Arab J Sci Eng (2016) 41:829–840
DOI 10.1007/s13369-015-1928-y

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Case Retrieval Algorithm Using Similarity Measure and Adaptive Fractional Brain Storm Optimization for Health Informaticians

Poonam Yadav[1]

Received: 8 May 2015 / Accepted: 13 October 2015 / Published online: 26 October 2015

Abstract Managing and utilizing health information is recently a challenging task for health informaticians to provide the highest quality healthcare delivery. Here, storage, retrieval, and interpretation of healthcare information are important phases in health informatics. Accordingly, the retrieval of similar cases based on the current patient data can help doctors to identify the similar kind of patients and their methods of treatments. By taking into consideration this as an objective of the work, a hybrid model is developed for retrieval of similar cases through the use of case-based reasoning. Here, a new measure called parametric-enabled similarity measure is proposed and a new optimization algorithm called adaptive fractional brain storm optimization by modifying the well-known brain storm optimization algorithm with inclusion of fractional calculus is proposed. For experimentation, six different patient datasets from UCI machine learning repository are used and the performance is compared with existing method using accuracy and F-measure. The average accuracy and F-measure reached by the proposed method with six different datasets are 89.6 and 88.8 %, respectively.

Keywords Case-based reasoning · Case retrieval · Optimization · Similarity · Fractional calculus

1 Introduction

Health information technology (HIT) has been used by medical practitioners as well as medical students for several years. HIT, which was initially used as a tool for accessing data and to perform information retrieval rarely, has now turned out to be a ubiquitous and helpful tool in health care and supports medical diagnosis as well as treatment in a number of ways. The quick progression in the utilization of HIT and the advancements in the science behind biomedical and health informatics have caused the physicians and the medical students to utilize the HIT tools rather than spending much time on medical education. Though the features of biomedical informatics were not included in the syllabus of few medical schools, attention is paid towards the training on how the basic tasks like accessing knowledge sources can be performed or to understand facts like why the electronic health record is used [1]. Once the electronic health records are used for storing the medical information, the patient information can be retrieved to plan the treatment or to predict a similar patient's behaviour.

Patient information retrieval [2–7] is an extension of document retrieval or image retrieval. But, it completely relies on the features possessed by the patients and hence, it should tackle the issues related to the gap that exists between the similarity measures used and the high-level semantics that a user is trying to locate. Though plenty of medical information retrieval systems have been put forth, it is not widely used in real-world medical applications. The performance of the information retrieval systems in health information technology can be optimized, if both the visual and textual retrieval [8] in conjunction with similar patient information retrieval are employed. It is not an easy task to find the appropriate electronic health record always because the resulting articles may be dedicated largely to a class of diseases, medical practices, or organs and hence cannot be applied directly to a certain medical issue [9].

The searching for medical records poses severe challenges [10–12]. If keyword-based approaches are utilized

✉ Poonam Yadav
poonam.y2002@gmail.com

1 D.A.V College of Engineering and Technology, Kanina, Haryana, India

for searching the medical information, medical entities that are termed with different names could not be accessed. For instance, 'heart attack' and 'myocardial disorder' are similar in meaning, but the keywords do not match. To overcome these kinds of issues arising from keyword-based approaches, concept-based retrieval approaches have been proposed [13]. Accordingly, semantic measure and optimization-based neural network are combined for retrieval of patients' cases in this paper. The input for the proposed system is patient information stored as health records, which is directly given to PESM measure along with query posted by doctor. The query is matched with stored health records to obtain the similar cases of patients. Similarly, AFBSO neural network is trained with history data as neighbour patients as output neurons. For the input query, neural network can predict its neighbour patient through its algorithmic procedure. Finally, hybrid measure combines these two results and produces the more suitable information to the doctor who can diagnose even better by analysing the similar report done for those patients.

The paper is organized as follows: Sect. 2 presents the motivating scenario and discusses the contributions of the paper. Section 3 presents the proposed retrieval method for case-based reasoning using similarity measure and fractional brain storm optimization. Section 4 discusses the experimental results, and conclusion is given in Sect. 5.

2 Motivating Scenario

The primary principle of case-based reasoning (CBR) [14–19] is that experience in the way of past cases can be leveraged to answer new challenges. A human being's practice is called a case, and its collection is stored in a case base. Naturally, every case is depicted by a problem description and the equivalent solution description. Among the four classical phases in CBR (i.e., retrieval, reuse, revise, and retain), retrieval is a important phase in CBR, since the success of CBR systems is greatly reliant on the performance of retrieval, which have the objective of retrieving useful or relevant cases that can be effectively utilized to solve a target problem. If the retrieved cases are not helpful, CBR systems may not ultimately provide an appropriate solution to the problem.

The retrieval of similar cases based on the current patient data can help doctors to identify the similar kind of patients and their treatments done along with sensitive information. Also, the history of old patients and their current health information may help doctors to decide the medical prescription. These are the main motivations behind developing the proposed method. The overall architecture of the proposed method is given in Fig. 1. Here, patient information is stored in the patient case database and doctor utilizes medical diagnosis support system to extract the similar cases to analyse their information.

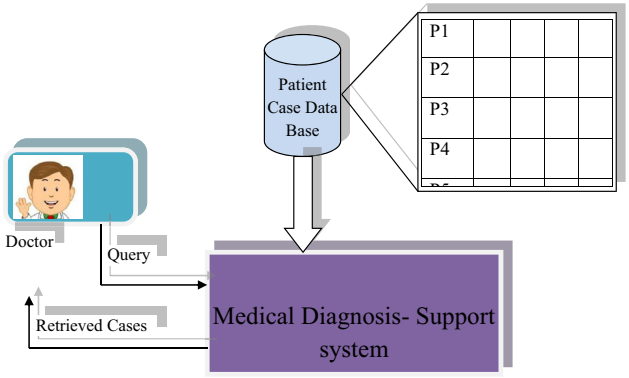

Fig. 1 General architecture

Let C be the repository of patient information where each case is represented as $C_i; 1 \leq i \leq n$. Here, n represents the number of cases stored in the database. Here, each case is represented as an attribute–value pair like the representation used in case-based reasoning. The attribute–value pair for the CSR is expressed as follows:

$$C_i : \{A_j, a_j\}; 0 \leq j \leq m \tag{1}$$

The objective is to retrieve k similar cases by finding the most similar ones to the input query Q_c from 'n' cases stored in C. In order to accomplish this task, three main contributions are given in this paper.

The main contributions of the paper are as follows:

- A new measure called PESM is proposed to match two patient cases using four different parameters with the assumption of occurrence and non-occurrence probability. Then, this measure is applied for case retrieval case.
- A new optimization algorithm called AFBSO by modifying the well-known algorithm BSO with the inclusion of fractional calculus is used. Along with, AFBSO algorithm is applied for training of neural network which is then utilized for retrieval of neighbour cases for the query.
- Hybrid model is proposed newly by integrating PESM measure and AFBSO neural network for effectively retrieving patient case to easily identify the similar cases for the input query.

