American Journal of Education and Information Technology
2021; 5(2): 93-96
http://www.sciencepublishinggroup.com/j/ajeit
doi: 10.11648/j.ajeit.20210502.15

Disease Prediction Through Syndromes by Clustering Algorithm

Raihana Zannat[1], Shammir Hossain[2], Shakawat Al Sakib[2], Sumaya Akter[2], Khadija Tut Tahera[2], Ohidujjaman[2]

[1]Department Software Engineering, Daffodil International University, Dhaka, Bangladesh

[2]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

Email address:
zannat.swe@diu.edu.bd (R. Zannat), shammir15-1641@diu.edu.bd (S. Hossain), shakawat15-1650@diu.edu.bd (S. Al Sakib), sumya15-1648@diu.edu.bd (S. Akter), khadijatuttahara15-2640@diu.edu.bd (K. T. Tahera), tuhin.iu31@gmail.com (Ohidujjaman)

To cite this article:
Raihana Zannat, Shammir Hossain, Shakawat Al Sakib, Sumaya Akter, Khadija Tut Tahera, Ohidujjaman. Disease Prediction Through Syndromes by Clustering Algorithm. *American Journal of Education and Information Technology*. Vol. 5, No. 2, 2021, pp. 93-96. doi: 10.11648/j.ajeit.20210502.15

Received: September 25, 2021; **Accepted**: October 12, 2021; **Published**: October 28, 2021

Abstract: There are numerous machine learning methods that capable to develop smart automated algorithms to examine high-dimensional and multi-modal biomedical dataset. This paper emphases on through clustering algorithms to advance revealing and analysis of human diseases. The mass population's disease assessment was not ever been familiar and nevertheless is an intricate procedure and necessitates a great level of competence. Numerous assessment support methods established encouraging diagnostic representations however merely an insufficient have been properly estimated in clinical surroundings. Moreover stand-alone decision support systems rely on profoundly on a massive volume of dataset. This research deploy unsupervised clustering approach such as K-means algorithm to build a proficient system to recognize human diseases by assessing syndromes to progress the superiority of health issues. The medical professionals and practitioners can use this smart system to corroborate the diseases diagnosis. The study is significant in health sector to reduce all kinds of diagnosis expenses.

Keywords: Clustering Approach, Data Mining, K-means, Machine Learning, Symptoms

1. Introduction

Data mining is the procedure of determining information and hidden patterns from random and complex data space. The use of data mining or computer-based system is growing in this contemporary in medical diagnosis because of the quantity and complication of data [6]. In this article, K-means unsupervised learning is deployed to evaluate its performance and efficiency in the field of disease analysis and prediction. The primary approach is to analyze and predict 19 individual common diseases. There have been used a medical survey data in this research.

2. Data Preparation

The real-world information is extremely sensitive, sound, absent, and inconsistent. Therefore, pre-processing of data is very important [11]. The utilized data in this study are taken from a survey and a few steps are taken to prepare the dataset to work with as much efficiency as possible. The steps are as follows:

s1: Data encoding

There is used 'One-Hot Encoding' to encode important features. The reason is to choose "one-hot encoding" over label encoding is that there is no need the model to unnecessarily prioritize certain attributes.

s2: Dummy attribute elimination

There have eliminated one dummy attribute to avoid the dummy variable trap.

s3: Discretization

The necessary nominal data are changed into numeric to work with machine learning equations.

s4: Feature scaling

There have been scaled the values of necessary features into the same range so that no feature will unnecessarily dominant over another in the calculation. It is always a crucial thing to take into consideration while working with distance metrics like 'Euclidian distance'.

3. K-means Algorithm

K-means clustering calculates the centroids and iterates up to it finds optimum centroid. K-means algorithm is a partition-based clustering procedure. This is the modest unsupervised clustering algorithm utilized to resolve the cluster-based problems [7]. The data sets are divided into a static number of clusters in K-means clustering method. The representation of K-means algorithm is as follows:

```
Input: data points N, number of cluster k
Output: Data points with cluster memberships
s1: Initialize k centroids randomly
s2: Associate each data point in N with the nearest centroid
This will divide the data points into k clusters
s3: Recalculate the position of centroids
s4: Repeat steps 2 and 3 until there are no more changes in the membership of the data points, otherwise end the process.
```

Merits of k-means clustering algorithm:

i). The foremost merits of k-means is easiness and the speed of k-means allows to run large datasets.

ii). In contrast, k-means is faster than hierarchical clustering algorithm while k is small.

iii). It might produce close-fitting clusters in compare to hierarchical algorithm when the clusters are spherical.

