Hybrid AFS Algorithm and k-NN Classification for Detection of Diseases

Logapriya S
II ME(CSE)
Department of Computer Science and Engineering
Dr. Mahalingam college of Engineering and Technology, Pollachi

Dr.G.Anupriya
Associate Professor
Department of Computer Science and Engineering
Dr. Mahalingam college of Engineering and Technology, Pollachi

***Abstract*— k-Nearest Neighbor (k-NN) is a widely used classification algorithm in pattern recognition and data mining. k-NN based classification algorithm is dependent on the weight tuning methods and distance metrics. A major challenge is the issue of how to explore the optimal weight values of the features and how to measure the distance between the neighbors affecting the classification accuracy of the k-NN. In this paper, AFS (Artificial Fish Swarm) algorithm is used for the optimization of the parameters. A similarity measurement method, which is called the fuzzy distance metric, is used to measure the similarities between the test and training observations. The development of these hybrid methods requires additional tasks, such as determining the fuzzy membership function and membership degrees for each dataset which is very time-consuming process. To improve the performance of the classification model, an automatic approach will be developed to search the optimal bounds of fuzzy membership functions heuristically for every dataset. Moreover, exploring the optimal bounds may decrease the error rates of the proposed system and also increase the accuracy. Experimental results are to be performed on real world classification problem obtained from the UCI machine-learning benchmark repository.**

Keywords*— *Classification, Genetic Algorithm, Artificial Fish Swarm Algorithm, Fuzzy distance metric, Hybrid k-NN classifier

I. INTRODUCTION

Data mining techniques have been widely used to mine knowledgeable information from medical data bases. In data mining, classification is a supervised learning technique that can be used to design models describing important data classes, where class attribute is involved in construction of the classifier. K-Nearest Neighbor search is one of the most popular learning and classification techniques introduced by Fix and Hodges [1], which has been proved to be a simple and powerful recognition algorithm. Weight tuning procedure and similarity measurement method has an important role in k-NN classification algorithm. In recent years, heuristic methods such as artificial neural network, genetic algorithm, simulated annealing, ant colony optimization, particle swarm optimization, differential evolution and artificial bee colony have been used for solving the classification problem. The weighting of the features is not a new method. Although there are many successful optimization techniques to weigh the observation features, the number of misclassified observation may still high. In this paper, AFS (Artificial Fish Swarm) algorithm will be used for the optimization of the parameters. It is a kind of intelligent optimization method based on fish behavior [2]. Main component of AFS algorithm is exploitation and exploration process. This algorithm is inspired by the collective movement of the fish and their various social behaviors. This algorithm has many advantages including high convergence speed, flexibility, fault tolerance and high accuracy.

In this paper, efficient weight tuning procedure by applying the AFS based heuristic approach is combined with the genetic based weight tuning procedure. Finally, weight tuning procedure and fuzzy distance metric is combined with the k-NN algorithm to improve the accuracy and stability of the classification problem. The effectiveness of the performance of the proposed system is compared with classic approach.

The paper is organized in the following order. The related study is given in section II. Section III deals with the methodology used to perform classification. Section IV discusses the performance metrics and also the expected outcome.

II. RELATED STUDY

k-NN is considered a lazy learning algorithm that classifies the datasets based on their similarity with neighbors. ‘k’ stands for number of data set items that are considered for the classification problem. Optimization of the parameters using hybrid models and measuring the distances between the test data and each of the training data are used to decide the final classification output. Various hybrid models have been used to increase the accuracy and the stability of k-NN classification algorithm. [3].

Akhil Jabbar et al. [4] developed the GA based k-NN for effective classification. GA perform global search in complex large and multimodal landscapes and provide optimal solution. This approach is not suitable for irrelevant and redundant attributes in some datasets.GA based hybrid k-NN classification algorithm is one of the most popular approaches to improve the classification performance.

Nebojsa Bacanin et al. [5] have integrated Genetic Algorithm with ABC algorithm to improve exploitation process. Genetically inspired ABC improves the performance of the ABC algorithm by applying uniform crossover and mutation operators from genetic algorithms.

Suguna and Thanushkodi [1] introduced a version of GA based k-NN classification, where initially reduced features set is constructed by using rough set-based bee colony optimization and then by GA, the k-number of samples are chosen for similarity measurement. It reduces the response time rather than improving its classification accuracy. Weight tuning of the features is one of the practically used methods to improve the classification performance.

k-NN algorithm gives the same importance to every neighbor, assuming that the boundaries between classes are perfectly defined, which is not always true. Hui-Ling Chen et al. [6] uses k-NN classifier with concepts of Fuzzy logic to assign degree of membership to different classes while considering the distance of its k-NN where exploring the optimal bounds may increase the error rate.