3 Retrieval Methods for Case-Based Reasoning Using Similarity Measure and Adaptive Fractional Brain Storm Optimization

This section presents the proposed method for case-based reasoning using new similarity measure and AFBSO algorithm. Figure 2 shows the block diagram of the proposed

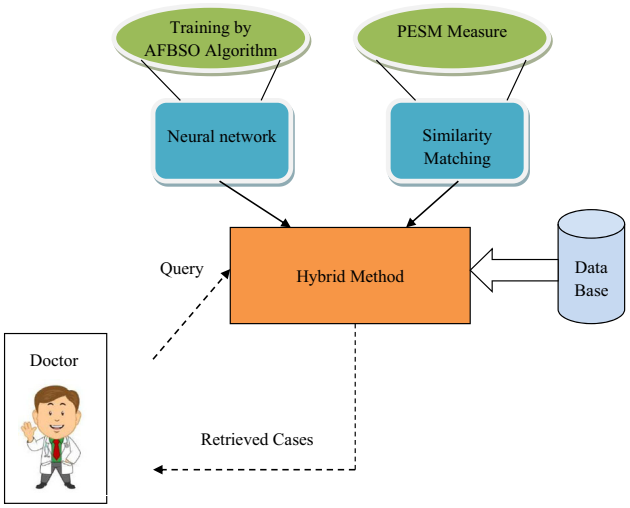

Fig. 2 Block diagram of the proposed retrieval method

retrieval method which has two major components: (i) PESM measure and (ii) designing of AFBSO algorithm. At first, input database of patient cases is given as input to the proposed method. Then, two retrieval methods such as PESM measure and AFBSO-based neural network are applied individually to get two sets of cases after querying with query case. The outputs are effectively combined to obtain the final retrieval of cases to easily identify the existing diagnosis of similar patients for a doctor.

3.1 PESM for Case Retrieval

The proposed similarity measure called PESM measure is used to retrieve the patient cases. Let us assume that parameters of PESM measure x and y are initialized as $x = 0$, $z = 1$. C_i and Cj are two cases from the data repository and are represented as

$$C_i = \left[C_i^{(1)} C_i^{(2)} \ldots C_i^{(m)}\right] \tag{2}$$

$$Cj = \left[C_j^{(1)} C_j^{(2)} \ldots C_j^{(m)}\right] \tag{3}$$

Based on two cases, four parameters $P1\,(q)$, $P2\,(q)$, $P3\,(q)$, and $P4\,(q)$ are computed using the mutual occurrence of attributes in both cases. The occurrence of values in the attributes decides the similarity level of the two cases. The parameter $P1\,(q)$ provides maximum value if any one of the concerned attributes' values of both cases is equivalent to the initial assignment of x. The second parameter $P2\,(q)$ provides the maximum value if the concerned attributes' values of both cases are equivalent to x. The third parameter $P3\,(q)$ can have a higher degree if the values of concerned attributes are not equivalent to x. The final parameter $P4\,(q)$ provides the maximum similarity if the values of concerned attributes should equal but not to be x.

$$P1\,(q) = \begin{pmatrix} z; & \text{if } C_i^{(q)} = x \;||\; C_j^{(q)} = x \\ x; & \text{else} \end{pmatrix}; 0 \le q \le m-1 \tag{4}$$

$$P2\,(q) = \begin{pmatrix} z; & \text{if } C_i^{(q)} = C_j^{(q)} = x \\ x; & \text{else} \end{pmatrix}; 0 \le q \le m-1 \tag{5}$$

$$P3\,(q) = \begin{pmatrix} z; & \text{if } C_i^{(q)} \ne C_j^{(q)} \ne x \\ x; & \text{else} \end{pmatrix}; 0 \le q \le m-1 \tag{6}$$

$$P4\,(q) = \begin{pmatrix} z; & \text{if } C_i^{(q)} = C_j^{(q)} \ne x \\ x; & \text{else} \end{pmatrix}; 0 \le q \le m-1 \tag{7}$$

The parameters computed for every values of attributes are then combined based on the following set of equation.

$$S1 = \sum_{q=1}^{m} P1\,(q) \tag{8}$$

$$S2 = \sum_{q=1}^{m} P2\,(q) \tag{9}$$

$$S3 = \sum_{q=1}^{m} P3\,(q) \tag{10}$$

$$S4 = \sum_{q=1}^{m} P4\,(q) \tag{11}$$

Based on the above values, the similarity degree called PESM is defined as follows. Here, m is the number of attributes given in the case repository.

$$PESM\left(C_i, C_j\right) = \frac{1}{m}\left[\frac{z^* S_1}{4} + \frac{z^* S_2}{2} + x^* S_3 + z^* S_4\right] \tag{12}$$

For the retrieval of 'k' cases to the input case of Q_c, Q_c is matched with all the cases in the repertory using the PESM measure and then cases are sorted in descending order. Finally, top 'k' cases are selected to study the similar diagnosis.

3.2 AFBSO for Neural Network Training

3.2.1 Adaptive Fractional Brain Storm Optimization

Brain storm optimization is modified with mathematical theory called fractional calculus (FC) [20] to improve the solution searching in the predefined search space. In AFBSO algorithm, ideas are represented as a solution which is updated every iteration. The advantage of the proposed

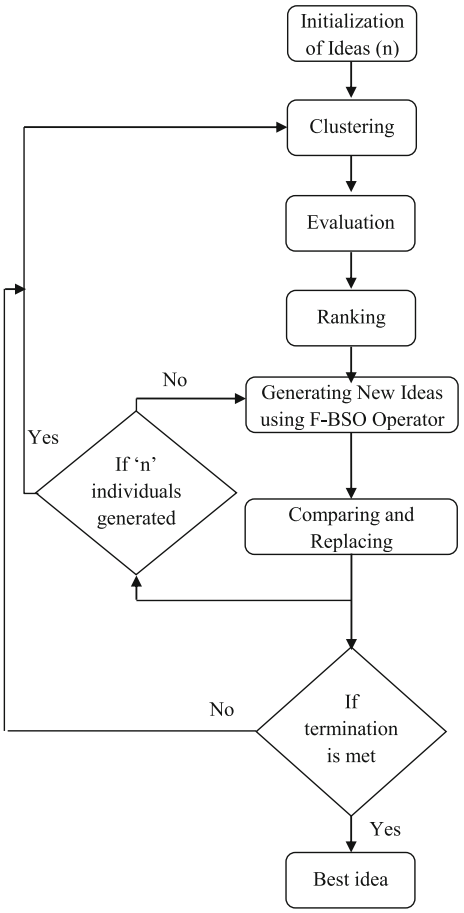

Fig. 3 Flow chart of the fractional brain storm optimization

AFBSO algorithm is the better utilization of global information and further improvement in the evolutionary diversity. The flow chart of the AFBSO algorithm is shown in Fig. 3.

Initialization: Let us assume that n ideas are randomly initialized within the search space as $I_i = [I_{i1}, I_{i2,} \ldots, I_{iD}]$, where $i = 1,2,\ldots n$ and D is the dimension of the solution which signifies the variable taken for optimization.

Grouping: After initialization, ideas are grouped into two sets of ideas based on k-means clustering algorithm, where k is the number of clusters required.

Evaluation: Every idea is then evaluated with fitness function.

Selection: Three different probability values are utilized to select the clusters, selection of one or two clusters or another idea selection as per the probability values, like P_{5a} P_{6b} P_{6b3}.