4. Related Works

The authors Sellappan Palaniappan and Rafiah Awang published an article in which established a prototype Intelligent Heart Disease Prediction System (IHDPS) [1]. In this research coronary heart disorder prediction is executed through data mining strategies such as naive bayes, decision trees and neural network. IHDPS can answer complex “what if” queries which traditional decision support systems are not capable to do. It is carried out on the. NET platform. This method is applied for coronary heart sickness prognostic [1].

Shsudhab Adam Pattekari and Asma Parveen had proposed the research which goal is to develop an Intelligent System by data mining modeling system named Naive Bayes [2]. This study recovers concealed data from warehoused database and compares the user values with trained dataset. This model can answer complicated queries for analyzing heart disease and thus assists medical practitioners to prepare intelligent clinical decisions where outdated DSS is failed to do. The research effects on reducing treatment costs [2].

S. Vijayarani and S. Sudha published a research work focuses on predicting diseases from the hemogram blood test dataset through data mining procedure [3]. This study developed a new clustering algorithm named as weight based k-means algorithm for identifying complicated diseases from the hemogram blood test samples dataset. The new algorithm’s performance is assessed in comparing with k-means and hierarchical algorithm [3].

Rebecca Hermon and Patricia A H Williams present a research work named “Big data in healthcare: What is it used for?” and it clarified the convenient components of the huge data approach to health sector. The research conveys a reference line to evaluate the rapid increase of the usage of big data in healthcare and be able to contribute in the thoughtful the extensiveness of big data applications [4].

Jyoti Soni et al. introduced predictive data mining for medical diagnosis. Equal dataset predicted the test results are executed to compare the overall performance of statistical mining methods, and the result reveal that the decision tree outperforms and the Bayesian of the category contains the same accuracy as the decision tree. However thee KNN, the neural networks, classification based clustering are not performing well [5].

M. Umamaheswari and P. Isakki Devi proposed an article indicating the principal objective of the study is to forecast the myocardial infraction utilizing data mining clustering methods. The method can distinguish and select concealed intellect from historical myocardial infraction data sets. This paper has been observed the estimate of heart attack with further amount of input features. The organization exploits medical terms such as gender, age, blood pressure, cholesterol and so more for forecasting the patients who got heart disease [9].

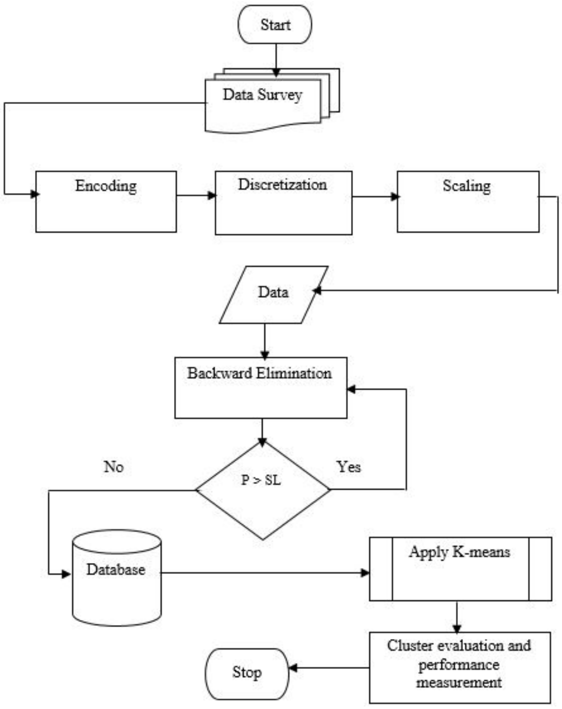

Figure 1. *Proposed Model.*

5. Methodology

In this research work, there are analyzing various kind of syndromes to identify the corresponding diseases. Data mining technique is used to find out the patterns in syndrome dataset. Data is processed and the corresponding diseases are identified with the desired efficiency based on the outcome. To apply the prediction method, features were developed in a model to determine the characteristics of a disease in terms of several health criteria. The proposed model is shown in figure 1. An approach that has taken to create the learning model is as follows:

a) Feature Selection:

To build an optimized machine learning model it is always vital to determine which independent attributes or features have more statistical significant. To find out highly noteworthy features there are four wrapper methods available such as backward elimination, forward selection, bidirectional elimination and score comparison. A machine learning model with all attributes is not always efficient and contains a high probability to build a 'garbage in - garbage out' model. This paper has used 'Backward Elimination' to find highly significant features because it is easy to implement and faster than all other methods available.

