



**HAL**  
open science

## Rule-based diagnostic system fusion

Abdelhamid Zemirline, Laurent Lecornu, Basel Solaiman

► **To cite this version:**

Abdelhamid Zemirline, Laurent Lecornu, Basel Solaiman. Rule-based diagnostic system fusion. Fusion 2007: 10th International Conference on Information Fusion, July 9-12 juillet Québec, Canada, Jul 2007, Québec, Canada. 10.1109/ICIF.2007.4408205 . hal-02121388

**HAL Id: hal-02121388**

**<https://hal.science/hal-02121388>**

Submitted on 6 May 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Rule-Based Diagnostic System Fusion

Zemirline A.  
ITI Department  
ENST Bretagne  
LATIM, INSERM-U650  
29285 Brest, France  
abdelhamid.zemirline@enst-bretagne.fr

Lecornu L.  
ITI Department  
ENST Bretagne  
LATIM, INSERM-U650  
29285 Brest, France  
laurent.lecornu@enst-bretagne.fr

Solaiman B.  
ITI Department  
ENST Bretagne  
LATIM, INSERM-U650  
29285 Brest, France  
basel.solaiman@enst-bretagne.fr

**Abstract**—In this work, we present a new fusion method that uses fuzzy set theory. This method is applied to the diagnostic system rule bases. It aims at combining all the rule bases into only one rule base and then taking into consideration the characteristics of this base. The fusion method is characterized by a hybrid fusion which combines rule fusion approach with knowledge fusion approach. Knowledge fusion relies on the distortion measure of various bases. This distortion measure is integrated into the rule fusion process in order to generate one rule base for improving the diagnostic system performance. It is defined as the confidence degrees associated to each rule base parameter. The confidence degrees are then integrated into prediction procedure of the new diagnostic system.

**Keywords:** Data mining, Data fusion, Rule fusion, Knowledge-based systems, Diagnostic systems, Fuzzy set theory.

## I. INTRODUCTION

Nowadays, several institutions and organizations combine homogeneous data coming from different sources and/or produced at different instants. This situation is faced, in particular, by medical centers which have to store a new set of data each year. On the one hand, the data have to be exploited to extract new information while on the other hand, it have to be amalgamated with its older versions. The fusion procedure is important to tackle this kind of problems.

In this work, we have the rule bases coming from different harmonized sources, i.e. the rule bases have a significant number of parameters in common. The objective of our fusion is not limited only to augment the rule base but also to allow to the diagnostic system to have more precise and more relevant results for the new case prediction. This is possible by taking into account the conflict of the merged rule bases and then integrating it into prediction procedure of the new diagnostic system.

There are three main types of fusion methods which can be distinguished according to the conceptual level of information [1]: data fusion, decision fusion and model fusion.

- Data fusion is a fusion process operated on the first conceptual level of information. It consists of combining raw data resulting from several sources or various primitive levels extracted from only one source in order to deduce perfect and less noisy data.
- Decision fusion is the solution of the problem modeling applied on a specific data set. Several data sources and

several types of processing can respectively provide a decision for the same problem. Thus, when several sensors observe the same scene or where several independent approaches make it possible to provide a solution then, the decision fusion consists of confronting the solutions suggested by various systems in order to choose only the most realistic one, or to combine these decisions in order to choose a more powerful, more reliable or more careful decision.

- Model fusion is a concept which combines data processing and artificial intelligence. A model characterizes and represents in a more or a less complex way the knowledge that composes the advanced system. The model fusion either builds the new knowledge model or adopts a compromise of the precedents.

In our case, the decision fusion is usually applied and is based on the data source confidences and relevance [2]; however, we have to know the exact information about each source. Many methods such as that of the vote [3], that of the classifiers collaborate [4] and others could be used. Nevertheless, even when these methods are efficient, they are unable to get the synthesis knowledge provided by each source.

The fusion type that we propose in this study is a hybrid fusion combining two types of fusion: rule fusion and knowledge fusion. Besides, our fusion method is a model fusion because it combines knowledge bases and rule bases of several diagnostic systems in order to build a new diagnostic system.

This rule fusion consists of grouping several sources of rule bases to obtain only one rule base. The knowledge fusion consists of attributing the confidence degree to each parameter of the new rule base diagnostic system. The confidence degree is produced by measuring the conflict of several sources of the knowledge base.

In this work, we have tackled the regularly confronted problem of integrating the new case bases into a diagnostic system. As an example, we take a diagnostic system which is applied to medical bases. These medical bases are of the same type and contain descriptions of endoscopic lesions. However, each one of them has its own features when describing the lesions.

This article is organized in the following way: in the second section, we describe the diagnostic system on which our

method of fusion is applied and the architecture on which our system is based. In section three, we describe the fusion method. In the section fourth, we analyze the results obtained through our method. We finish finally with a conclusion in the fifth section.

