Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 314-318

http://www.krishisanskriti.org/Publication.html

K-means: Its Various Enhancements and Integration with Classification Techniques in Healthcare Industry

Sandeep Kaur[1] and Sheetal Kalra[2]

[1,2]Department of Computer Science Guru Nanak Dev University, RC Jalandhar
E-mail: [1]kaursandeep116@gmal.com, [2]sheetal.kalra@gmail.com

Abstract—*Healthcare industry generates and utilizes ample volumes of data on day-to-day basis. So, in order to extract only the useful data, data mining became requisite to predict various diseases. Clustering is an important step in data mining for analyzing the data in an efficient manner by dividing it into a set of subclasses known as clusters. The most frequently used clustering algorithm is K-means that partitions the given data into k number of clusters based on the distance between the data item and the centroid of each cluster. But, k-means has some limitations like predefined number of clusters, random initial centroid selection and number of iterations etc. that reduces the performance of the algorithm. To improve the performance of K-means in disease prediction, various enhancements are done in the existing algorithm. This paper reviews the various improved k-means algorithms used in healthcare industry to better predict the chances of various diseases like breast cancer, heart disease, diabetes, lung cancer, blood infection etc. K-means algorithm can also be integrated with some classification techniques to obtain better results. This paper is also a survey of various models in which K-means algorithm is integrated with classification techniques like Naïve Bayes, Support Vector Machines, C4.5 etc. The various techniques discussed in this paper are compared by calculating the improvement ratio for predicting different diseases.*

Keywords*: Data mining; K-means; Advanced K-means; Foggy K-means; Hybridized K-means; C4.5, Support Vector Machines.*

1. INTRODUCTION

Data mining is referred to as extracting the useful amount of information from a very vast amount of data in the database. It is also referred to as 'knowledge mining from data'. Now a day, healthcare industry generates and utilizes large volumes of data on regular basis. So, in order to extract only the useful data, data mining became important step to predict various diseases**.** Data mining helps to find patterns and knowledge in order to predict the disease in an efficient manner [1]. Data mining is used to predict various diseases like cancer, diabetes, and heart disease etc. Algorithms are designed for large amount of data in order to produce only the patterns that are interesting or needed to the user and these algorithms also extract the hidden patterns of the disease [2]

Data mining is an iterative process having the following steps:

1. State the problem.
2. Collect the data related to the problem.
3. Perform preprocessing on the data.
4. Mining the useful data by studying and assessing the model of prediction
5. Adjust the model as needed and draw conclusion.

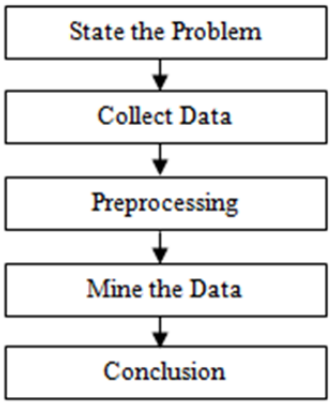

Fig. 1: Data mining process

1.1 Clustering

Clustering is a technique used in data mining for grouping a set of data items into subsets or clusters in such a way that similar type of data items are placed in one type of cluster. There is central point in each cluster that represents that cluster. These cluster centers determine if there is a similarity between clusters [3].

There are basically three types of clustering algorithms: partitioning based, hierarchical and density based clustering. In partition Based Clustering, the dataset containing n objects is partitioned into clusters based on the Euclidean distance. The main aim is to minimize the distance as much as possible. Examples of this type are K-means and K-medoid algorithms. In hierarchical clustering, the nested clusters are formed in the form of a tree. The main advantage of this is that it does not

assume any specific number of clusters. It is of two types: Agglomerative (bottom up approach), in which the clusters have sub-clusters and those sub-clusters have further subclasses and so on; divisive (top down approach), which is in the opposite direction of Agglomerative technique. In density based clustering, the clusters are formed based on the density. It form arbitrary number of and produce clusters as long as the density reaches some threshold value. The most common example for density based clustering is DBSCAN (Density Based Spatial Clustering of Application with Noise) algorithm [4].

