World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:8, 2015

Semi-Automatic Method to Assist Expert for Association Rules Validation

Amdouni Hamida, Gammoudi Mohamed Mohsen

***Abstract*—**In order to help the expert to validate association rules extracted from data, some quality measures are proposed in the literature. We distinguish two categories: objective and subjective measures. The first one depends on a fixed threshold and on data quality from which the rules are extracted. The second one consists on providing to the expert some tools in the objective to explore and visualize rules during the evaluation step. However, the number of extracted rules to validate remains high. Thus, the manually mining rules task is very hard. To solve this problem, we propose, in this paper, a semi-automatic method to assist the expert during the association rule's validation. Our method uses rule-based classification as follow: (i) We transform association rules into classification rules (classifiers), (ii) We use the generated classifiers for data classification. (iii) We visualize association rules with their quality classification to give an idea to the expert and to assist him during validation process.

***Keywords*—**Association rules, Rule-based classification, Classification quality, Validation.

I. Introduction

THE major problems related to the association rules extraction are their redundancy, their large number and their relevance level. Several studies tried to solve the problem of association rule's relevance by proposing quality measures in the objective to help the expert to validate them. Two types of measures were proposed: objective measures and subjective measures [33], [26].

The objective quality measures depend only on the nature of the input data from which the rules are extracted. This evaluation technique is based on the following principle: (i) the expert choose the measure and fix the associated threshold. (ii) Rules having a value greater or equal to the threshold are retained [18].

We observe that there are three problems related to the use of objective quality measures such as:

1) The quality measures are expressed by formulas which aren't generally easy to be understood by experts.
2) The arbitrary choice of their thresholds may not cover the expert interest domain.
3) The number of rules remains important even using objective quality measures.

For these reasons, the subjective measures are proposed in [35], [18], [12], [5], [4], [25], [14], [11]. In fact, in these works, authors notice that subjective measures include two subcategories.

The first one is to provide interactive exploration systems of rules for the expert [23], [6]. In this case two limitations have been observed. The first one is related to the way of displaying rules. In fact, textual display of association rules makes their interpretation difficult especially when the number of rules is high [33]. The second one is related to the limited number of objective measures offered to the expert, indeed, in [18], the authors mention that it is useful to choose the objective measure with respect to input data, since each one has specific characteristics and the right choice leads to good results. Thus, to facilitate the expert mission to make the right choices of association rules for validation, the researchers [35], [4], [11] consider that visualization systems, which are the second subcategory of subjective measures, can be one of the most suitable solutions. However, it is very difficult to assimilate the visualization of all rules, especially if the number of their attributes is important. In addition, most of these systems require that the expert must understand the statistical tools to interpret the results of the visualization. To contribute for solving these problems, we propose in this paper, a method of semi-automatic validation exploiting the rule-based classification [27], [24], [10], [26], [8]. In fact, it is proved in several works such as: [8], [22] that the classification based on association rules has high classification precision and strong flexibility to handle unstructured data compared to traditional classification methods. Thus, the quality of classification can support expert opinion in order to avoid their evaluation and validation one by one.

The rest of this paper is organized as follow: In the second section we recall some basic concepts necessary to understand our method. In the third section, we present a related works of rule-based classification method. The fourth section is devoted to the presentation of our contribution. In the fifth section, we illustrate our method by a training set used by the domain community. In sixth section, we present our prototype and some experiments for validate the proposed method. Finally, we end with a conclusion and some perspectives.

II. Basics Notions

In this section, we present some preliminary notions related to Formal Concept Analysis [13], association rules and classification rules.

A. Formal Concept Analysis

1) *Formal context*: let (O, I, R) be a triplet with O and I are respectively sets of objects, sets of items and $R \subseteq O \times I$ is

Amdouni Hamida is with ESEN, University of Manouba, Tunisie (e-mail: hamdouni.ecri@gmail.com).
Gammoudi Mohamed Mohsen is with ISAMM, university of Manouba (e-mail: mohamed.gammoudi@fst.rnu.tn).

a binary relation between objects and items.

2) *Galois connection*: let K = (O, I, R) be a formal context, For every set of objects A ⊆ O, the set f(A) of attributes which characterize the objects of A using relation R is defined as:

$$f\left(A\right) = \{i \in I \mid \forall o \in A, \ \left(o,i\right) \in R\} \quad (1)$$

Dually, for every set of attributes B ⊆ I, the set g(B) of objects which are characterized by the attributes of B is defined as:

$$g\left(B\right) = \{o \in O \mid \forall i \in B, \ \left(o,i\right) \in R\} \quad (2)$$

The two functions f and g are called Galois connection. The operator $f°g(B) = \varphi$is called the closure operator.

3) *Formal concept*: a formal concept is a maximal objects-attributes subset where objects and attributes are in relation. More formally, it is a pair (A, B) with A ⊆ O and B ⊆ I, which verifies f(A) = B and g(B) = A.A is the extent of the concept and B is its intent.
4) *Partial order relation between concepts* ≤: The partial order relation called ≤ is defined as follow: for two formal concepts (A_1, B_1) and (A_2, B_2): (A_1, B_1) ≤ (A_2, B_2) ⇔ A_2⊆ A1 and B_1⊆ B_2.
5) *Meet/Join*: for each concepts (A_1, B_1) and (A_2, B_2), it exists a greatest lower bound (resp. a least upper bound) called Meet (resp. Join) denoted as ((A_1, B_1) ∧ (A_2, B_2) (resp. (A_1, B_1) ∨ (A_2, B_2)) defined by:

$$\left(A_1,\ B_1\right) \wedge \left(A_2,\ B_2\right) = \left(g\left(B_1 \cap B_2\right), \left(B_1 \cap\ B_2\right)\right) \quad (3)$$

$$\left(A_1,\ B_1\right) \vee \left(A_2,\ B_2\right) = \left(\left(B_1 \cup B_2\right), f\left(B_1 \cup\ B_2\right)\right) \quad (4)$$

6) *Galois lattice*: Let T is a set of formal concepts extracted from a formal context K=(O,I,R), (T,≤) is called Galois lattice associated to K. Its representation is done by Hass diagram.
7) *Iceberg lattice*: A partial ordered structure of frequent closed itemsets having only the join operator. It's considered a superior semi-lattice.

B. Association Rules

The following notions are defined in several works such as: [30], [2], [3], [15].