Updation: The selected idea based on the above selection method is then updated with the following equation. The equation utilized in BSO algorithm is as follows,

$$I_{t+1} = I_t + \xi N(\mu, \sigma) \tag{13}$$

The above equation can be written as,

$$I_{t+1} - I_t = \xi N(\mu, \sigma) \tag{14}$$

The left side $I_{t+1} - I_t$ is the discrete version of the derivative of order $\alpha = 1$, leading to the following expression,

$$D^{\alpha}[I_{t+1}] = \xi N(\mu, \sigma) \tag{15}$$

The order of the velocity derivative can be generalized to a real number $0 \leq \alpha \leq 1$, if the FC perspective is considered, leading to a smoother variation and a longer memory effect. Here, α is adaptively changed using the following formula.

$$\alpha = \frac{t_{ct} - t_{\min}}{t_{\max} - t_{\min}} \tag{16}$$

Therefore, the above equation can be written by considering the first $r = 4$ terms of differential derivative.

$$I_{t+1} = \alpha I_t + \frac{1}{2}\alpha\ I_{t-1} + \frac{1}{6}\alpha(1-\alpha) I_{t-2} + \frac{1}{24}\alpha(1-\alpha)(2-\alpha) I_{t-3} + \xi N(\mu, \sigma) \tag{17}$$

where I_t is idea selected from the last iteration and I_{t+1} is to be newly generated idea.

$$I_t = \begin{cases} I_i j & ; \text{one cluster} \\ w_1\ I_{i1,}\ j + w_2 I_{i2,j} & ; \text{two cluster} \end{cases} \tag{18}$$

where $N(\mu, \sigma)$ is the Gaussian random value with mean μ and variance σ and w_1, w_2 are weight values of the two ideas. ξ is an adjusting factor slowing the convergence speed down as the evolution goes.

$$\xi = r \log \text{sig}\left(\frac{Nc_{\max}/2_{-Nc}}{K}\right) \tag{19}$$

where r is a random value between 0 and 1. $Nc_{\max}$ and Nc denote the maximum number of iteration and current number of iteration, respectively. K adjusts the slope of the logsig function.

Crossover: After the new idea is created, a crossover between the new one and the old one is conducted. The two ideas generated by crossover, together with the old one and the created one, are evaluated using the fitness function, and the old one is replaced with the best of the four.

Termination: The process above repeats until m ideas are updated. Thus, one generation is finished. The iteration goes until terminal requirement is met. Then, the best idea is output as the optimal solution to the problem.

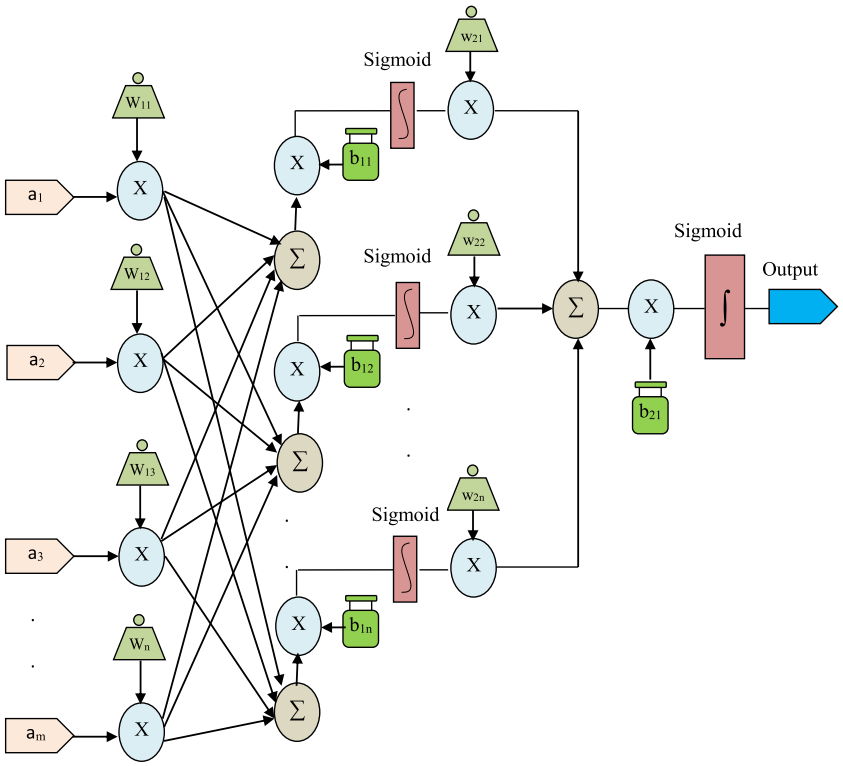

Fig. 4 Neural network architecture

3.2.2 *Feed-Forward Neural Network*

A feed-forward neural network (FFNN) [21,22] is a biologically inspired learning algorithm which consists of a (probably large) number of simple neuron-like processing units, arranged in layers. Each unit in a layer is connected with all the units in the earlier layer.

These connections are not all equal; every connection obtains various strengths or weights. The weights on these connections carry the knowledge of a network. Frequently, the units in a neural network are also called nodes. In NN architecture, input attributes come into input nodes and processes through the network, layer by layer, until it comes at the outputs. The processing of operation contains multiplication and addition as per the architecture shown in Fig. 4. Here, two weights are utilized: neuron weight and bias weight. The weights are processed as per the mathematical operation defined in Fig. 4, and the final output can be obtained from the output neuron.

Assume that input attributes are entered into input layer and it is multiplied with weights of input layer as follows:

$$L_1^{(j)} = a_j{}^* w_{1j}; \quad 0 \le j \le m-1 \tag{20}$$

where m is number of input attributes or number of input neurons. The output of the input layer is the given for summation process and bias weight is multiplied with it.

$$H_1^{(l)} = \left(\sum_{j=1}^{m} L_1^{(j)} \right) {}^* b_{1l}; \quad 0 \le l \le h \tag{21}$$

where h is number of hidden neurons in the hidden layer 1. Then, the output of hidden layer 1 is given to sigmoid function to regularize the data space.

$$O_{H1}^{(l)} = \frac{1}{1 + e^{-H_1(l)}} \tag{22}$$

This process is repeated for all the hidden layers until it reaches the final output layer. h1 is the number of neurons in hidden layer 2.

$$L_2^{(j)} = O_{H1}^{(l)} {}^* w_{2j} \ ; \ 0 \le j \le h1 \tag{23}$$

$$H_2^{(l)} = \left[\sum_{j=1}^{h1} L_2^{(j)} \right] \ ; \ 0 \le j \le h1 \tag{24}$$

$$O_{H2}^{(l)} = \frac{1}{1 + e^{-H_2(l)}} \tag{25}$$

The final output for the neural network given in Fig. 4 is obtained based on the following equation. Here, N is the number of hidden layers.

$$O_d = O_{H_N}^{(l)} \tag{26}$$

3.2.3 Neural Network Training by AFBSO Algorithm

Neural network training is the process of identifying the weights for neurons, suitable for the input cases taken as an input. Training is an iterative process of identifying weights by changing weights in every iteration. Here, AFBSO algorithm is adapted to do neural network training by finding the optimal weights as output fixed as [23].

