Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia

J. Loveline Zeema *, D. Francis Xavier Christopher

School of Computer Studies, Rathnavel Subramaniam College of Arts & Science, Coimbatore, Tamilnadu, India

ARTICLE INFO

Article history:
Received 1 March 2019
Revised 9 August 2019
Accepted 14 September 2019
Available online xxxx

Keywords:
Dyslexia
Neutrosophic C-means
Artificial bacterial foraging
Vagueness
Imprecision
Indeterminacy

ABSTRACT

Precise prediction of risk for dyslexia among children's in earlier stages is a significant long-term aim in the field of cognitive computing. Producing such accurate results for detection of dyslexia from a dataset which consist of low-quality dataset and the presence of vague information is the toughest challenge among researchers. This paper aims at developing an evolving model to handle the impreciseness in the detection of dyslexia more intelligently. In this work, each instance is described in a neutrosophic domain by defining a membership degree of truthiness, indeterminacy, and falsity. These instances are neutrosophically clustered by applying Neutrosophic C-Means clustering (NCM) which forms four different clusters namely dyslexia, no dyslexia, control/revision and hyperactivity or other issues. The outlier and noise are the special categories of indeterminacy which often occurs in real datasets are promptly discovered and clustered. NCM is optimized by introducing Artificial Bacterial Foraging (ABF), especially when there is vagueness or imprecision in the selection of cluster centroids. With the merits of global searching, ABF selects more promising clustering during cluster re-computation. The interpreted results confirm that the role played by proposed ONCMC-ABF algorithm produces better results in the prediction of dyslexia with the low-quality dataset.

1. Introduction

A significant impairment of reading and spelling skills of children's in spite of regular intellectual capabilities and educational prospects which is termed as developmental dyslexia (Shaywitz and Shaywitz, 2005). The prevalence of dyslexia in India is believed to be 15% as mentioned in (Minister for Science & Technology). There is an extensive agreement that dyslexia has a genetic basis (Franke et al., 2010; Galaburda et al., 2006). Although the majority percentage of children whom belongs to such family risk do not develop dyslexia, they still perform more poorly than naturally evolving children on actions like non-word reading, spelling and reading comprehension (Lyytinen et al., 2005). Empowerment in the field of software applications and computer technology has greatly influenced the educational environment. In Recent years, researchers started focusing on machine learning approaches that could play a vital role in motivating interest of dyslexic student's performance in their academics (Figs. 1 and 2).

By considering the merits of earlier detection of dyslexia will help these children in acquiring suitable services. Machine learning and cognitive computing are relevant tools to analyze and classify data to determine dyslexic and non-dyslexic children. But in reality, the dyslexia dataset consists of the imprecise dataset which cannot be handled precisely using conventional machine learning systems and there is a less proof on handling such an uncertain environment.

Hence this research work focuses on handling uncertainty by inferring the knowledge of the degree of truthiness, falsity, and indeterminacy to handle vagueness and indeterminacy in dyslexia dataset by developing an integrated approach for prediction of dyslexia using neutrosophically clustered pattern optimized by behavioral inspiration of bacterial foraging.

* Corresponding author.
E-mail addresses: j.lovelinezeema@gmail.com (J. Loveline Zeema), christopherd@rvsgroup.com (D. Francis Xavier Christopher).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.09.008
1319-1578/

2. Related works

This section discusses few of the previous studies done on dyslexia prediction using various approaches.

Puolakanaho et al. (2007) in their work modeled a logistic regression to predict dyslexia. They used language ability measures along with the status of Family Risk to predict dyslexia. But this type of models is related to correlational, and they are not predictive because they test their model only with specific samples instead of testing with outside the sample. This proves that still, the predictive model of dyslexia is missing highly.

Schnack et al. (Chen et al., 2017) proposed a model which used pattern recognition technique to discover regularities of the given input data and it determines the most interesting information in a bottom-up approach using machine learning which finds discriminative characteristics between various groups. This is especially applied to psychiatric disorders.

Vapnik (1999) in their work used Support vector machine based on a linear kernel which gains the knowledge about the symptom of dyslexia and predict its presence at an individual level to answer whether the child is under family risk or typically developing child using machine learning approach. The prediction is applied to new cases by training linear support vector machine on such high dimensional pattern detection.

Janousova et al. (2016) and Kassraian-Fard et al. (2016) in their work they used many classification models for diagnosis of dyslexia which indicates significant features may contribute more in prediction model using linear discriminative analysis, logistic regression for determining the performance of each model.

