Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic

Mohammed Benamina, Baghdad Atmani, Sofia Benbelkacem*

Laboratoire d'Informatique d'Oran (LIO), University of Oran 1 Ahmed Benbella (Algeria)

Abstract

In the medical field, experts' knowledge is based on experience, theoretical knowledge and rules. Case-based reasoning is a problem-solving paradigm which is based on past experiences. For this purpose, a large number of decision support applications based on CBR have been developed. Cases retrieval is often considered as the most important step of case-based reasoning. In this article, we integrate fuzzy logic and data mining to improve the response time and the accuracy of the retrieval of similar cases. The proposed Fuzzy CBR is composed of two complementary parts; the part of classification by fuzzy decision tree realized by Fispro and the part of case-based reasoning realized by the platform JColibri. The use of fuzzy logic aims to reduce the complexity of calculating the degree of similarity that can exist between diabetic patients who require different monitoring plans. The results of the proposed approach are compared with earlier methods using accuracy as metrics. The experimental results indicate that the fuzzy decision tree is very effective in improving the accuracy for diabetes classification and hence improving the retrieval step of CBR reasoning.

Keywords

I. Introduction

Decision support is not intended to replace the decision-maker by proposing solutions, but rather to guide him to decisions that he will have to take under his responsibility. Decision support can be found in many application domains such as economics, mathematics, computing, medicine, etc. In medicine, the decision is considered to be the center of the medical procedure. The medical decision-making process involves making a diagnosis, proposing treatment, a surveillance plan, etc. Thus, many applications of decision support have been developed in this field. These applications are intended to support healthcare personnel in their decision-making. This involves the use of various decision-support tools, such as case-based reasoning (CBR), which can be seen as a management tool for decision-making.

The CBR is a powerful methodology for designing intelligent systems. It is based on the re-use of solutions of similar cases in order to solve new encountered problems, by capitalizing acquired knowledge from past experiences. In daily life, we are frequently confronted with problems already experienced and which probably have similar solutions.

During their practice, practitioners use not only their theoretical knowledge but also their acquired experience. To this end, case-based reasoning should be exploited for medical decision-making and diagnosis of diabetics. Diabetes has become the fourth leading cause of death in developed countries and there is substantial evidence that it is reaching epidemic proportions in many developing and newly industrialized countries.

In this paper we propose a fuzzy diagnosis aid system based on data mining and case-based reasoning. CBR systems based on fuzzy logic are successfully applied in various fields including medical diagnosis. Our approach is applied to assist experts in the diagnosis of diabetes. The objective is to help experts in the field in the choice of diabetic surveillance plan.

The aim of this paper is to demonstrate the value of a fuzzy inference system guided by data mining in CBR modeling. It deals with the originality of the integration of different techniques derived from artificial intelligence (AI) in the stages of CBR and its ability to jointly use expertise and learning from data in the same context, especially that artificial intelligence tools become a part of the diabetes health care [43].

Case retrieval is the first step in the CBR cycle that requires using similarity measures to find similar cases to the given problem. Generally, the similarity measures used such as k nearest neighbors (k-nn) consist in calculating a distance between each case of the base and the case to be solved. However, when the case base is too large, similarity calculation will be expensive in computational time. To remedy this problem, we opted for the fuzzy decision trees in the retrieval of similar cases.

Retrieval is often considered the most important phase of CBR. In this paper, we integrate fuzzy logic and data mining to improve this step. The proposed Fuzzy CBR is composed of two complementary parts; the modeling part fuzzy realized by Fispro and the reasoning part realized by the platform JColibri. The use of fuzzy logic aims to reduce the complexity of calculating the degree of similarity that can exist between individuals who require different monitoring plans.

This paper is organized as follows. A state of the art about work using AI techniques in CBR in the medical field is presented in Section II. Section III is devoted to our FDT4CR approach. Finally, a conclusion and some perspectives are given in section IV.

* Corresponding author.

E-mail address: sofia.benbelkacem@gmail.com

II. State of the Art

Case-based reasoning is a general decision-making paradigm used in the medical field [1]. A lot of work on CBR has been done in this area of research.

Choudhury and Begum [2] presented a literature review about case-based reasoning systems used in the medical field over the past few decades. In this study the difficulties of implementing CBR in medicine have been discussed. Blanco et al. [44] presented a review study focused on papers published between 2008 and 2011 in the field of case-based reasoning applied to the health sector. The authors of this review study presented a new proposal to improve the adaptation step of CBR. They used association rules to reduce the number of cases and facilitate learning around adaptation rules. El-Sappagh and Elmogy [45] identified research papers on CBR published between 1999 and 2015. Integrated evaluation metrics have been used for the analysis. The results show that there is a need for more comprehensive improvements in medical CBR particularly in diabetes systems. Montani et al. [3] proposed a case-based decision support tool, designed to assist physicians in the diagnosis of type 1 diabetes. The authors have put in place a two-stage procedure; first, it finds the classes to which the new case might belong; then it lists the most similar cases, using the k-nn method. Similarly, Schmidt and Gierl [4] used case-based reasoning techniques to speed up the process to propose appropriate treatment recommendations for patients with complications due to infection. Doyle et al. [5] presented a decision-support system using case-based reasoning for the treatment of the bronchiolitis. The system provides guidance based on previous cases. Montani et al. [6] presented a case based-reasoning system for the management of patients with end-stage renal disease and undergoing hemodialysis. This system helps practitioners to develop treatment plans for dialysis. Marling et al. [7] presented a CBR decision-making approach for the management of diabetic patients. To prevent serious complications of the disease, patients must constantly monitor their blood sugar levels and keep it as close to normal as possible, but maintaining good blood sugar control is a difficult task for patients and their doctors. A prototype system for CBR decision support has been designed to help with this task. Azar and Bitar [46] proposed to compare three artificial intelligence-based algorithms, CBR, C4.5, and Genetic Algorithms, to treat the problem of Diabetes Mellitus. The three algorithms have been evaluated on a Diabetes data set extracted from the UCI Machine Learning Repository. The results demonstrated the effectiveness of all three techniques. De Paz et al. [8] presented a CBR decision-support system for the diagnosis of different types of cancer.