## II. DIAGNOSIS SYSTEM

The diagnostic system is composed of two bases: the internal knowledge base and the external knowledge base. The internal knowledge base is a rule base which enables the diagnostic system to make a diagnosis. This rule base is generated from a case base.

The external knowledge base is the representation of the case base characteristics. It is considered as an interface that permits to compare the rule bases of various systems.

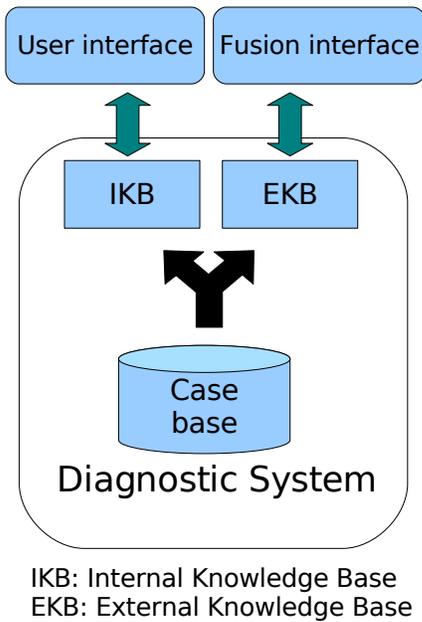


Figure 1. Diagnostic system architecture.

Figure 1 represents the system architecture on which fusion procedure is applied.

### A. Internal knowledge base

The internal knowledge base is generated from the case base which is a normal relation table consisting of  $N$  elements, and each element is described by its own item set. An item is composed of a discrete value. Each case of this case base belongs to a single class label. The rules of internal knowledge base are composed of item set and class label as the database element. So the internal knowledge base is the rule base.

The rule is an implementation of the form  $X \rightarrow c$ , where  $X$  is the item set, and  $c$  is a label class. This rule is kind of class-association rule and it is generated by the associative classification method [5], [6].

The internal knowledge base is used by the diagnostic system in order to predict the class of the new instance (the

new element) whose class label is undefined. The prediction procedure used by the diagnostic system is described as in the following way: Given a rule set from the internal knowledge base which covers the new instance. This rule set contains the rules of each class and the diagnostic system uses the best rules of each class for prediction [7], with the following procedure:

- 1) Select all the rules whose conditions are satisfied by the new instance.
- 2) From the rules selected in step (1), select the best  $k$  rules for each class (i.e. the rules have the high weight).
- 3) Compare the expected average accuracy of the best  $k$  rules of each class and choose the class with the highest expected accuracy as the prediction class.

The diagnostic system uses multiple rules in prediction where (a) the accuracy of rules cannot be precisely estimated, and (b) one expects that any single rule can perfectly predict the class label of every new instance satisfying its body. Moreover, the diagnostic system uses the best rules instead of using all the rules because there are different number of rules for different classes and the diagnostic system doesn't need to use low ranked rules in the prediction when there are already enough rules to make a prediction [6], [7].

### B. External knowledge base

The external knowledge base is generated from the case base like the rule base. It designed to be a fusion interface to a diagnostic system. It is used to estimate the divergence of different sources of a diagnostic system. The external knowledge base attributes a linguistic value of uncertainty to each item that composes the rule base in order to define the importance of these items in the characterization of different classes of a rule base.

The linguistic characterization is, in general, less specific than the numerical one and more significant.

Before going to the presentation of the method external knowledge base construction, we illustrate an example of divergence between two rule bases. We take a item  $t$  which is contained in a rule base  $A$ . It has a high frequency of appearance in the representation of a class  $c_i$  which can be expressed by a strong degree of membership of the "very-frequent" term of the class  $c_i$ . In the rule base  $B$ , the same item has a very low frequency of appearance in the representation of the same previously mentioned class which is expressed by a strong degree of membership of the "exceptional" term of the class  $c_i$ . In this example, we can remark a great disparity in the consideration of the item  $t$  in the definition of the same class of two distinct sources. So, the item  $t$  can be regarded as an ambiguous one for defining the class  $c_i$ .

The procedure of setting the external knowledge base is composed of 3 steps (Figure 2): the first step regroups the items according to their class label, then it generates a group of items for each class label. The second, measures the occurrence number of the items for each group (i.e. for each class). In the last step, for each class, the membership functions for several frequency terms are generated. These membership functions attribute a degree of membership of the specific

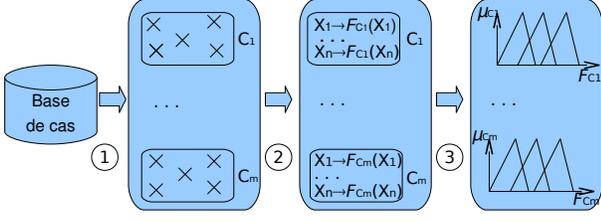


Figure 2. The procedure of setting the external knowledge base

frequency term to an item. This step assigns, for each item, a frequency term to each class.

