European Journal of Business & Social Sciences
Available at https://ejbss.org/
ISSN: 2235-767X
Volume 07 Issue 6
June 2019

An Ensemble Big Data Classification for Healthcare Data Predication Analysis

VR. NAGARAJAN[1] & Dr. D. VIMAL KUMAR[2]

[1]Research Scholar, Nehru Arts and Science College, and Assistant Professor, PG & Research Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India, vrnag74@gmail.com

[2]Professor & Head, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India, vimal1519@yahoo.co.in

ABSTRACT

Due to tremendous growth in the volume of healthcare data, data analytics in a healthcare information system can extract valuable information. Nowadays, healthcare organizations have been moving towards digitization of the massive volume of healthcare data. Such data are handled by a latest emerging technology called Big data. A firefly optimization algorithm with different classifier method was used for prediction of different diseases. Initially, a healthcare data was collected and splitted into partitions. Each partition was loaded into mappers of MapReduce and the process of firefly was carried out in each mapper. It returned subgroups which had best features for disease prediction. The selected best features were used in Naïve Bayes, C4.5 and Random Forest (RF) which predicted the leukemia cancer, lung cancer and heart disease effectively. This work is extended in this paper by using classifier in a distributed environment. It explored the use of Fuzzy Rule Based System (FRBS) as a model where the rules are generated for prediction of diseases. For the rule generation, different classifier such as RF, Bayesian Tree and NeuroTree are used. The generated rules are combined either in homogenous or heterogeneous way. In homogenous way of disease prediction, the results of RF from different mappers are aggregated in single reducer whereas in the heterogeneous way of disease prediction, the results of RF, Bayesian Tree and NeuroTree are aggregated in single reducer. Hence, the learning accuracy and prediction accuracy is improved by using different classifiers in mappers. The effectiveness of this method is tested in experiments in terms of accuracy, precision, recall and F-measure.

Keywords: Healthcare data, Big data, disease prediction, Fuzzy Rule Based System, Random Forest, Bayesian Tree, NeuroTree.

1. INTRODUCTION

Generally, Big data (Wu, X. et al. 2014) refers to the huge volume of structured and unstructured data that overflow the organization. If the overflowed data is used in the proper way it leads to meaningful information. Comparatively, the big data includes a large number of data which needs more processing in real time. It also provides opportunities to discover new values, to understand an in-depth knowledge from hidden values. Data mining is a process of identifying relevant and significant information from large dataset.

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

Nowadays, health organizations are capable of generating and collecting a large amount of data. This increase in volume of data automatically requires an efficient technique to retrieve the significant data when it is needed. With the help of data mining techniques (Ţăranu, I. 2016), it is possible to extract the knowledge and determine interesting and useful patterns. By extracting knowledge from the healthcare data, it is more useful to predict a disease and to provide proper diagnosis for patient. The number of death is increased due to the late prediction of diseases. Data mining techniques were widely used for early prediction and diagnosis of diseases.

A firefly optimization algorithm with different classifier method (Nagarajan, VR. 2018) was introduced for early prediction of different diseases like lung cancer, leukemia cancer and heart disease. In this method, healthcare data were collected and it was splitted and assigned to each map in MapReduce. Then firefly optimization algorithm was employed in each map to discover the subgroups which select the most relevant features from the dataset. The most relevant features were collected from the reducers. The selected features were used in Naïve Bayes, C4.5 and Random Forest (RF) classifier to predict lung cancer, leukemia cancer and heart disease. However, only the feature selection process is carried out in distributed environment.

In this paper, the firefly optimization algorithm with different classifier method is improved by applying classifier in distributed environment. The healthcare data are collected and the most discriminative features are selected by applying firefly optimization algorithm in each mapper of MapReduce. Then a Fuzzy Rule Based System (FRBS) is developed where the rules are generated and combined either in homogenous way or in heterogeneous way. In each mapper, Random Forest (RF) classifier is applied to generate the rules and it is aggregated in single reducer which is called as homogeneous way based disease prediction. In order to improve the learning and disease prediction accuracy, different classifier such as RF, Bayesian Tree and NeuroTree are applied in mapper and the generated rules are aggregated in single reducer. It predicts lung cancer, leukemia cancer and heart disease effectively with high prediction accuracy.