1) *Itemset*: let I a set of items, the itemset is a nonempty subset of items. An itemset containing k elements is called k-itemset.
2) *Support of an itemset*: the frequency of simultaneous occurrence of an itemset (I') in the set of objects. It is called supp(I').
3) *Frequent itemset (FI)*: FI is a set of items where their support ≥ minsup. Minsup is a user-specified threshold. All its subsets are frequent. The set of all frequent itemsets are called SFI.
4) *Association rule*: each association rule has the following form: A → B, where A and B are disjoint itemsets. A is its antecedent (condition) and B is its consequent.
5) *Confidence*: The confidence of an association rule R (A → B) measures how often items in B appear in objects that contain A. It is computed by:

$$Confidence\left(R\right) = Supp(A \cup B) \ / \ Supp\left(A\right) \quad (5)$$

where: Supp(A∪B) is the number of objects which are shared by itemsets A and B. Supp(A) is the number of objects that contain A. Based on the degree of confidence, there are three kinds of association rules: (i) Exact association rule which has a confidence = 1, (ii) Approximate association rule which has confidence < 1 and (iii) Valid association rule which has a confidence ≥ minconf. Minconf is a threshold specified by the expert.

6) *Frequent Closed Itemset (FCI)*: An itemset I' is called closed if I' = φ(I'). In other words, an itemset I' is closed if the intersect of the objects to which I' belongs to I' and it is frequent if its support ≥ minsup.
7) *Minimal Generator*: An itemset c ⊆ I is a closed itemset generator I' iff φ(c) = I'. c is a minimal frequent generator if its support is ≥ minsup. The set of frequent minimal generators of I' is called GMFI' and defined as follow:

$$GMF_{I'} = \{c \subseteq I \mid \varphi\left(c\right) = I' \wedge \neg\exists c_1 \subset c \wedge \varphi\left(c_1\right) = I'\} \quad (6)$$

8) *Generic base of exact association rules and Informative base of approximate association rules*: Generic base of exact association rules (GBE) is a base composed of non-redundant generic rules having a confidence ratio equal to 1. Given a context (O, I, R), the set of frequent closed itemsets (SFCI) and the set of minimal generators GMF_k:

$$GBE = \{R \mid R:\ c \rightarrow \left(I' \setminus c\right), I' \in SFCI \wedge c \in GMF_k \wedge \varphi\left(c\right) = I' \wedge c \neq I'\} \quad (7)$$

while informative base of approximate association (GBA) rules is defined as:

$$GBA = \{R \mid R: c \rightarrow \left(I_1 \setminus\ c\right), I_1, I_2 \in SFCI \wedge c \in GMF_{I_2} \wedge I_2 \subset I_1 \wedge Conf\left(R\right) \geq minconf\} \quad (8)$$

The union of those bases constitutes a generic base without losing information.

C. Classification Rules

1) *Classification rule (Classifier):* it has the form: A →ck where the premise A is an itemset and the conclusion ck is an instance of the attribute class which is denoted C [8].
2) *Recall / Precision*: before defining these two notions, we recall the terms of True positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The terms positive and negative refer to the classifier's

prediction or the expectation, and the terms true and false refer to whether that prediction corresponds to the external judgment or observation. Let be TP a Correct result, FP an Unexpected result, FN a Missing result and TN a correct absence of result, the Recall and the Precision are defined as follow [28]:

- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)

III. Related Works: Rule-Based Classification

The main objective of the classification is to discriminate the maximum of classes by grouping objects which are similar in each class. It has two steps, the first one is to use a training set to define classifiers and the second step is to use these classifiers on a test set to determine the class of each new object. To evaluate the quality of the classification method, statistical measures are used such as: recall and precision [28].

In the case of a rule-based classification, classifiers are a set of classification rules. According to [8], this method involves two steps. The first one is to generate a classifier in the form of classification rules from a training set, while the second is the use of these rules to predict the class of new data.

To determine the set of classification rules or classifier, three approaches have been proposed. The first one is to extract classification rules from a decision tree built using a training set such as: C4.5rules algorithm [31]. The second approach generates the classification rules directly from the training set as in the case of Prism [9], GRAND [29], LEGAL [19], RULEARNER [32], CIBLe [27], IPR [24], CITREC [10] and BFC [26]. The third approach extracts classification rules by exploiting association rules. Several algorithms have adopted this approach to build classifiers such us: CBA [20], HARMONY [34], GARC [7] and CAR-Miner [22].

Once the classification rules are determined, they will be used to classify new objects. To be classified, each object is assigned to the most weighted rule to be triggered to determine the associated class. Otherwise, the object will not be classified because no rule has been triggered.

Among the rule-based classification methods already mentioned, there are those who exploit the Formal Concept Analysis (FCA) to classify new data. These methods contain two phases. The first one is the training which consists on organizing the data of the initial context under the form of a lattice. The second phase is the classification which allows determining the class of new objects by exploiting the lattice [26].

The first classification algorithms like GRAND [29] and RULEARNER [32] are based on a complete lattice to classify new objects. However, a major limitation related to the exponential complexity of the lattice construction phase which is a handicap in classification process. To resolve this problem, several algorithms have been proposed such as LEGAL [19] and CIBLe [27]. They build a sub-lattice for the training process which reduces the complexity of the treatment. In fact, it increases simultaneously the run time and the number of generated concepts.

Some other algorithms have used the cover concepts which belong to the family of sub-lattices such as: IPR [24] and BFC [26]. Indeed, a sub-lattice contains only the most relevant formal concepts from a lattice. The determination of the relevant concepts is based on a measurement function such as entropy, gain, etc.

In this context, we can mention the work of [21] which is based on Formal Concept Analysis and the notion of consensus to generate classification rules. In fact, the proposed work consists on: (i) building a random forest from an initial training set in order to generate classification rules. (ii) From the result obtained in (i) they generate a subset of rules having a consensus degree without loose of information. They are called consensual classification rules.

IV. Our Contribution

In [1], an algorithm of association rules extraction called CondClose was proposed. To evaluate association rules extracted by this algorithm, we introduce a method which exploits a rules-based classification to validate them. The choice of rules-based classification is founded on the results of some research works in [8] and [22]. They proved that classification rules are precise for classification.

Before detailing our validation method of association rules, we recall the principle of CondClose algorithm. In fact, to extract non redundant rules without losing information, CondClose is based on three steps. The first one allows extracting the frequent minimal generators as well as the positive border by condensing the initial context in the objective to minimize the run time. The second one uses the minimal generators and the condensed context results to construct a frequent minimal generators lattice. The third step determines the generic base of exact and approximate association rules related to frequent minimal generators lattice.