In our study, we focused on the retrieval of cases stage, the main step in the case-based reasoning process. This step is usually dealt with k-nn [9]. Retrieval by k-nn is costly in computing time and its learning base generally requires a large memory space. Data mining could be useful to improve case retrieval. Data mining uses several techniques of computer science, statistics and artificial intelligence in various fields, especially for decision support. In the following, we will cite some works which encouraged us to contribute in this field, especially in the medical field for the classification of diabetes. Quellec et al. [10] presented a CBR approach about the diagnosis of diabetic retinopathy. They proposed to proceed to the adaption of the decision tree method in the context of remembering. The authors propose to adapt decision trees, generally used for classification, to image indexing context and therefore extend CBR to cases containing images. Burke et al. [11] calculated the similarities between the cases using decision trees. The case base refers to a decision tree where cases are the graph attributes. Huang et al. [21] suggest a model for a prognostic and diagnostic system for chronic diseases that integrates data mining and case-based reasoning. They adopted data mining techniques to discover the implicit meaningful rules from the health examination data and used the extracted rules for the specific prognosis of chronic disease.

In this work, case-based reasoning is used to support the diagnosis and treatment of chronic diseases. Kushwaha and Welekar [52] used a data mining technique to develop an approach of feature selection for image retrieval based on content. The experimental result shows that the selection of features using Genetic Algorithm reduces the time for retrieval. Houeland [22] developed a RDT (random decision tree) algorithm implemented in a CBR framework. RDT algorithm has been combined with a simple similarity measure. They used two measures of similarity. A local subset of cases is selected using the first similarity measure. Then, the subset is reduced with the second measure of similarity by using a decision tree approach. RDT lets to define the similarity between two cases where trees are fully developed binary trees. The results have been evaluated in the area of palliative care of cancer.

Remembering cases must face a certain degree of vagueness and uncertainty that are almost always encountered while dealing with complex applications of the real world. A recent survey [12] presents the role of fuzzy sets and fuzzy logics in the various stages of CBR. This study showed that the integration of fuzzy logic in CBR resulted in successful hybrid systems. Boyen and Wehenkel [23] describe a new algorithm capable of deducing fuzzy decision trees in domains where most of the input variables are numeric and the output information is referred to as a fuzzy set. It includes three complementary steps: growth to select relevant attributes and fuzzy thresholds; Pruning to determine the appropriate tree complexity; the redevelopment to adjust the parameters of the tree in a global way. Begum et al. [24] proposed a case-based decision support system to assist clinicians in the diagnosis of stress. Case-based reasoning is applied as the main methodology to facilitate the re-use of the experience and the explanation of the decision by retrieving similar previous profiles. Other fuzzy techniques are also used and incorporated into the case-based reasoning system to address imprecision, inherent uncertainty in clinicians' reasoning, and inaccuracy of characteristic values. The work of Barrientos and Sainz [25] supports the decision-making on resource planning of an emergency call center in order to achieve its obligatory quality of service. This is done by searching the activity data collected by an emergency call center. A linguistic prediction, a categorization and a description of the days based on the activity and the information of the call center make it possible to know the workload of each category of day. This was generated by a fuzzy version of an unsupervised decision tree, merging decision trees and clustering. Levashenko and Zaitseva [26] proposed a decision support system based on fuzzy logic for the diagnosis of oncology diseases. The decision-making procedure corresponds to the classification of the new case by analyzing a set of instances for which the classes are known. Solved cases are defined as fuzzy classification rules that are formed by different fuzzy decision trees. Three types of fuzzy decision trees are considered in the document: unordered, ordered and stable. The induction of these fuzzy decision trees is based on cumulative estimates of information. Adidela et al. [27] proposed a hybrid classification system to predict the onset of diabetes. The system adopts three phases. In the first phase, the data are grouped together using the EM algorithm. The second phase performs the classification of the individual clusters obtained using fuzzy ID3. From the second phase of the process, adaptation rules are obtained. These rules are essential to the prediction of diabetes. In the third phase, the test tuple is provided to the rules to predict the class tag. Tahmasebian et al. [47] tried to use the information documented in the summary of patient medical records to measure their similarity rate. First, the effective parameters were extracted from the structured discharge summary prepared for patients. Then, the weights of the parameters were extracted using data mining methods. Finally, the fuzzy system has been used to compare similarities between the current case and other cases for patients with chronic kidney disease.

The proposed similarity measure has been compared with other methods in CBR systems. As a result, employing fuzzy method to measure the similarity can lead to a higher flexibility compared to other methods. El-Sappagh and Elmogy [48] proposed a CBR fuzzy ontology (CBRDiabOnto) for the diagnosis of diabetes, specifically for DM diagnosis. The resulting ontology supports many kinds of reasoning, such as crisp, semantic data, and fuzzy reasoning. These different data types improved case representation and case retrieval. The evaluation of CBRDiabOnto shows that it is consistent, accurate and covers logic and terminologies of diabetes mellitus diagnosis.

This literature review confirms that CBR systems based on fuzzy logic are successfully applied in various fields including medical diagnosis. The aim of our contribution is to demonstrate the value of a fuzzy inference system guided by data mining in CBR modeling. The use of fuzzy logic aims to reduce the complexity of calculating the degree of similarity and hence the retrieval step of CBR reasoning.

III. Proposed Approach FDT4CR

The main purpose of the proposed approach is to improve the accuracy of Diabetes classification. We integrate fuzzy decision tree to improve cases retrieval step of CBR process. The main contributions of FDT4CR are given as follows:

- Fuzzy decision tree classifier is used for the generation of a crisp set of rules.
- Fuzzy modeling is used to deal with uncertainty of medical reasoning.
- The proposed approach is a new combination of different methods which performs classification of diabetic patients using CBR, fuzzy modeling and data mining techniques.
- Improvement of the retrieval step of JColibri CBR process.

The proposed Fuzzy Inference Mechanism model reaches high prediction and classification accuracy, which improved both specificity and sensitivity. Numerous research works were carried out on the Pima Indian Diabetes Dataset (PIDD) with the problem of developing a prediction model. Karegowda et al. [28], [29] discussed the accuracy of the classification of 65 classfiers. They developed a combined model cascaded decision tree and k-means, which has a classification accuracy of 93.33% for categorized data without missing data [30].