2. LITERATURE SURVEY

A neural network was developed (Zeinalzadeh, A., et al. 2017) to differentiate the liver patients as high-risk groups and low-risk groups using genomic data. It employed neural network in the developed method to classify big datasets. Before training the Neural Network models, the data were pre processed. Then, the data were expanded using wavelet analysis and it was compressed by mapping the wavelet coefficients onto a new scaled orthogonal coordinate system. Finally, the neural network was used to train the data and it was used to classify the patients as high risk patients and low risk patients. However, this method predicts only the lung cancer patients.

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

A prototype lung cancer prediction system (Bharathi, H. & Arulanath, T. S. 2017) was developed using data mining classification techniques. This system extracted useful knowledge from a historical ling cancer database. In explanatory phase of data mining, Self Organizing Map (SOM) was introduced. It considers as an excellent tool which project the input space on the prototype of a low dimensional regular grid that was effectively used for visualization and explored the properties of the data. The lung cancer prediction system was processed in two stages. In the first stage SOM was utilized to produce the prototypes and clustered the prototype which was done in the second stage. The clustered data returns the absence of lung cancer and presence of lung cancer. However, this system involves high time complexity.

A prediction and diagnosis method (Daqqa, K. A. A. et al. 2017) was proposed for prediction and diagnosis of leukemia using classification algorithms. The main intention of this method was to determine the relation of blood properties and leukemia with health, gender and age status in patients using classification algorithms. Initially, more than 4000 patients were taken from blood test laboratory from European Gaza Hospital at Gaza Strip. The three efficient classification algorithms such as Decision Tree (DT), k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) was used to identify of leukemia cancer. However, the accuracy of this method is low.

An automatic detection of lymphocyte leukemia method (Purwanti, E., & Calista, E. 2017) was proposed for detection of lymphocyte leukemia. This method classifying the lymphocyte cells as normal lymphocytes and abnormal lymphocytes. Initially, lymphocyte cell were obtained from peripheral blood smear single cell. This method was comprised of two objectives are to extract the most discriminative feature cells and to classify the lymphocyte cells into two classes are normal lymphocytes and abnormal lymphocytes. The feature selection process was carried out using shape feature and histogram features and the classification was done by using k-Nearest Neighbor (k-NN) with k variation is 1,3,5,7,9,11,13 and 15. However, the efficiency of this method is low.

A modified differential evolution algorithm (DE) (Vivekanandan, T., & Iyengar, N. C. S. N. 2017) was proposed for optimal feature selection for heart disease prediction. In the modified DE algorithm, Differential Evolution (DE), a target vector was selected at random (rand), four vectors were selected and the weighted differences of the pairs were added to form the mutant vector (2-wt) and the exponential crossover (exp) was used and these are all represented as DE/rand/2-wt/exp strategy. A vector to be perturbed was randomly selected which was denoted as rand and the crossover was performed on the certain number of variables in one loop until it was within the CR bound. The selected features were used in the fuzzy Analytic Hierarchy Process (fuzzy AHP) and feed forward neural network to predict the heart disease.

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

hybrid model (Mustaqeem, A. et al. 2017) was proposed for disease prediction and medical recommendations to cardiac patients. This model was consisted of two parts are prediction model and recommendation model. In the prediction model, the data was splitted randomly and Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset. Then the data was undergoes a cleansing process which removed ambiguities and noise. The optimal features in the dataset were selected by using attribute evaluation and ranker search algorithm and finally in the prediction model Support Vector Machine (SVM), Random Forest (RF) and Multi Layer Perceptron (MLP) was applied which the data as silent ischemia, non cardiac chest pain, myocardial infarction and angina. The second part provided general medical recommendations to patients.

A weighted fuzzy rule based clinical decision support system (CDSS) (Anooj, P. K. 2012) was proposed or diagnosis of heart disease. This system was automatically obtained knowledge from the patient's clinical data. The proposed clinical decision system was comprised of two phases are automated approach for the generation of weighted fuzzy rules and developing a fuzzy rule based decision support system. In the first phase of CDSS, mining techniques such as feature selection and attribute weightage method was utilized to obtain the weighted fuzzy rules. Then the fuzzy system was constructed in accordance with the weighted fuzzy rules and chosen attributes. However, the specificity of this method is low.