Our validation method has as a main objective to assist expert during the validation of the association rules. Five steps are necessary in our method (see Fig. 1). We describe them in the following subsections.

A. Step1: Association Rules Generation

The objective of this step is to generate association rules from a training set. The training set is represented as a matrix where columns are attributes and rows are instances of these attributes. The last column (attribute) of this matrix is a class where instances are nominal type. The other attributes can have values of different types: nominal, continuous, ordinal, etc.

To extract association rules, we use the training set without the class attribute. Thus, we can extract association rules that describe the correlations between attributes which aren't a class attribute.

In order to apply our mining association rules algorithm, we transform the training set without the class attribute in a formal context. The generated association rules by CondClose are the result of this step. We denote association rules by R_{AR}. Kinds of obtained rules are either exact (R_{EX}), or approximate or exact and approximate.

B. Step2: Classification Rules Generation

In this step we use a rule-based classification algorithm to generate classification rules. The input of the chosen algorithm is the same training set used in the step of generating association rules. The obtained classification rules are denoted R_{CR} and will be with R_{AR} the input of the next step in order to generate new classifiers.

C. Step3: Generation of New Classifiers

In this step, we generate new classifiers from association rules extracted by applying one of the following mappings which is defined as follows:

- *Complete mapping*: let $R_i \in R_{AR}$ and $R_j \in R_{CR}$, if antecedent(R_j) = consequent(R_i) then we said that there is a complete mapping between R_i and R_j.
- *Partial mapping*: let $R_i \in R_{AR}$ and $R_j \in R_{CR}$, if antecedent(R_j) $\subset$ consequent(R_i) then we said that there is a partial mapping between R_i and R_j.

The result obtained applying these two cases is a new classifier denoted R_{Result} which contains a set of rules having the form: antecedent(R_i) $\rightarrow$consequent(R_j).

D. Step4: Validation of New Classifiers

We use R_{Result} as new classifier and apply it to the same training set in order to determine its classification quality. The classification quality includes the rates of recall, precision, correctly classified instances (CCI), incorrectly classified instances (ICI) and unclassified instances (UI).

E. Step5: Association Rules Validation

In this step, our method offers the possibility to visualize the classification quality of the new classifiers in order to assist expert to validate association rules used to generate them.

V. Illustration of Our Method

To illustrate our method, we use a Weather.nominal training set [17], see Table I. The values of their attributes have a nominal type. To apply the first step of our method, we use a binarization method implemented in Weka Software [16] to obtain a formal context K = (O, I, R). O is the set of instances (or objects), I is the set of attributes and R is a binary relation between O and I.

TABLE I
Description of the Training Set

Number of instances (objects)	Number of attributes	Number of binary attributes	Number of classes
14	4	8	2

A. Results of the Step Association Rules Generation

After applying the CondClose algorithm to the binarized training set with minsup = 10% and minconf = 10%, R_{AR} contains 5 exact rules (R_{EX}) and 49 approximate rules (R_{AP}). Table II presents the R_{EX} and the R_{AP} results with their confidence rates denoted Conf.

TABLE II
Association Rules Result (R_{AR}) After Applying CondClose

	R_{AR} (Association rules)	Conf.
R_{EX} (Exact rules)		
R_1 :	IF [temperature=cool] THEN [humidity]	100%
R_2 :	IF [outlook=rainy, temperature=cool] THEN [humidity]	100%
R_3 :	IF [temperature=hot, outlook=overcast] THEN [windy]	100%
R_4 :	IF [windy, outlook=overcast] THEN [temperature=hot]	100%
R_5 :	IF [windy, temperature=cool] THEN [humidity]	100%
R_{AP} (Approximate rules)		
R_1 :	IF [temperature=hot] THEN [windy]	72%
R_2 :	IF [humidity, outlook=rainy] THEN [windy]	67%
R_3 :	IF [humidity, outlook=rainy] THEN [temperature=cool]	67%
R_4 :	IF [windy, temperature=mild] THEN [outlook=rainy]	67%
R_5 :	IF [windy, outlook=rainy] THEN [humidity]	67%
R_6 :	IF [windy, outlook=rainy] THEN [temperature=mild]	67%
R_7 :	IF [windy, temperature=hot] THEN [outlook=overcast]	67%
R_8 :	IF [temperature=mild, outlook=rainy] Then [windy]	67%
R_9 :	IF [outlook=sunny] Then [windy]	58%
R_{10} :	IF [humidity] Then [windy]	58%
R_{11} :	IF [humidity] Then [temperature=cool]	58%
R_{12} :	IF [outlook=rainy] Then [humidity]	58%
R_{13} :	IF [outlook=rainy] Then [windy]	58%
R_{14} :	IF [outlook=rainy] Then [temperature=mild]	58%
R_{15} :	IF [windy] Then [humidity]	51%
R_{16} :	IF [temperature=mild] Then [windy]	49%
R_{17} :	IF [temperature=mild] Then [outlook=rainy]	49%
R_{18} :	IF [outlook=overcast] Then [humidity]	48%
R_{19} :	IF [outlook=overcast] Then [windy, temperature=hot]	48%
R_{20} :	IF [temperature=hot] Then [outlook=sunny]	48%
R_{21} :	IF [windy, humidity] Then [outlook=rainy]	48%
R_{22} :	IF [windy, humidity] Then [temperature=cool]	48%
R_{23} :	IF [temperature=cool] Then [humidity, outlook=rainy]	48%
R_{24} :	IF [temperature=cool] Then [windy, humidity]	48%
R_{25} :	IF [temperature=hot] Then [windy, outlook=overcast]	48%
R_{26} :	IF [humidity] Then [outlook=rainy]	42%
R_{27} :	IF [outlook=sunny] Then [humidity]	39%
R_{28} :	IF [outlook=sunny] Then [temperature=mild]	39%
R_{29} :	IF [outlook=sunny] Then [temperature=hot]	39%
R_{30} :	IF [outlook=rainy] Then [windy, humidity]	39%
R_{31} :	IF [outlook=rainy] THEN [humidity, temperature=cool]	39%
R_{32} :	IF [outlook=rainy] THEN [windy, temperature=mild]	39%
R_{33} :	IF [windy] THEN [temperature=mild]	37%
R_{34} :	IF [windy] THEN [outlook=sunny]	37%
R_{35} :	IF [windy] THEN [outlook=rainy]	37%
R_{36} :	IF [windy] THEN [temperature=hot]	37%
R_{37} :	IF [temperature=mild] THEN [humidity]	33%
R_{38} :	IF [temperature=mild] THEN [outlook=sunny]	33%
R_{39} :	IF [temperature=mild] THEN [windy, outlook=rainy]	33%
R_{40} :	IF [humidity] THEN [outlook=sunny]	28%
R_{41} :	IF [humidity] THEN [temperature=mild]	28%
R_{42} :	IF [humidity] THEN [outlook=overcast]	28%
R_{43} :	IF [humidity] THEN [windy, outlook=rainy]	28%
R_{44} :	IF [humidity] THEN [outlook=rainy, temperature=cool]	28%
R_{45} :	IF [humidity] THEN [windy, temperature=cool]	28%
R_{46} :	IF [windy] THEN [humidity, temperature=cool]	25%
R_{47} :	IF [windy] THEN [temperature=hot, outlook=overcast]	25%
R_{48} :	IF [windy] THEN [humidity, outlook=rainy]	25%
R_{49} :	IF [windy] THEN [temperature=mild, outlook=rainy]	25%