Hence, the external knowledge base is composed of membership functions to each class label. To Define these membership functions, we denote:

- $B = \{X_1, \dots, X_n\}$  is a set of  $n$  cases.
- $\Omega = \{\omega_1, \dots, \omega_m\}$  is a set of  $m$  items which compose the cases of  $B$ .
- $C = \{c_1, \dots, c_p\}$  is a set of  $p$  class labels which also compose the cases of  $B$ .
- A case  $X_i \in B$  is the  $k+1$ -tuple and its representation is  $X_i = \{x_1, \dots, x_k, c\}$  where  $x_i \in \{x_{i1}, \dots, x_{il}\}$  or  $x_i \in \emptyset$  and  $x_{ij} \in \Omega$  and  $x \in C$ .
- $f_{c_i}$  is the function which gives the standardized frequency of the  $\omega_j$  to the label class  $c_i$  :

$$f_{c_i}(\omega_j) = \frac{g_{c_i}(\omega_j)}{\max_{\omega_k \in \Omega} (g_{c_i}(\omega_k))}$$

where:

$$g_{c_i}(\omega_j) = \frac{|\{X \in B | c_i \in X \text{ and } \omega_j \in X\}|}{|\{X \in B | c_i \in X\}|}$$

The linguistic variable is characterized by a quintuple [8]–[10]:  $(x, T(x), U, G, M)$

- $x$  is the frequency and is the name of the linguistic variable.
- $T(x)$  is the set of terms associated with the linguistic value, in which the frequency is represented according to the following set  $\{Never, Exceptional, Rare, Usual, Frequent, Very Frequent, Always\}$ .
- $U$  is a universe of discourse and  $U = \{\omega_j \in \Omega | f_{c_i}(\omega_j)\}$ .

The terms of  $T(x)$  are characterized by fuzzy subsets defined by the following functions of membership [11]:

- $K$ : is the set of centroids of fuzzy sets obtained by the algorithm of fuzzy c-means (FCM) [12] which is applied to  $U$  such as  $K = \{0, \dots, C_{i-1}, C_i, C_{i+1}, \dots, 1\}$ .
- $\nu_{i,\alpha}$  : corresponds to the membership function in the linguistic term  $\alpha$ . It is built from a set of instance

frequencies which belong to the class label  $i$ .

$$\nu_{c_i,\alpha}(f_{c_i}(\omega_j)) = \begin{cases} 1 - (f_{c_i}(\omega_j) - C_i)/(C_i - C_{i-1}) & \text{if } C_{i-1} < f_{c_i}(\omega_j) \leq C_i \\ 1 - (f_{c_i}(\omega_j) - C_i)/(C_{i+1} - C_i) & \text{if } C_i \leq f_{c_i}(\omega_j) < C_{i+1} \\ 0 & \text{otherwise.} \end{cases}$$

- $\mu_{c_i,\alpha}$  : corresponds to  $\mu_{c_i,\alpha}(\omega_j) = \nu_{c_i,\alpha}(f_{c_i}(\omega_j))$ .

The functions of membership of the set  $T(x)$  are presented in Figure 3.

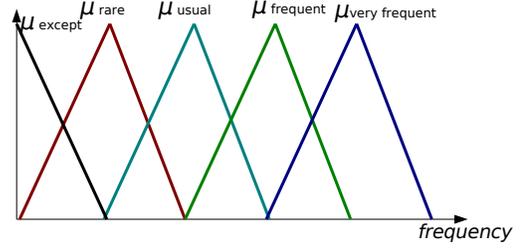


Figure 3. Linguistic values of the variable "frequency".

### III. FUSION PROCEDURE

The fusion of the rule sets which are produced from different data sets is little studied in the data fusion literature contrary to the artificial intelligence and more precisely in: non-monotonic logic, belief merging, belief revision, etc. which propose many combination operators of knowledge bases [13].

There are some applications combining the rule sets which are produced from the case bases coming from several sources. The merging operator proposed by these methods selects only the rules which are shared by their sets. It can also regroup all association rules of these sets.

Other applications defined by Zhang et al. [14] and Cheung et al. [15] propose to integrate the new rules in order to update the rule set which is considered as knowledge base.

Another one [16] compares the generated rules with the existing knowledge in order to find out what is right and what is wrong about this knowledge and it detects what has changed since the last learning.

Fromont et al. [17] investigate the problem of learning by inductive logic programming, symbolic rules that characterize cardiac arrhythmia from multisource data such as electrocardiograms or arterial blood pressure measures. They propose a two-step strategy that uses monosource learning to automatically bias and reduce the search space for multisource learning. This approach is considered as monitoring search rules.

