COMPARATIVE ANALYSIS OF THYROID DISEASE USING FUZZY C AND K MEANS ALGORITHMS

First Author
MELISA TARISA MAKOTAMO, (melisa_makotamo.scsebca@galgotiasuniversity.edu.in),
School of Computing Science and Engineering

Second Author
WESLEY TANAKA GEMU, (wesley_gemu.scsebca@galgotiasuniversity.edu.in), School of
Computing Science and Engineering

Guide
MS HEENA KHERA (Assistant Professor), (heena@galgotiasuniversity.edu.in), Galgotias
University

Abstract

Data mining is a process which involves sorting large data sets and spots patterns and relationships which will help solve problems through data analysis. Data mining techniques plus tools enable big enterprises to forecast future coming trends and is then able to make better decisions make more-informed business decisions hence this study. Thyroid disease is a medical problem that occurs when one's thyroid fails to produce enough hormones. This disease is known to affect everyone of all ages and all genders.in order to identify these disorders are detected by blood tests are taken, and are however difficult to analyze because of the vast amounts of blood samples of data to be forecasted. because of this barrier this study allows us to compare two algorithms to determine the best in results output enabling us to have a quick reaction to these disorders. Thyroid diseases have become the most common especially amongst African in the African continent with continent with 68-72%population affected while 4-6% affected yearly are women between the age of 18-25. The causes of thyroid diseases are different which further leads to different types of thyroid diseases and resulting disorders from just the popular known goiter to a cancerous goiter. The diseases are further classified into two the normal thyroid and the abnormal thyroid. This paper will be of comparison analysis of thyroid diseases using the unsupervised algorithms k means and fuzzy c in the African continent. In addition, the imagining in medical systems for thyroid diseases has a lot of research today. Effects caused by thyroid diseases are known to be uncomfortable and when managed well they may result positively. Hence when it is a simple goiter is can be cured naturally, but when it becomes cancerous then it may result in diseases like myxema coma. In order to cure this the measures like k means or fuzzy c

keywords cluster, fuzzy, kmeans, Africa, thyroid, diseases.

I. INTRODUCTION

Thyroid diseases are the diseases that affect the thyroid of a person's. The thyroid is assessed into different categories which are hyperthyroidism and hypothyroidism. These are often further classified into thyroxine (T4), triiodothyronine

(T3) these come to be an outcome thanks to the hormones during a human which are called thyroid stimulating hormones (TSH). Clustering is that the method of putting some things into two or more different distinct groups. In clustering the paper are going to be that specialize in fuzzy c and k means algorithms. Fuzzy c is an unsupervised algorithm whereby the weather can slot in quite one group. Furthermore, Fuzzy c has many differing types there's fuzzy c means, possibility fuzzy c means clustering, fuzzy possibility c means but, this paper will mainly specialize in Fuzzy c means. The K means algorithm is an unsupervised algorithm that put elements into two distinct groups with every element falling in either A or B groups. the two clustering algorithms during this study to supply the simplest results reliable on thyroid diseases within the African continent. The following k means algorithm and the fuzzy c algorithm shall be run side by side with the provided datasets, their performance are going to be compared based on effectiveness using clustering output. The numbers of datasets and number of clusters are the factors upon which the behavior patterns of both the algorithms are analyzed. Fuzzy c outputs similar as the k means results and continues to use more arithmetic steps when it runs. Diagnosis methods of the thyroid glands involve of 4 stages which are preprocessing of input image, feature selection, feature extraction and have classification, of these stages will assistance is in diagnosis of the disorders because sometimes they're thought to be simple goiters and may later become cancer which can need us to completely remove the thyroid

II. LITERATURE SURVEY

Manish Verma, Mauly Srivastava, Neha …" A Comparative Study of Various Clustering Algorithms in Data Mining”

The author did a research on several algorithms in data Ming and he concluded that the Kmeans algorihm is actually faster than many other algorithms and has been found to produce reliable and better output from the vast datasets

K. Aswathi and A. Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis," 2018

The author proposed the use of a support vector machine which can then be used to classify the thyroid dieses

Nora El-Rashidy, Tamer Abuhmed, Louai Alarabi, Hazem M. El-Bakry, Samir Abdelrazek, Farman Ali, Shaker El-Sappagh, "Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning", Neural Computing and Applications,

The author writes about a model of using non sorting genetic algorithm here he managed to find out how thyroid diseases can be found at a more faster rate by implementing row reduction and selected attributes’ being selected with atleat 3 data mining techiques

Veenita,Kunwar Sai and Sabitha. “A comparative study on thyroid disease detection using K-nearest neighbor and Naïve

The author suggested using many techiques the support vector machine and naïve baise are used and the result of the experiment showed that the k nearest is the best compared to than the naïve