Maillo and Luengo [7] developed effective improvement of k-NN that alleviates the issue by using fuzzy sets, named Fuzzy k-NN. To do so, Fuzzy k-NN has two different phases. First, it changes the class label for a class membership degree. After that, it calculates the k-NN with the membership information, achieving higher accuracy rates in most classification problems. The major challenge in fuzzy logic is the construction of the membership function.

Zadeh [8] proposed a series of membership functions that could be Triangular function, Trapezoid function, Gaussian function, Singleton function, Gamma function.

Kemal Polat [9] proposed the fuzzy c-means clustering based feature weighting and aim is to decrease the variance between classes and to improve the classification accuracy by the transformation from a linearly non-separable dataset to a linearly separable one.

Jamsandekar and Mudholkar [10] proposed fuzzy classification system by generating the membership function using k-means clustering technique. The automatic creation of the fuzzy classification system from the membership function generated reduces the hassle of every time creating a new fuzzy Inference system to perform classification on change of data input.

Titov et al. [11] proposed automatic tuning of membership functions of fuzzy sets in accordance with the training sample. Membership function is pre-defined and the application does not require conventional numerical methods (gradient search and so on) with a large computational complexity and the implementation of the algorithm in the software environment Matlab showed a low percentage of misclassified vectors of the test sample.

III. Methodology

In this work, Artificial Fish Swarm Algorithm and Genetic Algorithm based weight-tuning procedure are used as a searching approach and fuzzy logic based-distance measurement are used to create distance arrays representing distances between test observation and training observations. The distance arrays are used to predict the class of the test observations in the k-NN classification. The effectiveness of the proposed approach is to be proven by comparing their performances with classic approach on the real-world classification problem obtained from UCI machine learning repository. The details of the proposed approach are given in the following sections.

k-NN classification includes parameter optimization and similarity measurement. Parameter optimization is a process to adjust the weight values of the features. Artificial Fish Swarm Algorithm is a population-based approach used for optimizing the features. AFSA algorithm is used as a hybrid algorithm. Fuzzy distance metric is used for similarity measurement with automated generation of fuzzy membership function. In proposed system, Artificial Fish Swarm Algorithm will be used for parameter optimization process. Best weight values of the features can be derived by optimization. Genetic algorithm is combined as a hybrid algorithm for weight tuning procedure along with AFS Algorithm. The proposed hybrid algorithm is expected to improve the accuracy of the classification problem. The fuzzy distance metric with an automatic approach may be developed to search the optimal bounds of fuzzy membership functions heuristically for every dataset. Thus, proposed approach may improve the accuracy and stability of the classification. Initially the datasets are pre-processed before being fed into the classifier. Weight tuning of features leads to the success of classifier. The fuzzy distance metric is extended to measure the similarity between the test and training observations. The block diagram of the proposed system is given in Fig **1.**

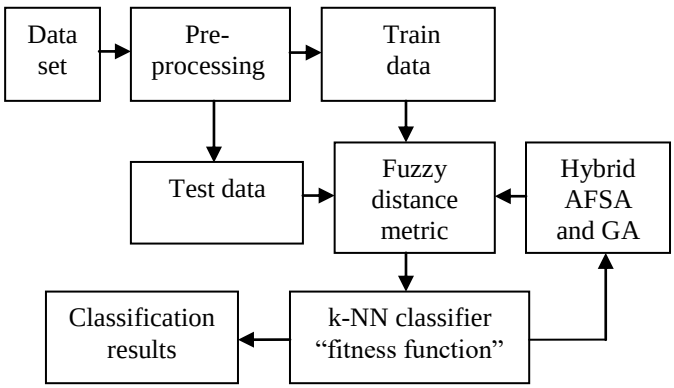

Fig 1: Block diagram of the proposed system

The steps involved are:

- *Pre-processing*
- *Hybrid GA and ABC Algorithm for weight tuning of parameters*
- *Hybrid GA and AFS Algorithm for weight tuning of parameters*
- *Fuzzy distance metric for similarity measurement between test and training samples.*
- *k-NN Classification*
- *Performance evaluation*

A. Pre-processing: In the pre-processing stage, duplicate values are removed in each data set in the training and testing data to avoid redundancy. After duplicate removal, next step is to handle missing values in the data. The preprocessing technique identifies the missing feature values and then they are replaced by the mean value for that feature. The results of the pre-processing is given in Fig 2.