Herein, an efficient and powerful knowledge fusion method is proposed. This process merges rule bases in order to produce a single one. Thereafter, in order to simplify the explanation,

an example of our fusion method is given where the fusion procedure is applied on two rule bases (Figure 4). This method merges two knowledge bases produced from two diagnostic systems.

This fusion method consists of three steps: the first measures the conflict between knowledge bases of several diagnostic systems. The second step searches the similar rules between the rule bases of several diagnostic systems. The third step builds the new rule bases from the last rule bases by adding, removing and changing rules of the last rule bases.

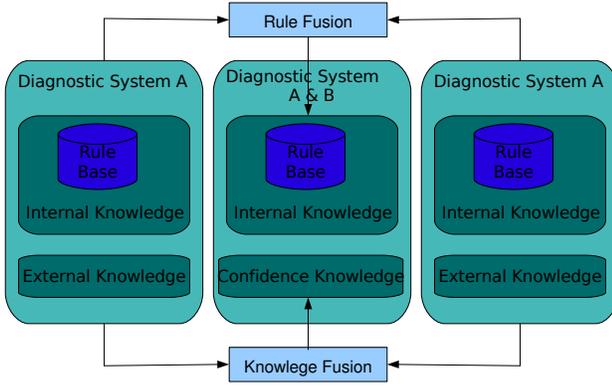


Figure 4. Fusion process integrated in a diagnostic systems

In order to search similar rules in two rule bases, an operator of computing similarity is defined. This operator takes into account the distortion of the case bases from where these rule bases are produced by measuring the conflict of these case bases. In order to measure the conflict, we need to define external knowledge.

#### A. Conflict measure of several sources

We propose a global measurement operator of conflict between  $p+1$  sources. This operator applies to  $p$  knowledge bases (BCE) deduced from  $p$  sources and to another knowledge base which is deduced from the coupling of the aforementioned  $p$  bases. This knowledge base is called  $F$  for simplification.

Our operator works in the following way: for an item  $t$  and a class  $c$ , it recovers the uncertainty term  $\alpha$  from  $F$  of which the item  $t$  has the greatest degree of membership for the class  $c$ . The linguistic term  $\alpha$  is taken as a reference mark for the calculation of disparity of  $p$  knowledge bases. Then for each  $p$  base, the operator calculates the membership degree of the item  $t$  to the linguistic term  $\alpha$  and to the other linguistic terms which are the direct neighbors of  $\alpha$  for the class  $c$  and we recover the highest degree of membership of one of these linguistic terms. For example, the usual linguistic terms for direct neighbors are 'rare' and 'frequent'. Thereafter, we keep the lowest value of the degrees of membership obtained from

$p$  bases. In this context, the goal is to find a base in the  $p$  bases which gives the level of representation of the parameter  $t$  for the class  $c$  which is the farthest one from the average of  $p$  bases. Then we deduce the degree of reliability (degree of confidence) of a parameter after the fusion of  $p$  bases.

The whole above-mentioned procedure is followed by the measurement of conflict between  $p$  sources ( $p$  bases). This measurement is inspired by the method proposed by Dubois and Prade [18]. In this work, a conflict is defined as the distance that separates the classification of a parameter between the new base (base resulting from fusion) and  $p$  bases (bases which were amalgamated). We propose a global measurement operator for the conflict between  $p$  sources ( $p$  knowledge bases defined previously) of a class  $c$ .

In the following sections, we present how the new knowledge base is set up, thereafter; we define how the referent linguistic term is selected. Then we define the conflict operator and finally we present the measurement of confidence.

#### Definition of the new external knowledge base which is deduced from the $p$ source coupling:

- $B = \{B_1, \dots, B_p\}$  is set of  $p$  case bases.
- $B_i = \{X_1^i, \dots, X_n^i\}$  is the case base of  $n$  cases.
- $\Omega = \{\omega_1, \dots, \omega_m\}$  is a set of  $m$  items which compose the cases of  $B_i \in B$ .
- $C = \{c_1, \dots, c_l\}$  is a set of  $l$  class labels which composes also the cases of  $B_i \in B$ .
- A case  $X_j^i \in B_i$  is the  $k+1$ -tuple and its representation is  $X_j^i = \{x_1, \dots, x_k, c\}$  where  $x_i \in \{x_{i1}, \dots, x_{il}\}$  or  $x_i \in \emptyset$  and  $x_{ij} \in \Omega$  and  $x \in C$ .
- $f_{c_i}$  is the function that gives the standardized frequency of the  $\omega_j$  to the label class  $c_i$  in all case bases of  $B$ :

$$f_{c_i}(\omega_j) = \frac{g_{c_i}(\omega_j)}{\max_{\omega_k \in \Omega} (g_{c_i}(\omega_k))}$$

where:

$$g_{c_i}(\omega_j) = \frac{|\{X \in B_1 \mid c_i \in X \text{ and } \omega_j \in X\}|}{|\{X \in B_1 \mid c_i \in X\}|} + \dots + \frac{|\{X \in B_l \mid c_i \in X \text{ and } \omega_j \in X\}|}{|\{X \in B_l \mid c_i \in X\}|}$$

The membership functions of this new external knowledge base have the same definition as given in section II-B; nevertheless, some notations are modified in order to integrate the notion of multisource.