3. PROPOSED METHODOLOGY

In this section, the proposed Fuzzy Rule-Based Classification Systems (FRBCSs) is described in detail. Initially, the leukemia dataset, lung cancer dataset and heart disease dataset is collected and it is splitted into number of partitions. The partitioned data are distributed to number of mapper nodes in MapReduce. Each mapper analysis the healthcare data and discovered sub groups by firefly optimization algorithm. The selected features are process by fuzzy rule based classification system in each mapper where the leukemia cancer, lung cancer and heart disease are predicted by different classifiers are Random Forest, Bayesian Tree and NeuroTree.

3.1 Fuzzy Rule Based Classification System

Generally, Fuzzy Rule-Based Classification Systems (FRBCSs) are one of the most popular tools used to solve the classification problems. It consisted of two components are knowledge base and Fuzzy Reasoning Method (FRM). The knowledge base is composed of database and the rule base, where the membership functions and rules used to model the linguistic labels are stored respectively. FRM is the mechanism used to classify the examples using the information stored in the knowledge base. A fuzzy rule learning algorithm is applied to generate the knowledge base. It used training set T_d composed of N labeled examples i.e., healthcare data with label which is presence of disease or absence of disease $x_n = (x_{n1}, \dots, x_{nm})$ with $n \in \{1, 2, \dots N\}$, where x_{ni} is the value of the i^{th} attribute

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

$(i \in \{1,2,\dots,h\})$ of the n^{th} training example. Each example belongs to a class $y_n \in C$, where C (presence of disease or absence of disease) is the class. The learning process in FRBCS is carried out by classifiers such as Random Forest (RF), Bayesian Tree and NeuroTree which are described as follows.

Random Forest

Random Forest is used for generating the number of random decision tree called forest. It is used for reducing the dimensionality of the input variables. For each tree in forest, the bootstrap sample is selected from the training data. Then, at each node of the tree, only few subsets of features are randomly split. The node then splits on the best features in subset of features for next iteration. Finally, the leaf nodes represents the distribution of values that the output variables for corresponding path through the decision tree. During the prediction process, the input variables are regularly monitored and the new values of metrics are provided every decision tree to reach the specific leaf node. Depending on the distribution of the output variable at the leaf node, the tree is voted to predict the disease. By traversing the nodes from the leaf node to root node rules are generated. The following algorithm describes the process of RF classifier.

Random Forest Algorithm

Input: Healthcare Training Data $\mathrm{D} = (\mathrm{x}_1,\mathrm{y}_1), \dots, (\mathrm{x}_n,\mathrm{y}_n)$, Features F, and number of trees in forest T.

//x_i is the list of attribute and y_i is their corresponding class label

Output: Set of rules

1. function RandomForest(D,F)
2. $\mathrm{G} \leftarrow \emptyset$
3. For $\mathrm{k} \in 1, \dots, \mathrm{T}$ do
4. $\mathrm{D}^k \leftarrow$ Bootstrap sample from D
5. $\mathrm{g}_k \leftarrow \mathrm{RandomizedTreeLearn}(\mathrm{D}^k,\mathrm{F})$
6. $\mathrm{G} \leftarrow \mathrm{G} \cup \{\mathrm{g}_k\}$
7. End for
8. Return G
9. End function
10. function RandomizedTreeLearn(D,F)
11. At each node do
12. $\mathrm{f} \leftarrow$ small subset of F
13. Split on best feature in f
14. Return learned tree
15. End function

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

Bayesian Tree

The original Bayesian Tree model is the Bayesian Classification and Regression Tree (BCART) model. The two basic components of this model consist of prior specification and stochastic search. The basic idea is to have the prior induce a posterior distribution that will guide the stochastic search toward more promising CART models. The Bayesian Tree uses a Metropolis-Hashing step with a transition kernel choosing randomly among four steps are grow, prune, change and swap. The growing process is determined by the specification of two functions are $p_{SPLIT}(\tau, Tree)$ and $p_{RULE}(\mu|\tau, Tree)$. For an intermediate tree $Tree$ in the process, $p_{SPLIT}(\tau, Tree)$ is the probability that terminal node τ is split and $p_{RULE}(\mu|\tau, Tree)$ is the probability of assigning splitting rule μ to τ if it is split. It is given as follows:

$$p_{SPLIT}(\tau, Tree) = \rho(1 + d_\tau)^{-\gamma}$$

Where $\rho < 1$, d_τ is the depth of the node τ and $\gamma \geq 0$. In the growing process, a terminal node is randomly picked and it is split into two new ones by randomly assigning it a splitting rule according to the p_{RULE} used in the prior. Then in the pruning process, a parent of two terminal nodes is picked and turns it into a terminal node by collapsing the nodes below it. In the changing process, an internal node is randomly picked and randomly reassign it a splitting rule according to p_{RULE} used in the prior. Finally in the swapping process, a parent-child pair is randomly picked which are both internal nodes. Swap their splitting rules unless the other child has the identical rule, in which case swap the splitting rule of the parent with that of both children. By restarting this algorithm repeatedly leads to a wide variety of different trees. In order to identify a good tree, evaluate posterior probability for all visited trees and obtain their relative probabilities. The tree which has the largest posterior probability is selected as a good tree. By traversing the tree from root node to leaf node, a set of rules are generated for prediction of lung cancer, leukemia cancer and heart disease.

NeuroTree

NeuroTree consists of two phases are pre-processing phase and prediction phase. The pre-processing phase generates a training set and creates a trained Neural Network Ensemble (NNE). The prediction phase constructs a NeuroTree and classifies the data as presence of disease or absence of disease. The pre-processing phase uses the bagging algorithm which trains the base classifiers generated by different bootstrap samples which is generated by uniformly sampling n instances from the healthcare dataset with replacement. The NN classifier is applied on the generated each bootstrap sample. The final classifier NNE is applied on bootstrap samples whose output is the class with majority of votes. In the prediction phase of NeuroTree, the generated training set is given as input to the extended C4.5 algorithm. According to the generated training set, gain ratios for each of the attributes have been calculated. Based on the gain ratio the tree is constructed. It consists of choosing an appropriate test attribute for each node and a corresponding class label for each leaf node. The tree thus obtained is NeuroTree. A ruleset is formed by following each individual path in NeuroTree.

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

NeuroTree Algorithm

Input: Healthcare training data $D = (x_1, y_1), \ldots, (x_n, y_n)$, extra data ratio μ, trials of bagging sampling B, number of records in the training set n.

Output: Set of rules.

1. Train the NN from D via Bagging. Call the procedure NNE = Bagging (D, NN, B).
2. Generated training set $D' = \emptyset$
3. Process D with the trained NNE and classify an instance x_i by counting votes for which NNE (x_i) represents the class with most votes.

   For $(i = 1; i < n; i++)$

   Begin

   a. Replace the class label (y_i) with those output from the NNE $y_i' = \text{NNE}(x_i : (x_i, y_i) \in D)$

   b. Include the new samples to the generated training set $D' = D' \cup \{x_i, y_i'\}$

   End
4. Read data from the NNE generated training set D'
5. Tokenize each record and save it in an array
6. Find the probability of occurrence for each value for each class
7. Find the entropy using following equation
   $Inf(P) = -(p_1 \times log(p_1) + p_2 \times log(p_2) + \cdots + p_n \times log(p_n))$
8. Calculate the information gain
   $Gain(X, D') = I(D') - I(X, D')$
9. Calculate the gain ratio using following equation

$$GainRatio(X, D') = \frac{Gain(X, D')}{SplitInfo(X, D')}$$

Where $SplitInfo(X, D')$ is the information due to the split of D' on the basis of the value of the categorical attribute X.

10. Construct extended decision tree with the highest $GainRatio$ attribute as the root node and the values of the attribute as the arc labels.
11. Repeat the steps 6 to 10 until categorical attributes or the leaf nodes are reached.
12. Derive rules following each individual path from root to leaf in the tree.
13. The condition part of the rules is built from the label of the nodes and the labels of the arcs, the action part is the classification (absence or presence of disease).

Procedure Bagging (D, NN, B)

Begin

for $(i = 1; i < n; i++)$

{

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

a. Create new training set of sizes n with replacements for each B trials D_i =bootstrap sample from D.
b. Create a classifier NN_i for each training set. Call procedure NeuralNetwork(D_i)

}
From NNE classifier by aggregating the $'B'$ classifiers
Return trained NNE
End
Procedure NeuralNetwork$(Training\ Set D_i)$
Begin

1. Get input D_i for training.
2. Read data from D_i
3. Initialize weights and bias to random values.
4. Compute the output for every neuron from the input layer, through the hidden layer to the output layer.
5. Compute the error at the outputs.
6. Use the output error to compute error value for hidden layer.
7. Use the calculated error value to compute the weight adjustments.
8. Apply the weight adjustments.
9. Repeat until the error value doesn't change in consecutive 5 iterations.
10. Return trained NN.

