Algorithms Used for Optimizing K-Means for Heart Disease Diagnosis

Shimpli Borkar[1], Sonali Dobhal[2], Priyal Dharmawat[3], Shloka Harne[4]
[1,2,3,4] *Student, B.Tech Computer Science Engineering, Narsee Monjee Institute of Management Studies, Mukesh Patel School of Technology, Management and Engineering, Shirpur, M.H, India*

***Abstract*- The heart is a significant organ of the human body. Life is reliant on the proper functioning of the heart. Most of the time is difficult to analyze a patient as a heart patient. For this purpose, data mining can be used to recognize a hidden clinical dataset. In this paper, we study to find ways to optimize the K-Means algorithm by overcoming its drawbacks which may help create a heart disease predicting system by applying it. Here, we present a study on the advanced data mining techniques and hybrid algorithms that could be used to optimize the K-Means and increase the prediction accuracy of the system of Heart disease prediction.**

***Index terms*- Heart disease, Data Mining, Clustering, K-Means.**

I.INTRODUCTION

At the age of 30 or above, cardiovascular disease is a typical problem that can be found in every single person. Due to change in our daily routine, there are numerous such factors, for example, smoking, liquor, cholesterol level, heftiness, hypertension, diabetes, and so forth which are accountable for the threat of having a heart disease. There is a survey conducted by WHO in 2016 which states that about 17.9 million people died worldwide due to Heart diseases and most of the people belonged to developing countries. Medical sciences and artificial intelligence we can be used to diagnose these kinds of diseases easily. Information mining is the procedure of extricating concealed data from a huge arrangement of the database. Machine Learning and data mining are the two techniques which are becoming more popular for analyzing data. It can be done by identifying important attributes in a dataset and then using them to recognize patterns in a clinical dataset which can be used to predict various heart diseases. Clustering is a process by which we can group data into clusters based on some similarity measure. Clustering recognizes patterns in large datasets and helps in decision making. Clustering depends upon various factors, for example, the likelihood of clustering, the number of clusters formed in the process of clustering, and the nature of clusters that are framed. K-Means Algorithm is one of the most simple and popular partitions based clustering algorithm. It was proposed by Mac Queen in 1967. In the beginning, centroids of clusters are chosen randomly. Clusters are then created based on the minimum distance between data points and centroids of the clusters. K-Means clustering algorithm has certain disadvantages such as the number of clusters that has to be specified in advanced and it gets stuck at local optima rather than global optima. This is caused due to centroids which are randomly chosen at the start. To overcome these disadvantages various hybrid algorithms have been introduced in the past few years. These hybrids have provided better results than K-Means. These hybrid algorithms have been studied in this paper.

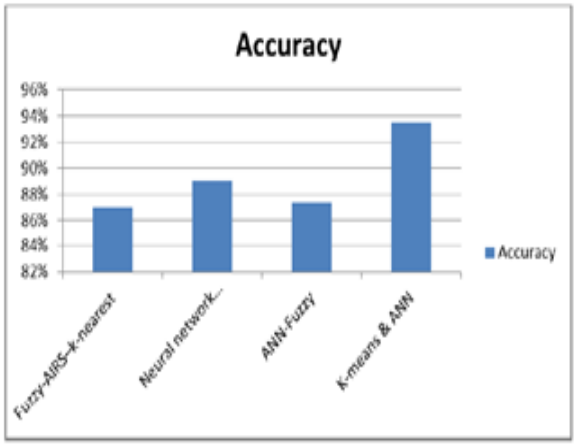
II. REVIEW

There have been several papers which have studied various algorithms for predicting heart diseases. Next few paragraphs describe those papers:

Reetu Singh, E.Rajesh [1], in their paper have used data mining algorithms to identify patterns in a clinical dataset and predict the occurrence of heart diseases. Various algorithms are applied and compared with each other to find an algorithm with the highest accuracy. K-Means algorithm is used for data clustering and its one of the major drawbacks includes a random selection of centroid which highly affects its clusters. Here, the author has enhanced the Euclidean distance formula to overcome k-means drawback and improving the quality of clusters. And this can be done Normalization. Normalization is a pre-processing technique which calculates best

Euclidean Distance between data points in a dataset that results in more accurate clusters which in turn increases the accuracy and efficiency of clusters. In this paper, two attributes from the dataset are plotted Patients cholesterol and age. The dataset is divided into two clusters and the K-Means algorithm is applied. Further, Logistic Regression is used for better classification of data points and after this Normalization is used to compute the best Euclidean distance for the K-Means algorithm. Accuracy of Simple K-Means is 70.58% while the accuracy of K-Means with Normalization is 84.84%. This showed that Normalization not only enhances the quality of clusters but also increases the accuracy of classification.