$\nu_{c_i, \alpha}^l$ : represents the function of membership in the linguistic term  $\alpha$  for the class  $c_i$  of the knowledge base  $l$ .

$\mu_{c_i, \alpha}^l$ : corresponds to  $\mu_{c_i, \alpha}^l(\omega_j) = \nu_{c_i, \alpha}^l(f_{c_i}(\omega_j))$ .

$F$ : corresponds to the knowledge base deduced from  $p$  knowledge bases.

#### Selecting the referent linguistic term

As we defined previously, the linguistic term  $\alpha$  is considered

as a reference mark of the base  $F$  to measure the disparity between  $p$  bases.

The linguistic term of frequency  $\alpha$  for the parameter  $\omega_j$  in the class  $c_i$  is selected as referent if  $\mu_{c_i,\alpha}^f(\omega_k)$  presents the greatest value such as:

$$\mu_{c_i,\alpha}^F(\omega_k) = \max_{j \in T(x)} (\mu_{c_i,j}^F(\omega_k))$$

### Operator of the conflict measure:

The operator of conflict seeks in  $p$  bases the lowest value of membership degrees of  $\omega_j$  to the linguistic reference  $\alpha$  for the class  $c_i$ :

$$h_{c_i}^\alpha(\omega_j) = \mathcal{I}(\mu_{c_i,\alpha}^1(\omega_j), \dots, \mu_{c_i,\alpha}^p(\omega_j)) \quad (1)$$

We note that this operator is strict, i.e., if one base of  $p$  bases has the value of membership of  $\omega_j$  to the linguistic referent  $\alpha$  equal to zero, we can induce that the value given by operator of conflict is also equal to null, even if this base has the membership value of  $\omega_j$  to the linguistics neighbor at the linguistic referent  $\alpha$  which is different from zero.

In certain cases, it is necessary to be less strict so we modified the conflict operator in order to integrate the tolerance parameter that is parametrized according to our needs.

$$h_{c_i}^\alpha(\omega_j) = \mathcal{I}(\max_{\alpha-d \leq k \leq \alpha+d} (\mu_{c_i,k}^1(\omega_j)), \dots, \max_{\alpha-d \leq k \leq \alpha+d} (\mu_{c_i,k}^p(\omega_j)))$$

- $d$  is the tolerance index. It is used by the conflict operator in order to take into account the membership of  $\omega_j$  to the referent linguistics term  $\alpha$  and also the  $d$  linguistic terms close to the referent linguistics term.  $d$  is an integer and  $\in [0, |T(x)|/2 - 1]$ .
- $\alpha - d$  indicates the linguistic that is at position  $d$  on the left of the referent linguistic term  $\alpha$ . Example: if  $\alpha$  is equivalent to the linguistic *frequent* and  $T(x) = \{\text{never, exceptional, habitual, very frequent, always}\}$  then  $\alpha - 2$  indicates the linguistic term “rare”.
- $\alpha + d$ : indicates the linguistic that is at position  $d$  on the right of the referent linguistic term  $\alpha$ .

### Confidence measure:

We define a confidence measurement function of all parameters in the new base:

$$\mu_{conf,c_i}(\omega_j) = \begin{cases} 1 & \text{if } h_{c_i}^\alpha(\omega_j) \in [0, \varepsilon] \\ h_{c_i}^\alpha(\omega_j)/(1 - \varepsilon) & \text{if } h_{c_i}^\alpha(\omega_j) \in [\varepsilon, 1[ \end{cases}$$

where :

$\varepsilon \in ]0, 1[$ , which is used as threshold in order to estimate the confidence of a parameter from the conflict value.

$\mu_{conf,c_i}(\omega_j)$  considers the agreement measurement and is completely reliable if the agreement degree is higher than a certain threshold, i.e. the parameter is reliable if its appearance frequencies in the various bases belong to the same linguistic term or a close linguistic with the referent linguistic which is obtained from  $F$ .

### B. Rule similarity measure

The fusion procedure needs to measure the similarity rules in order to aggregate them. Hence, the number of rules is reduced. The similarity measure of rules should take into account the confidences of parameters which compose the rules.