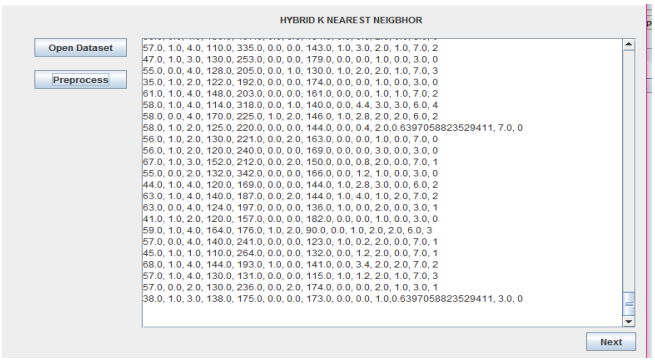

Fig 2: Output of Pre-processing

B. Hybrid ABC Algorithm and GA for weight tuning of parameters: Artificial Bee Colony is a relatively new swarm intelligence based metaheuristic [3 5]. Initially, weight-tuning is performed by applying the ABC-based heuristic searching approach to adjust weight of the features and this output is fed to the Genetic Algorithm. ABC algorithm searches the best weight values of the attributes of the objects in dataset by creating bee colony. The colony is created in three steps. In the first step, the bees are created. For every bee, a food source is randomly produced in the second step. In the last step, the colony parameters are to be determined. Consequently, the colony is created as weight values (the weighted features). The weight values are the input of the Genetic Algorithm in the classification unit, where Genetic Algorithm searches the best weight values of the attributes and these best weighted feature approach is adopted to calculate the distances among the observations.

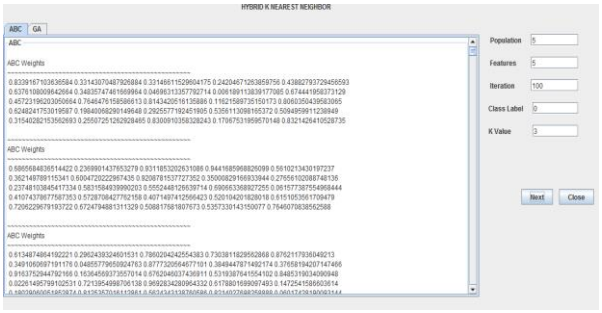

Fig 3: Artificial Bee Colony Algorithm output

C. Hybrid GA and AFS Algorithm for weight tuning of parameters: A hybrid algorithm based on Genetic Algorithm and Artificial Fish Swarm Algorithm is proposed. The hybrid algorithm takes advantage of their complementary ability of global and local search for optimal solution. GA is applied as global search , which can explore more global optimal region. AFS Algorithm is used as a local search to obtain the final optimal solution for improving the exploitation capability of algorithm. In this proposed system, AFS and GA is expected to explores the best weight valuesof the attributes in the colony. The weight values of the features in a classification problem, which are to be explored by the proposed hybrid algorithms will be used in the fuzzy distance metric measurement. The artificial fish's behaviour is described as follows [12, 13]:

Fisher Discriminant Steps:

The main steps of the process parameter optimization are as follows.

- Initialize the parameters, such as the maximum number of iterations, the number of artificial fish (AF), Visual Distance, Step length (visual and step are very significant in view of the fact that artificial fish basically move based on these parameters. Large values of these parameters increase the capability of algorithm in global search, while small values improve the local search ability of the algorithm).
- Randomly select N number of Fishes as its starting position and randomly select a Fishes location as the bulletin board's initialization location.
- Each AF exhibits searching, swarming and following behaviors. Bulletin board is used to record the state of the optimal artificial fish.
- According to the behavior's action, move the fish to the right location.
- Search the best AF in colonies and compare with the AF in bulletin.
- If its location is better than the AF in bulletin, then replace the AF by it, otherwise, leave the bulletin unchanged.
- If current number of iterations reaches or exceeds the maximum number of iterations, best optimal solution is obtained.

- Otherwise, the value of iteration is increased by one and performs search behavior.
- Finally, best optimal solution from bulletin is obtained.

