Improving Retrieval Performance of Case Based Reasoning Systems by Fuzzy Clustering

F. Saadi[1]*, B. Atmani[2], F. Henni[3]

[1] Laboratoire d'Informatique d'Oran (LIO), University of Oran 1 (Algeria)
[2] Laboratoire d'Informatique d'Oran (LIO), University of Mostaganem (Algeria)
[3] Computer Science and new Technologies Lab (CSTL), University of Mostaganem (Algeria)

Received 20 July 2022 | Accepted 2 June 2023 | Early Access 5 July 2023

Abstract

Case-based reasoning (CBR), which is a classical reasoning methodology, has been put to use. Its application has allowed significant progress in resolving problems related to the diagnosis, therapy, and prediction of diseases. However, this methodology has shown some complicated problems that must be resolved, including determining a representation form for the case (complexity, uncertainty, and vagueness of medical information), preventing the case base from the infinite growth of generated medical information and selecting the best retrieval technique. These limitations have pushed researchers to think about other ways of solving problems, and we are recently witnessing the integration of CBR with other techniques such as data mining. In this article, we develop a new approach integrating clustering (Fuzzy C-Means (FCM) and K-Means) in the CBR cycle. Clustering is one of the crucial challenges and has been successfully used in many areas to develop innate structures and hidden patterns for data grouping [1]. The objective of the proposed approach is to solve the limitations of CBR and improve it, particularly in the search for similar cases (retrieval step). The approach is tested with the publicly available immunotherapy dataset. The results of the experimentations show that the integration of the FCM algorithm in the retrieval step reduces the search space (the large volume of information), resolves the problem of the vagueness of medical information, speeds up the calculation and response time, and increases the search efficiency, which further improves the performance of the retrieval step and, consequently, the CBR system.

Keywords

Case Based Reasoning,
Case Retrieval,
Classification,
Data Mining,
Decision Support,
Fuzzy Clustering,
Immunitherapy, Kmeans.

DOI: 10.9781/ijimai.2023.07.002

I. Introduction

Case Based Reasoning (CBR) is a problem-solving paradigm. Pantic [2] and Aamodt & Plaza [3] defined a reasoning cycle with four phases: Retrieve-Reuse-Revise-Retain. Instead of relying only on general knowledge of a problem domain, CBR relies on the retrieval of past and solved problems, called source cases, to solve a current problem, called a target problem. A new experience is maintained each time a problem has been solved, making it immediately available for future problems. That is why the retrieval of similar cases is a crucial phase in the CBR cycle.

Dependent on the process of medical situation resolution, it is clear that the doctor's reasoning is mostly based on the fact that the current situation is probably treated before, and so the doctor will propose a solution that is more or less identical to the one previously adopted. This reasoning resembles the CBR reasoning methodology. This has motivated a lot of research into this reasoning method in the medical field [4], leading to the creation of computerized tools for solving decision-making problems using only this reasoning method (CBR). This work has ramifications in various fields of artificial intelligence: knowledge representation, classification, similarity measures, etc., which has made it a complex but widely used reasoning mode in medical decision support. However, The application of classical CBR systems in the medical domain has limitations due to the increasing complexity of this domain since many healthcare applications are simply too complex and multifaceted to be processed using this methodology [5]. The case representation has become more complex than in history in several applications: medical information (case) may come in part in the format of time series, images, or free text, and may also be intrinsically high-dimensional, imprecision, vagueness, and uncertainties [6]. Indeed, the large volume of generated medical information (symptoms, diseases, and treatments) which is increasing slows down the similarity computation in the retrieval phase and it becomes very expensive in computation time, knowing that time is a very important factor not to be neglected in a medical diagnosis [7]. Therefore, a suitable retrieval algorithm must be chosen to solve these problems [8].

The richness of various reasoning methods or approaches may be integrated to exceed the limitations of the application of classical CBR and support it as a knowledge engineering methodology at each phase of its cycle [9]. This integration has been widely deployed in multi-modal reasoning systems and it has been shown to be well-adapted, in particular for work related to the medical domain [8]. Among the techniques integrated into the classical CBR, data mining methods have shown some advantages in particular for improving the retrieval step [10].

* Corresponding author.
E-mail address: saadi_fatima@hotmail.fr

Please cite this article in press as:
F. Saadi, B. Atmani, F. Henni. Improving Retrieval Performance of Case Based Reasoning Systems by Fuzzy Clustering, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.07.002

The objective of this work, which is an extension of research presented in [11], is to improve the CBR cycle by improving the performance of the most important step in the CBR cycle, retrieve, through the choice of the best similarity measure. The proposed approach is adopted to develop a CDSS that assists dermatology experts in predicting a patient is responding to immunotherapy therapy for warts.

In this paper, we propose a new approach that integrates one of the data mining techniques which is clustering (Fuzzy C-Means (FCM) and K-Means) in the CBR cycle. The integration of clustering aims to demonstrate the value of this technique in CBR to reduce the number of cases while reducing the complexity of the retrieval step and speeding up the time which is a very important aspect of medical diagnosis. Thus, to prove the purpose of fuzzy logic to model the vague, imprecise, and uncertain concepts of medical information, we chose to integrate the FCM technique in the retrieval step. Fuzzy CBR should be used for this reason, as well as for better decision support [12]. Section II in this paper focuses on some related works. The methodologies provided in this study are explained in Section III. Experiments are implemented in Section IV, and the results are interpreted and evaluated. Finally, Section V, ends with conclusion.