K-means: In this paper, we are only concerned about the K-means partition based clustering algorithm.
K-means is a partitioning based clustering algorithm having the data objects with similarities in one cluster based on the Euclidean distance. It is the simplest and most widely used algorithm for clustering. It is an unsupervised learning method that follows an easy procedure for cluster data set. This algorithm always forms *k* number of clusters and in every cluster there must be atleast one item. In the initial step, K-means chooses random centroid points and number of clusters. It then finds the distance of each data point from every centroid. The points are assigned to the centroid at minimum distance from the point. The centroids are recalculated by taking the arithmetic mean of the data points assigned to that cluster [5]. The performance of K-means increases as the number of clusters increases, hence it is advantageous to apply on large datasets.

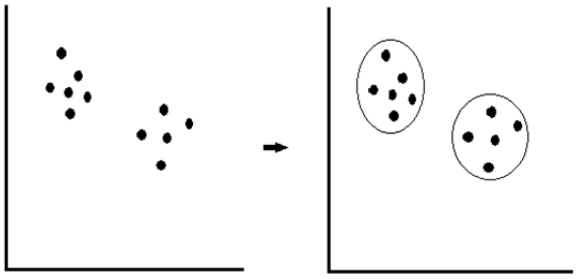

Fig. 2: Partition based K-means Clustering

Algorithm:
Input:
K is the random number of clusters,
D is the data set having n objects.
Output:
'*k*' clusters are formed.
Method:

1. Select k random data objects as the cluster centroids.
2. Assign cluster to each data object to its closest centroid.
3. Recalculate the new centroids for each cluster, by calculating the arithmetic mean of data objects in that cluster.
4. If atleast one data object changes its centroid, then move to step no. 2, else move to step no. 5
5. Output the final clusters.

K-means algorithm has following drawbacks [6]:

- Number of clusters needed to know in advance, but it is possible in real-world applications.
- K-means is sensitive to initial centres selection.
- It may meet to problem of local minima.
- It takes large amount of time to produce the required number of clusters in iterations.

2. RELATED WORK

Mrs. Sandhya G *et al* in 2013 proposed [7] an advanced K-means algorithm for the identification of microcalcifications in case of Breast cancer diagnosis. The improvement is done using Minimum Spanning Tree and the Kruskal's Algorithm. It helps to achieve better accuracy than K-means and also reduces the number of iterations.

A K Yadav *et al* in 2013 purposed [8] a Foggy K-means algorithm for the prediction of lung cancer using the attribute values. For experiment the real time dataset is used from SGPGI, Lucknow. This dataset is discussed with the domain experts to identify the impact of each attribute on lung cancer. The numbers of clusters are determined on the basis of the values of attributes in the dataset.

R Dash *et al* in 2010 proposed [9] a hybridized K-means algorithm that uses Principle Component Analysis (PCA) for the initial phase of K-means. PCA is a feature reduction technique that converts high dimensional data to low dimensional representation.

M. Nishara Banu *et al* in 2014 presented [10] a new model that uses three approaches i.e. K-means, MAFIA and C4.5 for heart disease prediction to obtain better results for prediction. K-means is used to cluster the data points in the dataset and to find the relevant data. C4.5 is used to classify the pattern that is obtained by MAFIA (Maximal Frequent Item set Algorithm).

R Shinde *et al* in 2015 proposed [13] a system that uses an integration of K-means with Naïve Bayes classification technique for heart disease prediction. K-means is used for clustering the heart patient data. Naïve Bayes classifies the data by finding the maximal similarity.

Bichen Zheng *et al* in 2013 proposed [15] a new model K-SVM that is the integration of K-means and Support Vector Machine (SVM) for Brest cancer prediction. K-means is used to find the hidden pattern of the tumor. SVM is machine learning classification technique that classifies the clustered data by drawing the hyperplanes.