Tamboer et al. (2016) examined children under 35 months of age with their vocabulary development to examine the presence of dyslexia. In this work, the support vector machine is used to discriminate among subjects with and without dyslexia.

Margaj and Purohit (2016) in their work discovered dyscalculia among children's studying primary school. By approaching positive psychology in the direction of detection learning and conducting intervention program for the learner is a critical feature which essentially necessitates besieged examination so that the right child should get the right assistance. This work uses three different classifiers to predict the different classes of dyscalculia.

3. Problem proclamation

In real time dyslexia dataset comprised of uncertain values in both output and input attributes. At the same time, each child has a chance to be assigned in more than one label, so that the desired output of the model may be moderately known. The inputs are represented using a range of linguistic values or numbers. While consulting with more than one expert, they do not agree on the same score, thus these situations signify that the dataset is incomplete. One of the main issues of machine learning is modeling uncertainty to solve the prediction of dyslexia in a precise manner. While collecting information about dyslexia children's they include primary uncertainties types like vagueness (nearly dyslexia), imprecision (range of dyslexia), ambiguity (percentage of dyslexia) and inconsistency (when the output of an individual falls more than one label). While using Fuzzy set it handles only vagueness and if intuitionistic fuzzy set is adopted it handles vagueness and imprecision. But if Neutrosophic logic is adopted it has the ability to deal with vagueness, imprecision, ambiguity and inconsistent information in the real-world dyslexia dataset.

In this present work, the presence of the missing value in dyslexia dataset is well handled using boosted decision tree based imputation model which produces complete dataset. This work contributes neutrosophic C means clustering to cluster the instances of dyslexia dataset based on three measures degree of truthiness, falsehood and indeterminacy. Unlike intuitionistic fuzzy they are independent of each other. The efficiency of the clustering approach is enhanced by utilizing bacterial foraging for selection of optimal cluster centroids during the clustering process. Here membership degrees to the ambiguity and outlier class of the instances are explicit, and these values are educated in the iterative clustering. Thus, the neutrosophic logic clustering is more immune to noise in dyslexia dataset and they correspond more closely to the belief of compatibility. The main contribution of this research work is to overcome the uncertainty problem in dyslexia by handling the outliers and the noise instances present in the dataset. Because failing to handle them as special cases lead to increase in false alarms. The incompleteness in dyslexia dataset is a very challenging problem in this work boosted decision tree is used for optimally imputing the missing value. The standard methodologies of fuzzy and intuitionistic fuzzy are not enough to handle the noisy and outlier instances thus in this work neutrosophic logic which is the generalization to both fuzzy and intuitionistic fuzzy which represents the truthiness, falsity and indeterminacy by their belongings or membership degree. Also the indeterminate cluster is further analyzed as outlier and ambiguity clusters. The main problem in clustering techniques is determining optimal cluster centroids which are involved in framing clusters. This paper introduces an artificial bacterial foraging algorithm to discover the optimal centroids of each cluster in every iteration. Thus the overall contribution with various enhancements in different dimensions of data quality enhancement and investigation the model classifies the presence or absence of dyslexia among children's in an effective manner.

4. Proposed Methodology of Neutrosophically Clustered pattern recognition in uncertain Dyslexia dataset

This work introduces a smart unsupervised learning approach by integrating different optimized techniques to well handle the vague dyslexia dataset. Fig. 1 represents the overall framework of neutrosophically clustered pattern recognition in uncertain dyslexia dataset. Once the dataset is collected from KEEL repository the data is found incomplete due to the presence of incomplete records. So the initial process of this proposal is to overwhelm the missing values by introducing a boosted decision tree based imputation method. Once the complete dataset is obtained which greatly influences the clustering accuracy then the similar patterns among records are obtained. For clustering neutrosophic logic based C-means is used to handling uncertain dataset which especially tackles the outlier or the instances lie in the border of the cluster.

The performance of the neutrosophic clustering is enhanced by introducing Bacterial Foraging Optimization for choosing centroids during each iteration of the processor when a new instance is encountered for clustering. After clustering all the instances their pattern of similarity is learned based on the category specified in dyslexia dataset.

5. Imputation using boosted decision tree

The vital process involved in the boosted decision tree for imputing missing values in dyslexia dataset is as follows. Fig. 2 explains the process of imputation using decision tree.