Idea encoding: In AFBSO algorithm, idea is encoded as shown in Table 1. The weights utilized in the neural network structure are put as vector which is an idea of the proposed FBO algorithm, so the size of an idea is equivalent to the number of weights in the neural network including neuron weights and bias weights.
Algorithmic procedure: At first, input weights of n vectors (ideas) are randomly initialized within the search space and k-means clustering algorithm is applied. The ideas are evaluated with fitness function. Based on three probabilities, ideas are selected and updated using the proposed updating equation and crossovers are applied. The two ideas generated by crossover, together with the old one and the created one, are evaluated using the fitness function, and the old one is replaced with the best of the four. This process is repeated until m ideas are updated. Thus, one generation is finished. The iteration goes until terminal requirement is met.
Fitness function: The fitness function is computed by giving training cases to neural network architecture, and the weights in taken idea for evaluation are filled in the neural network. The output for the taken idea is computed as per the mathematical model given in the neural network architecture. So, the output can be obtained for all the input cases which is then given for the fitness function.

$$fitness = \frac{1}{c}\left[\sum_{i=1}^{c} O_d - O_g\right] \tag{27}$$

where, O_d is output obtained from neural network, O_g is original neighbour index value from training cases, and c is number of training cases.

3.3 Hybrid Retrieval Method for Case-Based Reasoning

For a query case Q_c given by a doctor, PESM measure provides the similarity measure for every case stored in the training cases and neural network provides the neighbour index values by predicting it. The score value is then combined using the following equation, and the final score value is generated for all the cases. The top-k cases are then extracted to be given for doctor to analyse the medical reports and prescriptions.

$$H_{Q_c} = \frac{1}{2}\left[Q_c^{(\mathrm{PESM})} + Q_c^{(\mathrm{NN})}\right] \tag{28}$$

4 Results and Discussion

The experimental results of the proposed method are discussed in this section, and the performance of method is also discussed in detail with two different metrics.

4.1 Experimental Set-up

Platform: The proposed retrieval method is implemented using MATLAB 8.2.0.701 (R2013b) with a system configuration of 2GB RAM Intel processor and 32-bit OS. *Datasets utilized:* The datasets are taken from UCI machine learning repository [24], and the description of those datasets is given in Table 2. *Evaluation metrics:* Accuracy and F-measure are used for performance evaluation. ***Accuracy*** can be measured via computing the proportion of correctly classified instances overall the tested instances. That is, accuracy can be misleading when the testing data contain a disproportional number of cases with a certain solution class. Thus, we also use F-measure (FM) to overcome this problem. *Precision* is the

Table 1 Idea encoding for neural network training

w_{11}	w_{12}	…	w_{1n}	b_{11}	b_{12}	…	b_{1n}	w_{21}	w_{22}	…	w_{2n}	b_{21}	…

Table 2 Description of datasets

Dataset	Case number	Attribute number	Numerical data	Discrete data	Number of class
Breast Cancer (BC)	286	9		9	2
Breast Cancer Wins (BCW)	683	9	9		2
Breast Tissue (BT)	106	9	9		6
Pima Indian Diabetes (PID)	768	8	8		2
StatLog Heart Disease (SHD)	270	13	7	6	2
New Thyroid (THY)	215	5	5	12	3

Table 3 Parameter initialization

AFBSO parameters	n	P_{5a}	P_{6b}	P_{6b3}	$Nc_{\max}$	K	μ	σ
	5	0.2	0.8	0.4	2000	20	0	1
Neural network parameters	N		h		h_1		h_2	
	3		5		10		15	
Overall parameters	k							
	5							

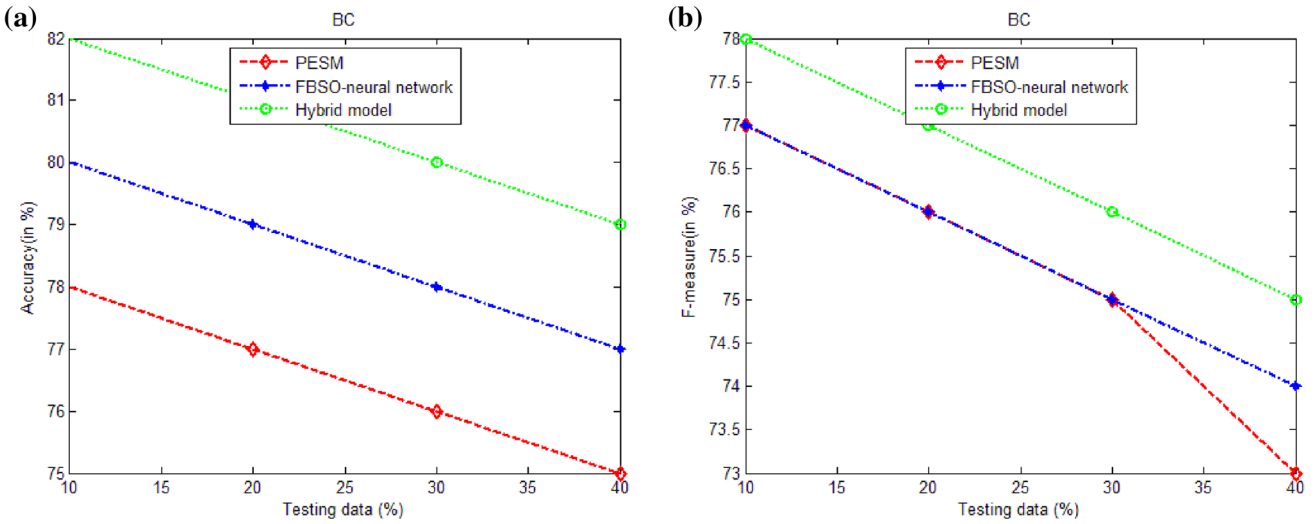

Fig. 5 Performance graph in Breast Cancer (BC)

fraction of retrieved cases that are relevant to the search, ***Recall*** is the fraction of the cases that are relevant to the query that are successfully retrieved, and the ***F-measure*** that combines precision and recall is the harmonic mean of precision and recall.

$$\text{Accuracy} = \frac{\text{Correctly classified insances}}{\text{tested insances}} \tag{29}$$

$$\text{Precision (P)} = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved}} \tag{30}$$

$$\text{Recall (R)} = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Relevant}} \tag{31}$$

$$\text{F-measure} = \frac{2^* P^* R}{P + R} \tag{32}$$

Parameter initialization: For the reason of validating the effectiveness and usefulness of the proposed retrieval method, a set of parameters are set naturally for the process of AFBSO, neural network, and core method as the values given in Table 3.

Existing algorithms: The first two algorithms taken for comparison are USIMSCAR-MV and USIMSCAR-WV [25] which leverage association knowledge (AK) in conjunction with similarity knowledge (SK). AK represents strongly evident, interesting relationships between known problem features and solutions shared by a large number of cases. USIMSCAR1 retrieved a combined set of both cases and rules relevant to a target problem, where the relevance is determined by quantification methods using an integration of SK and AK. The other two algorithms taken for comparison are artificial bee colony (ABC)-based neural network training [26] and cuckoo search-based neural network training [27]. These two algorithms were developed to improve the performance of neural network.