End

The rule structure used by Random Forest (RF), Bayesian Tree and NeuroTree is the following:

$$Rule_i: \text{If } x_1 \text{is } A_{i1} \text{ and... and } x_n \text{is } A_{in} \text{ then Class } = C_i \text{ with } RuleW_i$$

Where $Rule_i$is the label of the i^{th} rule, $x = (x_1, \ldots x_n)$ is n dimensional healthcare data, A_{ij} is a linguistic label modeled by a triangular membership function, C_i is the class label (presence or absence of disease) and $RuleW_i$ is the rule weight of i^{th} rule.

$$RuleW_i = \frac{\sum_{x_m \in C_i} \delta_{A_i}(x_m) - \sum_{x_m \notin C_i} \delta_{A_i}(x_m)}{\sum_{m=1}^{M} \delta_{A_i}(x_m)}$$

Where $\delta_{A_i}(x_m)$ is the matching degree of the example x_m with the antecedent part of the fuzzy rule $Rule_i$ which is computed as follows:

$$\delta_{A_i}(x_m) = \prod_{j=1}^{n} \delta_{A_{ij}}(x_{mj})$$

Where $\delta_{A_{ij}}(x_{mj})$ being the membership degree of the value x_{mj} of the fuzzy set A_{ij} of the rule $Rule_i$. To build the rule base, the following learning algorithm is applied which consisted of two steps are construction of linguistic labels and generation of fuzzy rule for

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

each example. In the construction of linguistic labels, fuzzy sets are built with the same triangular shape and equally distributed on the range of values. In the generation of a fuzzy rule for each example, the membership degree of each value x_{mj} to all different fuzzy sets of j^{th} variables are computed. Then, the linguistic label is selected for each variable which has the greatest membership degree. The rule which has the highest rule weight is kept for prediction of leukemia cancer, lung cancer and heart disease.

3.2 Homogeneous and Heterogeneous way of disease prediction

In each mapper, classifier is employed to predict the disease. In this manner multiple classifiers are concurrently obtained in different mappers. The learning process of each classifier is performed considering the selected features associated with the mapper. Subsequently, the rules and weights obtained in each classifier are generated using a subset of training set. At the end of this prediction, the map function stores all the input examples in the main memory and the cleanup function is where the RF algorithm is applied to build the rule base. Once all classifiers finished their learning phase as many rule base as mappers are obtained. After the generation of rule base, the generated rule bases are aggregated in a single reducer in order to provide final rule base. All the rules are added to the final rule base. When two or more roles share the same antecedent part, only the one having the highest rule weight is kept in the final rule. If the RF classifier is used in all mappers and then their rules are aggregated in the reducer, then it is called as homogenous way of disease prediction. If the different classifier such as RF, Bayesian Tree and Neurotree are used in mappers and their results are aggregated in reducers, then it is called heterogeneous way of disease prediction.

4. RESULT AND DISCUSSION

In this section, the performance of the proposed Fuzzy rule based system is tested in terms of accuracy, precision, recall and F-measure. For the experimental purpose, three datasets leukemia cancer, lung cancer and heart disease dataset are collected. The leukemia dataset was taken from a collection of leukemia patient samples reported by Golub. The dataset consisted of 72 samples: 25 samples of AML, and 47 samples of ALL. Each sample is measured over 7,129 genes. The lung cancer dataset contains 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32 of them, 16 MPM and 16 ADCA. The rest 149 samples are used for testing. Each sample is described by 12533 genes. Both leukemia cancer dataset and lung cancer dataset are available in http://cilab.ujn.edu.cn/datasets.htm link. The heart disease dataset is Cleveland database which consists of 76 attributes and it is available in http://archive.ics.uci.edu/ml/datasets/heart+Disease link.

4.1 Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among total number of cases examined.