Before presenting this function of similarity, some notations have to be specified:

- $r_a$  and  $r_b$  are two rules of class association (definition to see §II-A) which come respectively from the rule base of system  $A$  and the rule base of system  $B$ .
- $itemset(r_a \cap r_b)$ : common items in both rules.
- $itemset(r_a \cup r_b)$ : all items which compose both rules.
- $class\_label(r_a, r_b)$ : common class label in both rules.
- $w_{c_i}(r_a, r_b, \omega_j)$ : matching degree of the rules  $r_a$  and  $r_b$  to the item  $\omega_j$  for the class  $c_i$ .

$$w_{c_i}(r_a, r_b, \omega_j) =$$

$$\begin{cases} 1 & \text{if } c_i \in class\_label(r_a, r_b) \\ & \text{and if } \omega_j \in itemset(r_a \cap r_b) \\ 1 - \mu_{conf,c_i}(\omega_j) & \text{if } c_i \in class\_label(r_a, r_b) \\ & \text{and if } \omega_j \notin itemset(r_a \cap r_b) \\ & \text{and if } \omega_j \in itemset(r_a \cup r_b) \\ 0 & \text{otherwise} \end{cases}$$

We denote by  $sim_{c_i}(r_a, r_b)$  the degree of matching between  $r_a$  and  $r_b$  for the class  $c_i$ .  $sim_{c_i}(r_a, r_b)$  is the average of all  $w_{c_i}(r_a, r_b, \omega_j)$  derived from  $itemset(r_a, r_b)$ . This function measures the similarity between two rules of the same class and it takes into account the matching items of these rules and the reliability of these items.

$$sim_{c_i}(r_a, r_b) = \frac{\sum_{\omega_j \in itemset(r_a \cup r_b)} w_{c_i}(r_a, r_b, \omega_j)}{|itemset(r_a \cup r_b)|}$$

This function is used in the fusion in order to find the best similar rules between two rule sets coming from two different sources.

### C. Rule base fusion

In this section, it is described how several rule bases are merged into one rule base. The fusion procedure is composed of 3 phases, entitled: combining rule bases, clustering rules, detecting ambiguous.

- 1) **Combining rule bases:** All rule bases of several systems are combined into one base without any modification in rules.
- 2) **Clustering rules:** This phase consists of isolating the rules according to their similarity. The function  $sim_\alpha$  is used to measure the similarity between two rules. We proceed step by step and in each step, the rules have the value of the highest similarity which is higher than a specific threshold. Thereafter these two rules are amalgamated using the function  $equival(r_a, r_b)$  to

obtain one rule. This method of clustering is hierarchical ascending classification (Figure 5).

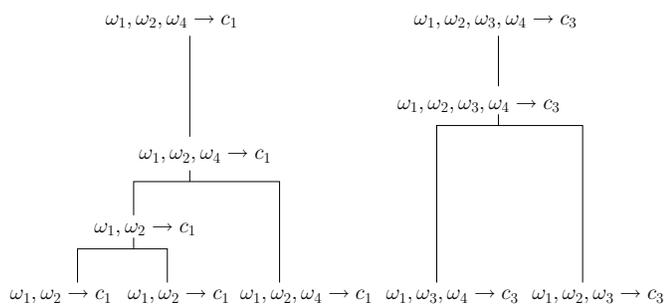


Figure 5. Example of rule fusion by hierarchical ascending classification.

The function  $equival(r_a, r_b)$  merges  $r_a$  and  $r_b$  to generate the rule  $r_{eq}$  such as:

- $sim_{c_i}(r_{eq}, r_a) \geq sim_{c_i}(r_a, r_b)$
- $sim_{c_i}(r_{eq}, r_b) \geq sim_{c_i}(r_a, r_b)$
- $itemset(r_a \cap r_b) \subseteq itemset(r_{eq} \cap r_a)$
- $itemset(r_a \cap r_b) \subseteq itemset(r_{eq} \cap r_b)$
- $itemset(r_{eq}) \subseteq itemset(r_a \cup r_b)$

The rule  $r_{eq}$  is the rule of the consensus of the rules  $r_a$  and  $r_b$ .

- 3) **Detecting ambiguous:** The ambiguous rule is the rule where confidences of its parameters are less than a certain threshold. All the ambiguous rules are marked as exception rules by the diagnostic system which will be set up.

After the rule base fusion procedure, the new rule base which is produced represents the new internal knowledge base which is used by the diagnostic system in order to predict the class label of the new instance. The diagnostic system prediction procedure, which is described in the section §II-A has to give the rules which cover the new instance, i.e. the rule should have all its body items included in the new instance items set. However, in this new rule base, the diagnostic system could select the rules which cover the new instance with less strictness than the procedure of section §II-A. The new rule base is associated with confidence knowledge and each item, which composes the rules of these bases, has its confidence value to each class label. Thus, the diagnostic system retrieves rules which cover the new instance by taking into account the confidence value of its items.