4.2 Performance Analysis

The performance of the proposed hybrid method is analysed with two other variants of the proposed method called PESM and AFBSO neural network. The size of the data is changed and the performance is analysed for six different datasets. Figure 5 shows the performance graph of proposed methods in BC datasets. Here, hybrid method shows the better performance as compared with other two methods. The maximum accuracy of 82 % is reached by the hybrid model when the testing data size is 10 %, and for the same data size, PESM method and AFBSO neural network obtains the value of 80 and 78 %, respectively. Similarly, hybrid model achieved 78 % for the BC datasets, whereas PESM and AFBSO neural network obtained the same value of 77 %. Figure 6a shows the performance graph in terms of accuracy in BCW dataset, and

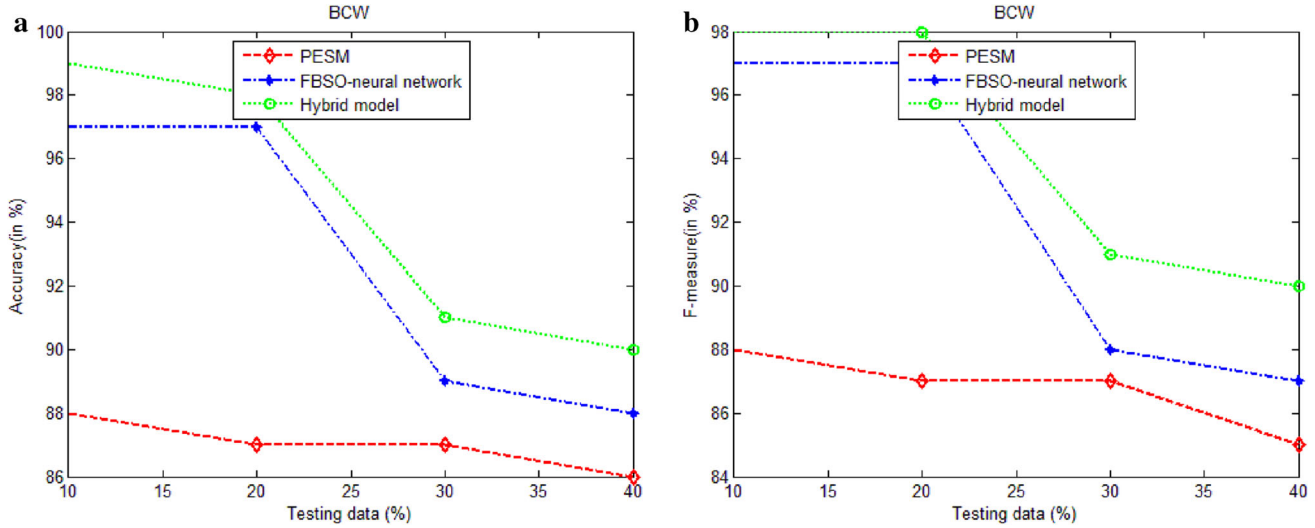

Fig. 6 Performance graph in Breast Cancer Wins (BCW)

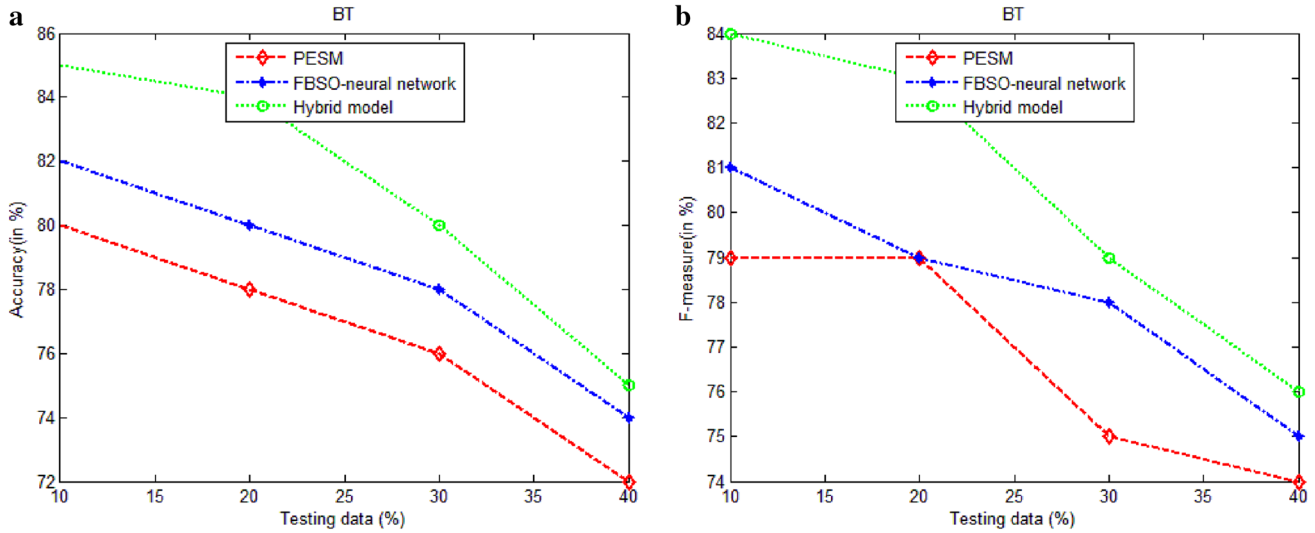

Fig. 7 Performance graph in Breast Tissue (BT)

Fig. 6b shows the performance graph in terms of F-measure in BCW dataset. From Fig. 6a, hybrid model attained about 99 % accuracy by comparing with other two methods for the data size of 10 %. The minimum accuracy for the hybrid model is 90 % as compared with other two methods for the data size of 40 %. Similarly, hybrid model attained about 98 % F-measure by comparing with other two methods for the data size of 10 %. The minimum accuracy for the hybrid model is 97 % as compared with other two methods for the data size of 40 %.

Figure 7a shows the accuracy graph of hybrid model, PESM model, and AFBSO neural network in BT datasets. When the size of the testing data is 10 %, hybrid model achieved 85 %, whereas PESM and AFBSO neural network obtained the value of 82 and 80 %, respectively. The minimum accuracy reached by the hybrid model is 75 %. Figure 7b shows the F-measure graph of hybrid model, PESM model, and AFBSO neural network in BT datasets. When the size of the testing data is 10 %, hybrid model achieved 84 %, whereas PESM and AFBSO neural network obtained the value of 81 and 79 %, respectively. The minimum F-measure reached by the hybrid model is 76 %. From Fig. 8a, it can be seen that hybrid model attained about 90 % accuracy in PID data by comparing with other two methods for the data size of 10 %. The minimum accuracy for the hybrid model is 86 % as compared with other two methods for the data size of 40 %. Figure 8b shows the F-measure graph of hybrid model, PESM model and AFBSO neural network in PID datasets. When the size of the testing data is 10 %, hybrid model achieved 90 %, whereas PESM and AFBSO neural network obtained the value of 88 and 85 %, respectively. The minimum F-measure reached by the hybrid model is 86 %.

