VALIDATION METHOD OF FUZZY ASSOCIATION RULES BASED ON FUZZY FORMAL CONCEPT ANALYSIS AND STRUCTURAL EQUATION MODEL

Imen Mguiris[1], Hamida Amdouni[2] and Mohamed Mohsen Gammoudi[3]

[1]Computer Science Department, FST-University of Tunis ElManar,
Tunis, Tunisia
[2]ESEN, University of Manouba, Manouba, Tunisia
[3]ISSAM, University of Manouba, Manouba, Tunisia

ABSTRACT

In order to treat and analyze real datasets, fuzzy association rules have been proposed. Several algorithms have been introduced to extract these rules. However, these algorithms suffer from the problems of utility, redundancy and large number of extracted fuzzy association rules. The expert will then be confronted with this huge amount of fuzzy association rules. The task of validation becomes fastidious. In order to solve these problems, we propose a new validation method. Our method is based on three steps. (i) We extract a generic base of non redundant fuzzy association rules by applying EFAR-PN algorithm based on fuzzy formal concept analysis. (ii) we categorize extracted rules into groups and (iii) we evaluate the relevance of these rules using structural equation model.

KEYWORDS

Fuzzy Association Rules Validation, Fuzzy Formal Concept Analysis, Structural equation model

1. INTRODUCTION

The extraction of association rules is one of the most known techniques of data mining [1]. It aims at discovering correlations between the properties (attributes) characterizing the objects saved in the databases. Correlations discovered allow decision-makers to make better judgments. Indeed, association rules have been used in several fields, including medical research [2], analysis of geographic data and biological data [3] and electronic commerce [4].

The integration of fuzzy logic into the extraction of association rules made it possible to solve the problem of discretization and to process the quantitative databases without loss of information. A fuzzy association rule was the object of several studies since the work of [5]. However, the major drawback of fuzzy association rule extraction algorithms is the large number of rules generated. As a result, it becomes very difficult to interpret and exploit these rules when making decisions. The expert is obliged to validate them manually. Several researchers have proposed various

Dhinaharan Nagamalai et al. (Eds) : NATL, CSEA, DMDBS, Fuzzy, ITCON, NSEC, COMIT - 2018
pp. 91–108, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80608

methods of assisting evaluation in order to make this task less time-consuming. However, these problems still persist.

In this context, we present a new method of validation of fuzzy association rules exploiting the structural equations Model (SEM). Our method contains three steps. The first step consists of applying EFAR-PN algorithm to extract generic bases of association rules. These bases contain a set of non-redundant association rules. This algorithm is based on Fuzzy Formal Concepts Analysis. The second step consists of categorizing the extracted rules into groups based on their items. This step provides a synthetic representation of the rules. The Final step allows evaluating the rule by using Structural Equation Model (SEM). We are using specifically one of the SEM techniques known as Partial Least Square (PLS).

The remainder of this article is organized as follows. In section 2, we introduce some basic notions necessary to better understand our work. In section 3, we present the different categories of fuzzy association rules algorithms. Section 4 surveys related work. In section 5, we detail the principle of our method by illustrating it with an example. Section 6 is devoted to evaluate our method by performing a series of experiments on three test bases used by the scientific community of the field. Finally, section 6 presents a conclusion and some future work.

2. BASIC NOTIONS

In this section, we present some basic concepts related to our work [6][7]

- *Association rules:* an association rule is written in the following form:

  $$R: A \rightarrow B$$

  Where $A \cap B = \emptyset$. A is called the premise of the rule and B its conclusion. Two measures are used when extracting association rules:

  - Support: It is the measure of the frequency of simultaneous appearance of an itemset AB in the set of objects, denoted Supp (AB). An itemset is said to be frequent if its support is greater than or equal to a minimal support (minsup).

  - Confidence: It is the probability of having the itemset B, knowing that we already have the itemset A. According to [8], this measure is equal to Conf ($R: A \rightarrow B$) = Supp (AB) / Supp (A). An association rule is said to be valid if, and only if, its confidence is greater than or equal to the threshold set by the user called minconf.

- *Fuzzy Formal Context*: It is a triplet K= (O, I, R) describing a set of objects O, a set of attributes I and a fuzzy binary relation $R \subseteq O \times I$. The value $u_R(o, i)$ with $o \in O$ and $i \in I$, is the association degree between o and i.

- *Fuzzy Galois Connection and Closure Operator*: K= (O, I, R) is a fuzzy formal context, for $X \subseteq O$ and $Y \subseteq I$. Operators Φ and Ψ are defined as follows:

  The fuzzy operator Φ is applied to a set of objects $X \subseteq O$ to determine a fuzzy set of items associated with all objects of X having the minimal degree

  Φ: P (O) → P (I)

  $$\Phi(X) = \{ i^{\alpha} \mid \forall o \in X, \alpha = \min \mu_{\tilde{R}}(o,i) \} \quad \textbf{(1)}$$

The fuzzy operator Ψ is applied on a fuzzy set of items Y ⊆ I providing a set of objects satisfying the constraint imposed by the input set.

Ψ: P (I) → *P (O)*

$$\psi = \{ \mathrm{o} \mid \forall \mathrm{i}, \mathrm{i} \in \mathrm{Y}, \mathrm{u}_{Y}(\mathrm{i}) \leq \mathrm{u}_{R}(\mathrm{o}, \mathrm{i}) \} \qquad \textbf{(2)}$$