3. VARIOUS DISEASE PREDICTIONS USING K-MEANS

K-means is the most frequently used algorithm in healthcare industry to predict various diseases. But, k-means has some limitations like predefined number of clusters, random initial centroid selection and number of iterations etc. that reduces

the performance of the algorithm. To improve the performance of K-means in disease prediction, various enhancements are done in the existing algorithm. In this paper, three enhancements done in K-means in disease prediction are discussed. The three techniques: Advanced K-means, Hybridized K-means and Foggy K-means, are discussed in this paper for predicting various diseased. To improve the performance of K-means in disease analysis, it is also integrated with various classification techniques in various researches. This paper also presents the models in which K-means algorithm is integrated with three classification techniques like Naïve Bayes, Support Vector Machines and C4.5 to obtain better results for disease analysis. These total six techniques of enhancements and integration with classification method are discussed below:

3.1 Advanced K-means

Mrs. Sandhya G *et al* in 2013 proposed [7] an advanced K-means algorithm for the identification of microcalcifications in case of Breast cancer diagnosis. The improvement is done using Minimum Spanning Tree and the Kruskal's Algorithm. It helps to achieve better accuracy than K-means and also reduces the number of iterations. The first step in this proposed algorithm is to calculate the distance of each data point from every other data point in the dataset. This distance is assigned as a weight to the edge between two data points. After assigning these weights for all the edges, Kruskal's algorithm is applied to obtain Minimum Spanning Tree (MST) having k-1 edges by sorting the weights in a decreasing order. The connected data points in MST are taken as the initial centroids. The image from the mammogram is given as an input to proposed architecture, and then the homographic filtering is applied on the image to remove the noise and to identify the effective points based on the frequency. The next step is point processing in which negative image is identified for mammogram image in order to clearly identify the clear image of tissue. After point processing, proposed advanced K-means based on MST is applied on data point and required output is generated. The comparison of K-means and proposed Advanced K-means algorithm is done on the basis of identification of microcalcifications, efficiency and noise detection. The proposed algorithm has shown better results as compared to simple K-means.

3.2 Foggy K-means

A K Yadav *et al* in 2013 purposed [8] a Foggy K-means algorithm for the prediction of lung cancer using the attribute values. There are basically two types of attributes, one is demographic like gender, age etc and the second is diagnosis attributes like smoking, tumor size etc. For experiment the real time dataset is used from SGPGI, Lucknow. This dataset is discussed with the domain experts to identify the impact of each attribute on lung cancer. The numbers of clusters are determined on the basis of the values of attributes in the dataset. For example, if the tumor size is more than 3 than there are chances of having cancer. On the basis of this, two clusters are formed one for having cancer and other for not having cancer. The cluster moves left or right according to the impact of the next attribute on the cluster. The experimental results had shown an improvement in the performance of lung cancer prediction.

3.3 Hybridized K-means

R Dash *et al* in 2010 proposed [9] a hybridized K-means algorithm that uses Principle Component Analysis (PCA) for the initial phase of K-means. PCA is a feature reduction technique that converts high dimensional data to low dimensional representation. Mathematically, it is an orthogonal linear transformation that uses variance for feature reduction. The highest variance data lie on first coordinate, then second highest on second coordinate and so on. It converts correlated variables into uncorrelated variables. These variables are called principle components. Principle components are calculated by using Eigen values of correlated matrix. In this proposed algorithm, the first step is to normalize the dataset by using Z-score values. The Principle Components are calculated from Principle Component Analysis. The unnecessary Principle Components are eliminated. At the last step, K-means is applied on these reduced Principle Components.