Kirubha.M, Prinitha.R, P.Preethika, A.Samyuktha, "Analysis of Thyroid Disease Using K Means and Fuzzy C Means Algorithm," 2019

These authors proposed a system having the following input image , preprocessing, selection, feature selection and feature extraction they used fuzzy algorithm and it provided the best result for overlapped data

III. THYROID DISEASES

Thyroid diseases are diseases that affect the thyroid gland of a person. The main cause of this problem is not really known by any person though some may say that it is due to iron and iodine. Thyroid diseases are classified into three main parts that is hyperthyroidism, hypothyroidism and normal. The normal thyroid can be easily worked on naturally. Hyperthyroidism is the overproduction of TSH. In this case the T3 and T4 they both increase. Under normal circumstances the normal thyroid T4 would range from 5.0 to 12.0ng/dL in adults and T3 will be from 80-220ng/dL . then Hyperthyroidism may also lead to diseases like the grave diseases which are n immune disorder caused by overproduction of TSH. Thehypothyroidism is underproduction of TSH. This results after the T3 and T4 fall under the normal level. If this is not attended to it may result into myxema coma this is a disease that is a loss of brain function as a result of hypothyroidism. The other diseases that may be caused named infertility, birth defects, heart problems just to name a few. The methods of diagnosing may include image segmentation, algorithms and many others. In this paper we will compare the two algorithms to fuzzy c and k means algorithm to cure and diagnose the problem of thyroid diseases in Africa.

IV. A.AFRICA

The African continent has been on the verge with the thyroid diseases as the Africans people are in the third world countries. The 6-8% of the women on the age of 18-25years are affected yearly with the 60% of the ones above 30years suffering from the thyroid diseases. The main problems that continent are many some are, technological illiteracy and lack of finance which leads the people which leads to lack of machinery and skilled personnel which in this paper will look at the fuzzy c and k means to see which one will be better considering the problems in Africa

V. C.CLUSTERING

Clustering is a method of classifying things into classes or significant groups. This is normally unsupervised aspect. In this paper we are going to be focusing on two algorithms namely k means and fuzzy cwhich will help us put the thyroid diseases into two different categories. There a number of algorithm which can be taken for comparison and analysis of disease The main and same thing between these things is that the algorithm first finds the centroid or in other words a center position and this position before any grouping take place. With the help of this paper we will see how the centroid and grouping is done in k means and fuzzy c clustering algorithm.

VI. D. **IMAGE PROCESSING**

The image processing in medical image segmentation is a very important phase before the application of any algorithm. The image processing consists of levels first the taking of image, preprocessing phase, selection of features phase and then extraction of the features.

In image taking the image is taken by the image using a microscope, microscope or radiology scan. After the taking of the picture the image is resized and reduced without the reduction of the quality and converted to grayscale in other words converted into black and white. At the stage the RAO and the LAO as well as anterior are performed. The RAO and LAO pictures are just taken to consideration to see the picture at different angles before analysis.

DWT is then performed in the selection of feature stage. The data in this stage is broken down into smaller bandwidths or finer frequency. The GLCM is then used for the next stage which is the extraction of features. Here the pixels of the image are calculated their tone or intensity on the gray scaled picture.

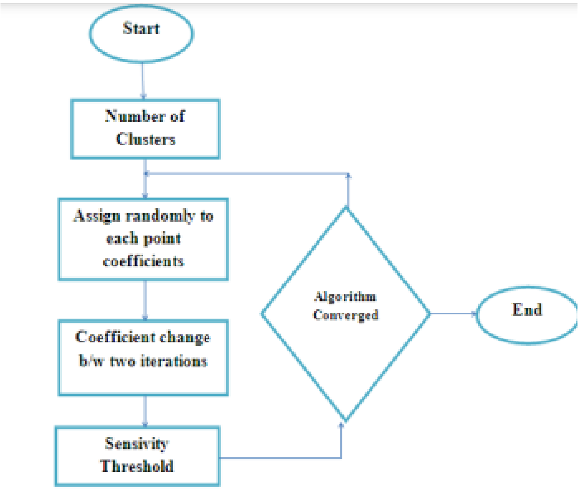
VII.FUZZY C

The fuzzy c is an unsupervised algorithm that works on grouping elements into different groups with some elements falling in either of the groups. Fuzzy c means is an unsupervised learning algorithms which helps in solving the well-known clustering problem. This method may be very useful if the data set proves to have so many behaviors. Such kind of a method is ideal within cluster analysis where information is in three groups or more in data mining. Fuzzy c-means outputs atleast three clusters in grouping just a single dataset. It works in this way it assigs membership to every data point one by one relating to each and every cluster center based on the space between each and every cluster dataset including data points. Fuzzy c function below:

Fuzzy algorithm formula :

$$J_{m}=\sum_{i=0}^{N}\sum_{j=0}^{C} u_{ji}^{m}\left|\left|x_{i}-c_{j}\right|\right|^{2}$$

Where,
n- original number greater than 1,
M -number of data,
C –cluster number,
Uij - membership level xi within the cluster j and i
xi – the ninth item in the data measured dimensionaly-d,
cj -dimension center of the cluster set


Start
Number of Clusters
Assign randomly to each point coefficients
Coefficient change b/w two iterations
Sensivity Threshold
Algorithm Converged
End

The fig above shows the flow of fuzzy c algorithm

Fuzzy clustering algorithm

Step a: Randomly choose a center of a cluster

Step b: Membership of the fuzzy is calculated 'µij' with the help of the Formula:

$$\mu_{ij}=1/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(\frac{2}{m}-1)}$$

Step c: work out the centers of the fuzzy 'vj

'Step d: go back the steps b and c until the smallest or 'J' result is obtained

at,

k is the loop step.

β is taken as the completion criteria amongst points [0, 1].

U = (µij)n*c' is the fuzzy membership matrix. J is the objective function

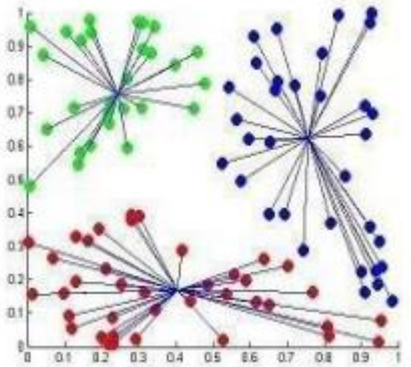


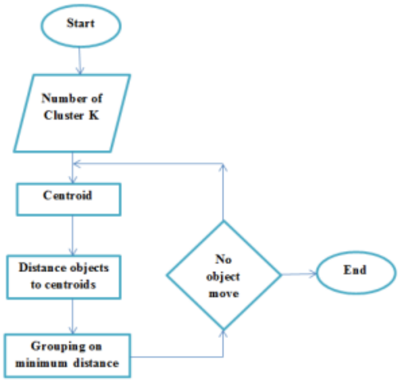
Figure above shows clustering of fuzzy algo

VIII. K MEANS

It is clustering algorithm that is not monitored. It is often used to categories the images and to solve clustering problems. This algorithm groups datasets which are not labelled into variety of different. The algorithm divides a data set into clusters or groups. At first k is taken at a point where its the clusters is equivalent to k and they are fixed throughout the problem. Where center k is defined means that a center is paired with a single cluster. Here the centers is supposed to be at a correct position else the results and positioning will be wrong or different. In order to avoid this crisis clusters should be far away from each other. Afterward we take dataset and match every point with a nearest center. First loop is considered complete where there is no pending point. After, then we to recalculate the new centers of k for the new clusters we got in the step before. A new iteration will have been resulted. Due to the outcome or answers of this iteration we will understand that the new centeroids of k will change the position in each and every loop of the iteration. The loop will keep on continuing s until no more changes are done. function is defined as

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\| x_i^{(j)} c_j \right\|^2$$

where,
||xi - vj|| represents gap between xi and vj.
ci refer to the data points amount happening in the group ith cluster.
c represents the total of centeroids in a cluster.



The fig above shows flow of k means algorithm K means or hard c algorithm

W= {w1,w2,w3,……..,xn} represents the group of data points

T= {t1,t2,…….,tc} represents the centroids in a set manner.

Step a : Randomly pick centeroids of the cluster c.

Step b: Gap between every cluster and datapoint is calculated.

Step c: Give data point to a center whose distance is smaller than any other cluster center .

Step d: now calculate again the new center of a cluster

$$V_i = (1/c_i)\sum_{j=1}^{c_i} X_i$$

using:

ci is representing the total of data sets.

Step e: calculate distances between every point of data and the newly found location of clusters.

Step f: If all the data points do not change then end,
otherwise go back to step c

ci- is considered the total of data points occurring in ninth cluster.

c -is the total of centers within a cluster.