- *Fuzzy Minimal Generator*: Let c be a fuzzy itemset, I' is FFCI, if I'= φ (c) and ∄ c1 ⊆ c such as φ (c1) = I', then c is a minimal fuzzy generator of I'. It's frequent if its support is greater than minsup.
- *Fuzzy Closed Itemset (FCI):* An itemset *I'* is an FCI iff I'= φ (I'). It's frequent if its support is greater than minsup.
- *Partial order relation between concepts <<:* Let (A1, B1) and (A2, B2) two fuzzy formal concepts: (A1, B1) << (A2, B2) ⇔ A2 ⊆ A1 and B1⊆ B2.
- *Meet/Join:* For each pair of concepts (A1, B1) and (A2, B2), there exists a greatest lower bound (resp. a least upper bound) called Meet (resp. Join) denoted as (A1, B1) ∧ (A2, B2) (resp. (A1,B1) ∨ (A2, B2)) and defined by

$$(A_1, B_1) \wedge (A_2, B_2)=(\Psi(B_1 \cup B_2),(B_1 \cup B_2)) \qquad \textbf{(3)}$$

$$(A_1,B_1) \vee (A_2, B_2)= ((A_1 \cup A_2),\phi(A_1 \cup A_2)) \qquad \textbf{(4)}$$

- *Iceberg Lattice:* It is a partially ordered structure of a frequent fuzzy closed itemset and having only a join operator.
- *Fuzzy Equivalence Class:* It is a set of frequent fuzzy itemsets having the same support and the same closure. The largest element of the equivalence class is a frequent fuzzy closed itemset called c and smaller ones are their minimal generators.
- *Frequent Fuzzy Minimal Generators Lattice:* It is a partially ordered structure where the nodes are equivalence classes.
- *Generic base of exact fuzzy association rules*

  Generic base of exact associative rules (GBEF) is a base composed of non-redundant generic rules having a confidence ratio equal to 1 [9].

  Let FG_k be the set of fuzzy frequent closed itemsets and FG_I the set of minimal generators of the itemset I. The generic base of exact fuzzy association rules is defined as follows:

$$GBEF = \left\{ R : g \rightarrow \left(I - g \right) | \mathrm{I} \in \left(FC_k \right) \wedge g \in \left(FG_I \right), g \neq I \right\}. \qquad \textbf{(5)}$$

- *Generic base of approximate fuzzy association rules*

  The generic base of approximate associative (GBAF) rules is defined as follows:

$$GBAF = \{ R : \mathrm{g} \rightarrow (I_1 - \mathrm{g}) \mid \mathrm{I}, \mathrm{I}_1 \in FC_k \ , \mathrm{g} \in FG_I \wedge I \subset I_1 \wedge, conf(R) \geq minconf \}. \qquad \textbf{(6)}$$

- *Generic base of transitive fuzzy association rules*

  The generic base of Transitive associative (RIF) rules is defined as follows:

$$RIF = \{ R : \mathrm{g} \rightarrow (I_1 - \mathrm{g}) \mid \mathrm{I}, \mathrm{I}_1 \in FC_k \ , \mathrm{g} \in FG_I \wedge I \subset I_1 \wedge \nexists I_2 \ \mathrm{s.t.} \mathrm{I} \subset \mathrm{I}_2 \subset \mathrm{I}_1 \wedge conf(R) \geq minconf \}. \quad \mathbf{(7)}$$

The exact fuzzy association rule is a relationship between the frequent fuzzy closed itemset FFCI and their minimal generators. However, the approximate rule is a relation of an FFCI with another FFCI that covers it and the transitive rule is a link between two FFCIs, one of which covers the other immediately.

3. Fuzzy Association Rules

Association rule is one of the most important unsupervised methods of data mining also called Market Basket Analysis. Several algorithms have been proposed in the literature to extract association rules. These algorithms deal only with binary contexts, whereas the real databases include not only binary data, but also quantitative data. In order to apply these algorithms, the quantitative databases must be converted into binary bases. This transformation causes several problems, namely the loss of information. This loss causes the non-coverage of the association rules of the processed database. To remedy this problem, fuzzy logic was introduced in the association rule extraction process to form a new category of association rules called fuzzy association rules. They convert numerical data into fuzzy data. This transformation maintains the integrity of the information conveyed by the numerical attributes. To extract fuzzy association rules, several algorithms have been introduced. These algorithms can be divided into two categories. The first category includes algorithms based on the extraction of frequent fuzzy itemsets [10], [1], [11], [12], while the second category includes algorithms based on extracting frequent fuzzy closed itemsets[6], [7], [13]. In the following section, we present these two categories of algorithms.

3.1 Algorithms based on frequent fuzzy itemset extraction

In this category, the fuzzy association rule has the following form:

If X is A, then Y is B.

With (X is A) is his premise of the rule and (Y is B) is his conclusion. This rule is noted (X, A) → (B, Y) where X = {x_1... x_p} and Y = {y1, ..., y_n} are two disjoint itemsets. A = {a_1.... a_p} and B = {b_1, .., b_n} are the sets of fuzzy subsets associated with X and Y.

Example:

If the age is young and the account balance is small then the loan is moderate. This rule is represented as follows: {Age, account balance} {Young, small} → {Loan} {Moderate}

The algorithms of this category are based on two phases:

1) Find all frequent itemsets
2) Generate all fuzzy association rules between frequent fuzzy itemsets having a confidence at least equal to *minconf.*

The first algorithms [5], [12], [14] of fuzzy association rules have been proposed to adopt the Apriori algorithm [15] in fuzzy contexts. They focused on reformulating rule validation measures. They offer formulas for support and confidence measures using fuzzy operations and implications.

They adopt the "test and generate" strategy where the algorithm browses the transactional database by level. The principle of these algorithms is to generate the frequent fuzzy itemsets iteratively.

They generate k-itemset (itemset having k items) then determinate the k+1-candidate itemsets by joining the k-itemsets obtained in the previous iteration and preserving only the frequent itemsets whose supports are greater than or equal to the minsup. This step is repeated until there are no candidate itemsets. The extraction of frequent itemsets step is based on the anti-monotonicity constraint. This constraint means that if we have two itemsets I1 and I2 with, I1 included in I2, then the support (I1)> support (I2). Indeed, all the over-itemsets of a non-frequent itemset are not frequent. This constraint reduces the number of candidates and the search space.