Accuracy can be calculated from formula given as follows

European Journal of Business & Social Sciences
Available at https://ejbss.org/
ISSN: 2235-767X
Volume 07 Issue 6
June 2019

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positive + False\ negative}$$

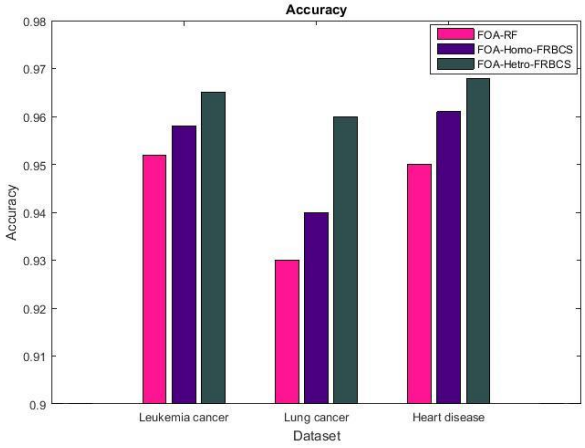


Figure.1 Comparison of Accuracy

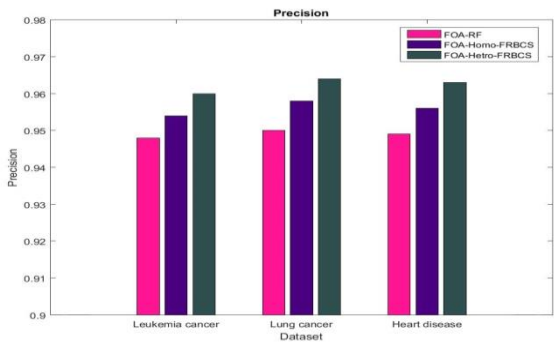
Figure 1, shows the comparison between existing Firefly Optimization Algorithm with Random Forest (FOA-RF) and proposed FOA with Homogeneous Fuzzy Rule Based Classification System (FOA-Homo-FRBCS) and FOA with Heterogeneous Fuzzy Rule Based Classification System (FOA-Hetro-FRBCS) in terms of accuracy. Three different dataset such as leukaemia cancer, lung cancer and heart disease dataset are taken in X-axis and accuracy is taken in Y-axis. From the figure 1, it is proved that the proposed FOA-Hetro-FRBCS based disease prediction has better accuracy than the other methods.

4.2 Precision

Precision value is evaluated according to the prediction of disease at true positive prediction, false positive.

$$Precision = \frac{Truepositive}{(Truepositive + Falsepositive)}$$

Figure 2, shows the comparison between existing FOA-RF and proposed FOA-Homo-FRBCS and FOA-Hetro-FRBCS in terms of precision. Three different dataset such as leukemia cancer, lung cancer and heart disease dataset are taken in X-axis and precision is taken in Y-axis. From the figure 2, it is proved that the proposed FOA-Hetro-FRBCS based disease prediction has better precision than the other methods.



ISSN: 2235-767X
Volume 07 Issue 6
June 2019

Figure.2 Comparison of Precision

4.3 Recall

The Recall value is evaluated according to the prediction of disease at true positive prediction, false negative.

$$Recall = \frac{Truepositive}{(Truepositive + Falsenegative)}$$

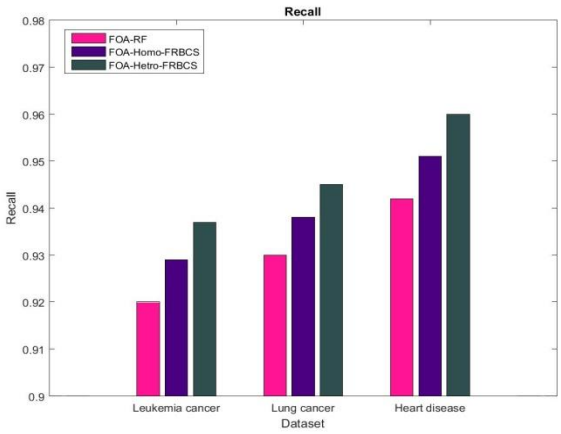


Figure.3 Comparison of Recall

Figure 3, shows the comparison between existing FOA-RF and proposed FOA-Homo-FRBCS and FOA-Hetro-FRBCS in terms of recall. Three different dataset such as leukemia cancer, lung cancer and heart disease dataset are taken in X-axis and recall is taken in Y-axis. From the figure 3, it is proved that the proposed FOA-Hetro-FRBCS based disease prediction has better recall than the other methods.