1. Initially split the full dyslexia dataset into two sub-datasets. First sub-dataset comprised of instances without missing values denoted as CD and the second sub-dataset comprised of instances with missing values signifies as MD

Please cite this article as: J. Loveline Zeema and D. Francis Xavier Christopher, Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.008

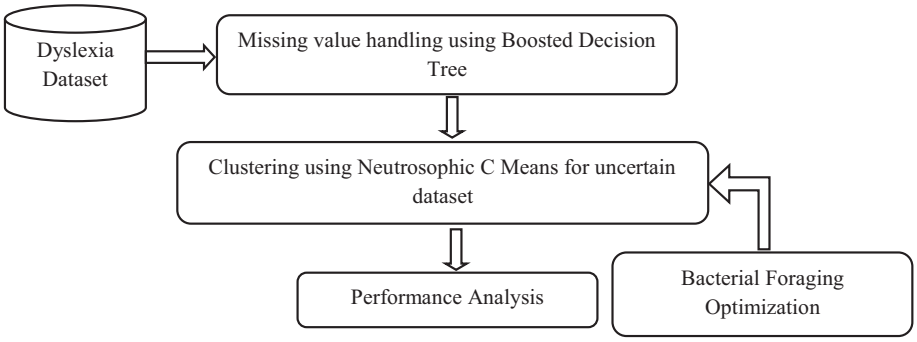

Fig. 1. Overall Framework of Neutrosophically Clustered pattern recognition in uncertain Dyslexia dataset using behavioral optimization approach.

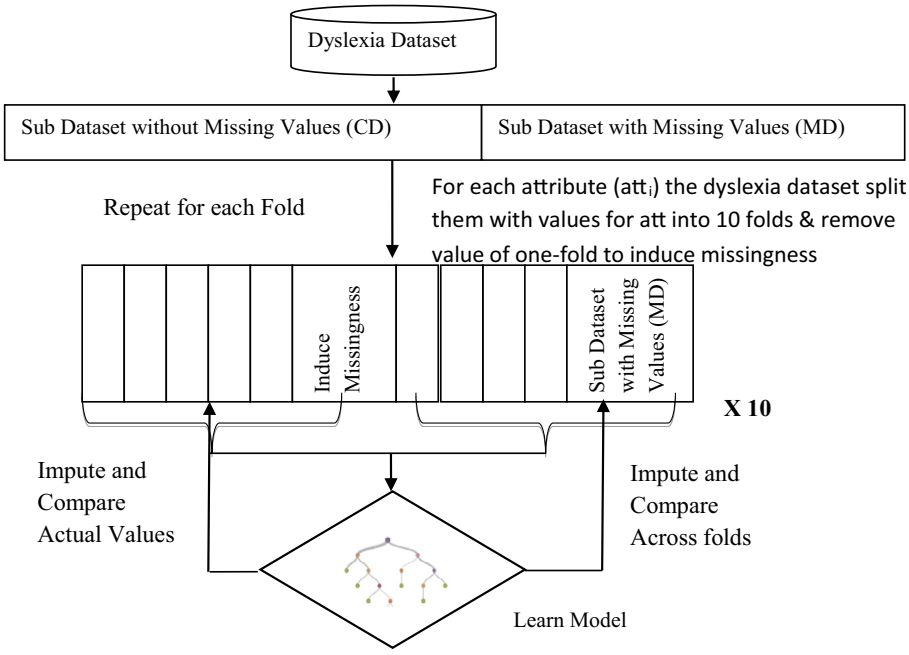

Fig. 2. Imputation using boosted Decision Tree.

2. Construct a set of decision trees on complete sub-dataset CD by considering the attributes which hold missing values in MD, as the class attributes.
3. Allocate each instance of MD to the leaf where it falls in for the tree that considers the attribute that has a missing value for the record as the class variable. If the record has more than one attributes as missing values then it will be assigned to more than one leaves
4. Impute numerical missing values using Newton Raphson algorithm and categorical missing values using majority class values within the leaves.
5. Combine records to form a completed data set CD without any missing values

6. Preamble of neutrosophic logic

The neutrosophic logic is a multivalve approach introduced by Smarandache (2002) in this the truth value of an instance is denoted by degree of truthiness, falsity, and indeterminacy (Smarandache, 2002). Neutrosophic logic maintains consistency with both classical and fuzzy logic with special case when the amount of truthiness, indeterminacy, and falsity where T + I + F = 1. It turns to intuitionistic logic while T + I + F < 1 while handling incomplete information. It also handles paraconsistency when the event is T + I + F > 1.

Hence the merit of NL with the property of nonstandard analysis it distinguishes relative falsehood signified by 0 and absolute falsehood signified by $^{-}0$. Likewise, relative truth and absolute truth signified by 1 and 1^{+} respectively.