These algorithms require a considerable computation time due to an iterative access to the database and they generate huge number of rules, most of them are redundant rules, and often considered irrelevant. The resolution of this problem has been the subject of several studies [10], [11], [16–21]

3.2. Algorithms based on frequent closed fuzzy itemset Extraction

In this category, the fuzzy association rule has the following form:

$$\tilde{r}: \tilde{I}1 \Rightarrow \tilde{I}2$$

Where $\tilde{I}1, \tilde{I}2 \sqsubseteq \tilde{I} = \tilde{I}\{\frac{\alpha_1}{a_1}, \frac{\alpha_2}{a_2}, \ldots . \frac{\alpha_p}{a_p}, \frac{\alpha_q}{a_q}, \ldots . \frac{\alpha_n}{a_n}\}$, $\tilde{I}1 = \{\frac{\alpha_1}{a_1}, \frac{\alpha_2}{a_2}, \ldots . \frac{\alpha_p}{a_p}\}$, $\tilde{I}2 = \{\frac{\alpha_q}{a_q}, \ldots . \frac{\alpha_n}{a_n}\}$. $\tilde{I}1$, $\tilde{I}2$ are called, respectively, the premise and the conclusion of the fuzzy rule r. The value α_i, i = 1,. . ., n, is called the local weight of the element.

Example: R: $A^{0.2}$, $B^{0.3}$, $C^{0.1} \rightarrow D^{0.8} E^{0.5}$

If the attributes A, B and C, respectively, have at least the values 0.2, 0.3 and 0.1 **then** D and E, respectively, have at least the values 0.8 and 0.5.

This category includes the algorithms for extracting frequent fizzy closed itemsets [6], [7], [13]. An itemset is said to be closed if it does not have any superset with the same support. These algorithms use the fuzzy formal concept analysis FFCA in the process of extracting association rules. Its principle is to extract the set of frequent closed itemsets, from which a subset of rules is generated. This set of generic rules covers the entire extraction context, which ensures the non-loss of information [7]. These algorithms comprise two steps: the extraction of frequent closed itemsets based on the AFC and then the deduction of the generic bases of the association rules.

4. RELATED WORK

Since the setting of the algorithm of [5], many algorithms for extraction fuzzy association rules have been proposed. The major problems of these algorithms are their redundancy, their large number and finally, their degree of relevance. Indeed, several works have tried to deal with these problems. We categorized them into three categories: those that use quality measures, those that remove redundant rules, and those that reduce the extraction context and use ontology. In the following, we present these different categories

4.1. Use of quality measures

In order to deal with the problem of the relevance of the fuzzy association rules, some works propose to use quality measures. These measurements help the expert to validate them. In the following, we will list some of them.

Indeed, [12] proposes a method which consists of applying two measures: the factor of certainty and the concept of a very strong rule. The certainty factor (CF) is defined as follows:

$$\text{CF } (A \longrightarrow C) = \frac{conf(A \longrightarrow c) - supp(C)}{1 - supp(C)} \text{ if conf } (A \longrightarrow c) > \text{supp } (C) \qquad \textbf{(8)}$$

$$\text{CF } (A \longrightarrow C) = \frac{conf(A \longrightarrow c) - supp(C)}{supp(C)} \text{ if conf } (A \longrightarrow c) < \text{supp } (C) \qquad \textbf{(9)}$$

The value of the certainty factor is between -1 and 1. It is positive when the dependence between the premise and the consequence is positive. It is negative when the dependence is negative; finally it is zero when the premise and the consequence are independent.

The second measure used in [12] is the new concept of a very strong rule. Indeed, a fuzzy association rule is said to be very strong if the two rules A→ C and ¬A→ ¬C are valid. According to [12], the extraction of very strong rules probably leads to the generation of relevant knowledge.

In 2006, [22] used the correlation measure to evaluate the interest of a rule. A rule is not presented to the decision maker if its interest is <1.

$$\text{Fcorr } (<X{:}\ A><Y{:}\ B>) = \frac{\text{supp } (Z{:}\ C)}{(\text{supp } (X{:}\ A)\ .\text{supp } (Y{:}\ B))} \qquad \textbf{(10)}$$

[16] extended three measures of intensity of classic association rules to use them as part of the fuzzy association rules. These three measures are lift, conviction and leverage. All these measures are based on the assumption of independence.

The lift is a measure of quality that represents the relationship of independence between the premise and the conclusion of the rule. It is the ratio between the observed support and the expected support under the independence hypothesis. The lift is a symmetrical non-implicative measure. It is sensitive to the size of the data: it is a statistical measure. The values of lift ∈ [0;+1].

Conviction also measures independence but between counterexamples. It is the ratio between the number of counterexamples under the assumption of independence and the number of counterexamples observed. It is a non-symmetrical and implicative measure, unlike Lift measurement. However its values ∈ [0; +1] as in the case of the lift.

Leverage measures the difference between the observed support and the expected support under the independence assumption.

The author has proved that the simple substitution of the binary medium by the fuzzy support in the classical formulas of these measurements does not give a correct definition and generates erroneous results. In order to propose a correct definition of the measures, the author has defined the expected support and the expected confidence under independence hypothesis as follows
Let ⊗ be the t-norm, X and Y are fuzzy attributes; the expected support is then equal to:

$$\widehat{fsupp}(X \to Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{X(o_i) \otimes Y(o_j)}{n^2} \quad \textbf{(11)}$$

The expected confidence is equal to

$$\widehat{conf}(X \to Y) = \frac{fsupp(X \to Y)}{fsupp(X)} \quad \textbf{(12)}$$

The author has proposed the definition of the three measures as follows:

$$flift(X \to Y) = \frac{fsupp(X \to Y)}{\widehat{fsupp}(X \to Y)} \quad \textbf{(13)}$$

$$flever(X \to Y) = fsupp(X \to Y) - \widehat{fsupp}(X \to Y) \quad \textbf{(14)}$$

$$fconv(X \to Y) = \frac{\widehat{fsupp}(X \to \neg Y)}{fsupp(X \to \neg Y)} \quad \textbf{(15)}$$

4.2 Removal redundant rules

In order to reduce the enormous number of extracted rules, other works have proposed to remove the redundant rules. Each offers its definition of redundant fuzzy association rules.