II. Background

A. Use of CBR

CBR is a general decision-making methodology used in the medical field [7]. Several studies have used CBR in this area. Sharma and Mehrotra [13] applied the retrieval of cases by similarity measurement to develop a CBR implementation for the diagnostic of chronic renal failure. Demigha [14] designed a generic eLearning application for radiologists and other hospital staff. They developed this instrument using the CBR method. Mansoul and Atmani [10] proposed using Multi-Criteria Analysis (MCA) with the standard CBR retrieval to aid in finding the ideal solution. They included this method in a clinical decision support system. Gu et al. [15] implemented a CBR system for breast cancer diagnosis and used it in two experiments, one on benign/malignant tumor prediction and the other on secondary cancer prediction. Benfriha et al. [16] developed a new approach for case acquisition in CBR based on multi-label text categorization applied in a child's traumatic brain injuries dataset. El-Sappagh et al. [17] demonstrated that non-clinical CBR systems have made more progress than clinical CBR systems. In addition, when contrasted to other diabetes healthcare systems, CBR systems achieve the smallest gains. These studies show that clinical CBR, especially in diabetic systems, requires more thorough improvements.

B. Use of Data Mining

Data mining was crucial in the development of intelligent healthcare systems [18], [19]. We provide a list of research papers that employ data mining techniques in the healthcare industry. Dewan Sharma [20] created a tool that can identify and retrieve new heart illness information from a previous heart disease database record. They applied data mining algorithms including Neural Networks, Decision Tree, and Naive Bayes for their proposal. The idea of this research is to handle complex requests in diagnosing heart disease, allowing health doctors to improve medical judgment more than standard decision support systems could. Chen et al. [21] conducted research on localized chronic cerebral infarction and presented a new illness prediction method that has multiple modes and based on convolutional neural networks. Kumar & Sahoo [22] introduced a novel approach that uses Naive Bayes as well as genetic algorithms to enhance cardiovascular disease prediction. The outcomes of their research demonstrate that this approach enhances the efficiency of heart disease detection. Chaurasia & Pal [23] developed prediction models for heart disease survivability using a large dataset and applied three data mining techniques including decision trees and rule-based classifiers. Hachesu et al. [24] presented a method for determining and predicting heart patient length of hospital stay, they adopted in this research the decision trees, Support Vector Machines, and Artificial Neural Networks data mining algorithms. Martín et al. [25] created An algorithm for semi-supervised clustering. The technique, which is based on an ensemble of dissimilarities, has been used to identify tumor samples using gene expression patterns.

Clustering is a crucial unsupervised data mining technique used to find some underlying structure in a collection of patterns or objects [26]. A cluster maximizes the similarity of these objects and minimizes the similarity of objects not belonging to it. To do this, the data mining process uses distance functions. These functions evaluate the existing similarities (distances) between the entities to be grouped. In order to uncover the dataset's underlying natural cluster patterns, the choice of similarity measure is crucial [1]. Many distance functions are available in the literature. Saadi et al. In this work, we adopted the standard version of K-Means and FCM that uses the Euclidean distance, but there are many works that improve these techniques by using other distances. Kapil & Chawla [27] used Manhattan distance with C-means clustering. Sharma et al. [28] integrated the S-distance and the Euclidean distance with the C-means clustering algorithm. Karlekar et al. [26] added the S-distance to the traditional fuzzy K-means method in place of the Euclidean distance. Seal et al. [29] developed Fuzzy c-means clustering using a novel similarity metric based on Jeffreys-divergence.

1. Combining CBR With Data Mining Techniques

Some works that integrate data mining techniques in each step of the CBR cycle has illustrated in Table I.

TABLE I. Medical Systems Integrated Data Mining Techniques in CBR Cycle

References	CBR process and methods	Application domain
[30]	Retrieve: Euclidean, Manhattan, or Hamming distance Adaptation decision rules	medical
[31]	Retrieve: KNN Adaptation ANN	representations of human organs
[32]	Retrieve: KNN Adaptation: Rule-Based Reasoning	Cancer diagnosis
[23]	Retrieve: Dissimilarity Measurementsrevises and reuses genetic algorithm	Medical Diagnosis
[21]	Retrieve: Decision treeAdaptation: Decision Rule extracted from Decision tree	cardiovascular disease
[33]	case acquisition: Multi-label Retrieve: + KNN	child's traumatic brain injuries

Retrieve is often regarded as the most important step in CBR systems. The similarity measurement is the main task of this step. From the case base, the process will select similar cases that will be deemed the most relevant to begin the process of determining the solution for the medical situation. Several research papers have looked at the use of data mining algorithms to improve the efficiency of the retrieval stage. Gu et al. [34] and Jung et al. [35] proposed CBR-based models which combined the naive Bayes and KNN approaches for similarity measurement. Benbelkacem et al. [31], Chen et al. [21], and Saadi et al. [36] integrated the decision trees and the KNN in the retrieval step for the similarity calculation. Mansoul & Atmani [7] and Khussainova &