B. Results of the Step Classification Rules Generation

In this step, we use BFC algorithm to generate classification rules (see Table III) because it's based on a sound mathematical foundation on what is based CondClose.

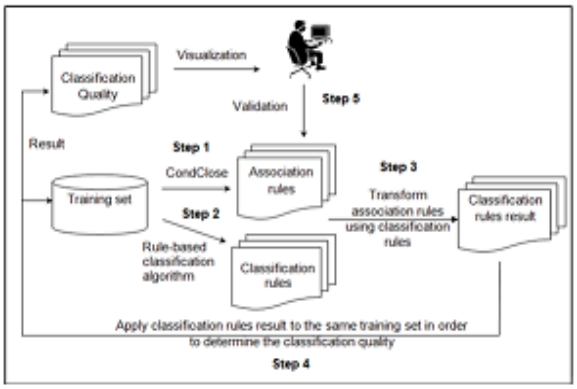

Fig. 1 General architecture of our method

TABLE III
CLASSIFICATION RULES(R_{CR}) OBTAINED BY BFC

	R_{CR} (Classification rules)
R_1 :	IF [outlook=overcast] Then [Yes]
R_2 :	IF [humidity] THEN [Yes]
R_3 :	IF [outlook=sunny, temperature=hot] THEN [No]
R_4 :	IF [outlook=sunny] THEN [No]
R_5 :	IF [outlook=overcast, temperature=hot, windy] THEN [Yes]
R_6 :	IF [windy] THEN [Yes]
R_7 :	IF [temperature=mild, humidity] THEN [Yes]
R_8 :	IF [outlook=sunny, temperature=hot, windy] THEN [No]
R_9 :	IF [outlook=overcast, temperature=mild] THEN [Yes]
R_{10} :	IF [outlook=sunny, windy] THEN [No]
R_{11} :	IF [outlook=rainy] THEN [No]
R_{12} :	IF [outlook=overcast, temperature=cool, humidity] THEN [Yes]

C. Results of the Step Generation of New Classifiers

We use the set R_{CR} to transform R_{EX} and R_{AP} rules. For this, two cases can be handled: the complete mapping and the partial mapping.

1) *Example of complete mapping:* the results of the complete mapping between (exact association rule R_1 and the classification rule R_2) is shown in Table IV.

TABLE IV
EXAMPLE OF COMPLETE MAPPING

	Rule result after complete mapping
R_1 : IF [temperature=cool] THEN[***humidity***] R_2 : IF [***humidity***] THEN [Yes]	R_{Result}: IF [temperature=cool] THEN [Yes]

2) *Example of Partial mapping:* The partial mapping between exact association rule R_1 and the classification rule R_7 is shown in Table V.

TABLE V
EXAMPLE OF PARTIAL MAPPING

	Rule result after partial mapping
R_1 : IF [temperature=cool] THEN[***humidity***] R_7 : IF [temperature=mild, ***humidity***] THEN [Yes]	R_{Result}: IF [temperature=cool] THEN [Yes]

Using the same principle illustrated by these two examples, we handle the R_{EX}, the R_{AP} and R_{AR} and we obtain six kinds of classifiers. Each kind of classifier depends on the type of association rules (exact, approximate or (exact and approximate)) and the kind of mapping (complete or partial).

D. Results of the Step Validation of New Classifiers

We apply the six obtained classifiers to the same training set in order to determine rates of recall, precision, correctly classified instances (CCI), incorrectly classified instances (ICI) and unclassified instances (UI).

Table VI presents the details of different rates after the mapping of R_{EX}, R_{AP} and R_{AR} which express the quality of our association rules after their use as a classifier.

TABLE VI
THE CLASSIFICATION QUALITY OF THE SIX CLASSIFIERS

	BFC Rates	$R_{Resultat}$					
		Case1 : Complete mapping			*Case2 : Partial mapping*		
		R_{EX}	R_{AP}	R_{AR}	R_{EX}	R_{AP}	R_{AR}
Recall	0.929	0.833	0.714	0.714	0.833	0.714	0.714
Precision	0.936	0.694	0.802	0.802	0.694	0.802	0.802
CCI	92.857	35.714	71.429	71.429	21.429	71.429	71.429
ICI	7.143	7.143	28.571	28.571	21.429	28.571	28.571
UI	0	57.143	0	0	57.142	0	0

We observe in Table VI that using only exact rules as a classifier, the precision rate is the lowest one and the recall rate is the highest comparing with recall and precision rates obtained when we use separately R_{AP} and R_{AR} in the two kind of mapping. However, when we use at the same time exact and approximate rules (R_{AR}), we get the best rate of precision and recall.

We observe also that when we use R_{AR}, the CCI is greater than using exact rules.

E. Results of the Step Association Rules Validation

The classification quality will be displayed to the expert according to his choice such as: mapping type and the type of association rules to be used in classification. Having these information's, the expert could start the validation of association rules used.

VI. EXPERIMENTAL STUDY

After using an illustrative example to explain the different steps of our method, we implemented an interactive prototype called S2A2RV (System of Semi-Automatic Association Rules Validation) to evaluate our method. We present in this section the general architecture of our prototype and the results of the different experiments using training sets on medical domain.