While comparing intuitionistic fuzzy set, neutrosophic can distinguish among absolute membership of an element ($T = 1^{+}$), relative membership ($T = 1$) and partial membership signified by $0 < T < 1$ (Smarandache, 2002). In addition, while summing membership components of NL it need not be 1 as like fuzzy or intuitionistic fuzzy components but may be any number ranges from 0 to 3. The NL is trying to measure the truth, indeterminacy, and falsehood.

7. Neutrosophic C-means clustering

In conventional clustering, similar samples are grouped into the same cluster. To describe in detail let us assume that X be a dataset consisting of N instances represented as $\{x_i, i = 1,2,3,\ldots N\}$, where x_i denotes a single instance of the dataset. To partition among dataset

Please cite this article as: J. Loveline Zeema and D. Francis Xavier Christopher, Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.008

there are two categorize of clustering they are hard and fuzzy clustering approaches (Baraldi and Blonda, 1999a,b). In hard clustering, while a data instance is assigned to a specific cluster, then it cannot be contained within other clusters. But in contrast, the fuzzy clustering each instance may belong to more than one clusters with different membership degrees.

This work uses Neutrosophic C-means Clustering which computes each dyslexia dataset instance's degree towards determinant and indeterminate clusters are well defined. Here T denotes the degree of determinant clusters and other two membership F and I is used to determinate two different types of indeterminate clusters namely outlier cluster and ambiguity cluster for each dyslexic record respectively. When the data instances are very far from the centroids of each cluster then it is included in the outlier cluster which allows rejecting such instances. The ambiguity cluster permits to consider the data instances that are lying near the boundaries of the clusters.

During clustering iterations, both these clusters are involved and not during decision processing. Because the degree of membership towards ambiguity and outlier class of an instance are explicit and their values are erudite during iterative clustering. Thus, applying these memberships function the methodology is more immune noise and they relate more closely to the concept of compatibility. Hence, the inability in FCM and IFCM which fails or lacks to detect such unusual data instances can be solved precisely using NCM.

8. Bacterial foraging optimization

With the inspiration of Bacterial foraging behavior, Passino (2002) introduced the Bacterial foraging optimization algorithm (BFOA). The artificial bacterial foraging optimization mimics the four basic mechanisms which are observed in a real bacterial system, namely chemotaxis, reproduction, swarming and elimination dispersal to solve non-gradient optimization issue (Vipul Sharma and PattnaikTanuj Garg, 2012). An artificial bacterium may act as a search agent which moves on the functional surface to locate the global optima and thus in this work it is used in clustering of dyslexia dataset more optimally with neutrosophic C-means clustering to discover the more prominent cluster center during re-clustering on each iteration by obtaining local optima.

9. The step-by-step process of BFO algorithm

Algorithm 1: Procedure for Bacterial Foraging Algorithm

Step 1. Initialize parameters N, NB, CS, SS, RS. EDS, PE, RL(i) = (i = 1, 2, . . ., S), θi, where
N: dimension of the search space (dyslexia dataset),
NB: the number of bacteria,
CS: chemotactic steps,
SS: swimsteps,
Rs: reproductive steps,
EDS: elimination and dispersal steps,
PE: probability of elimination,
RL(i): the run-length unit _i.e., the chemotactic step size during each run or tumble.
θi : data point in a population
Step 2. Loop for Elimination-dispersal: l = l + 1.
Step 3. Loop for Reproduction: k = k + 1.
Step 4. Loop for Chemotaxis: j = j + 1.
4.1. For i = 1 = 1, 2, . . ., NB, chemotactic step for a bacteria i is as proceeds

(*continued*)

Algorithm 1: Procedure for Bacterial Foraging Algorithm

4.2. calculate the fitness function of Fit(i, j, k, l) using the formula
Fit(i, j, k, l) = Fit(i, j, k, l) + Fitcc(θi(j,k,l), Pop(j,k,l)) (1)
Where Fitcc is the objective function which has to be added to the actual objective function inorder to present a time fluctuating objective function.
4.3. Let Fit_{last} = Fit(i, j, k, l) storing this value to discover better value during further run
4.4.Peform Tumble operation by generating a random vector $\Delta(i) \in R^n$ with each element Δm(i), m = 1, 2, . . .,NB , a random number on [−1, 1].
4.5. Move:

$$\theta^i(j+1,k,l) = \theta^i(j,k,l) + c(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}} \quad (2)$$