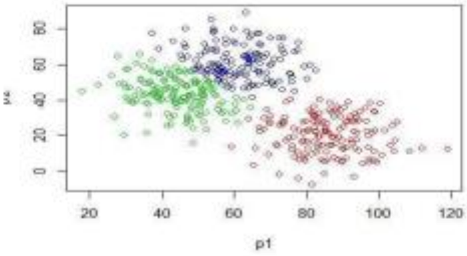
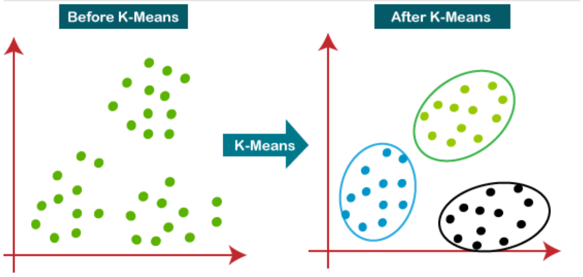


fig shows above shows means clustering



K means algorithm output before and after

.A.DISCUSSION AND COMPARISON

Comparison between the two algorithms fuzzy c means and K Means algorithm

The advantages shown by K-Means clustering

method are as follows simplicity, efficiency or effectiveness, and self-organization. And It is also used as initial process in many other algorithms. The disadvantages are: it is a linearly separating algorithm and hence time consuming. How the algorithm performs is based the centroids placing in the beginning. However, due to many reasons like storage for example the algorithm does not the most conducive solution.

The advantages of the fuzzy clustering algorithm;

clustering methods gives best result for overlapped data set and comparatively better than the K-Means algorithm. Unlike the K-Means where the data point must definitely exclusively belong to one cluster center here the data point is assigned. Membership to each and every cluster center as a result of which data point may be belongi8ng to many cluster center.

IX. CONCLUSION

In the medical department thyroid diseases has been recorded as part of the biggest diseases in the sector and based of this factor, we take the a very serious approach in grouping of thyroid diseases. Our research paper, shows that thyroid disease medical diagnosis results data sets is grouped using the fuzzy c clustering algorithm, and the clustering algorithms are known to have been made in MATLABand their role being for analysis and comparing of the thyroid disease. Based on the research the algorithm which has proven to be a better solution in the analysis of thyroid disease is the k means clustering algorithm. The reason behind is that the algorithm has the ability to identify problems fast and give authentic results. We also found out that the fuzzy c algorithm gives the solutions which are just similar to the other algorithm, as well the fuzzy c is known use many steps and in the analysis of thyroid disease and uses more arithmetic steps than the k means.

These types of algorithms are liable to more complex issues which maybe as follows like vulnerable to many external sources, also having too many steps to follow for a certain process, they also have minimum storage which might not be enough for some of the big calculations. Based on time complexity the time the k means is found to still be the best as compared to the fuzzy c, where k means clustering is 0(nidi) while that of fuzzy c is 0(ndc2i).

Reason to why K means is found to be better than fuzzy c is that it gives reliable results and the analysis of the thyroid disease does not take long because this algorithm uses a few steps and arithmetic procedures, while the fuzzy C clustering algorithm is just giving similar solutions ask means at a slower rate because it involves more steps and arithmetic procedures. the k means clustering has shown to be more efficient and better than the fuzzy in speed since they just give the same results

REFERENCE

[1] Kirubha.M, Prinitha.R, P.Preethika, A.Samyuktha, "Analysis of Thyroid Disease Using K Means and Fuzzy C Means Algorithm" *2019*

[2] Anil and Antony, “An Intelligent System for Thyroid Disease Classification and Diagnosis” IEEE-2018.

[3] Veenita,Kunwar Sai and Sabitha. “A comparative study on thyroid disease detection using K-nearest neighbor and Naïve 2017

[4]Bayes classification techniques” Springer-February 2017.

[5] AiyunZhou,Lili Zhang, Cheng Zhang, “Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images”Springer 2017.

[6] Geetha,SanthoshBaboo., “An Empirical Model for ThyroidDisease Classification using Evolutionary Multivariate Bayseian

[7]Prediction Method”, Global Journals Inc. (USA) 2016.

[8] Yuanyuan Zhang, Min Zuo, “Improved Ensemble Classification

[9]Method of Thyroid Disease Based on Random Forest.”, IEEE -2016.

[10]FerreiraaCarvalho, “Kernel fuzzy c-means with automaticvariableweighting ”, Springer, 2014.

[11] Dai Y, Ru B “ Feature selection of high-dimensional biomedicaldata using improved SFLA for disease diagnosis” IEEE, 2015.

[12]Nora El-Rashidy, Tamer Abuhmed, Loui Alarabi, Hazem M. El-Bakry, Samir Abdelrazek, Farman Ali, Shaker El-Sappagh, "Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning", Neural Computing and Applications, 2021.

[13]Jamil Ahmed Chandio, M. Abdul Rehman Soomrani, “TDTD: Thyroid disease type diagnostics”, Intelligent Systems Engineering, 2016 International Conference

[14]Prasad, T. Sreenivasa Rao, M. Surendra Prasad Babu “Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms”, Springer, 2015

[15]A. Asuncion and D. J. Newman, UCI Machine Learning RepositoryIrvine, CA: University of California, School of Information and computer Science, 2013.

[16] A. K. Jain, M. N. Murty and P. J. Flynn, “Data Clustering: A review”,ACM Computing Surveys, vol. 31, no. 3, 1999.