A. General Architecture of Our Prototype S2A2RV

S2A2RV integrates all steps of our method and includes five modules which are: preprocessing module, association rules generation module, classification rules generation module, generation of new classifiers module, validation of new classifiers module and association rules validation

module. Fig. 2 presents the functional architecture of our prototype and Fig. 3 presents its general architecture.

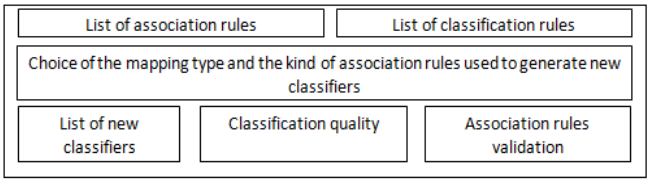

Fig. 2 Functional architecture of S2A2RV

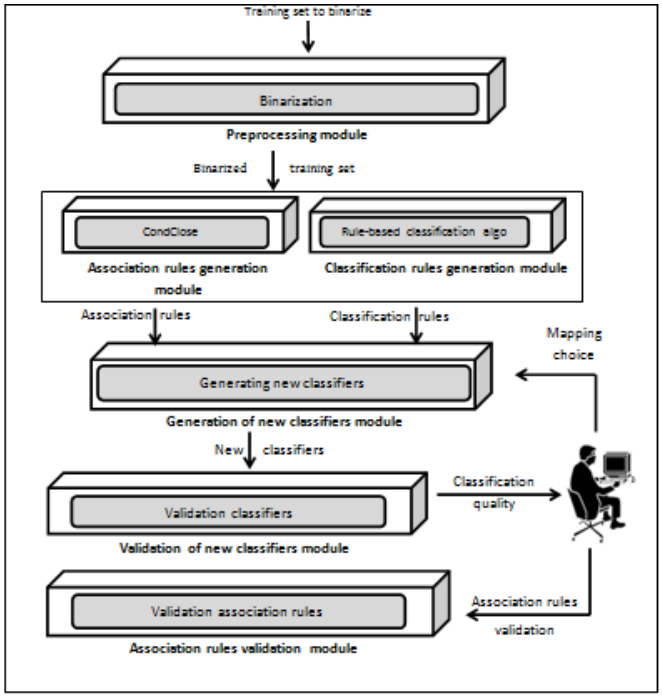

Fig. 3 General architecture of S2A2RV

B. Evaluation of Our Method

To evaluate our method, we have chosen a medical domain and especially the ENT field. We made some experiments using a real training set which contains 127 instances (X-Ray) of the nasal bones. This X-Rays are collected from El Amen Clinic and Hospital La Rabta in Tunisia. They are classified into three classes: healthy (class A), small fracture (class B) and big fracture (class C). Each nose is characterized by 12 attributes which are: ExtentL, ExtentR, BoundingL, BoundingR, PerimeterL, PerimeterR, AreaL, AreaR, SolidityL, SolidityR, CompactnessL, CompactnessR. These attributes are specified by the doctors and their values have a numeric type.

To apply the first step of our method, we use a binarization method for numeric values which is implemented in Weka to obtain a formal context having 60 binarized attributes. We summarize all these information in Table VII.

TABLE VII
DESCRIPTION OF THE REAL TRAINING SET

Number of instances (objects)	Number of attributes	Number of binary attributes	Number of classes
127	12	60	3

To generate new classification rules, we use two different rule-based classification algorithms such as: BFC [26] and Reduced Random Forests [21] in order to study the impact of the chosen algorithms on the association rules quality classification.

Table VIII describes the classification quality (precision, recall, CCI, ICI and UI) of BFC and Reduced Random Forest algorithms using the training set described in Table VII.

TABLE VIII
CLASSIFICATION QUALITY RATES OF BFC AND REDUCED RANDOM FORESTS

		BFC	Reduced Random Forests
% CCI		**72.441**	**94.488**
% ICI		**1.575**	**5.5118**
% UI		**25.984**	**0**
Precision	Class A	1	1
	Class B	0.950	0.952
	Class C	0.955	0.848
	Weighted Avg.	**0.979**	**0.947**
Recall	Class A	1	1
	Class B	0.950	0.889
	Class C	0.955	0.933
	Weighted Avg.	**0.979**	**0.945**

We observe that Reduced Random Forest algorithm classifies all instances. In fact, 94.488% of instances are correctly classified. However BFC algorithm classifies only 74.016% where 1.575 aren't correctly classified.

1) Description of the Functionalities of Our Prototype

After running our prototype using the binarized training set with for example a support value = 10% and confidence value = 50%, we obtain the result as it is shown in Fig. 4.

The two boxes in the top represent the list of association rules generated by the algorithm CondClose and the list of classification rules extracted using Reduced Random Forest algorithm. The middle part of the interface allows the expert to select the type of mapping and the association rules used to generate new classifiers.

In our case, the expert has chosen to apply complete mapping of exact and approximate rules. New classifiers are displayed. The expert can be applied them on the same binarized training set to obtain the recall, precision, CCI, ICI and UI rates. The visualization of these rates allows the expert to have an idea about the quality of exact and approximate rules used to generate the classifiers. This step assist expert to validate easily association rules.

In fact, our prototype provides an opportunity for the expert to affirm or reject the validity of the association rules selected to view their classification quality.

2) Results of Different Experiments

In this section, we present the different results of experiments using the training set described in Table VII. These experiments are executed using a PC with CPU Intel Atom N2600 1.6 GHz, 2 GB memory and 1 MB cache.