Jagannathan [37], Koo et al. [38] and Yadav [39] improved the retrieval step by decreasing the search space using Clustering techniques. In faced with complex real-world applications, retrieving cases must deal with uncertainties [30]. Demigha [14] focused at the function of the fuzzy system in the various stages of CBR and found that integrating fuzzy logic with CBR resulted in efficient hybrid approaches. Geetha et al. [40] presented a fuzzy CBR strategy for deciding the urgency of COVID-19 sick people. Ibrahim & Odedele [41] developed a system for detecting and diagnosing infectious diseases, COVID-19, using fuzzy CBR. Choudhury et al. [42] used fuzzy-rough nearest neighbor to enhance the retrieval step of the CBR system's performance and efficiency. The experimental findings reveal that this combination beats KNN for classification by a large margin, effectively boosting case retrieval efficiency and performance. Yamin et al. [43] offered a case-matching process that uses two algorithms to find similar cases the case similarity algorithm in the case when the case database is small and FCM secondary retrieval algorithm in the case when the case base is large. Banerjee & Chowdhury [44] used the (FCM) algorithm in the CBR system to classify the most prevalent anomalies in retina images caused by maturity-level eye disease and diabetes. Ekong et al. [6] developed a clinical decision support system based on CBR, neural networks, and fuzzy logic for diagnosing depressive illnesses. Begum et al. [30] created a fuzzy CBR approach that categorizes healthy and distressed people. Benamina et al. [12] developed a fuzzy CBR technique to ensure a better diagnosis for diabetics, which includes fuzzy-decision trees with CBR to increase reaction speed and recovery accuracy of similar cases. In previous work [11], a medical decision support system has been proposed; this system is guided by case reasoning and the clustering technique. The research aimed to enhance the retrieving process by integrating the FCM method in the similarity calculation. The results of the experiments performed in this paper were encouraging since they improved the most critical phase in the CBR cycle (retrieve). In this work, we propose an extended version of the article proposed in [11].

III. Contribution

The authors of this research develop a new approach that integrates the clustering algorithms(K-Means and FCM) in the retrieval step. this approach aims to speed up the similarity calculation which accelerates the retrieval phase and improves its performance. Fig. 1 summarizes the essential phases of the proposed approach. Our approach's primary steps are as follows:

- Identify the number of clusters using the elbow approach, then apply K-means techniques to create the clusters;
- Define the membership degree matrix using the FCM algorithm;
- Using the jCOLIBRI platform, construct the base case;
- The start of the CBR cycle begins with the arrival of a new case by launching the retrieval phase. In this phase, the user chooses the K-means or FCM for the retrieval step:
  - Retrieval with K-means: retrieval of the best cluster and similar cases with the KNN algorithm.
  - Retrieval with FCM: To identify the ideal clusters where the KNN algorithm technique measures the similarity, determine the new case's membership degree in each cluster.

A. Clustering Techniques

These techniques are used as part of a solution-finding strategy that helps you to pick the optimal answer from a smaller number of options. We chose the K-means and FCM as unsupervised classification techniques for clustering, with the goal of structuring the case base, guiding, and speeding up the retrieval.

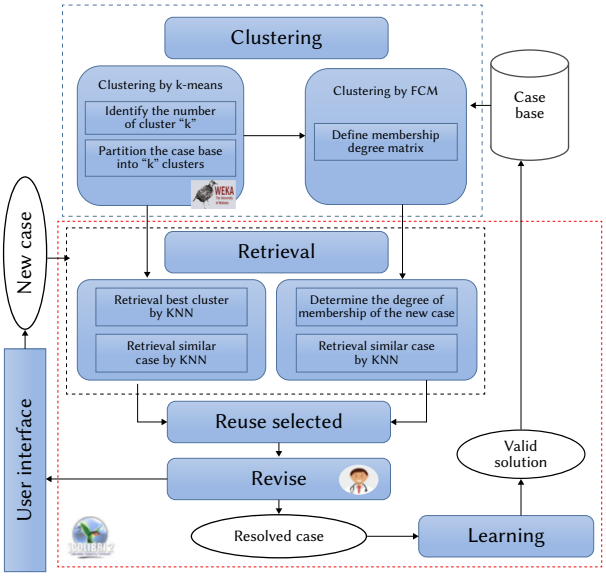

Fig. 1. The Proposed approach's architecture.

1. Clustering With K-means

The method consists in splitting the data into k clusters. It starts with a random clustering of the data (into k clusters), then assigning every element to the cluster that is nearest to it. Once the first iteration is completed, the averages of the clusters are calculated and the process is repeated until the clusters are stabilized. In this work, the clusters were generated from the data presented in a CSV file using the K-means algorithm implemented in the WEKA platform [45]. This algorithm has a fundamental drawback in that it requires the number of clusters, K, to be provided [46]. A good classification produces classes with a strong similarity within each class and a minor similarity between different classes. The distance between a point and its cluster center is called intra-class inertia. It's will be quantified as in (1) [46].

$$Intra - class = \sum_{i=1}^{k} \sum_{x \epsilon C_i} \| x - z_i \|^2 \quad (1)$$

Where k represents the total number of clusters, and C_i represents the cluster center. The distance between the clusters is the inter-class inertia. The distance between cluster centers is calculated, and the lowest of these numbers is utilized to determine it. The equation of inter-class inertia is defined in (2) [46].