[23] defined the redundant rule as follows: Let A, B, C be three itemsets, A→ B and A→C are redundant rules if there is a valid fuzzy association rule A →B $\cup$ C.

[20] is also among the researchers who presented a new fuzzy association rules extraction algorithm that eliminates the rules redundant. The procedure for pruning redundant rules is based on the idea that the confidence value of the rule should increase by increasing the number of elements in the premise.

He has defined the redundant rule as follows: let A, B, and C be three itemsets. A and B are disjoint itemsets and Q contains the subsets of A.

If $\max_{C \in Q}(\text{conf}(C \to B)) \geq \text{conf}(A \to B)$ then A→B is a redundant rule.

The author has also defined the concept of the strong redundant rule.

If $\min_{C \in Q}(\text{conf}(C \to B)) \geq \text{conf}(A \to B)$ then $A \to B$ is a very redundant rule

However, according to [24], the proposed algorithm fails because it sometimes eliminates non-redundant rules.

To overcome this problem, [21] propose an improvement of the algorithm presented by [20] by introducing a new notion called notion of equivalence of fuzzy association rules. Its role is to prevent the generation of redundant rules and to prune the redundant itemsets.

The equivalence rules are defined as follows:

Let F = $\{B_1, B_2, \ldots B_m\}$ be a fuzzy itemset, where B is a fuzzy item (label) defined on different attributes and m is the number of elements (m> 1) and q is a threshold equivalence fixed in advance and higher than the predefined value of the minconf.

If $\mathrm{conf}(\ \mathrm{U}_{i\neq s}\ \mathrm{B}_i \rightarrow \mathrm{B}_S) \geq q, \forall s \in \{1,2, \ldots\ldots m\}$ then the rules generated from F are equivalence rules.

The principle of using the notion of equivalence during pruning is as follows:

Let F = $\{B_1, B_2, \ldots B_m\}$ be a fuzzy equivalence itemset, G = $\{B_1, B_2, \ldots B_n\}$ a fuzzy itemset (F includes G), where B is a fuzzy item (label) defined on different attributes and m and n are respectively the number of elements of F and G (m> 1, n> 1, n> m). Let q be an equivalence threshold and R_F and R_G are association rules generated from F and G, respectively:

$$R_F : \bigcup_{i\neq s}^{m} B_i \rightarrow B_s, \exists s \in \{1,2,\ldots,m\}, R_G : \bigcup_{i\neq s}^{n} B_i \rightarrow B_s$$

If confidence (R_F)> q, confidence (R_G)> q then R_G rules are redundant rules.

4.3. Context Reduction and Use of Ontology

This category includes works that proposed to solve the problem of the number and the relevance of fuzzy association rules by reducing the context and using ontology such as [17]. [17] proposes a method that includes two phases. The first is a preprocessing step. It consists of finding the attributes that have similar behaviors and merging them. Indeed, to measure the similarity of behavior between attributes, the authors use Chi-square test X^2. The second phase consists to use of the ontology to reduce the candidate itemsets. Indeed, the ontology contains taxonomic relations related to each concept, and the semantic relations between them. To generate frequent itemsets, they examine only the relationships between items related to a concept or items related to different concepts having semantic relations in the ontology.

4.4. Discussion

To overcome the problem of relevance of the fuzzy rule, [12], [16], [23], [25], [26] have used a measure to test the validity of extracted rules. This evaluation technique is based on the following principle: following the choices of the measure as well as the threshold of validity by the expert, only the rules having a value higher than or equal to the set threshold are retained. However, several problems can arise. The first problem concerns the arbitrary setting of the threshold which may not cover the desired domain. The second problem is related to the number of extracted rules that can be numerous. In these different cases, the expert finds difficulties during the validation of the rules.

In [17], the authors tried to derive the context by measuring the similarity between the attributes. They build several contingency tables which greatly increases the execution time. This approach also uses the ontology to reduce the number of candidate rules which requires finding or constructing ontology for each context.

[20], [21] have proposed algorithms to remove redundant rules. The deletion of the rules is based on the deletion of the itemsets during the extraction process. However, these algorithms suffer from some problems namely: the suppression of non-redundant rules [20], the fixation of the value of the parameter q (equivalence threshold) [21], because any variation of this value can produce different results. In addition, the performance of these two algorithms is not valid because in their experiments the authors used medium-sized test bases. So, we cannot predict how the proposed algorithm will behave in case of large database.

Despite these efforts, the problem of the pertinence and the huge number of fuzzy association rules still persists. In order to resolve these problems, we propose a new validation method able to extract a generic basis of fuzzy association rules and validate them automatically based on fuzzy formal concepts analysis and structural equation model.

5. Validation Method of Fuzzy Association Rules Based on SEM

In order to validate generated fuzzy association rules and present to the decision maker only useful rules, we propose a new method entitled VMFAR-SEM (Validation Method of Fuzzy Association Rules based on Structural Equation Model). It consists in validating the fuzzy association rules by exploiting the structural equation model. In the following, we present the principle of our method and we illustrate it with an example

5.1. Structural Equation Model

Structural equation model is a powerful, versatile, multi-varied and very general analysis technique used to evaluate the validity of hypotheses with empirical data [27]. The structural equation model offers the flexibility to research and interpret theory and data. It also makes it possible to simultaneously estimate several dependency relationships.

There are two types of variables in this model.

- Manifest variable is a directly collected variable (observed, measured).
- Latent variable is a variable that cannot be directly measured. These variables can be estimated from overt variables.