The architecture of the proposed approach is shown in Fig. 1. It is composed of two complementary parts; the fuzzy modeling part using Fispro and the Case based reasoning part using JColibri platform.

FisPro has been used for various modeling projects [15].

JColibri [17] is an object-oriented tool used to develop CBR applications. It is an open-source platform developed in Java that defines a clear architecture to design case-based reasoning systems. The user can customize methods and classes of the platform as required. The JColibri platform has been widely used to develop CBR applications in the medicine field. Sharma and Mehrotra [49] used the jColibri framework to develop a CBR application for diagnosis of chronic kidney disease while Kiragu and Waiganjo [50] developed a prototype for treatment and management of diabetes using JColibri.

A. Architecture of Fuzzy Case-Based Reasoning

There are several studies found in the literature that have used various techniques on the Pima Indian Diabetes dataset to train and test data. The architecture of the proposed Fuzzy CBR applied on diabetes is shown in Fig. 1. FDT4CR is the process by which the retrieval phase of a CBR is modeled not in the conventional form of mathematical equations, but in the form of fuzzy rules base.
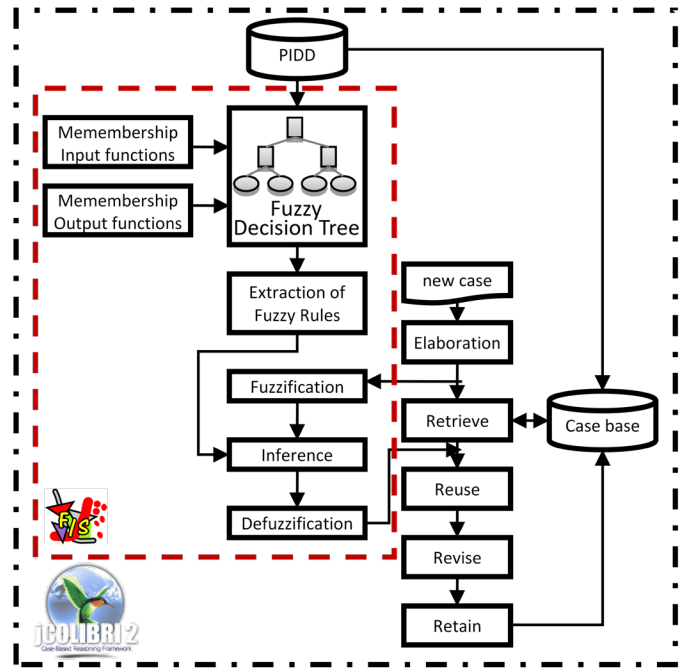

Fig. 1. The architecture of the proposed Fuzzy CBR.

B. Pima Indians Diabetes Database

The Pima Indian Diabetes Dataset (PIDD) has been taken from the UCI Machine Learning repository (http://archive.ics.uci.edu/ml/). The input variables are X_2: Plas (Plasma glucose concentration in 2-hours), X_5: Ins (2-hour serum insulin), X_6: Mass (Body mass index), X_7: Pedi (Diabetes pedigree function) and X_8: Age; the output variable is Y: DM (Diabetes Mellitus) (Table I). The data used to test the Fuzzy Inference Mechanism Framework has been taken with the age group from 25-30. The experiment compares the accuracy of the fuzzy classification task with results of studies involving the PIDD [19], [20].

TABLE I. PIDD Attributes

Exogenous Var	Abbreviation	Semantic
X_1	Preg	Number of times pregnant
X_2	Plas	Plasma glucose concentration in 2-hours
X_3	Dias	Diastolic blood pressure
X_4	Tric	Triceps skin fold thickness
X_5	Ins	2-hour serum insulin
X_6	Mass	Body mass index
X_7	Pedi	Diabetes pedigree function
X_8	Age	Age
Y	DM	Diabetes Mellitus where 1 is interpreted as tested positive for diabetes

C. Fuzzy Modeling Using Fispro

The Algorithm 1 explains the process of fuzzy inference mechanism. First, the fuzzification phase consists of gathering a crisp set of input data and converting it into a fuzzy set using fuzzy linguistic variables, fuzzy linguistic terms and membership functions. Then, an inference is executed according to a set of fuzzy rules. Finally, the defuzzification step makes it possible to transform the resulting fuzzy output into a crisp output using the membership functions.

Algorithm 1. Fuzzy logic algorithm

1. Initialization
   - Linguistic variables and terms are defined.
   - Construction of the membership functions.
   - Construction of the fuzzy decision tree and extraction of the fuzzy rule base.
2. Fuzzification
   - This step converts crisp input data to fuzzy values using the membership functions.
3. Fuzzy inference
   - Evaluate the rules in the fuzzy rule base.
   - Combine the results of each rule.
4. Defuzzification
   - This step converts the output data to non-fuzzy values.

1) Fuzzification

The conversion from crisp to fuzzy input is known as fuzzification [31]. Crisp inputs are converted to their fuzzy equivalent using a family of membership functions [53]. Furthermore, an interface is proposed to validate and tune parameters of constructed fuzzy numbers.

Each parameter is set with the minimum value, the mean, the standard deviation and the maximum value. The membership function $\mu(x)$ of the triangular fuzzy number that we have used is given by:

$$\mu(x) = \begin{cases} 0, x \le a \\ (x-a)/(b-a), a < x \le b \\ (c-x)/(c-b), b < x \le c \\ 0, x > c \end{cases}$$

(1)

The Memberships graphic for the fuzzy values with Fispro are given in Fig. 2 and Fig. 3.
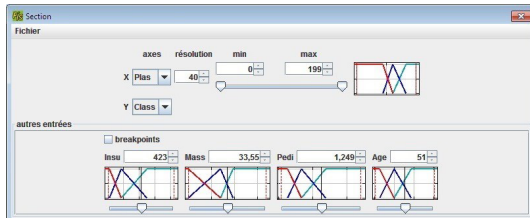

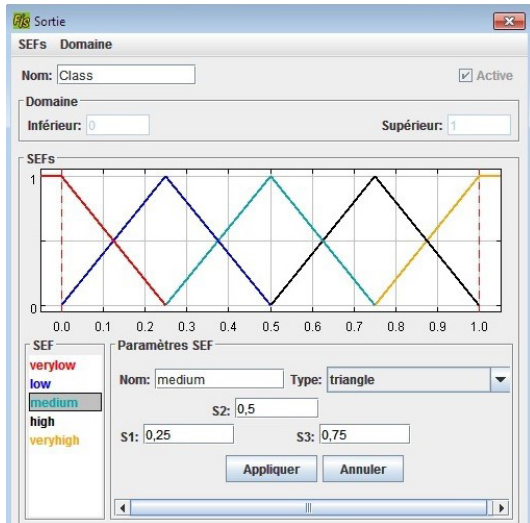Fig. 2. Membership graphics for the fuzzy values of Insu, Mass, Pedi and Age.