Where C (i) denotes the size of the step taken during tumble operation in the random direction for a specific bacteria i.
4.6. Calculate Fit(i, j + 1, k, l) with θ^i(j + 1, k, l).
4.7. Swim
Assign m = 0 (swim length counter)
while m < Ns (While it is not climbed to long)
(a) let m = m + 1
(b) if Fit (i, j + 1, k, l) < Fit_{last}, then Fit_{last} = Fit(i, j + 1, k, l), and compute

$$\theta^i(j+1,k,l) = \theta^i(j,k,l) + c(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}} \quad (3)$$

and using θ^i(j + 1, k, l) generate new fit(i, j + 1, k, l) as in 4.2
(c) else let m _ Ns {end of while statement}
4.8. Go to next bacterium (i + 1): if i ≠ NB go to step 4.2 for processing next bacteria.
5. If j < CS, go to Step 4, it means to continue the chemotaxis steps till the bacteria life is not over.
6. Process for Reproduction.
6.1. For the assigned *k* and *l*, and for each *i* = 1, 2, . . ., *NB*,
Assign $Fit^i_{health} = \sum_{j=1}^{N_{c+1}} \mathrm{Fit}(i,j,k,l)$,
be the health of the bacteria. Sort bacterium in order of ascending values (Fit_{health}) which means higher cost with lower health
6.2. The bacteria SB_r with the value of highest Fit_{health} will die, and other *Sr* bacteria which contain best values split, and their copies that are located at the same place as their parent.
7. If *k* < *RS* go to Step 3. During in this event if the number of predefined reproduction step is not next then start the next generation in the chemotactic loop.
8. Process *for* Elimination-dispersal:
for *i* = 1, 2, . . ., *NB*, with probability PE, eliminate and disperse each bacterium, which consequences in maintaining the number of bacteria in the population as constant. To perform this, if a bacterium is eliminated, just disperse one to a random location on the optimization area.
If *l* < *EDS*, then go to Step 2, else end.

10. Neutrosophic C-means clustering for predicting dyslexia

Inspired by neutrosophic set, this proposed work not only considers degree belonging to determinate clusters but in addition it also considers the degree belonging to the indeterminate clusters. For instance, a new set D has been defined as the union of both determinant and indeterminate clusters, which is signified as

Please cite this article as: J. Loveline Zeema and D. Francis Xavier Christopher, Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.008

$$D = C_i \cup B \cup Y, i = 1...C \quad (4)$$

where C_i represents an indeterminate cluster, B signifies as clusters in boundary regions, Y is related to noisy data and $\cup$ is the union operation. Here B and Y are two types of indeterminate clusters. In Neutrosophic clustering, the degree of determinant clusters is represented as T or μ, the degree of boundary clusters are denoted as I or γ and Degree belonging to the noisy clusters are represented using F or τ. To handle indeterminacy by clustering this work uses a new objective function which is defined as follows

$$J(\mu, \gamma, \tau, C) = \sum_{i=1}^{N}\sum_{j=1}^{C}(\Omega_1\mu_{ij})^m \| x_i - c_j \|^2 + \sum_{i=1}^{N}(\Omega_2\gamma_i)^m \| x_i - \bar{c}_{imax} \|^2 + \sum_{i=1}^{N}\delta^2(\Omega_2\tau_i)^m \quad (5)$$

$$\bar{c}_{imax} = \frac{C_{fi} + C_{gi}}{2}, fi = \underset{j=1,2\cdots C}{\text{argmax}}(\mu_{ij}); \; gi = \underset{j\neq f_i \cap j=1,2\cdots C}{\text{argmax}}(\mu_{ij}) \quad (6)$$

where fi and gi are the cluster numbers with the biggest and second biggest value of T (μ). M is a constant value, when the value of fi and gi are determined then $\bar{c}_{imax}$ value is computed and its value is considered as a constant number of each data instances I, which will not change anyway. μ_{ij}, γ_i and τ_i is the membership values which belongs to determinate clusters, boundary and noisy instances of the dataset. $0 < \mu_{ij}, \gamma_i, \tau_i < 1$, which satisfies the subsequent formula

$$\sum_{i=1}^{N}\mu_{ij} + \gamma_i + \tau_i = 1 \quad (7)$$

Since the partitioning of dyslexia dataset is done through an iterative manner of the objective function, the membership of μ_{ij}, γ_i, τ_i and cluster centers Cj are updated by equations (Galaburda et al., 2006; Alcala Fdez et al., 2011; Janousova et al., 2016; Kassraian-Fard et al., 2016) respectively during iteration.