Amita Malav, Kalyani Kadam and Pooja Kamat[2] in their paper have combined K-Means and Artificial Neural Network techniques of data mining to achieve higher prediction accuracy for predicting heart diseases. In this model, clustering is done with the help of K-Means followed by classification done by the help of Artificial Neural Network. Multilayer Perceptron type of neural network is used by the authors. In this paper, the author has used K-Means and ANN data mining technique to analyse UCI Heart Disease dataset. In the proposed method, Cleveland dataset is loaded into the system from which important attributes which pertain to the risk of heart diseases are identified. Then this data is sent for pre-processing in which those important attributes are set aside. The drawback of ANN leading to longer training times was overcome by K-Means since it converges faster as compared to any other ways. Using this combination, the authors could achieve considerably good precision and sensitivity.



Majhi and Biswal[3], in this paper, proposed a hybrid approach which consists of K-Means and Antlion Optimizer which is a nature-inspired algorithm. Its performance is further compared with the performance of various clustering algorithms such as K-Means, KMeans-FA, Revised DBSCAN, KMeans-PSO and DBSCAN. Antlion Optimizer is one of the most popular nature-inspired algorithms which is inspired by the hunting behaviour of antlion larvae. To catch its prey, antlion digs ups a conical shape hole with its enormous jaw. Then it waits inside that hole for its prey. As soon as the prey gets trapped in that pit antlion throws sand out so that its prey gets slipped into the hole and eats it. Here, antlions are the centroid and ants are the data points. Based on their fitness value antlions are selected as elites using Roulette Wheel Selection method. Random movements of ants around antlions and elite helps in updating the positions of ants. The fitness of ants are also calculated and if its fitness is more than any antlion then it gets replaced by it. Also if the fitness value of an antlion is more than an elite then also replacement takes place. This hybrid approach is applied to different datasets and its performance is calculated in terms of intracluster distance and f-measure. In best quality clusters intracluster distance should be minimum and f-measure should be maximum. Its performance is compared with other algorithms such as K-Means, KMeans-FA, Revised DBSCAN, KMeans-PSO and DBSCAN and results show that KMeans-ALO performs better than them in various situations.

V.R.Geetha, Dr.K.Rameshkumar [4] state that it is known that the K-Means algorithm is efficient for clustering but it does have some drawbacks, one of them being the random initializing of centroids. When random centroids are initialized during each run in the K-Means algorithm, different clusters are obtained each time for the same dataset. For practical applications, this cannot be relied upon. In this paper, the authors propose a procedural method for the initialization of the centroids, to begin with, the algorithm of K-Means. The goodness of the clusters is measured using the intra-cluster distance. Reduced intra-cluster distance means higher the goodness of the clusters i.e. tighter clusters. This improves the quality of clusters.

The proposed algorithm is as follows-

PROPOSED ALGORITHM:

1. Input the Dataset.
2. Choose the number of clusters.
3. Find the Distance between every data point and the origin. The distance is calculated using the Euclidean Distance or Manhattan Distance.
4. Sort the data points based on the above calculated Distance.
5. Divide the sorted data points into K number of clusters chosen.
6. Pick the middle data point from each resulting cluster and make it the initial cluster heads.
7. Find the distance between each data points to the above chosen centroids.
8. Assign data points to the cluster whose distance is minimum.
9. Re-compute the Cluster mean once allocated.
10. Repeat the above process for all other data points in the set.

The authors have applied this on 2 datasets namely Diabetes dataset and Cars dataset. A graph has been plotted to observe and compare the k-means and the modified K-Means effect on the intra-cluster distance reduction. By using this method there has been proven reduction in the intra-cluster distance i.e. improved "goodness of clusters".