D. Fuzzy logic unit: the weighted distance metric and automated membership function: A classification problem is represented by the features and target class as shown in equation (1) where q_x denotes sample observation , 'c' is the target class and 'v' is the number of features.

$$qx \equiv [fq_{x[1]}, fq_{x[2]}, fq_{x[3]}, \ldots, fq_{x[v]}, c] \quad (1)$$

Fuzzy logic-based similarity measures are used in various applications [3]. The fuzzy logic similarity measure is used to represent the distance between test obsevations and training obsevations/ samples. The main components of fuzzy logic are difference, fuzzification, the rule base, decision making and defuzzification. The purpose of difference component is to measure the distance between test samples ($fq_{x[i]}$) and train samples ($fq_{y[i]}$) as given in equation (2) and (3).

$$Ifq_{x[v]} \equiv \forall^{v}_{i=1} \mid fq_{x[i]} - fq_{y[i]} \mid \quad (2)$$

To rewrite above equation, we can use weight values or weighted features $W_{f[i]c}$, which are searched by the GA and AFS Algorithm

$$Ifq_{x[v]} \equiv \forall^{v}_{i=1} \mid W_{f[i]c} \, (fq_{x[i]} - fq_{y[i]}) \mid \quad (3)$$

The $\mu(If_{qx})$ membership degrees of these inputs are calculated using equation (4) by the "fuzziness" component based on the triangular membership functions as given in Fig 4.

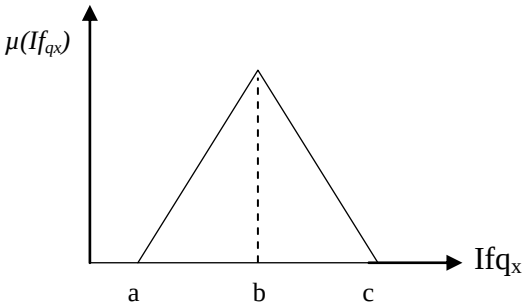

Fig 4: The calculation of the membership function

$$\mu(If_{qx}) = \begin{cases} \dfrac{Ifq_x - a}{b - a} & a \le Ifq_x \le b \\ \dfrac{c - Ifq_x}{c - b} & b \le Ifq_x \le c \\ 0 & Ifq_x < a \text{ or } Ifq_x > c \end{cases} \quad (4)$$

The membership functions are prepared depending on the number of the input and the output parameters of the problem. The interrelation between the inputs and the outputs are defined in the form of *if-then* rules and processed by the "decision making" component.
In the "defuzzification" component, the center of gravity method will be used to obtain the output. The array will be produced which represents the distance between the test observation and the training samples.

E. k-NN classifier: In k-NN algorithm all of the objects in the Q set correspond to the points in the v-dimensional space. The nearest neighbor of the q_x object example are defined in terms of fuzzy distance metrics. In this study, k-Nearest Neighbor is proposed as a fitness function in order to evaluate the candidate solutions of the searching algorithm. Fitness values will be obtained in the form of equation (3). In proposed approach, the weight values of the AFS and GA will be used in the similarity measurements between the test and the training observations. The k-NN finds the class of the test instances based on the distance metric. The error rate and accuracy are widely parameters to measure the performance of classifiers.

IV. Results and Discussions

In this section, expected results are discussed. One of the aims of the proposed algorithm is to improve the accuracy and stability of the k-NN classification algorithm. For this purpose, GA based weight tuning method is going to be combined with an AFS Algorithm. Fuzzy distance metrics in will be used to measure the similarity between the test observations and training observations. Finally, weight tuning methods and distance metrics are to be combined with the k-NN classifier to predict the chances of disease. The experiments are to be repeated for different datasets, different 'k' values, different distance metrics and three weight tuning methods.

Data description: Detection of disease classification system intends to identify the chances of the diseases. Medical datasets are collected from UCI machine repository. Medical datasets used are heart disease dataset and thyroid dataset. Heart disease dataset consists of 303 instances and 14 attributes including class label and Thyroid disease dataset consists of 215 instances and 6 attributes including class label. In all dataset, one third of the samples are to be used for testing and verification and two thirds have been used for training purposes. All of the datasets to be used in classification problems are labelled.

Evaluation metrics: The performance of the proposed k-NN classification system based on parameter optimization using AFSA is evaluated using two measures: Accuracy (Acc) and Stability(E). A classifier is trained to classify the chance of the diseases. The evaluation metrics such as Accuracy and

stability are assessed to determine the Accuracy and Error rateof the classification. The terms used in evaluation metrics are

False positive (FP): The number of incorrectly classified disease data

False negative (FN): the number of incorrectly classified health data

True positive (TP): The number of correctly classified disease data

True negative (TN): the number of correctly classified health data

The performance of the proposed system and the existing can be compared in terms of accuracy and Stability can be assessed in terms of Error rate (E).

The term error rate can refer to the degree of errors encountered during data classification. The higher the Error rate, less reliable the data classification.