$$\mu_{ij} = \frac{E}{\Omega_1}(x_i - c_j)^{-\frac{2}{m-1}} \quad (8)$$

$$\gamma_i = \frac{E}{\Omega_2}(x_i - \bar{c}_{imax})^{-\frac{2}{m-1}} \quad (9)$$

$$\tau_i = \frac{E}{\Omega_3}\delta^{-\frac{2}{m-1}} \quad (10)$$

$$E = \left[\frac{1}{\Omega_1}\sum_{j=1}^{C}(x_i - c_j)^{-\frac{2}{m-1}} + \frac{1}{\Omega_2}(x_i - \bar{c}_{imax})^{-\frac{2}{m-1}} + \frac{1}{\Omega_3}\delta^{-\frac{2}{m-1}}\right]^{-1} \quad (11)$$

$\bar{c}_{imax}$ is computed according to indexes of the largest and second largest value of μ_{ij} during each iteration. The iteration will stop once the criteria $|\mu_{ij}^{(k+1)} - \mu_{ij}^{(k)}| < \varepsilon$, the termination condition ε lies between 0 and 1, k represents iteration step.

To achieve optimization in Neutrosophic C Means clustering based dyslexia prediction this work introduced artificial bacterial foraging algorithm to discover the optimal centroids for each cluster in every iteration. With the ability of global optima bacterial foraging algorithm significantly discovers each cluster's center with its four basic mechanisms. The detailed process of bacterial foraging algorithm to optimize the neutrosophic C means clustering to detect dyslexia presence or absence is shown below.

Algorithm 2: Optimized Neutrosophic C Means clustering using behavioral inspiration of Artificial Bacterial Foraging Algorithm (ONCMC-ABF)

Load input clustering data {dyslexia dataset}
Discover number of attributes in the data
Initialize μ^0, γ^0, τ^0, C, m, $\Omega_1, \Omega_2, \Omega_3$
Initialize K = 0; {iteration step}
Compute Center's of each cluster's C_i using BFO (call procedure for BFO Algorithm) at iteration k
Calculate $\bar{c}_{imax}$ depending on the largest and second largest value of μ by a comparison process
$\bar{c}_{imax} = \frac{C_{fi}+C_{gi}}{2}$; fi = $\underset{j=1,2\cdots C}{\text{argmax}}(\mu_{ij})$;
gi = $\underset{j\neq f_i \cap j=1,2\cdots C}{\text{argmax}}(\mu_{ij})$
Update the value of $\mu^{(k)} to \mu^{(k+1)}$, $\gamma^{(k)} to \gamma^{(k+1)}$, $\tau^{(k)} to \tau^{(k+1)}$
If $|\mu_{ij}^{(k+1)} - \mu_{ij}^{(k)}| < \varepsilon$ then stop the process, else return to step 5
Assign each data into the class which has the biggest TM = [T, I, F] value
Where x(i) $\in$ kth class if k = $\underset{j=1,2\cdots C+2}{arg\,max}(TM_{ij})$
Output: Clustered dyslexia dataset

11. Results and discussions

In this work the data is clustered using Neutrosophic C-means clustering with Bacterial Foraging to detect the degree of dyslexia and simulation is done using MATLAB tool. The dataset is collected from dyslexic_12_4 data set available in KEEL dataset repository (Tamboer et al., 2016). The dyslexia dataset is represented as a low-quality dataset because it contains attribute values with crisp and vague values. The attribute of dyslexia dataset may be either discrete or continuous so that there is a need for converting it into neutrosophic value. The dataset consists of 65 instances with 12 attributes. The dataset also consists of missing values if such incomplete dataset is directly used, this may increase false clustering because elimination of those missing values may affect the clustering accuracy of the dyslexia dataset. The instances are classified using four different class labels namely No dyslexic, control and revision, dyslexic and intention, hyperactivity or other issues. Depending on their similarity and applying membership degree of truthiness (T), indeterminacy (I) and Falsity (F) obtained by neutrosophic clustering the vague, ambiguity and noise instances of low-quality dyslexia dataset is well handled. Further NCM based clustering is optimized by introducing bacterial foraging optimization. The BFO based searching strategy overcomes the partitioning problem by helping NCM. It selects the cluster centroid based on the fitness function and its ability of global optima the clustering accuracy is increased greatly.

Evaluation metrics

To evaluate the performance of the proposed Optimized Neutrosophic C Means clustering using behavioral inspiration of Artificial Bacterial Foraging ***(ONCMC-ABF)*** Algorithm to predict the presence or absence of dyslexia three other techniques are compared.