Fig. 3. Membership graphics of fuzzy values of Diabetes Mellitus.

Fuzzy numbers parameters are given in Table II. In this study other Fuzzy Triangular numbers are proposed inspired by the related work [32].

TABLE II. Triangular Membership Functions Parameters

Fuzzy variables	Fuzzy Numbers	Fuzzy Triangular numbers
Plas (X_2)	Low	[0, 88.335, 121.408]
	Medium	[88.335, 121.408, 166.335]
	High	[121.408, 166.335 ,199]
Ins (X_5)	Low	[0 ,17.276, 173.175]
	Medium	[17.276, 173.175, 497]
	High	[173.175,497, 846]
Mass (X_6)	Low	[0, 0, 27.792]
	Medium	[0, 27.792, 38.864]
	High	[27.792, 38.864, 67.1]
Pedi (X_7)	Low	[0.078, 0.272, 0.682]
	Medium	[0.272, 0.682, 1.386]
	High	[0.62, 1.386, 2.42]
Age (X_8)	Young	[21, 25.475, 40.537]
	Medium	[25.475, 40.537, 57.798]
	Old	[40.537, 57.798, 81]
DM (Y)	Very low	[0, 0, 0.25]
	Low	[0, 0.25, 0.5]
	Medium	[0.25, 0.5, 0.75]
	High	[0.5, 0.75, 1]
	Very high	[0.75, 1, 1]

2) Fuzzy Inference Engine

In our study, we make use of the FDT (Fuzzy Decision Tree) that is an extension of classical decision trees [33], [34]. The fuzzy decision tree proposed in FisPro (Fig. 4) is based on the algorithm [35]. The implementation of FisPro is based on a predefined fuzzy partition including *input* variables, which is left intact by the tree's growth algorithm.
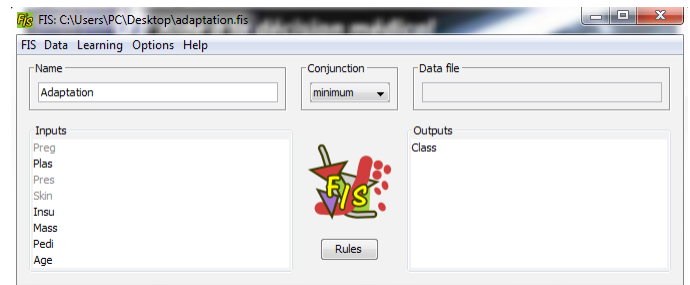

Fig. 4. Fispro interface for Decision Tree construction.

From a root node that includes all items of the data set, the fuzzy decision tree uses a recursive procedure to divide the nodes into M_j child nodes, where M_j is the fuzzy set number in the jth input variable partition which is selected for the split. The algorithm proceeds to the selection of the variable which maximizes the gain based on a discriminated criterion for each node:

$$G_n^j = H_n - \sum_{i=1}^{M_j} w_i H_{n,i}$$

(2)

where H_n is the entropy of the node n, $H_{(n,i)}$ is the entropy of the node corresponding to the SEF (Fuzzy Sub Set) of the variable j ($i=1,..,M_j$) and w_i is the relative weight of SEF.

The entropy of a node *n* is defined as:

$$H_n = -\sum_k P_k^n * \log(P_k^n)$$

(3)

where P_k^n is the fuzzy proportion of examples of *n* node belonging to class *K*; it will be calculated according to the degrees of membership as follows:

$$P_K = \frac{\sum_{i=1}^{N} \mu_k(y_i) * \mu_k(x_i)}{\sum_{c=1}^{K} \sum_{i=1}^{N} \mu_k(y_i) * \mu_k(x_i)}$$

(4)

where *K* is the number of classes, *N* is the number of examples, $\mu_K(y_i)$ and $\mu_K(x_i)$ are the degrees of membership.

Entropy is the global disorder of the system. For example, for two classes in the case of a conventional tree, entropy is maximum if the probability of belonging to a class is 0.5, and minimum where the probability of belonging to a class is 1 (and thus 0 for the other class). The purpose of a decision tree, fuzzy or classical, is to successively divide the data space in order to reduce the maximum entropy.

FisPro is an open source tool used to create fuzzy inference systems (FIS). It is dedicated for reasoning purposes, specifically for simulating a physical or biological system [15]. It includes many algorithms (most of them implemented as C programs) to generate fuzzy partitions and rules directly from experimental data. It offers FIS visualization methods and data with a user-friendly interface based on java. We make use of the Fuzzy Decision Trees (FDT) [36] algorithm provided by FisPro.

After the fuzzy decision tree induction, we built a classifier based on rules. A fuzzy rule is given in the form of a simple IF-THEN rule with a condition and a conclusion. Fig. 5 lists a sample about fuzzy rules applied for the Diabetes classification system.

Rule	Active	IF Plas	AND Insu	AND Mass	AND Pedi	AND Age	THEN Class
1	✓	low	low		low		low
2	✓	low	medium		low	young	verylow
3	✓	low	medium		low	medium	verylow
4	✓	low	high		low		verylow
5	✓	low	low		medium		low
6	✓	low	medium		medium		verylow
7	✓	low			high		veryhigh
8	✓	medium		low	low		veryhigh
9	✓	medium	low	medium	low	young	medium
10	✓	medium	low	medium	low	medium	veryhigh

Fig. 5. Fuzzy rules for the Diabetes classification.

The fuzzy rules evaluations and combining results of individual rules are performed by fuzzy set operations [42]. The fuzzy sets operations are different from those on non-fuzzy sets. Let *μA* and *μB* the membership functions for the fuzzy sets *A* and *B*. Possible fuzzy operations for AND and OR operators on these sets are given in Table III. max and min are respectively the mostly used operations for OR and AND operators. For complement (NOT A) operation, (1- *μA*) is used for the fuzzy sets.