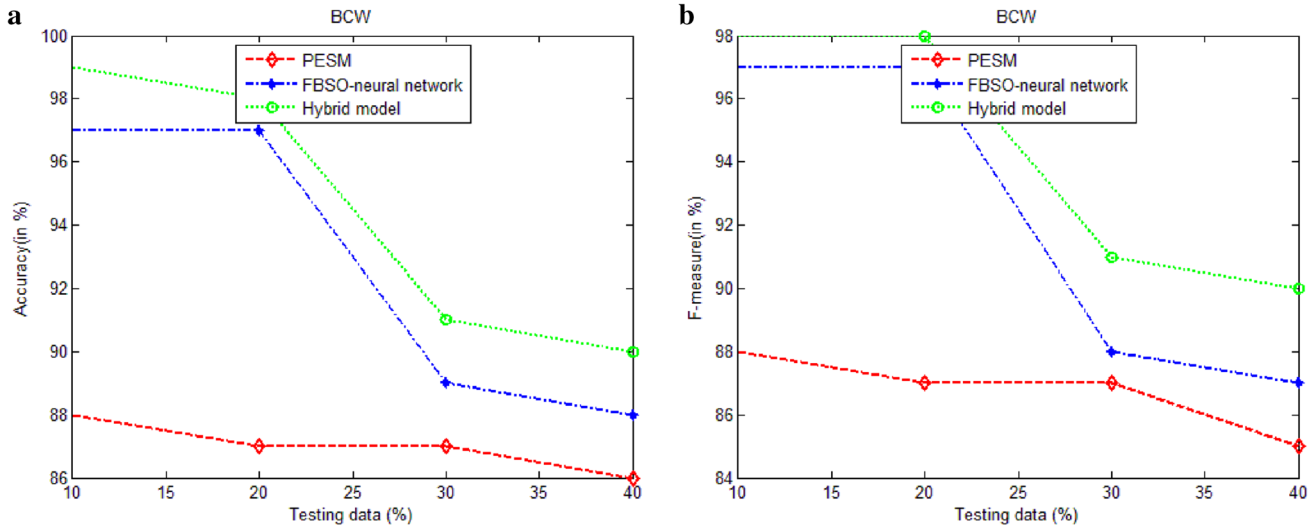

Fig. 8 Performance graph in Pima Indian Diabetes (PID)

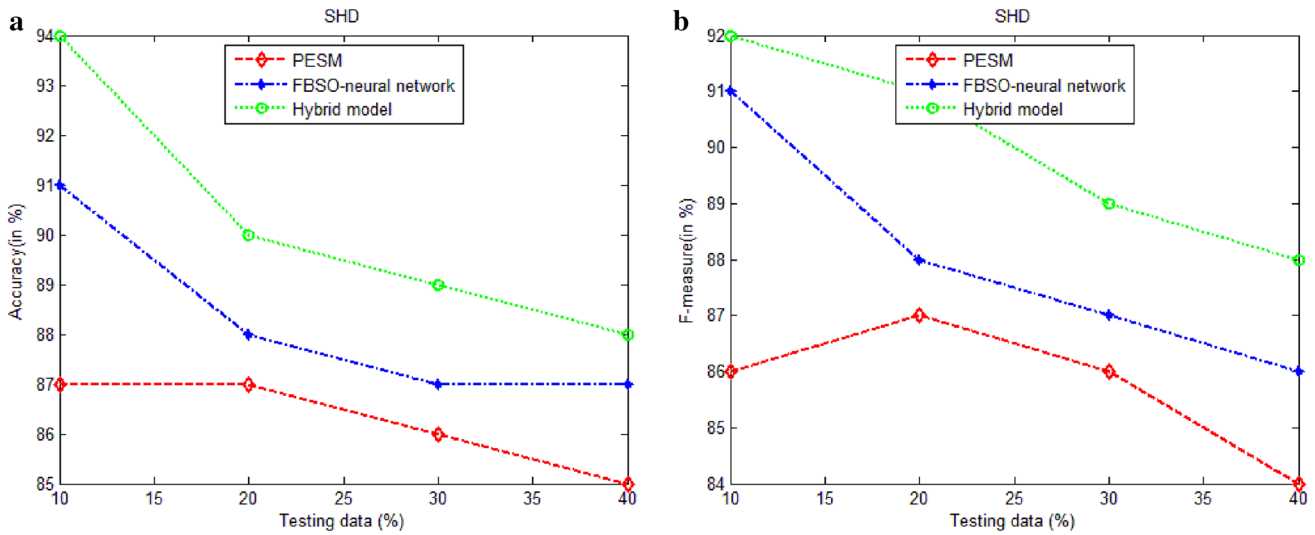

Fig. 9 Performance graph in StatLog Heart Disease (SHD)

Figure 9a shows the performance graph in terms of accuracy in SHD dataset and Fig. 9b shows the performance graph in terms of F-measure in SHD dataset. From Fig. 9a, it can be seen that hybrid model attained about 94 % accuracy by comparing with other two methods for the data size of 10 %. The minimum accuracy for the hybrid model is 88 % as compared with other two methods for the data size of 40 %. Similarly, hybrid model attained about 92 % F-measure by comparing with other two methods for the data size of 10 %. The minimum accuracy for the hybrid model is 88 % as compared with other two methods for the data size of 40 %. Figure 10 shows the performance graph in THY datasets. The maximum accuracy of 98 % is reached by the hybrid model when the testing data size is 10 % and for the same data size, PESM method and AFBSO neural network obtained the value of 95 and 88 %, respectively. Similarly, hybrid model achieved 98 % for the BC datasets, whereas PESM and AFBSO neural network obtained the values of 96 and 92 %, respectively.

When analysing the overall performance of the methods with different datasets, hybrid method shows the top performance in accuracy and F-measure as compared with other two methods: PESM and AFBSO neural network. Also, the size of testing data which is dependent on the size of training data influences the performance of the methods. When the training data size is increasing or testing data size is decreasing, the performance of the three methods is increasing, so we can say that size of the training data is directly proportional to the accuracy and F-measure.