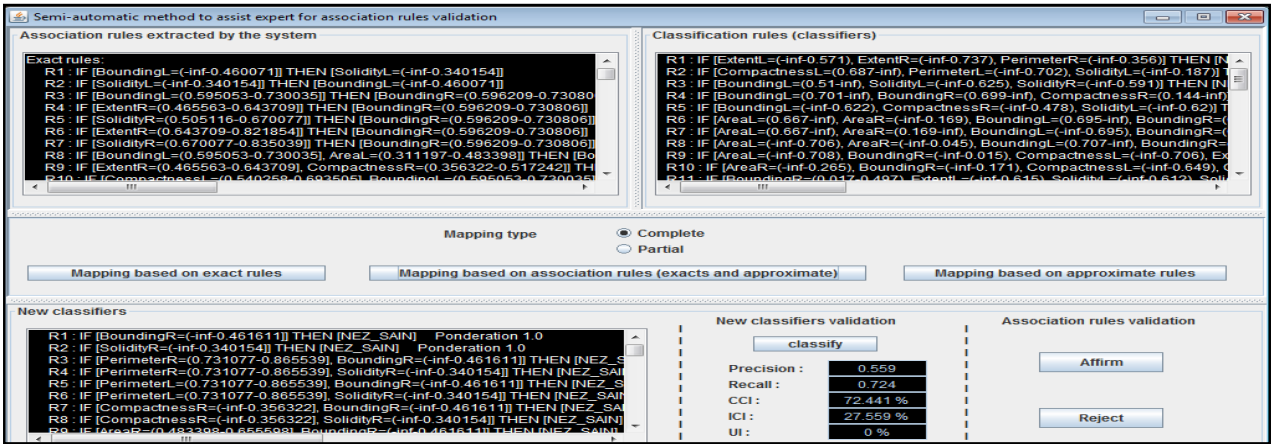

Fig. 4 Association rules with their classification quality in order to assist expert during validation

We can organize the experiments on two categories: evaluation of the complete mapping's quality classification and that of the partial mapping.

In order to evaluate the quality classification, we present the variation of precision, recall, CCI, ICI and UI rates (Weighted Avg.) associated to exact and approximate rules after mapping using BFC's and Reduced Random Forest's (RRF) classification rules. In Figs. 5-9, we present the results of the first category's experiments. We observe in Figs. 5 and 6 that the precision and recall rates associated to new classifiers resulting from complete mapping between exact rules and BFC's classification rules are better than the use of RRF's classification rules. However, using approximate rules having confidence between 10% and 50% with BFC classifiers, we observe that the precision and recall rates are lower than those with RRF classifiers. Furthermore, the quality classification of approximate rules having a confidence more than 50% is higher when they are mapped with BFC's classifiers.

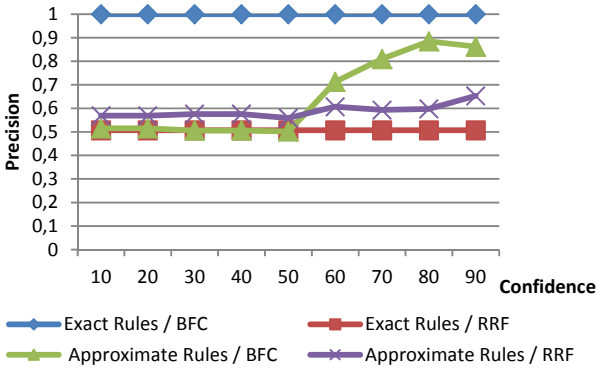

Fig. 5 Variation of precision rates associated to exact and approximate rules after complete mapping using BFC's and RRF's classifiers

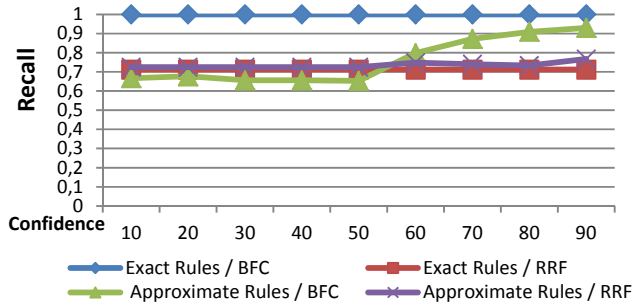

Fig. 6 Variation of recall rates associated to exact and approximate rules after complete mapping using BFC's and RRF's classifiers

Fig.7 shows that correctly classified instances (CCI) rates of exact and approximate rules using RRF's classifiers is better than using BFC's classifiers. This is related to the classification quality of the rule-based classification algorithm (see Table VIII). In fact, the CCI rate of RRF is better than BFC. The ICI and UI rates (see Figs.8and 9) vary in the same way as the precision and recall rates.

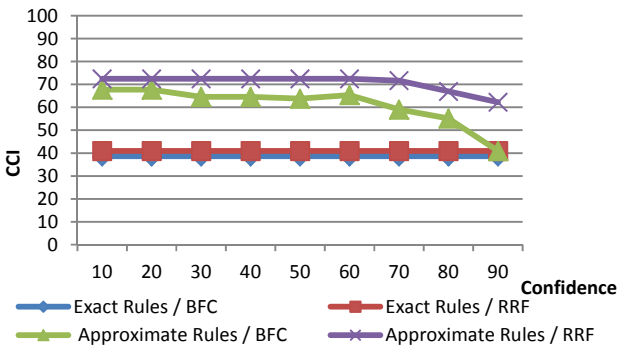

Fig. 7 Variation of CCI rates associated to exact and approximate rules after complete mapping using BFC's and RRF's classifiers

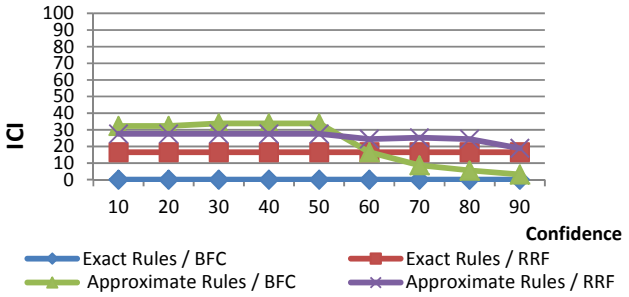

Fig. 8 Variation of ICC rates associated to exact and approximate rules after complete mapping using BFC's and RRF's classification rules

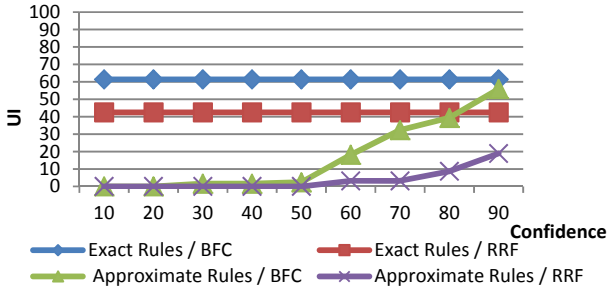

Fig. 9 Variation of UI rates associated to exact and approximate rules after complete mapping using BFC's and RRF's classification rules

Figs. 10-14 present the results of the partial mapping's experiments. We observe that precision and recall rates (see Figs. 10 and 11) of association rules mapped with BFC's classifiers are better than mapped with RRF's classifies because the BFC's precision and recall rates are higher than RRF's quality classification. In addition, the approximate rules quality classification mapped with BFC' classifiers are better than exact rules mapped with same classifiers.