Precision

It is a fraction of correct instances among those that the algorithm accepts as true belonging to the relevant class.

$$\text{Precision}\,(CLr, CSi) = Nri/Ni$$

where class CLr, whose size Nri, cluster CSi of its size is Ni, Nri data instance in CSi from the class CLr

Please cite this article as: J. Loveline Zeema and D. Francis Xavier Christopher, Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.008

Recall:

It is calculated as the fraction of actual instances that were identified.

$$\text{Recall}(\text{CLr}, \text{CSi}) = \text{Nri}/\text{Nr}$$

F-Measure:

It is considered as the harmonic mean of both precision and recall and it tries to produce a good combination of these two measures

$$\text{F}(\text{CLr}, \text{CSi}) = \frac{2 * \text{Recall}(\text{CLr}, \text{CSi}) * \text{Precision}(\text{CLr}, \text{CSi})}{\text{Recall}(\text{CLr}, \text{CSi}) * \text{Precision}(\text{CLr}, \text{CSi})}$$

The Table 1 and the Fig. 3 illustrates that the proposed work ONCMC-ABF produces the highest percentage of precision, recall, and f-measure.

The other three methods hold less value because the ability to determine instance which belongs to either ambiguity or noise (outlier) are not handled in FCM, IFCM and NCM. Furthermore, clustering approaches often face difficulty while partitioning the dataset, especially when there is a presence of uncertainty and with low quality due to vagueness and ambiguity of dyslexia Dataset. The boosted decision tree is used for improving the quality of dyslexia dataset which consist of missing values. In ONCMC-ABF it is well handled by integrating both neutrosophic values and bacterial foraging behavior to optimize the clustering process more precisely and better than the other approaches. And this work also produces the simulation analysis of four different clustering approaches and from the result it is observed that the proposed *ONCMC-ABF* produces highest percentage of correctly clustered instances of dyslexia dataset. It is caused by two reasons, the *ONCMC-ABF, defines each data instances into the neutrosophic domain and each of them has the membership degree of truthiness, indeterminacy and falsity for each defined cluster. And the functionality of Neutrosophic C means clustering is optimized by the searching behavior of bacterial foraging which discovers the best centroids during re-computing the clusters on each iteration. Thus, NCM overwhelms form the problem of partitioning where as other traditional approaches NCM, IFCM and FCM takes the remaining position of performance. These three methods fail to handle the low-quality dyslexia dataset more precisely.*

Table 1
Performance comparison based on Precision, Recall and F-measure.

Clustering Approaches	Precision	Recall	F-measure
FCM	72.4	73.1	72.7
IFCM	79.4	80.6	80.0
NCM	82.3	84.2	83.2
ONCM-BFO	96.9	98.7	97.8

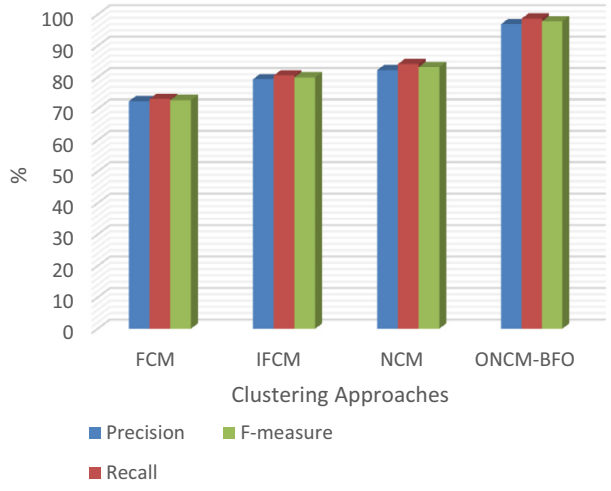

Fig. 3. Performance comparison based on Precision, Recall and F-measure.

12. Conclusion

This paper aims at developing an optimized clustering approach which overcomes the problem of handling ambiguity and outlier in dyslexia dataset which mainly affects the decision-making process in the prediction of the percentage of dyslexia. This work integrates neutrosophic concept and artificial bacterial foraging process to tackle the issue of partitioning the dyslexia dataset which consists of uncertainty and incompleteness. By imputing the missing values using boosted decision tree the quality of the dyslexia dataset increases by converting it into complete dataset because performing clustering with missing values may produce a high degree of false alarms. The bacterial foraging behavior-based search elects the cluster centers with the possibility of global optima further the representation of dataset in neutrosophic domain precisely handles the ambiguity and noise or outlier kind of instances in dyslexia dataset. Simulation results prove that by introducing membership degree of indeterminacy and falsity more accurately clusters the uncertain instances which belong to the category of dyslexia, no dyslexia, control with revision and hyperactivity and other issues. As future work, different nature-inspired algorithms can be used for clustering the dataset, deep learning models can be used as its extension and imputation methods with variants of a decision tree or other models can be developed.