3.4 K-means + C4.5

M. Nishara Banu *et al* in 2014 presented [10] a new model that uses three approaches i.e. K-means, MAFIA and C4.5 for heart disease prediction to obtain better results for prediction. K-means is used to cluster the data points in the dataset and to find the relevant data. C4.5 is used to classify the pattern that is obtained by MAFIA (Maximal Frequent Item set Algorithm). MAFIA is used to obtain the maximal frequent data set from the database. The dataset is generated by using depth first search and a tree is generated with is traversed down in steps to find the frequent data. It stores the database in vertical bitmap, in which each bitmap is an item set [11]. C4.5 is a classification technique that classifies the dataset by decision trees. C4.5 chooses the attribute at each node of tree which splits the set into subsets and this splitting criteria is based on normalized information entropy. The attribute having this highest information entropy is chosen to split the set into subsets. It is a weka data mining tool [12]. The experiment [11] using this integration of K-means with C4.5 classification is performed to predict heart disease and the results showed an accuracy of 89%.

3.5 K-means + Naïve Bayes

R Shinde *et al* in 2015 proposed [13] a system that uses an integration of K-means with Naïve Bayes classification technique for heart disease prediction. K-means is used for clustering the heart patient data. Naïve Bayes classifies the data by finding the maximal similarity. It gives the output of whether the patient is suffering from heart disease or not.

Naïve Bayes Classifier is based on Bayes' theorem with naïve independence between the features. It assumes that the value of one feature is independent of the other feature. E.g. a fruit may be considered as mango if it is yellow, oval and sweet in taste etc. This classifier assumes that the color, shape and taste features are independent from each other to predict that the fruit is mango. It is advantageous because it requires a small dataset to predict the required output [14]. The purposed system [13] for heart disease prediction using naïve bayes and K-means increased the efficiency of prediction in terms of ease of model interpretation and accuracy.

3.6 K-means + SVM

Bichen Zheng *et al* in 2013 proposed [15] a new model K-SVM that is the integration of K-means and Support Vector Machine (SVM) for Brest cancer prediction. The breast cancer is recognized by predicting the type of tumor: malignant (that lead to cancer) and benign (that can't be cancerous and can be removed). K-means is used to find the hidden pattern of the tumor. SVM is machine learning classification technique that classifies the clustered data by drawing the hyperplanes. Hyperplane separates positive example and negative examples with maximum margin. Margin is the distance between the hyperplane and the nearest positive or negative example. It produces better results if it is given a reduced dataset after feature extraction as an input. SVM helps to achieve good accuracy of prediction and is most widely used. K-SVM model reduces the elapsed time for prediction without degrading the accuracy. In this model, feature reduction technique is used to reduce the feature from total 32 features to six features. This model achieved the accuracy of 97%.

4. COMPARISON OF DISEASE PREDICTIONS TECHNIQUES

To compare the various techniques discussed in this paper, we have calculated the improvement ratio for each technique. Improvement ratio is defined as a ratio of value of the parameter after enhancement to the value of the parameter before enhancement. We have calculated the percentage of this ratio to better define the amount of improvement for various techniques discussed technique in this paper. This comparison is a survey to order find which research method achieved more improvement than the existing technique in that research. We have calculated the average improvement by considering all the parameters of existing and proposed techniques considered in that particular research.