The asymmetric similarity measure function  $select_{c_i}(e, r \rightarrow c_i)$  has been defined in order to measure the matching between the new instance and the rule by taking into account the item confidence value.

$$select_{c_i}(e, r \rightarrow c_i) = \frac{\sum_{k \in (r \cap e)} \sqrt{\mu_{conf, c_i}(\omega_k) \mu_{c_i}(\omega_k)}}{\sum_{\omega_k \in e} \sqrt{\mu_{conf, c_i}(\omega_k) \mu_{c_i}(\omega_k)}}$$

where:

- $e$  : corresponds to the new instance.
- $r \rightarrow c_i$  : corresponds to the rule with the label class  $c_i$ .

After describing the fusion procedure, we will pass to the evaluation of this procedure.

#### IV. EVALUATION

In this section, we evaluated our knowledge fusion approach on four databases from an endoscopic image analysis system [19]. This system is an assistant system for decision-making of the endoscopic lesion diagnosis. These bases are composed of endoscopic image description via symbolic terms which are defined by the minimal standard terminology of the SEGE (European Company of Gastro-Enterology). A case in a base represents a description (a set of parameters) of an endoscopic lesion. In all the bases, there are 206 parameters and 89 types of lesions, (i.e. 89 label classes).

One of the four bases is used as the test base and the others are used as the training bases. Each of these bases presents specific characteristics as some parameters are used more than the others or they are never used:

- The test database noted  $T$ : it is composed of 25 % of usual descriptions, 25 % the rare ones, 25 % the exceptional ones and 25 % the doubtful one.
- The first training database noted  $A$ : it is composed of 30 % of usual descriptions, 30 % the rare ones, 30 % the exceptional ones and 10 % the doubtful ones.
- The second training database noted  $B$ : it is composed of 60 % of usual descriptions, 24 % the rare ones, 12 % the exceptional ones and 4 % the doubtful ones.
- The third training database noted  $C$ : it is composed of 65 % of usual descriptions, 25 % the rare ones and 10 % the exceptional ones.

We have conducted an extensive performance study to evaluate accuracy and efficiency of the fusion procedure and compared it with naive multisource and monosource approach. The first involves learning directly from the aggregated databases. The experimentation takes the training databases and sets up the diagnostic system database from them. We use the test bases related before to derive the classifiers and the error rates in the experiments.

The following notation  $A \otimes B$  means the the fusion procedure is applied to the diagnostic system based on database  $A$  and the diagnostic system based on database  $B$  and then the generated system is evaluated by the test database  $T$ . The notation  $A \oplus B$  means the diagnostic system is applied to the aggregated databases  $A$  and  $B$  (naive multisource) and evaluated by the test database  $T$ . At last, the notation  $A_T$  means the diagnostic system is applied to the training database  $A$  (monosource approach) and evaluated by the test database  $T$ .

We note that the diagnostic system based on the training basis  $A$  has the higher rate of accuracy than other systems based on other training bases (in the monosource approach).

The prediction values of diagnostic system based on the monosource approach are regarded as thresholds that the prediction values of the diagnostic systems based on the fusion procedure must exceed if the procedure of fusion wouldn't be

Database	accuracy	number of rules	time (s)
$A_T$	93.14	62	33
$B_T$	77.14	45	27
$C_T$	76.14	37	24
$A \oplus B$	94	30	38
$A \otimes B$	97.71	96	40
$A \oplus C$	94.85	60	40
$A \otimes C$	96.57	91	39
$B \oplus C$	81.14	52	39
$B \otimes C$	78.85	71	42
$A \oplus B \oplus C$	93.71	52	38
$A \otimes B \otimes C$	96	100	63

Table 1  
CLASSIFICATION ACCURACY OF THE FUSION PROCEDURE.

regarded as unadapted for these systems. In the actual case, this appears in all the cases in our evaluation:  $(A \otimes B) > A_T$ ,  $(A \otimes B) > B_T$ ,  $(A \otimes C) > A_T$ ,  $(B \otimes C) > B_T$ ,  $(B \otimes C) > C_T$ .

The prediction values of the diagnostic systems based on the naive multisource approach are them reference thresholds. If these thresholds are exceeded by the fusion based diagnostic systems, that means that the fusion procedure is very well adapted. This appears in all the cases except in one case in our evaluation:  $(A \otimes B) > (A \oplus B)$ ,  $(A \otimes C) > (A \oplus C)$ ,  $(B \otimes C) < (B \oplus C)$ ,  $(A \oplus B \otimes C) > (A \oplus B \oplus C)$ .

We note also that the fusion procedure generates more rules than the naive multisource approach but the fusion procedure outperforms the naive multisource approach in terms of accuracy rate. Besides, the sum of rules number generated by monosource approach is always upper than that of the fusion approach. That proves that the fusion approach is efficient. And in certain cases, the the fusion procedure is faster than the multisource approach when generating the rule base. That proves also that the fusion approach is able to gain time.