Despite, RRF's CCI rate is better than BFC's rate (see Table VIII); the association rules mapped with BFC's classifiers have the best CCI and ICI rates (see Figs.12 and 13). However, the UC rates of association rules mapped with RRF is better than mapped with BFC.

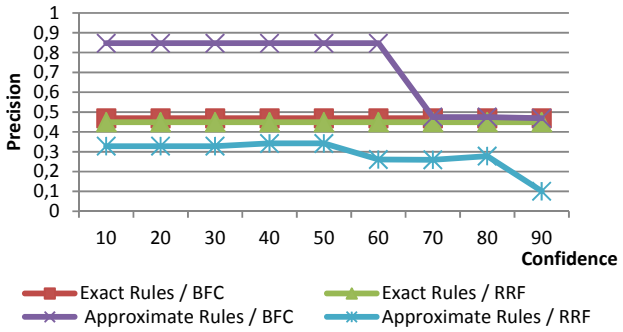

Fig. 10 Variation of precision rates associated to exact and approximate rules after partial mapping using BFC's and RRF's classification rules

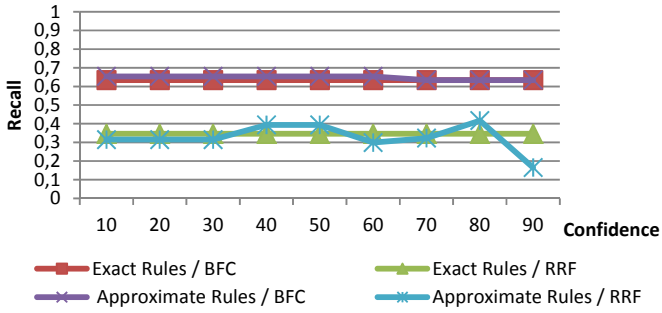

Fig. 11 Variation of recall rates associated to exact and approximate rules after partial mapping using BFC's and RRF's classifiers

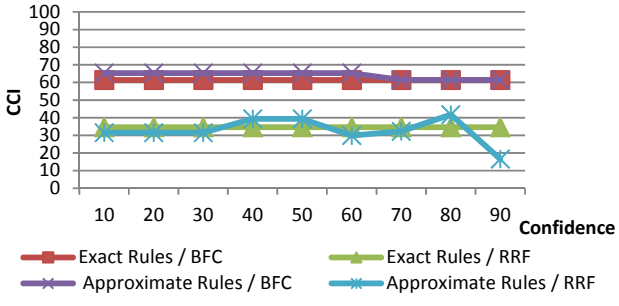

Fig. 12 Variation of CCI rates associated to exact and approximate rules after partial mapping using BFC's and RRF's classification rules

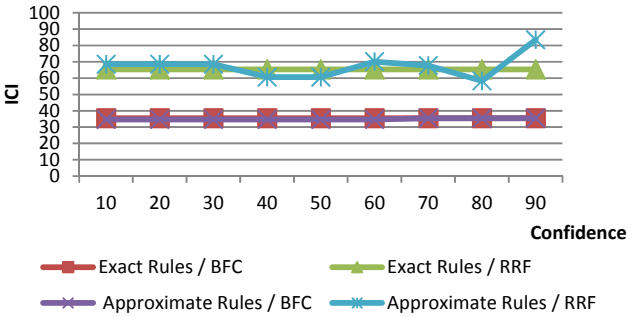

Fig. 13 Variation of ICI rates associated to exact and approximate rules after partial mapping using BFC's and RRF's classification rules

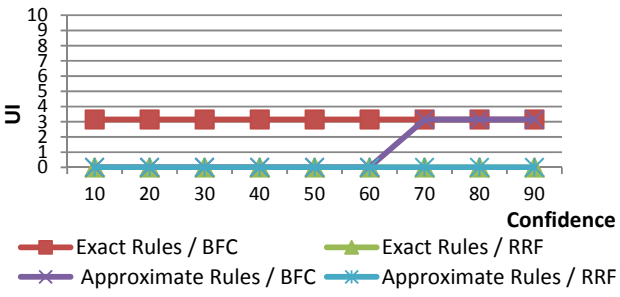

Fig. 14 Variation of UI rates associated to exact and approximate rules after partial mapping using BFC's and RRF's classification rules

VII. Conclusion

The validation of association rules by the expert is a hard task when their number is great. In this paper, we presented a semi-automatic method to help the expert during validation

task. The main idea of our method is to exploit rule-based classification systems by transforming association rules into classifiers and visualize them with their classification quality using our prototype. The display of the result provides an idea to the expert on the association rules quality (relevance). Thus he could validate them easily. As a perspective, we plan to perform our method to use it as a kernel of a recommender system in semantic web.