TABLE III. Fuzzy Set Operations

	OR (Union)		AND (Intersection)
MAX	Max { $\mu_A(x), \mu_B(x)$ }	MIN	Min { $\mu_A(x), \mu_B(x)$ }
ASUM	$\mu_A(x) + \mu_B(x) - \mu_A(x)\mu_B(x)$	PROD	$\mu_A(x)\mu_B(x)$
BSUM	Min {1, $\mu_A(x) + \mu_B(x)$}	BDIF	Max {0, $\mu_A(x) + \mu_B(x)-1$}

The result of each rule is evaluated. Then, the results should be combined to get a final result. This process is called inference. Results of individual rules can be combined in different ways. The maximum algorithm is generally used for accumulation.

3) Defuzzification

The aggregation result is converted into a crisp value for Diabetes Mellitus output. This is conducted by the defuzzification process [42]. A single number represents the fuzzy set outcome. Centroid method is used to convert the final combined fuzzy conclusion into a crisp value [20] as shown in Fig. 6.
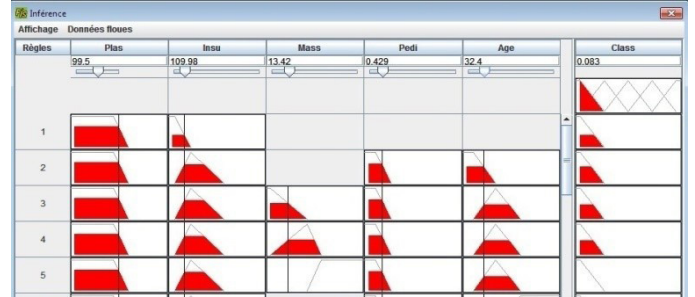

Fig. 6. Fispro interface for deffuzification.

4) Performance Evaluation

The proposed fuzzy inference mechanism has been implemented using Fispro. The data are taken from the Pima Indian Diabetes Dataset, the input variables are Plas, Ins, Mass, Pedi and Age; the output variable is DM (Diabetes Mellitus). To test the Fuzzy inference mechanism we took data including the age range 25-30 [19], [13]. The first experimental environment was constructed to evaluate the performance of the fuzzy classification task.

Well-defined MFs have been regulated and a fuzzy rule base has been produced. Then, the rules were extracted from a training data set of 768 instances including 500 no diabetics instances and 268 diabetics instances.

Specificity, sensitivity and accuracy are universal evaluation measures used to evaluate effectiveness of the data mining system, and particularly to compute how the assessment test is consistent and worthy. Sensitivity measure evaluates the diagnostic test correctly at detecting a positive disease. Specificity evaluates how the proportion of patients without disease can be properly ruled out. Accuracy can be concluded using specificity and sensitivity in the presence of prevalence. Accuracy is correctly defined by the diagnostic test by eliminating a given condition [18]:

$$Confusion\ Matrix = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$$

(5)

Where *FP*, *FN*, *TP* and *TN* are false positives, false negatives, true positives, true negatives respectively.

$$Sensitivity = \frac{TP}{TP + FN}$$

(6)

$$Specificity = \frac{TN}{TN + FP}$$

(7)

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

(8)

Different results of a two-class prediction are listed in Table IV. Accuracy is the proportion of the total number of predictions that were correct.

TABLE IV. Various Results of a Two-Class Prediction

Method	Accuracy (%)	Author
Our FDT framework for very young	91,67%	-
Fuzzy Expert System for Diagnosis of diabetes using Fuzzy Determination Mechanism	89,32%	Kalpana et al. (2011) [19][20]

```
1  create database travel;
2  use travel;
3  drop table travel;
4  create table travel(caseId INTEGER, Nbrenceinte INTEGER, Concentration INTEGER,
5   Tension INTEGER, Epaisseur INTEGER, Insuline INTEGER, IndexMass DOUBLE,
6  FctPure DOUBLE, Age INTEGER, VarDiag INTEGER);
7  insert into travel values(1 , 6 , 148 , 72 , 35 , 0  , 33.6  , 0.627 , 50 , 1 );
8  insert into travel values(2 , 1 , 85  , 66 , 29 , 0  , 26.6  , 0.351 , 31 , 0 );
9  insert into travel values(3 , 8 , 183 , 64 , 0  , 0  , 23.3  , 0.672 , 32 , 1 );
10 insert into travel values(4 , 1 , 89  , 66 , 23 , 94 , 28.1  , 0.167 , 21 , 0 );
11 insert into travel values(5 , 0 , 137 , 40 , 35 , 168, 43.1  , 2.288 , 33 , 1 );
```

Fig. 7. Extract from the case base in XML format.

D. Case Based Reasoning Using JColibri

Sometimes, we are confronted with similar problems that we try to resolve using our experience. CBR is a problem-solving approach based on the reuse by the analogy of past experience. We focus on the second step of the case-based reasoning: the retrieval phase. The aim of this step is to find cases that are similar to the given problem. In a previous work [14], a planning approach guided by case-based reasoning based on retrieval by decision tree has been proposed. In this paper, we propose a combination of methods to improve the retrieval step of this approach. To achieve this, we use fuzzy decision tree for the retrieval of similar cases.

The case-based reasoning consists in solving a new problem called target problem by using a whole of past solved problems and a source case is a past episode of solved problems. A case base consists of a set of source cases. Each source case has two parts: problem and solution. The problem part is described by a set of relevant characteristics called descriptors and the solution part is the plan to follow, it depends on the descriptors values. The process of case-based reasoning usually operates in five sequential phases: elaboration, retrieval, adaptation, revision and retain [16].

The elaboration step formalizes the problem using descriptors so as to build the target case. The cases retrieval step uses similarity measures to extract from the case base, the source cases which the problem part is similar to the target problem. The adaptation phase proposes one or more solutions to the target problem by adapting the proposed solutions. This step is often based on the use of knowledge in the field of the application. The objective of the revision step is to revise the proposed solutions by the previous phase according to certain rules and/or heuristics that depend on the field of application. This phase may be made by experts or in an automatic way. The last step, retain, is responsible for improving the experience of case-based reasoning system by enriching the case base with the new solved problems. Indeed, the resolved cases can be added to the case base and be used later to solve new problems.