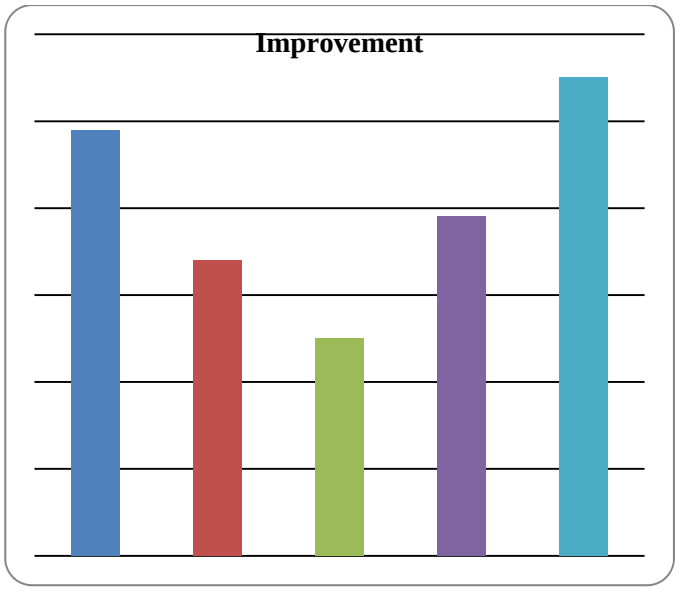

Fig. 3: Improvement graph

Table 1: Comparison of techniques

Method	Disease Predicted	Average Improvement
Advanced k-means	For detection of Breast Cancer tissue in mammography	49%
Hybridized K-means	Diabetes, Breast cancer, heart disease	34%
Foggy K-means	Lung cancer prediction using impact of attributes	25%
K-means + C4.5	Heart Disease, also use MAFIA to obtain maximal frequent data set	30%
K-means + SVM	Breast cancer prediction by also reducing features from 32 to 6.	55%

As the Table and the graph show that the maximum amount of accuracy is achieved by K-SVM algorithm. In K-SVM algorithm, K-means is used to find the hidden pattern of the tumor. SVM is machine learning classification technique that classifies the clustered data by drawing the hyperplanes.

5. CONCLUSION AND FUTURE WORK

Medical field contains large volumes of data, in order to find only the useful information; data mining become an important step in the field of healthcare industry to predict various diseases. K-means is the most frequently used clustering algorithm for data mining in healthcare industry to predict various diseases. In this paper, three enhanced k-means algorithm namely Advanced K-means, Hybridized K-means and Foggy K-means to improve its performance in disease prediction are discussed. This paper also reviewed the models in which K-means algorithm is integrated with three

classification techniques like Naïve Bayes, Support Vector Machines and C4.5 to obtain better results for disease analysis. The percentage of improvement achieved is calculated in each enhancement by considering all the parameters of existing and proposed techniques considered in that particular research. This survey shows that the percentage improvement achieved in case of K-means integration with Support Vector Machine is highest than all the other techniques discussed in this paper.

Our future work is to design a model which is a hybrid of K-means enhancement and Support Vector Machine. The results of the model will than compared with K-SVM model for Breast Cancer prediction.

REFERENCES

[1] G P Dineshgar & Mrs. L Singh (2016, February). A review on data mining for heart disease prediction. International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 5, Issue 2.

[2] Jain, N., & Srivastava, V. (2013). Data Mining techniques: A survey paper.IJRET: International Journal of Research in Engineering and Technology,2(11), 2319-1163.

[3] Joshi, A., & Kaur, R. (2013). A review: Comparative study of various clustering techniques in data mining. International Journal of Advanced Research in Computer Science and Software Engineering, 3(3), 55-57.

[4] Xiong, L. Data Mining: Concepts and Techniques.2008

[5] Teknomo, K. (2006). K-means clustering tutorial. Medicine, 100(4), 3.

[6] http://www.cs.uky.edu/~jzhang/CS689/PPDM-Chapter3. pdf

[7] Sandhya, G., Gowda, S., Swamy, L. N., Raju, G. T., & Vasumathi, D. (2013, December). Automated detection of cancer tissues in mammograms using advanced K-Means clustering with homomorphic filtering. In Circuits, Controls and Communications (CCUBE), 2013 International conference on(pp. 1-4). IEEE.

[8] Yadav, A. K., Tomar, D., & Agarwal, S. (2013, July). Clustering of lung cancer data using Foggy K-means. In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on (pp. 13-18). IEEE.

[9] Dash, B., Mishra, D., Rath, A., & Acharya, M. (2010). A hybridized K-means clustering approach for high dimensional dataset. International Journal of Engineering, Science and Technology, 2(2), 59-66.

[10] Banu, N., & Gomathy, B. (2014, March). Disease Forecasting System Using Data Mining Methods. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 130-133). IEEE.

[11] http://himalaya-tools.sourceforge.net/Mafia/

[12] https://en.wikipedia.org/wiki/C4.5_algorithm

[13] R Shinde, S Arjun, P Patil & Prof. J Waghmare (2015) . An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm. International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 637-639

[14] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[15] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

https://www.d.umn.edu/~gshute/arch/improvements.xhtml