REFERENCES

[1] Amdouni H. and Gammoudi M. M. "CondClose: A new algorithm of association rules extraction". IJCSI (International Journal of Computer Science Issues), Volume 8, Issue 4, July (2011).
[2] Bastide Y., Pasquier N., Taouil R., Lakhal L. and Stumme G. "Mining minimal non-redundant association rules using frequent closed itemsets". Proceedings of the Intl. Conference DOOD'2000, LNCS, Springer-verlag, p. 972-986 (2000).
[3] Ben yahia S., Latiri C., Mineau G.W. and Jaoua A. "Découverte des règles associatives non redondantes – application aux corpus textuels". In M.S. Hacid, Y. Kodrattof and D. Boulanger, editors EGC, volume 17 of Revue des Sciences Technologies de l'Information – série RIA ECA, pages 131-144. Hermes Sciences Publications (2003).
[4] Ben Yahia S. and Mephu Nguifo E. "Visualisation des règles associations : vers une approche méta-cognitive", dans Actes conférences INFORSID, Hammamet, pp 735-750, 30 Mai - 1 Juin (2006).
[5] Blanchard J. "Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association". Thèse de doctorat à l'Ecole Polytechnique de l'Université de Nantes, soutenue le 24 novembre (2005).
[6] Blanchard J., Guillet F. and Briand H. "Interactive visual exploration of association rules with rule-focusing methodology". Knowledge and Information Systems, vol. 13, num. 1, p. 43-75, Springer (2007).
[7] Bouzouita I., Elloumi S. and Ben Yahia S. "Garc: a new associative classification approach". In A. M. Tjoa and J. Trujillo, editors, Proceedings of 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2006), Springer-Verlag, 66LNCS 4081, Krakow, Poland, pages 554-565, 4-8 September (2006).
[8] Bouzouita I. and Elloumi S. "Generic Associative Classification Rules: A Comparative Study". International Journal of Advanced Science and Technology Vol. 33, August (2011).
[9] Cendrowska J. "PRISM: An Algorithm for Inducing Modular Rules. International Journal of Man-Machine Studies 27(4):349-370 (1987).
[10] Douar B., Latiri C. and Slimani Y. "Approche hybride de classification supervisée à base de treillis de galois : application à la reconnaissance de visages". In Conference Extraction et Gestion des Connaissances, pp. 309-320. RNTI, Sophia-Antipolis (2008).
[11] Fernandes L. A.F. and Garcia A. C. B. "Association Rule Visualization and Pruning through Response-Style Data Organization and Clustering". J. Pavon et al. (Eds.): IBERAMIA 2012, LNAI 7637, pp. 71–80, 2012. Springer-Verlag Berlin Heidelberg (2012).
[12] Fule P. and Roddick J. F. "Experiences in building a tool for navigating association rule result sets". In CRPIT'04: Proceedings of the second Australasian workshop on information security, data mining, web intelligence, and software internationalization (J. Hogan, P. Montague, M. Purvis & C. Steketee, éds.), Australian Computer Society, Inc., p. 103–108 (2004).
[13] Ganter B. and Wille R. "Formal Concept Analysis". Mathematical Foundations, Springer, (1999).
[14] Hahsler M. and Chelluboina S. "Visualizing association rules in hierarchical groups". In Computing Science and Statistics, Vol. 42, 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011). The Interface Foundation of North America, June (2011).
[15] Han J., Kamber M. and Pei J. "Data Mining: Concepts and Techniques, 3rd edition", Morgan Kaufmann (2011).
[16] http://www.cs.waikato.ac.nz/ml/Weka
[17] http://www.ics.uci.edu/~mlearn/MLRepository.html
[18] Lenca P., Meyer P. and Vaillant B. "Evaluation et analyse multicritère des mesures de qualité des règles d'associations". National Journal of Information Technologies (RNTI), France, pp.219-246 (2004).
[19] Liquiere M. et Mephu Nguifo E. "Legal: learning with galois lattice". In Proceeding of the 5ieme Journée sur l'Apprentissage (JFA'90), Lannion, FRANCE, Avril (1990).
[20] Liu B., Hsu W. and Ma Y. " Integrating classification and association rule mining". InKDD'98 (1998).
[21] Louizi Mehdi and Gammoudi Mohamed Mohsen, "Method for Classification of Images in the Medical Field: The Nose Case", International Journal of Information and Electronics Engineering, Vol. 2, No. 5, September 2012.
[22] Loan T. T. N., Bay V., Tzung-Pei H. and Hoang Chi T. "CAR-Miner: An efficient algorithm for mining class-association rules". Expert Syst. Appl. 40(6): 2305-2311 (2013).
[23] Ma Y., Liu B. et Wong C. K. "Web for data mining: organizing and interpreting the discovered rules using the web". SIGKDD Explorations 2, no. 1, p. 16–23 (2000).
[24] Maddouri M. "Towards a machine learning approach based on incremental concept formation". Intelligent Data Analysis, 8(3): 267-280 (2004).
[25] Maddouri M. and Gammoudi J. "On Semantic Properties of Interestingness Measures for Extracting Rules from Data". B. Beliczynski et al. (Eds.): ICANNGA 2007, Part I, LNCS 4431, pp. 148–158, Springer-Verlag Berlin Heidelberg (2007).
[26] Meddouri N. and Maddouri M. "Boosting Formal Concepts to Discover Classification Rules". In Proceeding of the 22rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE'09), Tainan, TAIWAN (2009).
[27] Njiwoua P. and MephuNguifo E. "Améliorer l'apprentissage à partir d'instances grâce à l'induction de concepts". Revue d'intelligence artificielle, 13(2): 413-440 (1999).
[28] Olson D. L. and Delen D. "Advanced Data Mining Techniques". Springer, 1st edition, page 138, ISBN 3-540-76916-1, (2008).
[29] Oosthuizen G. D. "The use of a lattice in knowledge processing". PhD thesis, Glasgow, Scotland, UK (1988).
[30] Pasquier N. "Data mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données". Thèse de doctorat, Université de Clermont-Ferrand II, (2000).
[31] Quinlan J.R. "C4.5: Programs for Machine Learning". Morgan Kaufman Publishers (1993).
[32] Sahami M. "Learning classification rules using lattices (extended abstract)". In Proceedings of the 8th European Conference on machine learning (ECML'95), Heraclion, Crete, GREECE (1995).
[33] Vaillant B., Menou S., Moga S., Lenca P. and Lallich S. "Qualité des règles d'association : étude de données d'entreprise". In Atelier Data mining dans la banque, l'assurance et la finance (associé à la conférence Extraction et Gestion des Connaissances (2007)), pp.45—54, Namur, Belgique (2007).
[34] Wang J. and G. Karypis. "HARMONY: Efficiently mining the best rules for classification". In SIAM'05 (2005).
[35] Zaki M. and Phoophakdee B. "MIRAGE: A framework for mining, exploring and visualizing minimal association rules". Technical report, July 2003, Rensselaer Polytechnic Institute, Computer Sciences Department, USA (2003).

AmdouniHamida is currently an Assistant at ESEN Manouba University. She received her Master degree in Computer Science at FST-Tunisia in 2005. She also obtains her PhD at the Faculty of Sciences of Tunis in 2014. Her main research contributions concern: data mining, Formal Concept Analysis (FCA) and Customer Relation Management. She is member of SCO-ECRI in Research Laboratory RIADI.

Gammoudi Mohamed Mohsen is currently a full Professor at ISAMM, University of Manouba, Tunisia. He is responsible of SCO-ECRI team in Research Laboratory RIADI. He obtained his habilitation to Supervise research in 2005 at the Faculty of Sciences of Tunis. He got his PhD in September 1993 in Sophia Antipolis Laboratory I3S/CNRS. Professor Gammoudi's professional work experience began in 1992 when he was assigned as an assistant at the Technical University of Nice. Then he was hired as a visiting professor between 1993 and 1997 at Federal University of Maranhao, Brazil. Since, he supervised several PhD and masterthesis.