1) Construction of the Case Base

We applied our approach to the Pima Indian Diabetes Database (PIDD), available on the official website of the UCI (machine learning repository). This database consists of a collection of medical diagnostic reports including 768 women, 268 who are diabetic and 500 non diabetic, over 21 years of age. Each case is described by 9 attributes. The solution part of cases represents the diagnosis of diabetes (0 non-diabetic patient, 1 otherwise). To establish the right diagnosis, the result of our fuzzy approach is a value between 0 and 1 (a percentage of having diabetes disease).

After discussing with experts in the field and consulting the site http://www.diabetes.co.uk/, we identified the type of diabetes and the best monitoring plan to be followed based on various factors including glucose, insulin, body mass, age.

In the jCOLIBRI tool, the database is given as an XML file. Thus, we need to convert the case base to an XML file. An excerpt of the case base in XML format is shown in Fig. 7.

To help the patient, we have proposed a monitoring plan and a printed notebook to take the blood sugar every day. The solution part refers to a plan of treatment Y that takes its values from the set of diabetes care plans C = {Plan1, Plan2, Plan3, Plan4}, where Plan1 = 'Ace-inhibitor therapy'; Plan2 = 'low fat diet'; Plan3 = 'determine exercise regime' and Plan4 = 'diabetes prevention program'.

2) Retrieval by Fuzzy Inference Mechanism

The studies cited above have led us to propose a new approach FDT4CR (Fuzzy Decision Tree For Cases Retrieval). The use of FDT for Retrieval optimizes the response time and avoids running whenever the k-nn algorithm. The objective of the fuzzy classification model consists in assigning a surveillance plan to the new case entry. Instead of using k-nn to find a plan, the retrieval by fuzzy decision tree is used in order to benefit from past experience. The classification model is responsible of the classification of the target case according to the descriptors values in order to find a solution.

The procedure of the fuzzy decision tree generation in FisPro needs both a Fuzzy Inference System configuration file and the PIDD data file. The prediction is based on one output only, even if the FIS has several outputs. This output is chosen by the user according to four possibilities:

1. A Fuzzy output, with the classification option.
2. A Fuzzy output, without the classification option.
3. A Crisp output, with the classification option.
4. A Crisp output, without the classification option

3) Performance Evaluation

The learning data are taken from the Pima Indian Diabetes Dataset and will serve to build the fuzzy decision tree. The second set of data including the age range 25-30, is used to test the validity of the classification procedure.

The experimental environment used JColibri platform to evaluate the performance of the proposed framework;

The proposed FDT approach analyzes the PIDD data and generates a corresponding plan based on Fuzzification, Fuzzy Inference Mechanism and Defuzzification for very young parameter [19].

The first experiment compared with the results of [19] is listed in Table V, indicating that the proposed approach automatically supports the analysis of the data.

TABLE V. Result Obtained from Fispro

	Glucose (mg/dl)	INS (mu U/ml)	BMI (Kg/m2)	DPF	Age
Data	172	579	42.4	0.702	28
Our FDT4CR Crisp output with classification option	The Decision statement justifies that the possibility of suffering from diabetes for this person is medium (53,03%)				
Medical Practitioner	The person is diabetic				

To compare FDT4CR with other techniques, we applied k-nearest neighbours (k-NN), decision tree and our fuzzy decision tree for cases retrieval (FDT4CR) on the same case base. Table VI lists the different accuracy results. These results show that FDT4CR gets the highest accuracy for very young surveillance plans.

TABLE VI. Results of Experimentation

JColibri	Weka	Fispro
k-NN	Decision tree	Fuzzy DT
66%	73%	81%

IV. Conclusion and Perspectives

Multiple competing motivations allowed to define a Fuzzy model for case-based reasoning knowledge base systems. Effectively, we did not just want to experiment with a new approach of case indexing by FDT, but the aim is also to improve the modeling of uncertain and vague natural language concepts.

Planning guided by case-based reasoning is described through the following steps:

1. Construct the fuzzy Decision Tree by symbolic learning using Fispro platform and extract fuzzy rules;

2. Import fuzzy knowledge base using the JColibri platform.

3. Build the base of cases by JColibri;

4. Facing a new problem (target case), the CBR process using JColibri begins with the stage of retrieval. The retrieval is to find, among the source cases, the most similar cases to the target case. This step is treated using the FDT (Fispro). If the suggested solution does not satisfy the target case constraints, adaptation rules are used to adjust the solution. In the revision step, first the expert verifies the validity of the outcome. Then he alters or confirms the solution. Finally, in the retain step the system checks if the case base does not contain the new case before adding it [41].

k-nearest neighbors (k-NN) is typically used to calculate similarity in the retrieval step (cases indexing). We compared our Fuzzy Inference Mechanism with decision tree and k-nearest neighbours. We noticed that case indexing for the selection of a diabetes surveillance plan is considerably better with our Fuzzy Inference Mechanism (FDT4CR). Compared to retrieval by k-nn which is costly in computing time, retrieval by FDT provides one major advantage, it optimizes the response time. Furthermore, fuzzy reasoning combined with data mining for retrieval presents several advantages. First, it reduces the complexity of similarity calculation between individuals. Second, it presents an improved retrieval in the CBR process of JColibri. However, FDT4CR has some disadvantages related to the complexity of the domain of Diabetes Diagnosis. Building the cases-base from diabetic patient databases, the encoding of case base knowledge with standard medical files and the adaptation of vague data are examples of these challenges. The case structure used in FDT4CR is quite simple. The part problem of cases has been described by Body mass index (BMI), 2-hour serum insulin (INS), Plasma glucose concentration in 2-hours OGTT (Glucose), Age (Age) and Diabetes pedigree function (DPF). The solution part is defined by a monitoring plan corresponding to a crisp output, with the classification option.

In this paper, FDT4CR has been applied to improve the classification of diabetic patients. As future work, we suggest to expand FDT4CR approach to support the different phases of CBR life cycle. And to optimize response time and complexity we propose a Boolean modeling of induction and rule inference [37], [38], [39], [40].