## V. CONCLUSIONS

We proposed a new fusion method which is applied on the rule based diagnostic system. This fusion method combines two fusion types: the knowledge base fusion and the rule base fusion. The knowledge bases refer to the diagnostic system rule bases and are used in order to measure the conflict of the diagnostic system rule bases. Thereafter, the conflict measure is integrated in the rule base fusion. The rule base fusion combines the rules of several bases into one rule base and then the most similar rules are merged into one rule. This fusion approach is new and is very fast and efficient. The fusion method evaluation outperforms the naive multisource and monosource approach.

Our fusion method is quite interesting because unlike the other methods, it is adapted to any domain.

As future work, we aim to adapt the fusion method in order to apply it on heterogeneous rule bases. The fusion method should integrate the similarity measure method in order to detect the parts of base rules which are homogeneous or heterogeneous. The homogeneous parts should be applied to

fusion approach in order to take into account the distortion of these parts. The heterogeneous parts should be applied to aggregation approach in order to cover the entire event space.

## REFERENCES

- [1] B. SOLAIMAN and B. Dasarathy, "Information fusion concepts. from information elements definition to the application of fusion approaches," in *SPIE proceedings series (SPIE proc. ser.) International Society for Optical Engineering proceedings series*, vol. 385, 2001, pp. 205–212.
- [2] L. Cholvy, "Information evaluation in fusion: a case study," in *In Proc of ECSQARU-03 Workshop "Uncertainty, Incompleteness, Imprecision and Conflict in Multiple Data Sources"*, July 2003.
- [3] D. Dubois, M. Grabisch, H. Prade, and P. Smets, "Using the transferable belief model and a qualitative possibility theory approach on an illustrative example: the assessment of the value of a candidate," *International Journal of Intelligent Systems*, vol. 16, no. 11, pp. 1245–1272, novembre 2001, bB.
- [4] P. Gancarski and C. Wemmert, "Collaborative multi-strategy classification: application to per-pixel analysis of images," in *MDM '05: Proceedings of the 6th international workshop on Multimedia data mining*. New York, NY, USA: ACM Press, 2005, pp. 15–22.
- [5] W. Li, J. Han, and J. Pei, "Cmar: Accurate and efficient classification based on multiple class-association rules," in *ICDM, 2001*, pp. 369–376. [Online]. Available: citeseer.ist.psu.edu/li01cmar.html
- [6] X. Yin and J. Han, "CPAR: Classification based on predictive association rules," in *Proceedings of 2003 SIAM International Conference on Data Mining*, San Francisco, CA, 2003.
- [7] A. Zemirline, L. Lecornu, and B. Solaiman, "Nouvelle méthode d'extraction de règles de classification multi-labels," in *Qualité des Données et des Connaissances (QDC 2007)*, 2007, pp. 25–35.
- [8] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning - ii," *Information Sciences (Part 2)*, vol. 8, no. 4, pp. 301–357, 1975.
- [9] —, "The concept of a linguistic variable and its application to approximate reasoning - ii," *Information Sciences (Part 1)*, vol. 8, no. 3, pp. 199–249, 1975.
- [10] —, "The concept of a linguistic variable and its application to approximate reasoning - ii," *Information Sciences (Part 3)*, vol. 9, no. 1, pp. 43–80, 1975.
- [11] —, "Fuzzy sets," *Informations and control*, vol. 8, pp. 338–353, 1965.
- [12] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition)*. Springer, July 1981. [Online]. Available: <http://www.amazon.co.uk/exec/obidos/ASIN/0306406713/citeulike-21>
- [13] S. Koniczny, "On the difference between merging knowledge bases and combining them," in *Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR'00)*, 2000, pp. 135–144.
- [14] M. Zhang, B. Kao, D. W.-L. Cheung, and C. L. Yip, "Efficient algorithms for incremental update of frequent sequences," in *PAKDD, 2002*, pp. 186–197.
- [15] D. W.-L. Cheung, K. Hu, and S. Xia, "An adaptive algorithm for mining association rules on shared-memory parallel machines," *Distributed and Parallel Databases*, vol. 9, no. 2, pp. 99–132, 2001.
- [16] B. Liu and W. Hsu, "Post-analysis of learned rules," in *AAAI/IAAI, Vol. 1, 1996*, pp. 828–834.
- [17] E. Fromont, R. Quiniou, and M. Cordier, "Learning rules from multi-source data for cardiac monitoring," in *AIME'05 (Artificial Intelligence in Medicine)*, ser. LNAI, S. Miksch, J. Hunter, and E. Keravnou, Eds., vol. 3581. Aberdeen, Scotland: Springer Verlag, July 2005, pp. 484–493.
- [18] D. Dubois and H. Prade, "Combination of fuzzy information in the framework of possibility theory," in *Data Fusion in Robotic and Machine Intelligence*, pp. 481–505, 1992.
- [19] J. Cauvin, C. L. Guillou, B. Solaiman, M. Robaszkievicz, P. L. Beux, and C. Roux, "Computer-assisted diagnosis system in digestive endoscopy," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, pp. 256–262, 2003.