4.4 F-measure

F-measure is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. It is defined as follows:

$$F - measure = 2.\frac{Precision\ .\ Recall}{Precision\ +\ Recall}$$

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

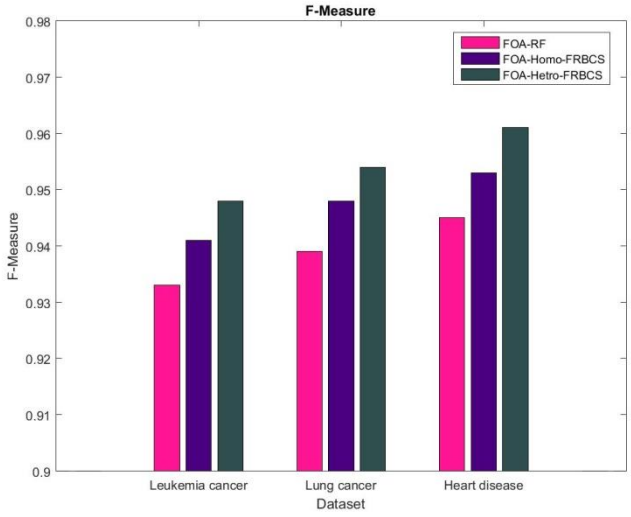


Figure.4 Comparison of F-Measure

Figure 4, shows the comparison between existing FOA-RF and proposed FOA-Homo-FRBCS and FOA-Hetro-FRBCS in terms of F-Measure. Three different dataset such as leukemia cancer, lung cancer and heart disease dataset are taken in X-axis and F-Measure is taken in Y-axis. From the figure 4, it is proved that the proposed FOA-Hetro-FRBCS based disease prediction has better F-Measure than the other methods.

5. CONCLUSION

In this paper, the disease prediction using Firefly Optimization Algorithm with different classifier is extended by using classifiers in distributed environment. It is achieved by exploring Fuzzy Rule Based Classification System where Random Forest, Bayesian Tree and NeuroTree are used to generate set of rules for prediction of leukemia cancer, lung cancer and heart disease. The classifiers are employed in each mapper of MapReduce and the rules are combined in the reducer of MapReduce. If the rules of multiple RF classifier are combined in reducer which is called as homogenous way of disease prediction and the rules of different classifiers (RF, Bayesian Tree and NeuroTree) are combined in reducer which is called as heterogeneous way of disease prediction. The experiments are carried out in three datasets are leukemia cancer, lung cancer and heart disease dataset in terms of accuracy, precision and recall and it is proved that the proposed method has better performance than the existing method.

References

1. Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.
2. Ţăranu, I. (2016). Data mining in healthcare: decision making and precision. *Database Systems Journal*, *6*(4), 33-40.

ISSN: 2235-767X
Volume 07 Issue 6
June 2019

3. Nagarajan.VR., Dr. D.Vimal Kumar, (2018) An Optimized Sub Group Partition based Healthcare Data Mining in Big Data. *International Journal for Innovative Research in Science & Technology*, Issue 10, Volume 4, 79 – 85.
4. Zeinalzadeh, A., Wenska, T., & Okimoto, G. (2017). A neural network model to classify liver cancer patients using data expansion and compression. In *American Control Conference (ACC), 2017* (pp. 2135-2139). IEEE.
5. Bharathi, H. & Arulanath, T. S. (2017). A Review of Lung cancer prediction System using Data Mining Techniques and Self Organizing Map (SOM). In *International Journal of Engineering Research, 12(10),* 2190-2195.
6. Daqqa, K. A. A., Maghari, A. Y., & Al Sarraj, W. F. (2017). Prediction and diagnosis of leukemia using classification algorithms. In *Information Technology (ICIT), 2017 8th International Conference on* (pp. 638-643). IEEE.
7. Purwanti, E., & Calista, E. (2017). Detection of acute lymphocyte leukemia using k-nearest neighbor algorithm based on shape and histogram features. In *Journal of Physics,* 853(1), 012011). IOP Publishing.
8. Vivekanandan, T., & Iyengar, N. C. S. N. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Computers in biology and medicine*, *90*, 125-136.
9. Mustaqeem, A., Anwar, S. M., Khan, A. R., & Majid, M. (2017). A statistical analysis based recommender model for heart disease patients. *International journal of medical informatics*, *108*, 134-145.
10. Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), 27-40.