The paper mainly focuses not on performance based on runtime but the achievement of good or quality clusters. The authors also suggest that the proposed method works good for small datasets and may be used for larger applications in the future.

In this paper [5], the author has proposed a new algorithm known as ABCGA (Adaptive Biogeography Clustering-based Genetic Algorithm) which is a combination of clustering-based genetic algorithm and ABPPO (Adaptive Biogeography based Predator-Prey Optimization). This algorithm provides better accuracy than the K-Means Algorithm as well as the Hybrid K-Means Algorithms. This algorithm is quite similar to a basic K-means algorithm but to achieve global optimum result ABPPO is used for selection. Further other steps of Genetic Algorithm like crossover, mutation and polygamy are performed. The performance of ABCGA is calculated in terms of computation time, purity and recall. Its performance is then compared with K-Means and other existing hybrid k-means algorithms and it can be seen that ABCGA is better than other hybrid algorithms in terms of computation time, purity and recall.

Simon Fong, Suash Deb, Xin-She Yang and Yan Zhuang [6], in their paper the drawback of K-Means getting stuck at local optima has been tackled with. The authors suggest integrating the nature-inspired algorithms with the K-Means to obtain a global optimum result. Nature-inspired optimization algorithms namely bat, firefly, cuckoo, and ant search are of swarming behaviour and have been explained in the paper as to how they can be used to improve K-Means and drag it closer to achieving a global optimum result. The authors show the results of the evaluation experiments that they have performed on these hybrid algorithms. These hybrids show outstanding performance. There are 2 sets of experiments performed in this paper. The first experiment includes application of these algorithms to 6 different datasets and 3 general-purpose multivariate datasets to measure different parameters such as CPU time, accuracy, fitness value of objective function etc of these algorithms. The results of all the algorithms are better than that of the K-Means For the second set of experiments, the algorithms are applied for image segmentation in which these algorithms perform better as compared to K-Means. The results that have been obtained from the performed experiments show clearly that these nature-inspired algorithms have a performance enhancement property that could be used for enhancement of K-Means.

Seyadali Mirjalili [7] proposed a nature-inspired algorithm known as Antlion Optimizer which is inspired by the hunting behaviour of antlions and ants getting trapped in the trap set by the antlion. The performance of this algorithm was standardized on 19

test functions in terms of convergence rate, exploration, fitness improvement of the population, exploitation, trajectory of antlions, local optima avoidance, the trajectory of antlions, and search history. During the optimization, the average fitness of antlions are calculated and it can be seen that ALO has improved the quality of initial solution also it gives us best result globally by performing efficiently in all parts of search space. The Antlion Optimizer was compared with other well-known algorithms such as SMS, FPA, FA, PSO and GA, BA and CS. The results show that ALO performed better than other algorithms in most of the test functions. Here, ALO is used to solve three major engineering problems and also challenged the CFD problem. The performance of ALO under these situations was better than various algorithms that are part of literature.

Sairabi H. Mujawar, P. R.Devale [8] have applied modified K-Means along with naive Bayes approach for the creation of a system for the detection of heart diseases. This approach has been applied to the Heart Disease data set from the UCI repository, consisting of 303 records. The modified K-Means proposed by the authors removes the randomness of initialization of the centroids by calculating the sum of attributes of each record and then initializing the values of the centroids as the two farthest points achieved by the calculation of sum i.e. maximum and minimum sum records. By this, the two most dissimilar clusters are created. There is no need for the input of the number of clusters by this technique. Naive Bayes rule helps in creating a model possessing predicting capabilities. This also identifies the characteristics of the patient who has heart disease. The total precision and recall of the model obtained are 91% and 75.83% respectively.

Algorithm	Advantages	Disadvantages
Initializing the centroids by alternative approach[4]	Reduces intra-cluster distance.	Works for small dataset. Uncertain about working on large dataset.
Naive Bayes and Modified K-Means[8]	If independence of attributes is true then works more efficiently than other algorithms and there is no need to input number of clusters.	Data might be independent rarely in real life so there may be a loss of accuracy.
K-Means and ANN [2]	Gives high accuracy.	It is equipment dependent.
C-bat, C-ACO, C-cuckoo,C-firefly hybrid with k-means respectively[6]	Enhance performance of K-means steering it to a global optimum.	Slow convergence Can be overcome by k-Means.
Hybrid K-Means and Ant Lion Optimizer[3]	Faster convergence and thus provide global optimum result.	In large scale problems local optimum may arise.
ABCGA (Adaptive Biogeography Clustering based Genetic Algorithm) [5]	Requires less computation time.	May or may not perform with large data sets
The Ant Lion Optimizer[7]	Gradient-free Algorithm and can be applied to real life problems.	It is unable to solve multi- and many-objective problems.