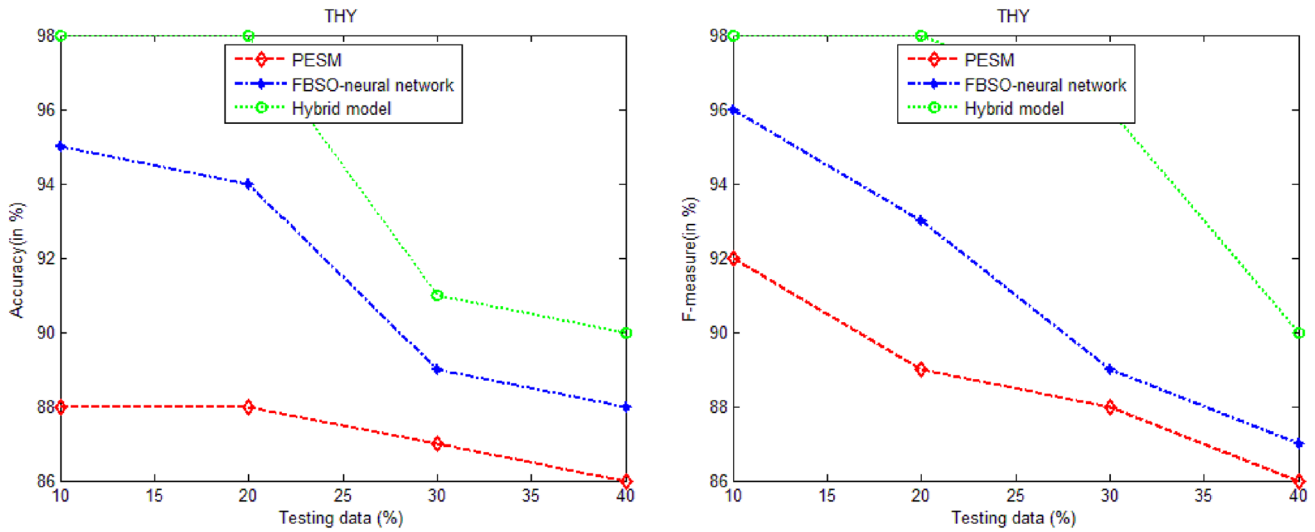

Fig. 10 Performance graph in New Thyroid (THY)

Table 4 Accuracy comparison

Dataset	BC	BCW	BT	PID	SHD	THY
USIMSCAR(MV) [25]	75.87	97.66	71.7	75.78	83.33	97.67
USIMSCAR(WV) [25]	79.02	97.66	78.3	87.5	**89.63**	97.67
ABC-NN [26]	77	96	78	85	89	95
CS-NN [27]	76	95	79	86	88	95
PESM	77	87	78	86	87	88
AFBSO-NN	79	97	80	87	87	94
Hybrid model	**81**	**98**	**84**	**88**	89	**98**

4.3 Comparative Analysis

The comparative analysis of the proposed methods is done with two existing methods given in [25]: artificial bee colony (ABC)-based neural network training [26] and cuckoo search-based neural network training [27] for the same datasets with tenfold cross-validation. Table 4 shows the comparative performance of the taken methods for the six different datasets using accuracy. In BC dataset, hybrid model outperformed by reaching the accuracy of 81 %, but the other two proposed methods are less than the existing method USIMSCAR (WV). Here, ABC-NN and CS-NN obtained the value of 77 and 76 %. Similarly, in BCW dataset, hybrid model outperformed by reaching the accuracy of 98 %, but the other two proposed methods are less than the existing method USIMSCAR (WV). The hybrid model achieved the accuracy of 84 % for the BT dataset as compared with the other proposed method AFBSO-NN. Similarly, 88 % accuracy is reached by the hybrid model, and it shows the top performance in PID datasets. In SHD, USIMSCAR (WV) shows better performance when compared with three proposed methods: 98 % accuracy is reached by the hybrid model in THY datasets, and 95 % accuracy is reached by ABC-NN and CS-NN.

Table 5 shows the comparative performance of the taken methods for the six different datasets using F-measure. F-measure of 77 % is reached by the hybrid model in BC datasets while ABC-NN and CS-NN reached the value of 73 and 74 %. In BCW datasets, hybrid model outperformed by reaching the F-measure of 98 %, but the other two proposed methods are less than the existing method USIMSCAR (WV). F-measure of 83 % is reached by the hybrid model, and it shows the top performance in BT datasets. The hybrid model achieved the F-measure of 88 % for the PID dataset as compared with the other proposed method AFBSO-NN. In SHD, USIMSCAR (WV) shows better performance when compared with three proposed methods, and hybrid model outperformed by reaching the F-measure of 98 % in THY datasets and the ABC-NN and CS-BB obtained the value of 95 and 96 %.

4.4 Statistical Test

Pair-wise statistical tests are conducted to evaluate the algorithmic performance by combining different algorithms. Here, statistical test is conducted on the accuracy values by combining two different algorithms. Table 6 shows the p values of different combinations of algorithms. For the

Table 5 F-measure comparison

Dataset	BC	BCW	BT	PID	SHD	THY
USIMSCAR(MV) [25]	68.74	97.42	71.18	72.21	83.08	96.88
USIMSCAR(WV) [25]	74.25	97.42	77.63	86.14	**89.55**	96.88
ABC-NN [26]	73	90	75	86	88	95
CS-NN [27]	74	89	79	87	87	96
PESM	76	87	79	85	86	89
AFBSO-NN	76	97	79	87	87	93
Hybrid model	**77**	**98**	**83**	**88**	89	**98**

Table 6 Pair-wise statistical test of algorithms on accuracy

	USIMSCAR(WV) [25]	USIMSCAR(WV) [25]	ABC-NN [26]	CS-NN [27]	PESM	AFBSO-NN	Hybrid model
USIMSCAR(MV) [25]	–	0.125	0.685	0.6875	0.6875	0.6875	**0.0313**
USIMSCAR(WV) [25]	0.125	–	0.0313	0.2188	**0.0313**	0.2188	0.2188
ABC-NN [26]	0.685	0.0313	–	1	0.6250	0.6875	0.0625
CS-NN [27]	0.6875	0.2188	1	–	0.3750	0.6875	**0.0313**
PESM	0.6875	0.0313	0.6250	0.3750	–	0.0625	**0.0313**
AFBSO-NN	0.6875	0.2188	0.6875	0.6875	0.0625	–	**0.0313**
Hybrid model	0.0313	0.2188	0.0625	0.0313	0.0313	**0.0313**	–

Table 7 Pair-wise statistical test of algorithms on F-measure

	USIMSCAR(WV) [25]	USIMSCAR(WV) [25]	ABC-NN [26]	CS-NN [27]	PESM	AFBSO-NN	Hybrid model
USIMSCAR(MV) [25]	–	0.1250	0.6875	0.6875	0.6875	0.6875	**0.0313**
USIMSCAR(WV) [25]	0.1250	–	**0.013**	0.6875	0.6875	1	0.2188
ABC-NN [26]	0.6875	0.013	–	0.6875	0.6875	0.6875	**0.0313**
CS-NN [27]	0.6875	0.6875	0.6875	–	0.3750	1	**0.0313**
PESM	0.6875	0.6875	0.6875	0.6875	–	0.1250	**0.0313**
AFBSO-NN	0.6875	1	1	0.6875	0.1250	–	**0.0313**
Hybrid model	0.013	0.2188	0.2188	0.0313	0.01313	**0.0313**	–

hypothesis testing, p value should be less than 0.1. From the table, we understand that the hybrid model rejects null hypothesis in most of the combinations by reaching the value of 0.0313. The table again shows that the statistical test almost always returns lower p values for the proposed hybrid model than for other algorithms and more often rejects the null hypothesis. Overall, it is known that the proposed hybrid model is more likely to reject the null hypothesis.

Table 7 shows the pair-wise statistical test of algorithms on F-measure. The algorithms considered for the comparison are analysed with the proposed hybrid model to find the performance deviation of the algorithms using statistical test. From the statistical test, we obtain the p value for all the combinations of algorithms. From Table 6, we note that the proposed hybrid model mostly rejects null hypothesis by reaching the p value 0.0313 as compared with the existing algorithms. The lower values of p suggest that the differences between classifiers are significant.