The structural equation model is decomposed into two sub-models:

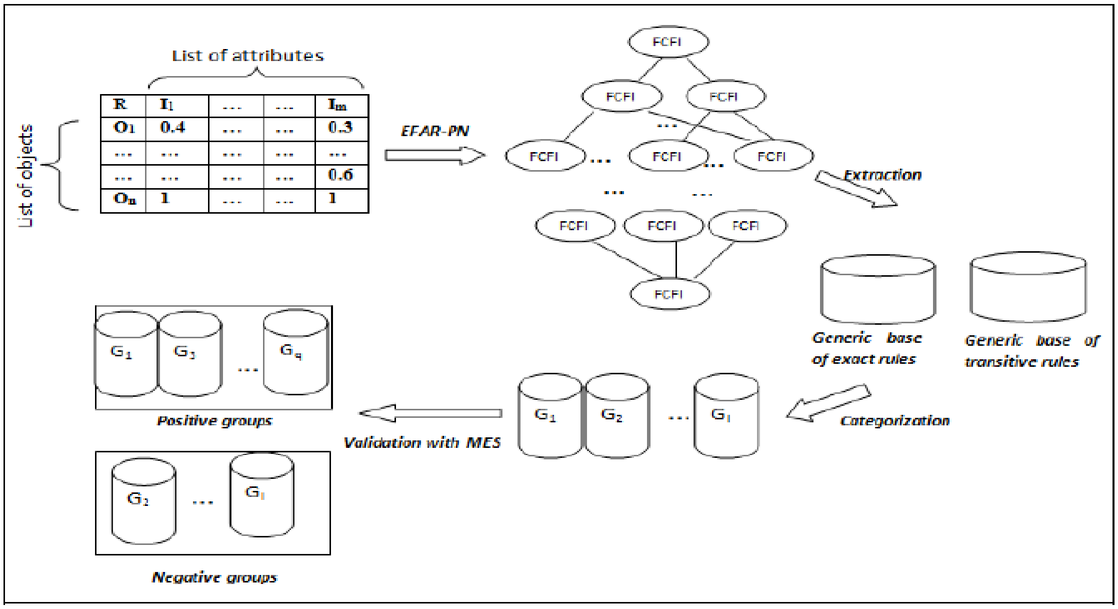
- Structural model or internal model is a subset of the complete model including relationships between latent variables.
- Measurement model or external model is a subset of the model

In order to estimate all the relations of the model (relation between the latent variable or relation between latent variable and its indicators), there are two approaches LISREL or PLS. In our work we will use the PLS approach.

The Partial Least Square (PLS) approach is one of the approaches to the structural equation model that comes from an earlier theory called least squares estimation [28]. This theory is based on simple and multiple regressions. Thus, it requires few assumptions and hence its name "soft modeling".

5.2. Principle of the proposed method

Our method consists in validating the fuzzy association rule. First we apply our algorithm EFAR-PN [13] to extract generic basis of fuzzy association rule. Then, we apply two steps: The first is to classify rules into groups, according to the items of the premises and conclusions. The second validates the rules by applying the PLS approach on the representative rule of each group. The architecture of our method is shown in Figure 1.



R	I_1	...	...	I_m
O_1	0.4	...	...	0.3
...	...	...	...	...
...	...	...	...	0.6
O_n	1	...	...	1

Figure 1. The general architecture of our method

In the following, we explain in detail the principle of each step.

5.2.1. Extraction of generic bases of fuzzy association rules

In this step, we apply our algorithm EFAR-PN. This algorithm makes it possible to extract non-redundant rules and without loss of information, while minimizing the execution time. In fact, it extracts frequent fuzzy itemsets without repetitive access to the extraction context and then determines the generic basis of the exact and approximate fuzzy association rules by constructing the iceberg lattice. The construction of the lattice is based on a system of encoding prime numbers. These databases provide the user with a reduced subset of Fuzzy Association Rules (RAFs) covering the entire initial retrieval context and with as much relevant and useful knowledge as possible.

Our EFAR-PN algorithm (Extraction of Fuzzy Association Rules based on the Prime Numbers) has three steps. The first step is to extract the fuzzy minimal generators (GMFFs) by performing a single access to the fuzzy formal context. The second step consists of building the fuzzy minimal generators lattice. In the third step, the ERAF-NP algorithm deduces the Iceberg lattice of frequent closed fuzzy itemsets (IFFF) and extracts the fuzzy association rules from the lattice of frequent fuzzy minimal generators. (for more details please refer to [13]).

5.2.2. Categorization of Fuzzy Association Rules

This step consists of classifying fuzzy association rules into groups. The rules having the same attributes in the premise and the same attributes in the conclusion belong to the same group. Each group has a representative rule. This grouping allows giving a synthetic view of the rules to the user which facilitates their interpretation.

For example

Let have following rules:

R1: $A^{0.5}, B^{0.6}, C^{0.2} \rightarrow D^{0.5}, E^{0.2}$
R2: $A^{0.6}, C^{0.3} \rightarrow D^{0.6}, F^{0.1}$
R3: $A^{0.3}, B^{0.2}, C^{0.8} \rightarrow D^{0.7}, E^{0.3}$
R4: $A^{0.2}, C^{0.7} \rightarrow D^{0.2}, F^{0.7}$

Then, we have two groups: G1 contains R1 and R3 and have A B C → D E as representative rule. G2 include two rules R2 and R4 and have A C → D as representative rule.

After classifying the rules into groups, we will apply the validation step.

5.2.3. Validation of rules using SEM

In this step, we apply the PLS approach to each representative rule cleared during the previous step. Each rule is considered a model of structural equations that contains two latent variables. The indicators of the first latent variable are the attributes of the premise and the indicators of the second are attributes of the conclusion as shown in Figure 2.

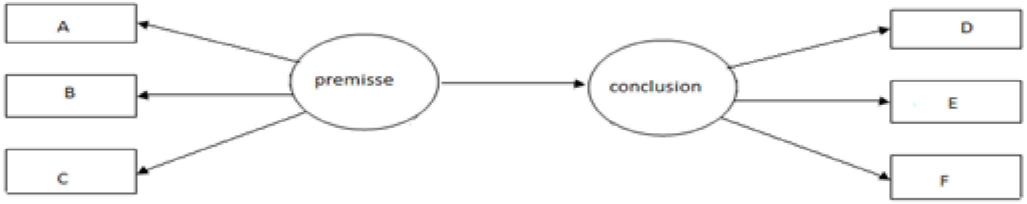


Figure 2. The structural equation model to Measure the ABC⟶ DEF

After creating the model, we apply PLS approach to estimate the different coefficients of the model. We first verify the validity of the model. PLS offers several indices measuring the intensity with which the model replicates the database. We will use the coefficient of determination R^2 and α of cronbach in our method

Then, we interpret the coefficients between the two latent variables (premise and conclusion) and the latent variable and its indicators (items).