$$Inter-class = min(\| z_i - z_j \|^2), i = 1,2,\ldots,k; j = 1,2,\ldots,k \quad (2)$$

The partition is excellent when the classes are homogeneous, the intra-class inertia is low, and the inter-class inertia is high. One of the existing strategies for identifying the clusters number is the elbow method. It is a visual method of varying the number of clusters and monitoring the evolution of a solution quality indicator (the proportion of inertia) in order to search for the "elbow" in the graphic. The idea is, to begin with $K = 2$, and keep increasing it in each step by 1, calculating the clusters and the intra-class inertia. The cost drops dramatically at some value for K and afterward, it reaches a plateau when you increase it further. This is the value we want for K. [47]. The K-means technique is applied to split the case base into K clusters after selecting the best value of K. As a result, we get a 0 and 1 Boolean matrix. The existence of cases in the cluster is indicated by a 1, whereas a 0 indicates the lack of this case.

2. Clustering With Fuzzy C-Means (FCM)

The FCM method, sometimes called the fuzzy K-means, was proposed by [48]. It is a fuzzy logic method in which the probability of belonging to the c groups for each observation is evaluated and presented as a membership matrix u of size n by c, where u_{ik} is the probability of observation k belonging to group i. The total of each row u equals one, like K-means, requires that the number of classes and iterations be fixed in advance; the initial class centers are also either randomly drawn or specified by the user. The class centers v are then iteratively updated with the matrix u. The membership probabilities are then re-evaluated using the new class centers. This method is performed till the given number of iterations has been attained or a convergence criterion is met. (3) is used to calculate the update of the classes and u:

$$u_{ik} = \frac{(\| x_k - v_i \|^2)^{\frac{-1}{(m-1)}}}{\sum_{j=1}^{c}(\| x_k - v_j \|^2)^{\frac{-1}{(m-1)}}} et v_i = \frac{\sum_{k=1}^{N} u_{ik}^m (x_k)}{\sum_{k=1}^{N} u_{ik}^m} \quad (3)$$

Where k and i represent an observation and a class, respectively. The main addition to the FCM method is the parameter m. It allows controlling the degree of fuzziness in the classification. If $m = 1$, the obtained classification is strict (the matrix u contains only 0 and 1 values); as m increases, the values of the matrix u decrease until they are perfectly homogeneous. Different types of distances (Euclidean, Chi-square, Mahalanobis, etc.) can be used as in the classical K-means, and the initial class centers, as well as the number of iterations, can potentially affect the classification results. It is common practice to experiment with various values for the number of classes until the ideal combination is found (More information, including an overview of the FCM algorithm, can be found in our previous work [11]. We applied the FCM approach in this step. The membership degree matrix determined from the K-means algorithm is the input. We get a membership degree matrix for each instance in each cluster as a result of this step.

B. CBR Process

1. Retrieval With K-Means Method

The jCOLIBRI platform is used for the retrieval stage. Once a new case is received, similar cases are chosen using the KNN algorithm to calculate similarity. Our search for similar cases is divided into two stages:

- The best cluster search: In this step, the similarity calculation is applied between the target case and the cluster centers (centroids) found via the K-means method.
- The search for similar cases: This step allows to filter the cases so that only the cases belonging to the best cluster are kept and that the similarity computation will be performed between the selected and filtered cases, instead of being computed between all of the case base's cases, which facilitates the retrieval step.

2. Retrieval With FCM Method

In this step, the FCM method is relaunched to calculate the membership degree of the new case in the cluster and takes from the resulting matrix the two clusters to which the new case has a high degree of membership and that is the advantage over the K-means. It is up to say instead of searching in one and only one cluster, with FCM, the search will be done in more than one cluster and this gives a high percentage to finding the best case similar to the new problem. The similarity is calculated only between the cases of the chosen clusters in the previous step and the KNN is used to select similar cases. To calculate the local similarity (similarity between attributes), the Euclidean distance is used because we only have attributes of numeric type; the formula of the Euclidean distance is defined by(4):

$$D(x,y) = \sqrt{\frac{x^2 - y^2}{x + y}} \quad (4)$$

After the computation of the local similarity, the global similarity is measured (the distance between two cases) using the technique most used in CBR systems KNN. The following equation is used to calculate similarity by KNN as in (5):

$$sim(C,S) = \frac{\sum_{f=1}^{n} w_f * sim(C_s, S_f)}{\sum(w_f)} \quad (5)$$

C stands for a new case, S for a saved case, w for an expert-defined weight, n represented case's attributes number, f for the attribute index and $Sim(C, S)$ is the local similarity for attribute f. Adaptation is not taken into account in our study. Because by definition it allows to partially or completely resume the solution that already exists in the database which is not our case. The revision entails validating the solution devised by the expert -doctor). The solution to the new problem has been discovered and validated, a new experiment is created, and it is saved in the case database to expand the case database and boost the potential for solving future situations.

IV. Experimentation and Results

A. Construction of the Case Base

The UCI Machine Learning Repository is used for evaluating the performance and efficiency of the proposed system. The authors specifically selected the "Immunotherapy" dataset which will be discussed further in the following. The UCI "Immunotherapy" dataset https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset has been obtained inside a clinical of dermatology in Mashhad city situated in Iran. It includes 90 instances that presented the patients who suffered from warts. Each instance consists of 08 attributes, as shown in Table II. Such as the class that signifies the patients' response to immunotherapy treatment (failure or success).