References

Alcala Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F., 2011. KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Multip.-Val. Log. Soft Comput. 17 (2–3), 255–287.

Baraldi, A., Blonda, P., 1999a. A survey of fuzzy clustering algorithms for pattern recognition—Part I. IEEE Trans. Syst., Man, Cybern. B, Cybern. 29 (6), 778–785.

Baraldi, A., Blonda, P., 1999b. A survey of fuzzy clustering algorithms for pattern recognition—Part II. IEEE Trans. Syst., Man, Cybern. B, Cybern. 29 (6), 786–801.

Chen, Ao, Wijnen, Frank, Koster, Charlotte, Schnack, Hugo, 2017. Individualized early prediction of familial risk of dyslexia: a study of infant vocabulary development. Front. Psychol. 8, 1–13.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Alzheimer's Disease Neuroimaging Initiative, 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50, 883–892.

Galaburda, A.M., LoTurco, J., Ramus, F., Fitch, R.H., Rosen, G.D., 2006. From genes to behavior in developmental dyslexia. Nat. Neurosci. 9, 1213–1217.

Janousova, Eva, Montana, Giovanni, Kasparek, Tomas, Schwarz, Daniel, 2016. Supervised, multivariate, whole-brain reduction did not help to achieve high classification performance in schizophrenia research. Front. Neurosci. 10. Article 392.

Kassraian-Fard, P., Matthis, C., Balsters, J.H., Maathuis, M.H., Wenderoth, N., 2016. Promises, pitfalls, and basic guidelines for applying machine learningclassifiers to psychiatric imaging data, with autism as an example. Front. Psychiat.

Lyytinen, P., Eklund, K., Lyytinen, H., 2005. Language development and literacy skills in late-talking toddlers with and without familial risk for dyslexia. Ann. Dyslexia 55, 166–192.

Minister for Science & Technology Dr. Harsh Vardhan to Release Assessment Tools for Dyslexia – 'A Learning Disorder' in Indian Languages, Press Information Bureau Government of India, Ministry of Science & Technology.

Passino, K.M., 2002. Biomimicry of bacterial foraging. IEEE Control Syst. Magz. 22, 52–67.

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P.H., 2007. T, Poikkeus A, Very early phonological and language skills: estimating individual risk of reading disability. J. Child Psychol. Psychiat. 48, 923–931.

Sampada Margaj, Purohit, Seema, 2016. Significance of data mining techniques in classifying dyscalculia. Int. J. Eng. Comput. Sci. 5 (08), 17799–17804.

Shaywitz, S.E., Shaywitz, B.A., 2005. Dyslexia. Biol. Psychiat. 57 (1301–1309), 2005.

Smarandache, F., 2002. Neutrosophy, a new branch of philosophy. Multiple-valued logic, An Int. J. 8 (3), 297–384.

Smarandache, F., 2002. A unifying field in logics: neutrosophic logic. Multiple-Valued Logic/An Int. J. 8 (3), 385–438.

Please cite this article as: J. Loveline Zeema and D. Francis Xavier Christopher, Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.008

Smarandache, F., 2002. Proceedings of the First International Conference on Neutrosophy, Neutrosophic Logic, Neutrosophic Set, Neutrosophic Probability and Statistics. University of New Mexico, Gallup Campus, Xiquan, Phoenix, p. 147.

Tamboer, P., Vorst, H.C.M., Ghebreab, S., Scholte, H.S., 2016. Machine learning and dyslexia: Classification of individual structural neuro-imaging scans of students with and without dyslexia. NeuroImage: Clinical 11, 508–514.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Networks 10, 988–999.

Vipul Sharma, S.S., PattnaikTanuj Garg, A., 2012. Review of bacterial foraging optimization and its applications, national conference on future aspects of artificial intelligence in Industrial Automation, Proceedings published by. Int. J. Comput. Appl. NCFAAIIA, 1–12.

Please cite this article as: J. Loveline Zeema and D. Francis Xavier Christopher, Evolving Optimized Neutrosophic C means clustering using Behavioral Inspiration of Artificial Bacterial Foraging (ONCMC-ABF) in the Prediction of Dyslexia, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.008