5 Conclusion and Future Scope

We have presented a case retrieval method using similarity measure and fractional brain storm optimization for health informaticians. In this paper, a new measure called PESM was proposed to match two patient cases using four dif-

ferent parameters with the assumption of occurrence and non-occurrence probability. Then, a new optimization algorithm called AFBSO was modified using the well-known algorithm BSO with the inclusion of fractional calculus. Also, a hybrid model was proposed newly by integrating PESM measure and AFBSO neural network for effectively retrieving patient case to easily identify the similar cases for the input query. At first, input database of patient cases is given as input to PESM method and AFBSO-based neural network. In PESM measure, query case is matched with historic data to identify similar cases. Also, AFBSO algorithm is applied to neural network trainings to identify the neighbour cases for the query patient case. Finally, both outputs are effectively combined to obtain the final retrieval of cases to easily identify the existing diagnosis of similar patients for a doctor. The experimentation is conducted with six different patient datasets from UCI machine learning repository, and the performance is compared with the existing method using accuracy and F-measure. The average accuracy and F-measure obtained by the proposed hybrid method over six different datasets are 89.6 and 88.8 %, respectively. The future work can be in the direction of including semantic and pragmatic concept in developing similarity measure for case retrieval. Also, multi-objective optimizations can be utilized to train the neural network to further improve the performance.

References

1. Hersh, W.R.; Gorman, P.N.; Biagioli, F.E.; Mohan, V.; Gold, J.A.; Mejicano, G.C.: Beyond information retrieval and electronic health record use: competencies in clinical informatics for medical education. Adv. Med. Educ. Pract. **1**(5), 205–212 (2014)
2. Wang, D.; Li, T.; Zhu, S.; Gong, Y.: Ihelp: an intelligent online helpdesk system. IEEE Trans. Syst. Man Cybern. B Cybern. **41**(1), 173–182 (2011)
3. Cassimatis, N.; Bignoli, P.; Bugajska, M.; Dugas, S.; Kurup, U.: An architecture for adaptive algorithmic hybrids. IEEE Trans. Syst. Man Cybern. B Cybern. **40**(3), 903–914 (2010)
4. Llorente, M.S.; Guerrero, S.E.: Increasing retrieval quality in conversational recommenders. IEEE Trans. Knowl. Data Eng. **24**(10), 1876–1888 (2012)
5. Susskind, J.; Blaisdell, J.M.; Iredell, L.; Keita, F.: Improved temperature sounding and quality control methodology using AIRS/AMSU data: the AIRS science team version 5 retrieval algorithm. IEEE Trans. Geosci. Remote Sens. **49**(3), 883–907 (2011)
6. Radosavljevic, V.; Vucetic, S.; Obradovic, Z.: A data-mining technique for aerosol retrieval across multiple accuracy measures. IEEE Geosci. Remote Sens. Lett. **7**(2), 411–415 (2010)
7. Pasolli, L.; Notarnicola, C.; Bruzzone, L.: Multi-objective parameter optimization in support vector regression: general formulation and application to the retrieval of soil moisture from remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(5), 1495–1508 (2012)
8. Muller, H.; Kalpathy-Cramer, J.: Analyzing the content out of context—features and gaps in medical image retrieval. Int. J. Healthc. Inf. Syst. Inf. **4**(1), 88–98 (2009)
9. Depeursinge, A.; Duc, S.; Eggel, I.; Müller, H.: Mobile medical visual information retrieval. IEEE Trans. Inf. Technol. Biomed. **16**(1), 53–61 (2012)
10. Castro, J.L.; Navarro, M.; Sánchez, J.M.; Zurita, J.M.: Introducing attribute risk for retrieval in case-based reasoning. Knowl. Based Syst. **24**(2), 257–268 (2011)
11. Ping, X.-O,; Tseng, Y.-J.; Lin, Y.-P.; Chiu, H.-J.; Lai, F.; Liang, J.-D.; Huang, G.-T.; Yang, P.-M.: A multiple measurements case-based reasoning method for predicting recurrent status of liver cancer patients. Computers in Industry, February (2015)
12. Mourão, A.; Martins, F.; Magalhães, J.: Multimodal medical information retrieval with unsupervised rank fusion. Comput. Med. Imag. Graph. **39**, 35–45 (2015)
13. Zuccon, G.; Koopman, B.; Nguyen, A.; Vickers, E.; Butt, L.: Exploiting Medical Hierarchies for Concept-Based Information Retrieval. ADCS'12, December 05–06 2012, Dunedin, New Zealand
14. Guo, Y.; Hu, J.; Peng, Y.: Research on CBR system based on data mining. Appl. Soft Comput. **11**(8), 5006–5014 (2011)
15. Park, Y.-J.; Choi, E.; Park, S.-H.: Two-step filtering datamining method integrating case-based reasoning and rule induction. Exp. Syst. Appl. **36**(1), 861–871 (2009)
16. Ahn, H.; Kim, K.-J.: Global optimization of case-based reasoning for breast cytology diagnosis. Exp. Syst. Appl. **36**(1), 724–734 (2009)
17. Pandey, B.; Mishra, R.: Case-based reasoning and data mining integrated method for the diagnosis of some neuromuscular disease. Int. J. Med. Eng. Inf. **3**(1), 1–15 (2011)
18. Chuang, C.-L.: Case-based reasoning support for liver disease diagnosis. Artif. Intell. Med. **53**(1), 15–23 (2011)
19. Huang, M.-J.; Chen, M.-Y.; Lee, S.-C.: Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Exp. Syst. Appl. **32**(3), 856–867 (2007)
20. Solteiro Pires, E.J.; Tenreiro Machado, J.A.; Moura Oliveira, P.B.de ; Boaventura Cunha, J.; Mendes, Luís: Particle swarm optimization with fractional-order velocity. Nonlinear Dyn. **61**(1-2), 295–301 (2010)
21. Feng, P.; Jie, C.; Xuyan, T.; Jiwei, F.: Multilayered feed forward neural network based on particle swarm optimizer algorithm. J. Syst. Eng. Electron. **16**(3), 682–686 (2005)
22. El-Melegy, M.T.: Random sampler M-estimator algorithm with sequential probability ratio test for robust function approximation via feed-forward neural networks. IEEE Trans. Neural Netw. Learn. Syst. **24**(7), 1074–1085 (2013)
23. Guezouli, L.; Kadache, A.: Information retrieval model based on neural networks using neighborhood. In: Proceedings of 2012 International Conference on Information Technology and e-Services, pp. 1–5 (2012)
24. Datasets from UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
25. Kang, Y.-B.; Krishnaswamy, S.; Zaslavsky, A.: A retrieval strategy for case-based reasoning using similarity and association knowledge. IEEE Trans. Cybern. **44**(4), 473–487 (2014)
26. Bullinaria, J.A.; AlYahya, K.: Artificial bee colony training of neural networks. In: Proceedings of Nature Inspired Cooperative Strategies for Optimization (NICSO 2013), vol. 512, pp. 191–201 (2014)
27. Valian, E.; Mohanna, S.; Tavakoli, S.: Improved Cuckoo Search Algorithm for Feedforward Neural Network Training. Int. J. Artif. Intell. Appl.**2**(3), July (2011)