TABLE II. Features Used in the Immunotherapy Dataset

N°	Attribute	Value
1	Gender(Sex)	Man or Woman
2	Age	From 15 to 56 years
3	Time passed before therapy (Time)	From 0 to 12 months
4	The amount of warts on the body (N° warts)	From 1 to 19 warts
5	Wart's type (Type)	common, plantar, or both
6	The biggest wart's surface area (Area)	From 6 to 900 mm2
7	The initial test's induration diameter (Ind- dia)	From 5 to 70 mm
8	The patients' response to treatment (Class)	Success or failure

B. Clustering Technique

1. Clustering With K-Means

In the beginning, the approach generates the clusters using the K-means algorithm implemented under the WEKA platform. To determine the cluster number, the elbow method is appliqued: we vary k and follow the intra-class inertia, Fig. 2 illustrates the variation in intra-class inertia as a function of the number of clusters selected. According to this result, we deduce that the partition in $K = 6$ is the last to induce a significant information gain (the curvature in Fig. 2 shows a clear peak for $K = 6$ clusters). The K-means algorithm restart and generates six clusters.

2. Clustering With FCM Method

The system imports the case base functionality and starts configuration after it has been launched. Furthermore, it builds the membership degree matrix using the FCM technique.

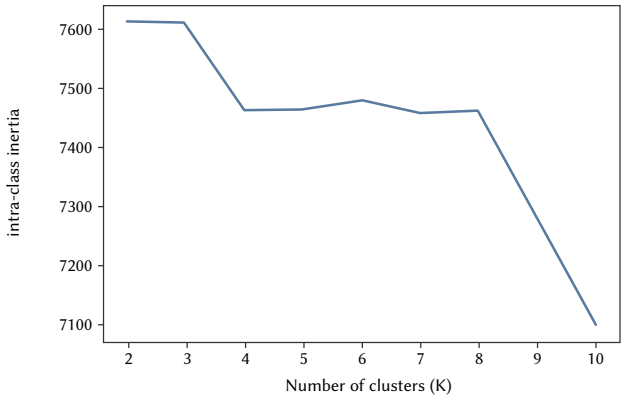

Fig. 2. The Intra-class inertia's variation.

3. Case Based Reasoning

The system consists of a simple interface via which a request may be launched. This request definitely illustrates a new patient with warts. The interface also provides the clinician the opportunity of specifying the amount of similar cases that he wants to retrieve. This option allows the doctor to define the number of cases to retrieve.

4. Retrieval With K-Means Method

The system calculates the similarity between centroids cases (cluster centers) and finds the best cluster that groups the nearest cases to the new case, then it filters the cases from the base so that only the best cluster cases are kept and it restarts the similarity calculation between the filtered cases to find the most similar cases requested.

5. Retrieval With FCM Method

The system moves on to the retrieval stage, where it re-runs the FCM technique to construct the two clusters in which the target case has a high degree of membership and search for the best similar case inside these two clusters rather than researching the whole case base. The clinician can go through similar cases and choose the one that he thinks is the most effective. The user can then make adjustments to the obtained solution after making this choice. Once the doctor has adjusted the solution, the system gives the user the option of retaining this target case and saving it with the cases of the case base or putting it in a temporary case base, where it may require more time before choosing whether to keep the target case.

C. Performance Evaluation

To evaluate the retrieval with FCM and K-means methods, we assessed them on a similar case base. These approaches' goal is to predict a patient with warts (target case) is responding to immunotherapy therapy (solution for the target case). In the following table Table III, five of target cases are randomly chosen from the cases from the case base, considering that the entire case base is used for the generation of the models. Then we execute the clustering using both methods and the retrieval using KNN ($K = 3$).

Table IV shows the results of experimentation obtained by two methods. Retrieving with K-means returns the best cluster, whereas retrieving with FCM provides the two best clusters with the of the target case's membership degree to these clusters, and the highest similar cases with the degrees of similarity and response of the target case to treatment.

TABLE III. Target Cases

Case	Sex	Age	Time	N° warts	Type	Area	Ind-dia	Y(CLASS)
Case 10	Women	32	12	6	3	35	5	Failure
Case 6	Man	15	5	3	3	84	7	Success
Case 81	Man	23	3	2	3	87	70	Success
Case 32	Man	30	1	2	1	88	3	Success
Case 89	Man	32	12	9	1	43	50	Failure

The quality of a diagnosis is crucial in providing medical treatment since a doctor's recommendations for medical therapy are based on diagnostic tests (medical tests, medical signs, symptoms, etc.). Fortunately, it is possible to measure the features of diagnostic tests. Based on these features, the ideal test may be selected for a particular illness condition. A diagnostic test is frequently described using the statistics of sensitivity, specificity, and accuracy. They are used in particular to measure a test's quality and dependability [49]. Several parameters are frequently used in conjunction with the definitions of