As a result of the literature review [4], there is a specific need for more comprehensive improvements in clinical CBR. We plan to apply our FDT4CR approach in the emergency field following the proposed scheme [51].

References

[1] B. Campillo-Gimenez, W. Jouini, S. Bayat, and M. Cuggia, "Improving case-based reasoning systems by combining k-nearest neighbour algorithm with logistic regression in the prediction of patients registration on the renal transplant waiting list," *PLoS One*, vol. 8, no. 9, 2013.

[2] N. Choudhury and S. A. Begum, "A survey on case-based reasoning in medicine," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 8, pp. 136-144, 2016.

[3] S. Montani, R. Bellazzi, L. Portinale, G. d'Annunzio, S. Fiocchi, and M. Stefanelli, "Diabetic patients management exploiting case-based reasoning techniques," *Computer Methods and Programs in Biomedicine*, vol 62, pp. 205–218, 2000.

[4] R Schmidt and L. Gierl, "Case-based Reasoning for Antibiotics Therapy Advice: An Investigation of Retrieval Algorithms and Prototypes," *Artificial Intelligence in Medicine*, vol. 23, no. 2, pp. 171-186, 2001.

[5] D. Doyle, P. Cunningham, and P. Walsh, "An evaluation of the usefulness of explanation in a CBR system for decision support in bronchiolitis treatment," *Computational Intelligence*, vol. 22, pp. 269-281, 2006.

[6] S. Montani, L. Portinale, G. Leonardi, R. Bellazzi, and R. Bellazzi, "Case-based retrieval to support the treatment of end stage renal failure patients," *Artificial Intelligence in Medicine*, vol.37, pp. 31-42, 2006.

[7] C. Marling, J. Shubrook, and F. Schwartz, "Case-based decision support for patients with type 1 diabetes on insulin pump therapy," *9th European Conference ECCBR 2008*, Springer-Verlag, Berlin, pp. 325-339, 2008.

[8] F. J. De Paz, S. Rodriguez, J. Bajo, and M. J. Corchado, "Case based reasoning as a decision support system for cancer diagnosis: A case study," *International Journal of Hybrid Intelligent Systems*, vol. 6,no. 2, pp. 97-110, 2009.

[9] R. Lopez de Mantaras, D. McSherry, D. Bridge, D. Leake, B. Smyth, S. Craw, B. Faltings, M.L. Maher, M.T. Cox, K. Forbus, M. Keane, A. Aamodt, and I. Watson, "Retrieval, reuse, revision and retention in case-based reasoning," *The Knowledge Engineering Review*, vol. 20, no. 3, pp. 215-240, 2006.

[10] G. Quellec, M. Lamard, L. Bekri, G. Cazuguel, B. Cochener, and C. Roux, "Multimedia medical case retrieval using decision trees," *International conference of the IEEE EMBS*, pp. 4536-4539, 2007.

[11] E.K. Burke, B.L. MacCarthy, S. Petrovic, and R. Qu, "Multiple-retrieval case based reasoning for course timetabling problems," *Journal of the Operational Research Society*, vol. 57, pp. 148-162, 2006.

[12] N. Choudhury and S. A. Begum, "The role of fuzzy logic in case-based reasoning: a survey," *Indian Journal of Computer Science and Engineering*, vol. 8, no. 3, 2017.

[13] D. K. Choubey, S. Paul, S. Kumar and S. Kumar, "Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System*, pp. 451-455, 2017.

[14] S. Benbelkacem, B. Atmani, and A. Mansoul, "Planification guidée par raisonnement à base de cas et datamining : Remémoration des cas par arbre de décision,"*aIde à la Décision à tous les Etages Aide@EGC2012*, pp. 62-72, 2012.

[15] S. Guillaume and B. Charnomordic, "Learning interpretable fuzzy inference systems with FisPro," *Information Sciences*, vol. 181, no. 20, pp. 4409-4427, 2011.

[16] T. Hastie, R. Tibshirani, and J. Friefman, *The Elements of Statistical Learning*, Data Mining, Inference and Prediction, 2nd ed., Springer, New York, 2001.

[17] J. Bello-Tomas, P. Gonzalez-Calero, and B. Diaz-Agudo, "Jcolibri: An object-oriented framework for building CBR systems," *Advances in Case-Based Reasoning, 7th European Conference on Case-Based Reasoning*, pp. 32-46, 2004.

[18] P. K. N. Anooj. "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules," *Central European Journal of Computer Science*, vol. 1, no 4, pp. 482-498, 2011.

[19] M. Kalpana and A. S. Kumar, "Fuzzy expert system for diagnosis of

diabetes using fuzzy determination mechanism," *International Journal of Computer Science & Emerging Technologies*, vol. 2, no.6, pp. 354-361, 2011.

[20] M. Kalpana and A. S. Kumar, "Fuzzy expert system for diabetes using fuzzy verdict mechanism," *International Journal of Advanced Networking and Applications*, vol. 3, no 2, pp. 1128-1134, 2011.

[21] M. Huang, M. Chen, and S. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis," *Expert Systems with Applications*, vol. 32, pp. 856-867, 2007.

[22] T. G. Houeland, "An efficient random decision tree algorithm for case-based reasoning systems," *24th International Florida Artificial Intelligence Research Society Conference*. May 2011.

[23] X. Boyen and L. Wehenkel, "Automatic induction of fuzzy decision trees and its application to power system security assessment," *Fuzzy Sets and Systems*, vol.102 , no.1, pp. 3-19, 1999.

[24] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and B. Von Schéele, "A case-based decision support system for individual stress diagnosis using fuzzy similarity matching," *Computational Intelligence*, vol. 25, pp. 180-195, 2009.

[25] F. Barrientos and G. Sainz, "Interpretable knowledge extraction from emergency call data based on fuzzy unsupervised decision tree," *Knowledge-based systems,* vol. 25, no.1, pp. 77-87, 2012.

[26] V. Levashenko and E. Zaitseva, "Fuzzy decision trees in medical decision making support system," *Proceedings of the IEEE Federated Confenrence on Computer Science & Information Systems,* pp.213-219, 2012.

[27] D. R. Adidela, D. G.Lavanya, S. G.Jaya, and A. R. Allam, "Application of fuzzy ID3 to predict diabetes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 4, pp.541-545, 2012.