The coefficient between the two latent variables is the slop of a simple line regression. This line has the following equation:

Conclusion = a * premise + b

With a is the slope of the line and b is the value at the origin.

If a is positive, then any increase of the premise leads to an increase of the conclusion. A high value of the slope indicates a significant influence of the premise in the conclusion. Indeed, a small change in the premise conducts to a big change in the conclusion.

5.3. Illustration of our method

In order to explain the progress of our method, we illustrate its different steps through an example. We will apply our method on the extraction context shown in Table 1.

Table1. Fuzzy extraction context

B	**C**	**E**	**M**
0.4	0.3	0.2	0.3
0.8	1	0.2	0.6
0.8	1	0.5	0.6
1	0.3	1	1

In the following, we apply the steps of our method:

First step:

We apply our algorithm EFAR-PN on the extraction context with a minsup=0.25 and minconf=0.5. The result of this step is shown in table 2. We have 23 association rules. The base of exact association rules contains 12 rules while the base of transitive fuzzy association rules contains 11 rules.

Table 2. Bases of exact and transitive fuzzy association rules

Base of exact fuzzy association rules		**Base of transitive fuzzy association rules**		
Rules	**Support**	**Rules**	**Support**	**Confidence**
$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.3}$	1	$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$	0.75	0.75
$C^{0.3} \rightarrow B^{0.4}, E^{0.2}, M^{0.3}$	1	$C^{0.3} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$	0.75	0.75
$E^{0.2} \rightarrow B^{0.4}, C^{0.3}, M^{0.3}$	1	$E^{0.2} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$	0.75	0.75
$M^{0.3} \rightarrow B^{0.4}, C^{0.3}, E^{0.2}$	1	$M^{0.3} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$	0.75	0.75
$B^{0.8} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$	0.75	$B^{0.8} \rightarrow C^{1.0}, E^{0.2}, M^{0.6}$	0.5	0.6
$M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$	0.75	$B^{0.8} \rightarrow C^{0.3}, E^{0.5}, M^{0.6}$	0.5	0.6
$C^{1.0} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$	0.5	$M^{0.6} \rightarrow B^{0.8}, C^{1.0}, E^{0.2}$	0.5	0.6
$E^{0.5} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$	0.5	$M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.5}$	0.5	0.6
$B^{1.0} \rightarrow C^{0.3}, E^{1.0}, M^{1.0}$	0.25	$C^{1.0} \rightarrow B^{0.8}, E^{0.5}, M^{0.6}$	0.25	0.5
$E^{1.0} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$	0.25	$E^{0.5} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$	0.25	0.5
$M^{1.0} \rightarrow B^{1.0}, C^{0.3}, E^{1.0}$	0.25	$E^{0.5} \rightarrow B^{0.8}, C^{1.0}, M^{0.6}$	0.25	0.5
$C^{1.0}, E^{0.5} \rightarrow B^{0.8}, M^{0.6}$	0.25			

Second step:

In this step, we divide the extracted fuzzy association rules into groups. The result of applying this step is shown in Table 3. This step allows summarizing the 23 rules in 5 groups.

Table 3. Fuzzy association rule groups

1	Exact fuzzy association rules	Transitive fuzzy association rules	Representative rule
G1	$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.3}$ $B^{0.8} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$ $B^{1.0} \rightarrow C^{0.3}, E^{1.0}, M^{1.0}$	$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$ $B^{0.8} \rightarrow C^{1.0}, E^{0.2}, M^{0.6}$ $B^{0.8} \rightarrow C^{0.3}, E^{0.5}, M^{0.6}$	B → C, E, M
G2	$C^{0.3} \rightarrow B^{0.4}, E^{0.2}, M^{0.3}$ $C^{1.0} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$	$C^{0.3} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$ $C^{1.0} \rightarrow B^{0.8}, E^{0.5}, M^{0.6}$	C → B, E, M
G3	$E^{0.2} \rightarrow B^{0.4}, C^{0.3}, M^{0.3}$ $E^{0.5} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$ $E^{1.0} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$	$E^{0.2} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$ $E^{0.5} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$ $E^{0.5} \rightarrow B^{0.8}, C^{1.0}, M^{0.6}$	E→ B, C, M
G4	$M^{0.3} \rightarrow B^{0.4}, C^{0.3}, E^{0.2}$ $M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$ $M^{1.0} \rightarrow B^{1.0}, C^{0.3}, E^{1.0}$	$M^{0.3} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$ $M^{0.6} \rightarrow B^{0.8}, C^{1.0}, E^{0.2}$ $M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.5}$	M → B, C, E
G5	$C^{1.0}, E^{0.5} \rightarrow B^{0.8}, M^{0.6}$		C, E → B, M

Third step:

In this step, we build a model of structural equations for each representative rule. We estimate these models using the approach PLS. We start with the representative rule B-> CEM. The result obtained is shown in Figure 3.

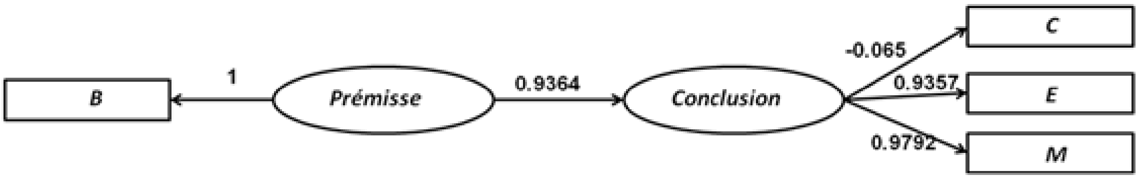


Figure 3. Structural Equation Model for B→ CEM

According to the first model of structural equations applied to the representative rule of the first group (B -> CEM), we find that the premise has a strong positive influence on the conclusion (coef = 0.936) and that the increase from B leads to:

- The decrease in C
- The increase of E
- The increase of M

Table 4 shows the result of applying this step on the rules representative of the rest of groups obtained in the first step.