TABLE IV. Results Obtained by Retrieving With K-means and FCM

Retrieving with K-means					Retrieving with FCM				
New case	No. cluster	No. Similar case	Similarity degree	Response to treatment	No. cluster	Membership degree	No. Similar case	Similarity degree	Response to treatment
Case N°10	1	Case23	0.32	Failure	5	0.28	Case 10	0.0	Failure
		Case 90	0.40	Success	2		Case 17	0.29	Success
		Case 11	0.41	Success			Case 45	0.37	Success
Case N° 6	2	Case6	0.00	Success	4	0.29	Case 6	0.0	Success
		Case 72	0.20	Success	3		Case 66	0.22	Success
		Case 66	0.22	Success			Case 36	0.23	Failure
Case N°81	2	Case21	0.34	Success	4	0.32	Case 81	0.0	Success
		Case 32	0.40	Success	3		Case 71	0.35	Success
		Case 6	0.47	Success			Case 42	0.34	Success
Case N°32		Case61	0.51	Success	4	0.28	Case 32	0.0	Success
		Case 63	0.58	Success	3		Case 4	0.25	Failure
		Case 74	0.59	Success			Case 6	0.27	Success
Case N°89		Case10	0.42	Failure	5	0.26	Case 89	0.0	Failure
		Case 27	0.45	Success	2		Case 21	0.25	Failure
		Case 18	0.45	Success			Case 59	0.27	Success

accuracy, sensitivity, and specificity. True positives *TP* , false positives *FP* , true negatives *TN*, and false negatives *FN* are the four parameters of the confusion matrix (Table V.). The outcome of the diagnostic test is regarded as a true positive if a disease is demonstrated to be present in a patient and the diagnostic test also demonstrates the existence of the disease. Similar to this, when a disease is absent in a patient and the diagnostic test indicates this is also the case, the test result is said to be a true negative (*TN*). True results, whether positive or negative, point to a correlation between the outcome of the diagnostic test and the established condition (also called the standard of truth). No clinical exam, though, is flawless. The test result is considered to be false positive if it shows that a patient has an illness when they actually don't (*FP*). Similar to this, a test result is false negative if it indicates that a patient who has a condition for sure does not have it (*FN*). The test findings are at odds with the real disease when they are both falsely positive and falsely negative.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

According to the equations (6),(7),(8), sensitivity is the percentage of true positives that a diagnostic test successfully identifies. It demonstrates how accurate the test is in identifying diseases. The percentage of true negatives that a diagnostic test accurately identifies is known as specificity. It demonstrates how well the test detects a normal (negative) situation. Accuracy is the percentage of real outcomes in a population, whether they are real positive or real negative. It gauges how reliable a condition-specific diagnostic test is. to assess the efficacy of our system, these performance measurements are calculated in Table IV.

TABLE V. Confusion Matrix

	Predicted Class	
Actual class	True Positives (TP)	False Positives (FP)
	False Negatives (FN)	True Negatives (TN)

In previous work [36], we applied KNN in the retrieval phase, we also tested K-means in the same way in order to compare the results, while [50] applied Fuzzy rule-based on the same case base that we used and for the same objective as ours, which is the prediction of the result of immunotherapy treatment.

In Fig. 3, We present the results of a comparison between our proposed approach FCM and other methods (KNN, Kmeans, and Fuzzy rule-based).

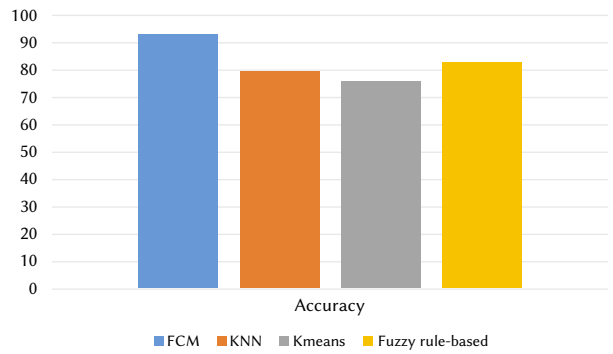

Fig. 3. Comparison of results of experimentation

D. Discussion of Results

As shown in Table IV, the retrieval step with FCM technique was always successful in finding the cluster containing the target case, however, the retrieval with K-means was only successful in finding the cluster containing case n°6. Thus, the similarity degrees provided by retrieval by FCM are higher than those of retrieval by k-means. We also observe that both techniques provide give the right result of treatment for all the target cases. This confirms that the integration of FCM clustering in the retrieval step has succeeded in improving the predictive accuracy. Table VI confirms these results whereas we observed in Fig 3, FCM retrieval offers good accuracy (93.33), sensitivity (92.59), and specificity (100). Even when comparing these results with other approaches, the proposed approach obtains the highest accuracy (93.33) compared to other techniques.

TABLE VI. Performance Evaluation

Accuracy	93.33%
Sensitivity	92.59%
Specificity	100%

In conclusion, the encouraging results obtained show that the integration of FCM clustering in the CBR cycle, precisely in the retrieval stage, allows the achieving of better performances and accelerates the similarity computation by reducing the search space, which leads to accelerated search time, improves the retrieval stage and consequently the CBR system, solves the problems of using the classical CBR system in the medical domain such as the large volume of generated medical data and the complexity and uncertainty of these data, and finally leads to a better medical decision making.

V. Conclusion

In this paper, we primarily offer a new approach that integrated the Clustering techniques (K-means and FCM) in the CBR cycle. This integration aims to enhance the retrieval step and consequently the CBR system in order to resolve the problems with applying the classical CBR system in the medical field. The proposed approach has been applied to the immunotherapy dataset in order to predict the response of patients with warts to the immunotherapy.