[28] A. G. Karegowda, V. Punya, M.A. Jayaram, and A.S. Manjunath, "Cascading K-means clustering and K-Nearest Neighbor classifier for categorization of diabetic patients," *International Journal of Engineering and Advanced Technology*, vol. 1, no. 3, pp. 147-151, 2012.

[29] A. G. Karegowda, V. Punya, M.A. Jayaram, and A.S. Manjunath, "Rule based classification for diabetic patients using cascaded K-means and decision tree C4.5," *International Journal of Computer Applications,* vol. 45, 2012.

[30] G. Kaur and A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, pp. 13-17, 2014.

[31] L. A. Zadeh. "Is there a need for fuzzy logic?" *Information Sciences*, vol. 178, no. 13, pp. 2751–2856, 2008.

[32] C. S. Lee and M. H. Wang, "A fuzzy expert system for diabetes decision support application," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol.41, no. 1, pp. 139-153, 2011.

[33] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," *Wadsworth International Group*, Belmont CA, 1984.

[34] J.R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[35] R. Weber, "Fuzzy-id3: A class of methods for automatic knowledge acquisition," *2nd International conference on fuzzy logic and neural networks*, pp. 265–268, 1992.

[36] J. M. Alonso and L. Magdalena, "Generating understandable and accurate fuzzy rule-based systems in a java environment," *International Workshop on Fuzzy Logic and Applications*, Springer, Berlin, Heidelberg, pp. 212-219, 2011.

[37] B. Atmani and B. Beldjilali, "Knowledge discovery in database: Induction graph and cellular automaton," *Computing and Informatics*, vol. 26, no. 2, pp. 171-197, 2012.

[38] B. Atmani, S. Benbelkacem, and M. Benamina, "Planning by case-based reasoning based on fuzzy logic," *International Conference of Artificial Intelligence and Fuzzy Logic*, 2013.

[39] F. Barigou, B. Atmani, and B. Beldjilali, "Using a cellular automaton to extract medical information from clinical reports," *Journal of Information Processing Systems*, vol. 8, no 1, pp. 67-84, 2012.

[40] M. Benamina and B. Atmani, "Définition d'un modèle booléen de raisonnement flou adapté à la planification," *EGC*, pp. 553-554, 2012.

[41] S. Benbelkacem, B. Atmani, and M. Benamina, "Treatment tuberculosis retrieval using decision tree," *International Conference on Control Decision and Information Technologies*, 2013.

[42] S. Mokeddem and B. Atmani, "Assessment of Clinical Decision Support Systems for Predicting Coronary Heart Disease," *Fuzzy Systems: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, vol. 184, 2017.

[43] M. Rigla, G. García-Sáez, B. Pons, and M. E. Hernando, "Artificial intelligence methodologies and their application to diabetes," *Journal of Diabetes Science and Technology*, 2017.

[44] X. Blanco, S. Rodríguez, J.M. Corchado, and C. Zato, "Case-based reasoning applied to medical diagnosis and treatment," In: S. Omatu, J. Neves, J. Rodriguez, J. Paz Santana, and S. Gonzalez (Eds.) *Distributed Computing and Artificial Intelligence. Advances in Intelligent Systems and Computing*, vol. 217, pp. 137-146, Springer, 2013.

[45] S. El-Sappagh and M. M. Elmogy, "Medical case based reasoning frameworks: Current developments and future directions," *International Journal of Decision Support System Technology*, vol. 8, no 3, pp. 31-62, 2016.

[46] D. Azar and M. Bitar, "AI-based methods for predicting required insulin doses for diabetic patients," *International Journal of Artificial Intelligence*, vol. 13, no 1, pp. 8-24, 2015.

[47] S. Tahmasebian, M. Langarizadeh, M. Ghazisaeidi, and M. Mahdavi-Mazdeh, "Designing and implementation of fuzzy case-based reasoning system on android platform using electronic discharge summary of patients with chronic kidney diseases," *Acta Informatica Medica*, vol. 24, no 4, pp. 266-270, 2016.

[48] S. El-Sappagh and M. Elmogy, "A fuzzy ontology modeling for case base knowledge in diabetes mellitus domain," *Engineering Science and Technology, an International Journal*, vol. 20, no 3, pp. 1025-1040, 2017.

[49] S. Sharma and D. Mehrotra, "Building CBR based diagnosis system using jcolibri," *7th International Conference on Cloud Computing, Data Science & Engineering – Confluence*, pp. 634-638, 2017.

[50] M. K. Kiragu and P. W. Waiganjo, "Case based reasoning for treatment and management of diabetes," *International Journal of Computer Applications*, vol. 145, no 4, 2016.

[51] S. Benbelkacem, F. Kadri, S. Chaabane, and B. Atmani, "A data mining based approach to detect strain situations in hospital emergency department systems," *International Conference on Modeling, Optimization and Simulation*, Nancy, France, 2014.

[52] P. Kushwaha and R. R. Welekar, "Feature selection for image retrieval based on genetic algorithm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no 2, pp. 16-21, 2016.

[53] A. Taibi and B. Atmani. "Combining Fuzzy AHP with GIS and Decision Rules for Industrial Site Selection." *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no 6, pp. 60-69, 2017.

Mohammed Benamina

Mohammed Benamina is currently a PhD candidat at the University of Oran 1 Ahmed Ben Bella and affiliated researcher in Laboratoire d'Informatique d'Oran, Algeria. He received his Master degree in Computer Science in 2010 from University of Oran, Algeria. His research interests include data mining algorithms, expert systems and decision support systems.

Baghdad Atmani

Baghdad Atmani received his PhD in computer science from the University of Oran (Algeria) in 2007. He is currently a full professor in computer sciences. His interest field is artificial intelligence and machine learning. His research is based on knowledge representation, knowledge-based systems, CBR, data mining, expert systems, decision support systems and fuzzy logic. His research are guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance, information retrieval, simulation, data integration and spatial data mining.

Sofia Benbelkacem

Sofia Benbelkacem is currently a PhD candidate at the University of Oran 1 Ahmed Ben Bella and affiliated researcher in Laboratoire d'Informatique d'Oran, Algeria. Her research interests include data mining, planning, case-based reasoning, medical decision support systems, machine learning.