Steps in 'Backward Elimination' are as follows:

s1: Choice a significant level (SL) to stay in the model

s2: Fit the full model with all possible predictors

s3: Consider the predictor with the peak P (Probability) value. If P > SL, go to step 4 otherwise finish the process

s4: Eliminate the predictor

s5: Fit the model with the rest of the predictors

b) Applying K-Means:

In this article, K-means is applied to predict the outcome of each point based on every independent selected features when datasets are ready with highly significant features.

c) Evaluation of clusters:

The clustering report is created in this study to assess the competency of the model and there have taken 'Accuracy Score' and 'f1-Score' into consideration as a numerical performance factor. The reason of taking 'f1-Score' over precision and recall is that the data have used in this study are imbalanced and "what f1-Score does" it combines precision and recall as a harmonic means which punishes extreme values and gives a better intuition about the performance of the operation [10].

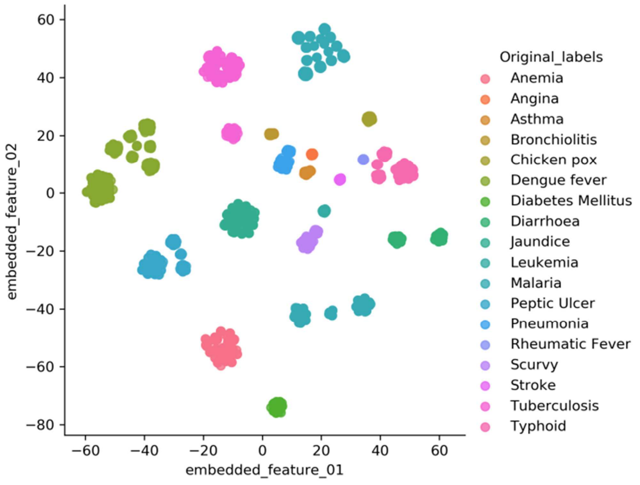

Figure 2. *Data before clustering.*

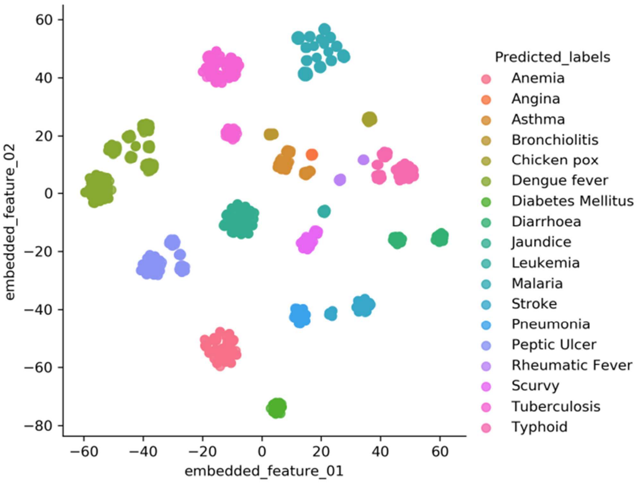

Figure 3. *Data after clustering.*

6. Result Discussion

In this paper, datasets are clustered and analyzed by the similarity measure between each data point and cluster centroid. Clusters are represented by the different colors to identify properly. The figures 1 and 2 show the data before clustering and after clustering respectively.

To plot the data into a two dimensional graph there is used t-SNE (t-Distributed Stochastic Neighborhood Embedding) [8]. A clustering report is presented in Table 1 to give the insight of numerical analytics such as number of observation, accurate predictions, wrong predictions, accuracy score and F1-score.