Experiments have demonstrated that the strategy based on CBR and fuzzy clustering (FCM) was successful in improving the performance of retrieval step such as the accuracy, case retrieval precision, and calculation time. It was discovered that this approach may greatly and effectively minimize the number of cases (research space), solve the problem of the complexity, and the uncertainty of medical information, speed up the similarity calculus, and increase the effectiveness of the search, allowing us to achieve our goal and resolve the problem of classical CBR and improve it.

This approach uses the standard version of k-means and FCM and adopts the Euclidean distance to generate clusters. However, there are several improved versions of these techniques such as [26], [1], [28], [29],... and they have achieved successful results. As a future work we aim to improve our approach by improving the similarity measures (distances) used in the clustering.

References

[1] A. Seal, E. Herrera Viedma, et al., "Performance and convergence analysis of modified c-means using jeffreys-divergence for clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, issue Regular Issue, no. 2, pp. 141–149, 2021, doi: https://doi.org/10.9781/ijimai.2021.04.009.

[2] A. Aamodt, E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications*, vol. 7, no. 1, pp. 39–59, 1994.

[3] M. Pantic, "Introduction to machine learning & case- based reasoning," *London: Imperial College*, 2005.

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

[5] S. Montani, "Exploring new roles for case-based reasoning in heterogeneous ai systems for medical decision support," *Applied Intelligence*, vol. 28, no. 3, pp. 275–285, 2008.

[6] V. E. Ekong, U. G. Inyang, E. A. Onibere, "Intelligent decision support system for depression diagnosis based on neuro-fuzzy-cbr hybrid," *Modern Applied Science*, vol. 6, no. 7, p. 79, 2012.

[7] A. Mansoul, B. Atmani, "Clustering to enhance case- based reasoning," in *Modelling and Implementation of Complex Systems*, Springer, 2016, pp. 137–151.

[8] R. Schmidt, L. Gierl, "Prognostic model for early warning of threatening influenza waves," in *1st German workshop on experience management: sharing experiences about the sharing of experience*, 2002, Gesellschaft für Informatik eV.

[9] I. Bichindaritz, C. Marling, "Case-based reasoning in the health sciences: Foundations and research directions," in *Computational Intelligence in Healthcare 4*, Springer, 2010, pp. 127–157.

[10] M. Abdelhak, A. Baghdad, "Combining multi-criteria analysis with cbr for medical decision support," *Journal of Information Processing Systems*, vol. 13, no. 6, pp. 1496– 1515, 2017.

[11] F. Saadi, B. Atmani, F. Henni, "Integration of fuzzy clustering into the case base reasoning for the prediction of response to immunotherapy treatment," in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 2019, pp. 192–206, Springer.

[12] M. Benamina, B. Atmani, S. Benbelkacem, "Diabetes diagnosis by case-based reasoning and fuzzy logic," *IJIMAI*, vol. 5, no. 3, pp. 72–80, 2018.

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

[14] S. Demigha, "A generic elearning tool for radiologists and hospital practitioners with cbr," in *European Conference on e-Learning*, 2015, p. 809, Academic Conferences International Limited.

[15] D. Gu, C. Liang, H. Zhao, "A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis," *Artificial intelligence in medicine*, vol. 77, pp. 31–47, 2017.

[16] H. Benfriha, B. Atmani, B. Khemliche, N. T. Aoul, Douah, "A multi-labels text categorization framework for cerebral lesion's identification," in *International Conference on Computing*, 2019, pp. 103– 114, Springer.

[17] S. El-Sappagh, M. M. Elmogy, "Medical case based reasoning frameworks: Current developments and future directions," *Virtual and Mobile Healthcare: Breakthroughs in Research and Practice*, pp. 516–552, 2020.

[18] C. Aflori, M. Craus, "Grid implementation of the apriori algorithm," *Advances in engineering software*, vol. 38, no. 5, pp. 295–300, 2007.

[19] K. Srinivas, G. R. Rao, A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *2010 5th International Conference on Computer Science & Education*, 2010, pp. 1344–1349, IEEE.

[20] A. Dewan, M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 704–706, IEEE.

[21] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[22] S. Kumar, G. Sahoo, "Classification of heart disease using naive bayes and genetic algorithm," in *Computational Intelligence in Data Mining-Volume 2*, Springer, 2015, pp. 269–282.

[23] V. Chaurasia, S. Pal, "Early prediction of heart diseases using data mining techniques," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208–217, 2013.

[24] P. R. Hachesu, M. Ahmadi, S. Alizadeh, F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare informatics research*, vol. 19, no. 2, pp. 121–129, 2013.

[25] M. Martín Merino, A. J. López Rivero, V. Alonso, M. Vallejo, A. Ferreras, "A clustering algorithm based on an ensemble of dissimilarities: An application in the bioinformatics domain.," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 6, 2022.

[26] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, "Fuzzy k-means using non-linear s-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

[27] S. Kapil, M. Chawla, "Performance evaluation of k-means clustering algorithm with various distance metrics," in *2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES)*, 2016, pp. 1–4, IEEE.

[28] K. K. Sharma, A. Seal, "Clustering analysis using an adaptive fused distance," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103928, 2020.