III. CONCLUSION

From all the above-studied papers it can be said that K-means alone cannot perform with good accuracy for detection of heart disease as it has a lot of drawbacks. The drawbacks include manually choosing the number of clusters, clustering of

outliers, randomly selecting centroids initially, high intra-cluster distance leading to clustering of dissimilar data points. K-means needs to be optimized using other algorithms so that its drawbacks can be overcome.

A lot of work has been studied in which the heart disease predicting system could be designed using different advanced data mining techniques along with K-Means. It has also been studied that the nature-inspired algorithms have been performing very well with K-Means and are an emerging field of algorithms these days. These algorithms have not only improved the accuracy but also the time complexity thus also improving the overall performance of K-Means. Nature-inspired algorithms have also proven to give outstanding results for other applications as well. As accuracy and precision are very crucial in the field of healthcare hence, we may use a Hybrid of K-Means and nature-inspired algorithms which may perform better than the other techniques which have already been implemented for the heart disease diagnosis system. This should be done to prevent any catastrophic events with diagnosis errors. Hence using the above-mentioned hybrid, heart disease detection could be fast and accurate.

REFERENCES

[1] Reetu Singh E. Rajesh," Prediction of Heart Disease by Clustering and Classification Techniques", International Journal of Computer Sciences and Engineering 2019.

[2] Kalyani Kadam, Amita Malav," A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K-means", International Journal of Engineering and Technology 2017.

[3] Santosh Kumar Majhi, Shubhra Biswal," Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer", Karbala International Journal of Modern Science 2018.

[4] DR.K.RameshKumar, V.R.Geetha," Initializing Centroids For K-Means Algorithm –An Alternative Approach", International Journal of Pure and Applied Mathematics 2018.

[5] Navreet Kaur, Shruti Aggarwal," Designing a New Hybrid K-Means Optimization Algorithm", International Journal of Advanced Research in Computer Science 2017.

[6] Simon Fong, Suash Deb, Xin-SheYang, Yan Zhuang," Towards Enhancement of Performance of K-Means Clustering Using Nature-Inspired Optimization Algorithms", The Scientific world journal 2014.

[7] Seyedali Mirjalili," The Ant Lion Optimizer", Elsevier 2015.

[8] Sairabi H. Mujawar, P. R. Devale," Prediction of Heart Disease using Modified K-means and by using Naive Bayes", Vol. 3, Issue 10, October 2015

[9] Sudhir Singh and Nasib Singh Gill, "Analysis And Study Of K-Means Clustering Algorithm", Vol. 2 Issue 7, July – 2013

[10] Mirpouya Mirmozaffari, Alireza Alinezhad, and Azadeh Gilanpour, "Heart Disease Prediction with Data Mining Clustering Algorithms", Vol. 4, Issue 1 (2017)

[11] Amandeep Kaur, Jyoti Arora, "Heart Disease Prediction Using Data Mining Techniques: A Survey", Volume 9, No. 2, March-April 2018

[12] Susmitha K, B. Senthil Kumar, "A Survey on Data Mining Techniques for Prediction of Heart Diseases", Vol. 08, Issue 9 (September. 2018)

[13] Bala Sundar V, T DEVI, N SARAVANAN, "Development of a Data Clustering Algorithm for Predicting Heart", Volume 48– No.7, June 2012

[14] Simon Fong, Suash Deb, Xin-SheYang, Yan Zhuang, "Integrating Nature-inspired Optimization Algorithms to K-means Clustering", Conference Paper • August 2012

[15] Amita Malav, Kalyani Kadam, Pooja Kamat, "Prediction Of Heart Disease Using K-Means and Artificial Neural Network as Hybrid Approach To Improve Accuracy", Vol 9 No 4 Aug-Sep 2017