Table 4. The coefficients of the model of the representative rules

Group	**Coefficients of the premise indicators**	**Coefficient between premise and conclusion**	**Coefficients of the conclusion indicators**
G2	1 **C**	-0.486664263392288	0.669890634808308 **B** 0.992430570086382**E** 0.847801045943832**M**
G3	1$^{\mathbf{E}}$	0.945169821893008	0.773122361529185**B** -0.43997389929851**C** 0.937682683668112**M**
G4	1**M**	0.954856975574631	0.84089923263202**B** -0.284435741070428**C** 0.988469944399833$^{\mathbf{E}}$
G5	0.346530943591408**C** -0.999259806037586**E**	-0.867319793400418	0.979881037935833**B** 0.991403087791832**M**

We can conclude from table 4 the following remarks:

- According to the second model C → BEM, we notice that the premise has a negative effect on the conclusion (coef = -0.48). Hence the increase of C leads to a decrease in the conclusion set B, E and M.
- According to the third model E →BCM, we synthesize that the premise has a positive and significant influence with a coefficient equal to 0.94. By Therefore, the increase of the premise (increase of E) leads to a increase of the conclusion (increase of B and M and decrease of C).
- According to the fourth model M → BCE, we conclude that the premise has a positive impact on the conclusion and that the increase of one generates the increase of the other. The increase in M leads to the increase of B, the decrease of C and the increase of E.
- According to the fifth model CE → BM, we find that the premise has a negative effect on the conclusion. The increase of C with the decrease of E leads to a decrease of all the conclusion attributes (B and M).

6. EXPERIMENTAL STUDY

In this section, we will evaluate our method through using three basics of Fars2008 test, Pendigits and Abalone.

- Abalone[1]: is a base that comes from the UCI Machine Learning Repository. This base represents the physical measurements of the shells. Each is described with 8 variables. This database contains 4177 instances.

- Fars-2008[2]: Data from this database come from the US FARS archive (Fatality Analysis Recording System) which aims at including all accidents in which there has been at least one death. The data concern automobiles where the front passenger seat was occupied, with one observation for each passenger.

[1] https ://archive.ics.uci.edu/ml/datasets.html

[2] https ://vincentarelbundock.github.io/Rdatasets/datasets.html

- Pendigits1: Pen-Based Recognition of Handwritten Digits Data Set, is a database available in the UCI Machine Learning Repository. This database is produced by collecting 250 examples of 30 writers, written on a pressure sensitive tablet that sent the pen location at fixed time intervals of 100 milliseconds.

The main characteristics of these databases are described in Table 5.

Table 5. Characteristics of the bases of test

Base	Number of objects	Number of attributes
Abalone	4177	8
Fars-2008	64881	24
Pendigits	7494	17

First step:

We apply our algorithm on each base of test. Table 6 shows the number of rules extracted in each database with a value of minsup equal to 0.8.

Table 6. Number of extracted rules

Base	Number of exact rules	Number of transitive rules
Pendigits	1988	6366
Abalone	1320	5526
Fars 2008	87	207

We notice that the number of rules can be high even with a high minsup (0.8).

Second step:

During the second step, we will divide the extracted rules into groups and identify the representative rule of each group. Table 7 shows the number of groups released for each test basis. We find that this step significantly reduces the number of rules. This step reduces 96% for Pendigits, 87% for Abalone and 77.78% for Fars2008. This reduction makes it easier for the user to explore generated knowledge.

Table 7. The result of the application of the categorization step on the three bases.

Base	Number of fuzzy association rules	Number of groups
Pendigits	8354	321
Abalone	6846	903
Fars2008	294	65

Third step:

After determining the different groups, we will apply the second step. The latter is the validation of the rules using PLS. For each group, we construct a structural equation model of its representative rule. The result of this step is shown in table 8.

Table 8. The result of the application of the validation step on the three bases.

Base	Positive groups	Negative groups
Pendigits	265(7933 règles)	56(521 règles)
Abalone	731(5581 règles)	172(1265 règles)
Fars2008	18(76 règles)	47(218 règles)

Positive groups contain rules with a positive PLS coefficient. In these rules, the premise has a positive influence on the conclusion.

Negative groups contain the rules with negative coefficient of degree. In these rules, the premise has a negative influence on the conclusion. We order the representative rules according to the coefficient. The more the coefficient is big and the impact is significant, the more the rule is relevant.

7. Conclusion

In this article, we have presented our method of validation of fuzzy association rules based on the structural equations model. This method has three steps. The first is to extract generic bases of fuzzy association rules using EFAR-PN algorithm based on fuzzy formal concept analysis. The second step is to classify the rules into groups according to their attributes and to determine a representative rule for each group. This provides a synthetic view of the rules. The third is to construct a model of structural equations from each representative rule. We applied the PLS approach to estimate model coefficients. The PLS coefficient makes it possible to check if the premise has a positive or negative influence on the conclusion. This information can be very useful in many areas. In the future work we plan to create an interactive prototype that integrates the various contributions and that allows the visualization of the rules for the expert to validate them easily. We also plan to test our VMFAR-SEM method on a real world application such as marketing or biology and validate it with a domain expert. For the validation of fuzzy gradual rules, we also intend to apply our VMFAR-SEM method to validate them in a semi-automatic way.

References

[1] S. Papadimitriou and S. Mavroudi, “The Fuzzy Frequent Pattern Tree,” in Proceedings of the 9th WSEAS International Conference on Computers, 2005, pp. 3:1–3:7.

[2] P. Rajendran and M. Madheswaran, “Novel Fuzzy Association Rule Image Mining Algorithm for Medical Decision Support System,” International Journal of Computer Applications, vol. 1, no. 20, pp. 87–94, 2010.

[3] N. Gupta, N. Mangal, K. Tiwari, and P. Mitra, “Mining quantitative association rules in protein sequences,” Data Mining, pp. 273–281, 2006.