[29] A. Seal, A. Karlekar, O. Krejcar, C. Gonzalo-Martin, "Fuzzy c-means clustering using jeffreys-divergence based similarity measure," *Applied Soft Computing*, vol. 88, p. 106016, 2020.

[30] S. Begum, M. U. Ahmed, S. Barua, "Multi-scale entropy analysis and case-based reasoning to classify physiological sensor signals," *Luc Lamontagne and Juan A. Recio-Garcıa (Editors)*, vol. 129, 2012.

[31] S. Benbelkacem, B. Atmani, M. Benamina, "Treatment tuberculosis retrieval using decision tree," in *2013 international conference on control, decision and information technologies (CoDIT)*, 2013, pp. 283–288, IEEE.

[32] X. Blanco, S. Rodríguez, J. M. Corchado, C. Zato, "Case-based reasoning applied to medical diagnosis and treatment," in *distributed computing and artificial intelligence*, Springer, 2013, pp. 137–146.

[33] H. Benfriha, B. Atmani, F. Barigou, F. Henni, B. Khemliche, S. Fatima, A. Douah, Z. Z. Addou, "Improving cbr retrieval process through multilabel text categorization for health care of childhood traumatic brain injuries in road accident," in *Proceedings of Sixth International Congress on Information and Communication Technology, 2022, pp. 721–731,* Springer.

[34] D. Gu, W. Zhao, Y. Xie, X. Wang, K. Su, O. V. Zolotarev, "A personalized medical decision support system based on explainable machine learning algorithms and ecc features: Data from the real world," *Diagnostics*, vol. 11, no. 9, p. 1677, 2021.

[35] Y.-G. Jung, B. Kim, H. Nam, M. Rhee, J.-S. Lee, "Effective diagnosis of coronary artery disease using case-based reasoning," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 449–457, 2021.

[36] F. Saadi, B. Atmani, F. Henni, "Integration of datamining techniques into the cbr cycle to predict the result of immunotherapy treatment," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–5, IEEE.

[37] G. Khussainova, S. Petrovic, R. Jagannathan, "Retrieval with clustering in a case-based reasoning system for radiotherapy treatment planning," in *Journal of Physics: Conference Series*, vol. 616, 2015, p. 012013, IOP Publishing.

[38] C. Koo, W. Li, S. H. Cha, S. Zhang, "A novel estimation approach for the solar radiation potential with its complex spatial pattern via machine-learning techniques," *Renewable energy*, vol. 133, pp. 575–592, 2019.

[39] P. Yadav, "Case retrieval algorithm using similarity measure and adaptive fractional brain storm optimization for health informaticians," *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 829–840, 2016.

[40] S. Geetha, S. Narayanamoorthy, T. Manirathinam, D. Kang, "Fuzzy case-based reasoning approach for finding covid-19 patients priority in hospitals at source shortage period," *Expert Systems with Applications*, vol. 178, p. 114997, 2021.

[41] H. D. Ibrahim, T. O. Odedele, "Covid19 infectious disease detection and diagnosis system using case- based reasoning and fuzzy logic inference model," in *International Conference on Intelligent and Fuzzy Systems*, 2021, pp. 162–170, Springer.

[42] N. Choudhury, S. A. Begum, "Neuro-fuzzy-rough classification for improving efficiency and performance in case-based reasoning retrieval," in *Computational Network Application Tools for Performance Management*, Springer, 2020, pp. 29–38.

[43] Z. Yamin, Z. Mengmeng, G. Xiaomin, Z. Zhiwei, Z. Jianhua, "Research on matching method for case retrieval process in cbr based on fcm," *Procedia engineering, vol. 174, pp. 267–274, 2017.*

[44] S. Banerjee, A. R. Chowdhury, "Case based reasoning in the detection of retinal abnormalities using decision trees," *Procedia Computer Science*, vol. 46, pp. 402–408, 2015.

[45] S. R. Garner, et al., "Weka: The waikato environment for knowledge analysis," in Proceedings of the New Zealand computer science research students conference, vol. 1995, 1995, pp. 57–64.

[46] S. Ray, R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," in Proceedings of the 4th international conference on advances in pattern recognition and digital techniques, 1999, pp. 137–143, Citeseer.

[47] T. M. Kodinariya, P. R. Makwana, "Review on determining number of cluster in k-means clustering," International Journal, vol. 1, no. 6, pp. 90–95, 2013.

[48] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," Journal of Cybernetics, vol. 3, no. 3, pp. 32–57, 1973, doi: 10.1080/01969727308546046.

[49] W. Zhu, N. Zeng, N. Wang, et al., "Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations," NESUG proceedings: health care and life sciences, Baltimore, Maryland, vol. 19, p. 67, 2010.

[50] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, S. Nahavandi, "An expert system for selecting wart treatment method," Computers in biology and medicine, vol. 81, pp. 167– 175, 2017.

Fatima Saadi

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

Baghdad Atmani

He is currently a Full Professor in Computer Science. His field of interest 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.

Fouad Henni

He is a teaching researcher in computer science at Mostaganem University, Algeria. His research interests are in the areas of semantic Web services, artificial intelligence, case-based reasoning, and deep learning, with a particular emphasis on applications in medical diagnosis. He is a member of the Computer Science and new Technologies Lab (CSTL).