Error rate (E) is calculated using the following equation (5)

$$Error\ rate = \frac{FP + FN}{P + N} \quad (5)$$

$FP+FN$ = total number of two incorrect predictions
$P+N$ = Total size of a dataset

Accuracy is how close a measured value is to the actual (true) value. The accuracy (Acc) of the classification performance is calculated using the equation (6)

$$Accuracy = \frac{FP + FN}{P + N} \quad (6)$$

TP+ TN = total number of two correct predictions.

Higher accuracy and lower error rate are expected in the proposed hybrid classification algorithm.

V. Conclusions

In this paper, a new hybrid method is introduced. Genetic Algorithm and Artificial Fish Swarm Algorithm will be used as a hybrid algorithm for weight tuning of the parameters. These hybrid algorithms are combined together with fuzzy distance metric to measure the similarity between the training samples and testing samples. The fuzzy distance metric and weight tuning procedures are combined with k-NN classification to do the detecting task of disease datasets. Finally, the fitness function of the classifier is used to find the nearest neighbor of the classification. It is expected that by including the automated setting fuzzy membership function for fuzzy distance metric, the stability of the k-NN classification will increase and by using the hybrid AFS Algorithm for the weight tuning procedure, accuracy of the k-NN classifier will increase.

REFERENCES

[1] N.Suguna and Dr.K.Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", International Journal of Computer Science Isuues, vol. 7, No. 4, pp.18-21, 2010.

[2] J. Jiang, Yuling Bo, Chuyi Song, and Lanying Bao, "Hybrid Algorithm Based On Particle Swarm Optimization And Artificial Fish Swarm Algorithm", Advances in Neural Network-ISNN 2012, Springer International publishing, vol.7367, pp.607–614, 2012.

[3] Hamdi Tolga Kahraman, "A Novel And Powerful Hybrid Classifier Method: Development And Testing Of Heuristic k-NN Algorithm With Fuzzy Distance Metric", Elsevier Transactions on Data and Knowledge Engineering, vol.103, pp.44-59, 2016.

[4] M. Akhil Jabbar, B.L Deekshatulua, and Priti Chandra, "Classification Of Heart Disease Using K- Nearest Neighbor And Genetic Algorithm", Elsevier Transactions on Procedia Technology, vol.10, pp.85 – 94, 2013.

[5] Nebojsa Bacanin and Milan Tuba, "Artificial Bee Colony (ABC) Algorithm for Constrained Optimization Improved with Genetic Operators", IEEE Transactions on Studies in Informatics and Control, vol.21, No.2, pp.137-146, 2012.

[6] H-L. Chen and Chang-Cheng Huang, "An Efficient Diagnosis System For Detection Of Parkinson's Disease Using Fuzzy K-Nearest Neighbor Approach", Elsevier Transactions on Expert Systems with Applications, vol.40, pp.263-271, 2013.

[7] Jesus Maillo, Juli´an Luengo, Salvador Garc´ıa, and Francisco Herrera, "Exact Fuzzy k-Nearest Neighbor Classification for Big Datasets", IEEE Transactions on Fuzzy Systems, vol. 17, pp.978-983, 2017.

[8] G. Ulutagay and S. Kantarci, "Classification with Fuzzy OWA Distance", International Conference on Fuzzy Theory and Its Applications, Kaohsiung, Taiwan, vol.14. pp.195-198, 2014.

[9] Kemal Polat, "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering", International Journal of Systems Science, vol. 43, No. 4,pp. 597–609, 2012

[10] S. Jamsandekar and R. Mudholkar, "Fuzzy Classification System by Self-Generated Membership Function Using Clustering Technique", International Journal of Information Technology, vol.6, No.1, pp.697-704, 2014.

[11] D.A. Titov, D. N. Klypin, and E.D. Bychkov, "Algorithms For Automatic Setting Membership Functions Of Fuzzy Sets", International Siberian Conference on Control and Communications (SIBCON), Russia, vol.15, pp.312-317, 2015.

[12] Chen Haifeng, Sun Xuebin, and Chen Dianjun, "Artificial Fish Swarm Algorithm in Industrial Process Alarm Threshold optimization", IEEE Transactions on Information and Engineering, vol.16, pp.691-694, 2016.

[13] M.A.Awad, M.Z.Rashad, "FAFSA: Fast Artificial Fish Swarm Algorithm", International Journal of Information Science and Intelligent System, vol.2, No.4, pp.60-70, 2013.

[14] Data set: "UCI Machine Learning Repository" [http://archive.ics.uci.edu/ml], Last accessed: August 2017.