[4] R. Natarajan and B. Shekar, “Interestingness of association rules in data mining: Issues relevant to e-commerce,” Sadhana, vol. 30, no. 2–3, pp. 291–309, 2005.

[5] K. C. Chan and W.-H. Au, “Mining fuzzy association rules,” Proceedings of the sixth international conference on Information and knowledge management, pp. 209–215, 1997.

[6] S. Ben Yahia and A. Jaoua, “Data Mining and Computational Intelligence,” J. Kacprzyk, A. Kandel, M. Last, and H. Bunke, Eds. Heidelberg, Germany, Germany: Physica-Verlag GmbH, 2001, pp. 167–190.

[7] S. Ayouni, “Etude et extraction de regles graduelles floues: définition d’algorithmes efficaces,” 2012.

[8] M. Kryszkiewicz, “Concise representations of association rules,” Pattern Detection and Discovery, pp. 92–109, 2002.

[9] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, “Mining minimal non-redundant association rules using frequent closed itemsets,” Computational Logic—CL 2000, pp. 972–986, 2000.

[10] A. Mangalampalli and V. Pudi, “FPrep: Fuzzy clustering driven efficient automated pre-processing for fuzzy association rule mining.,” in FUZZ-IEEE, 2010, pp. 1–8.

[11] C.-H. Chen, T.-P. Hong, and Y. Li, “Fuzzy association rule mining with type-2 membership functions,” Intelligent Information and Database Systems, pp. 128–134, 2015.

[12] M. Delgado, N. Marín, D. Sánchez, and M. Vila, “Fuzzy association rules: general model and applications,” IEEE Transactions on Fuzzy Systems, vol. 11, pp. 214–225, 2003.

[13] I. Mguiris, H. Amdouni, and M. M. Gammoudi, “An Algorithm for Fuzzy Association Rules Extraction Based on Prime Number Coding,” in 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2017, Poznan, Poland, June 21-23, 2017, 2017, pp. 182–184.

[14] C. M. Kuok, A. Fu, and M. H. Wong, “Mining fuzzy association rules in databases,” ACM Sigmod Record, vol. 27, no. 1, pp. 41–46, 1998.

[15] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, pp. 487–499, 1994.

[16] M. Burda, “Interest measures for fuzzy association rules based on expectations of independence,” Advances in Fuzzy Systems, vol. 2014, p. 2, 2014.

[17] Z. Farzanyar and M. Kangavari, “Efficient mining of Fuzzy Association Rules from the Pre-Processed Dataset,” Computing and Informatics, vol. 31, no. 2, pp. 331–347, 2012.

[18] J. C.-W. Lin, T.-P. Hong, and T.-C. Lin, “A CMFFP-tree Algorithm to Mine Complete Multiple Fuzzy Frequent Itemsets,” Appl. Soft Comput., vol. 28, no. C, pp. 431–439, Mar. 2015.

[19] R. Prabamanieswari, “Article: A Combined Approach for Mining Fuzzy Frequent Itemset,” IJCA Proceedings on International Seminar on Computer Vision 2013, vol. ISCV, pp. 1–5, Jan. 2014.

[20] T. Watanabe, “Fuzzy association rules mining algorithm based on output specification and redundancy of rules,” Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 283–289, 2011.

[21] T. Watanabe and R. Fujioka, “Fuzzy association rules mining algorithm based on equivalence redundancy of items,” Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, pp. 1960–1965, 2012.

[22] M. Kaya and R. Alhajj, “Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining,” Applied Intelligence, vol. 24, no. 1, pp. 7–15, 2006.

[23] Y. Gao, J. Ma, and L. Ma, “A new algorithm for mining fuzzy association rules,” in Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, 2004, vol. 3, pp. 1635–1640 vol.3.

[24] A. Roy and R. Chatterjee, “A survey on fuzzy association rule mining methodologies,” IOSR J. Comput. Eng.(IOSR-JCE), e-ISSN, pp. 2278–661, 2013.

[25] M. Kaya, R. Alhajj, A. Arslan, and others, “Efficient automated mining of fuzzy association rules,” in International Conference on Database and Expert Systems Applications, 2002, pp. 133–142.

[26] S. Lotfi and M. Sadreddini, “Mining fuzzy association rules using mutual information,” in Proceedings of the International MultiConference of Engineers and Computer Scientists, 2009, vol. 1.

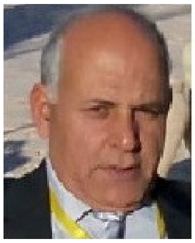
[27] W. W. Chin, “The partial least squares approach to structural equation modeling,” Modern methods for business research, vol. 295, no. 2, pp. 295–336, 1998.

[28] H. Wold, “Soft modeling: the basic design and some extensions,” Systems under indirect observation, vol. 2, pp. 589–591, 1982.

AUTHORS

Imen Mguiris received her Master degree in Computer Science at ISIMS-Tunisia in 2010. Now, she prepared her PhD at the Faculty of Sciences of Tunis. Her main research contributions concern: data mining, Fuzzy Association Rules, Formal Concept Analysis (FCA). She is member of Research Laboratory RIADI

Hamida Amdouni obtained her Ph.D. and DEA in Computer Science from the Faculty of Science of Tunis respectively in 2014 and 2005. She is currently Assistant Professor at the School of Digital Economy (ESEN), University from Manouba. Her main areas of research are Data Mining, Formal Concepts Analysis (CFA), CRM and Big Data. She is also a member of the SCO-ECRI team at the RIADI Research Laboratory.



Mohamed Mohsen Gammoudi is currently a full Professor in Computer Science Department at ISAMM, University of Manouba. He is the responsible of the research Team ECRI of RIADI Laboratory. He obtained his HDR in 2005 at the Faculty of Sciences of Tunis (FST). He received his PhD in 1993 at Sophia Antipolis Laboratory (I3S/CNRS) in the team of Professor Serge Miranda, France. He was hired as a Visiting Professor between 1993 and 1997 at the Federal University of Maranhao, Brazil.