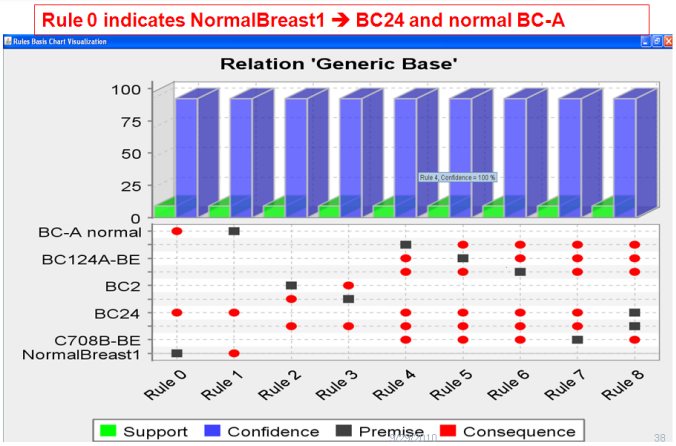

Fig.6. Illustration of the rules

C. *Results obtained*

To test our method, we developed a prototype where we examined data resulting from the micro arrays. The input of our prototype is the rough raw data files and the choice of the thresholds applied to the matrix, namely the profile of expression of genes or profile of expression of blades. The prototype discretizes the data, selects the data in which the specialist is interested then extracted the concepts by applying algorithm DCRB-INC and generates the association rules. Indeed, a concept includes genes which are over-expressed in the same experiments. If the appearing experiments in the extension share same properties, so again, we can make assumptions on genes possible. Moreover, one gene (or an experiment) can appear in several concepts (in opposition to what occurs the" clustering in the case of"). If the biologist is interested in a particular gene, we can thus study according to the biological situations which are the genes related to this one. Currently, the results which we obtained seem to be promising. In fact, our classification enabled us to raise what follows : the obtained classes by our method and organized in the lattice form the interpretation of our results make easier by the biologist thanks to the lattice structure. These results appear important the fact that such an information can be used in the clinical medium.

VI. Conclusion

In this paper, we presented our method of knowledge generation considered to be relevant by the expert starting from the data resulting from the microarrays. The suggested method generates conceptual graph either for the blades (individuals, or experiments), or for genes, as it can generate the whole of the association rules. To achieve this purpose, we selected data to which we applied a thresholding to release the binary context, and association rules relative to this concept lattice. We evaluated our work by making tests on real data files provided by Pasteur institute of Lille 2. Moreover, the interpretation of the results seems to be more reliable compared with former works.

Future works

This work permits to consider several prospects. Indeed, it would be interesting to finalize this work on other biological data by exploiting other diseases and while referring large number of accessible knowledge on the Web by experts, annotated by a several properties and this by considering scaling. Moreover, we propose to cover the subject of classification under-constraints to take account of the constraint types with the level of individuals such as the constraint" must-link" (two individuals must be in the same class) and the constraint"cannot-link" (two individuals should not be placed in the same class). Finally, to improve the quality of analysis, while considering the semantic aspect between the objects, what could play a big role in the classification and classes modeling to improve quality of the partition in clusters objects.

References

[1] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma,I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet,T. Tran, X. Yu, et al. , "Distinct types of 6di_use large B-cell lymphoma identi_ed by gene expression pro_ling. Nature,"403(6769) :503_511,2000.

[2] R. Bendaoud, Y. Toussaint, and A. Napoli., "Hié-rarchisation des règles d'association en fouille de textes. ,"Arxiv preprint cs/0510037, 2005.

[3] F. Bertucci, B. Loriod, R. Tagett, S. Granjeaud, D. Birnbaum, C. Nguyen, and R. Houlgatte. , "Puces à ADN : technologie et applications," Bull Cancer, 88(3) :243_252, 2001.

[4] S. Datta and S. Datta, " Comparisons and validationof statistical clustering techniques for microarray gene expression data,"Bioinformatics,19(4) :459, 2003.

[5] S. Dr ghici, "Data analysis tools for DNA mi-croarrays,"CRC Press, 2003.

[6] S. Dudoit and J. Fridlyand. , " A prediction-basedresampling method for estimating the number of clusters in a dataset,"Genome Biology, 3(7), 2002.

[7] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, " Cluster analysis and display of genome-wide expression patterns,"Procee-dings of the National Academy of Sciences,95(25) :14863, 1998.

[8] MM Gammoudi, "Décomposition conceptuelle des relations binaires et ses applications,"Ha-bilitation en Informatique, Faculté des Sciences de Tunis, 2005.

[9] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules,"Database Theory ICDT 99, pages398_416, 1999.

[10] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee,C.H. Yeang, M. Angelo, C. Ladd,M. Reich, E. Latulippe, J.P. Mesirov, et al. Multiclass, "cancer diagnosis using tumor gene expression signatures,"Proceedings of the National Aca-demy of Sciences of the United States of Ame-rica, 98(26) :15149, 2001.

[11] A. Savasere, E. Omiccinski, and S. Navathe, "An e_cient algorithm for mining association rules in large databases,"In Proceedings of the Inter-national Conference on Very Large Data Bases, pages 432_444. iteseer, 1995.

[12] D.K. Slonim, "From patterns to pathways : gene expression data analysis comes of age,"Nature genetics, 32 :502_508, 2002.

[13] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, " A comprehensive evaluation of multicategory classi_cation methods for microarray gene expression cancer diagnosis,"Bioinformatics, 21(5) :631, 2005.
[14] M. Steenman, G. Lamirault, N. Le Meur,M. Le Cun_, D. Escande, and J.J. Léger. Distinct, "molecular portraits of human failing hearts identi_ed by dedicated cDNA microarrays,"Eu-ropean Journal of Heart Failure, 7(2) :157, 2005.
[15] T. Trang, N.C. Chi, and H.N. Minh, "Data mining of gene expression microarray via weighted pre_x trees," Advances in Knowledge Discovery and Data Mining, pages 21_31, 2005.
[16] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo, "Validating clustering for gene expression data,"Bio-informatics, 17(4):309, 2001.

(1) Haifa Ben Saber, Is a member of the team of research : UTIC, prepares a subject of thesis at the high School of Sciences and Technologies of Tunis, University of Tunis. bensaberhaifa@yahoo.com

(2) Farah Harrathi is an associate Professor, Faculty of Sciences of Gafsa, University of Gafsa, member of RIADI Laboratory, Knowledge Extraction and Information Retrieval

(3) Mohamed Mohsen Gammoudi, is an associate Professor in Computer Sciences at ESSAIT, High School of Statistics and Information, November 7th University, Tunisia, member of RIADI Laboratory, Knowledge Extraction and Information Retrieval gammoudimomo@yahoo.com