Table 1. *A clustering report.*

Diseases	No. of observations	Accurate predictions	Wrong predictions	Accuracy (in %)	F1-scores
Anemia	101	101	0	100	1.00
Angina	11	11	0	100	1.00
Asthma	19	19	0	100	0.43
Bronchiolitis	14	14	0	100	1.00
Chicken pox	22	22	0	100	1.00
Dengue fever	202	202	0	100	1.00
Diabetes mellitus	37	37	0	100	1.00
Diarrhea	61	61	0	100	1.00
Jaundice	103	103	0	100	1.00
Leukemia	14	14	0	100	1.00
Malaria	256	144	112	56.25	0.72
Peptic ulcer	111	111	0	100	1.00
Pneumonia	50	0	50	0	0.00
Rheumatic fever	8	8	0	100	0.59
Scurvy	51	51	0	100	1.00
Stroke	11	0	11	0	0.00
Tuberculosis	143	143	0	100	1.00
Typhoid	101	101	0	100	1.00

The calculation summary is as follows:
Total no. of observation: 1315
Accurate predictions: 1142
Accuracy (weighted average) is calculated as below:
= (accurate predictions / total no. of observation)*100%
= (1142/1315)*100%=86.84%
The f1-Score (Weighted avg.) is calculated as follows:

$$F1\text{-score}=2(precision*recall)/\ precision +recall \tag{1}$$

Precision= TP/(TP+FP) =1142/(1142+61)=0.95
Pecall= TP/(TP+FN) =1142/(1142+112)=0.91
From the equation 1 F1-score is as follows:
F1-score=2*(0.95*0.91)/(0.95+0.91)=1.729/1.86=0.93
The terminologies are elaborated as below:
TP =number of true positive
FP= number of false positive
FN=number of false negative

7. Comparative Analysis

It is be concluded from Table 2 that K-means does a very good job in compare with hierarchical clustering containing almost 87% accuracy and 0.93 F1-score. However, the hierarchical clustering carries some bigger issues while working with big data. The obvious disadvantages of the hierarchical clustering are the high in time complexity, the order of the data has a great impact on the outcome and very sensitive to the outliers which is not suitable in this type of research.

Table 2. *Comparison.*

Methods	Precision	Recall	F1-score	Accuracy
K-Means clustering	0.95	0.91	0.93	86.84%
Agglomerative hierarchical clustering	0.94	0.79	0.86	78.71%

8. Conclusion

This article is to analyze the diseases in the medical sector to discover a novel variety of pattern and evidence through clustering algorithm. The proposed machine learning-based system diminishes the impact of humanoid error and enhances the effectiveness of the analysis. The K-means algorithm is fairly useful in predictive analysis of diseases and, it produces a promising outcome about 87% in the experimental session. To make K-means more effective it should be used in incorporation with additional algorithms to produce further accurate, relevant and useful results. However the proposed model in this study is handy enough for physicians and medical practitioners to effectively predict hazardous cases and evaluate accordingly. In the future the advanced clustering algorithm will be used for finding more accurate outcome.

References

[1] Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques", ACS/IEEE International Conference on Computer Systems and Applications, 2008.

[2] Shadab Adam Pattekari and Asma Parveen, "Prediction System For Heart Disease Using Naïve Bayes", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[3] S. Vijayarani and S. Sudha, "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples", Indian Journal of Science and Technology, vol. 8, pp. 1–8, Aug. 2015.

[4] Rebecca Hermon and Patricia A H Williams, "Big data in healthcare: What is it used for?" 3rd Australian eHealth Informatics and Security Conference. Held on the 1-3 December, 2014 at Edith Cowan University, Joondalup Campus, Perth, Western Australia.

[5] Jyoti Soni and et al., "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, vol. 17, pp. 43–48, Mar. 2011.

[6] K Rajalakshmi, S. S. Dhenakaran and N Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research, vol. 4, pp. 2697–2699, Jul. 2015.

[7] B. P. Shantakumar and Y. S. Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network" European Journal of Scientific Research ISSN 1450-216X Vol. 31 No. 4 (2009), pp. 642-656.

[8] B. Sundar V, T. Devi and N. Saravanan,"Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) Vol. 48– No. 7, 2012.

[9] M. Umamaheswari and P. Isakki Devi, "Prediction of myocardial infarction using K-medoid clustering algorithm", IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017.

[10] S. Shinde and B. Tidke, "Improved K-means Algorithm for searching Research Papers", International Journal of Computer Science & Communication networks, ISSN: 2249-5789, Vol. 4 (6), 197-202.

[11] Z. Muhammad, R. Imam, P. Yudi, "Log Classification using K-means Clustering for Identify Internet User Behaviours", International Journal of Compiler Applications, (0975-8887), Vol. 154-No. 3, November 2016.