Comparitive Analysis of K-Means and Fuzzy C Menas on Thiroid Disease

Medasani.Divya
M.Tech (Computer Science and technology),Department of Information Technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India.

Y.Sandeep
Asst. professor, Department ofInformation Technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India.

V. Sindhura
Asst. professor, Department of Information Technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India.

Abstract: To recognize in vast restorative and variation groups with unstructured information is dependably a major test furthermore hazard component to exhibit outside world in an organized configuration. To overcome dependably the information ought to be stemmed and sorted in grouped parts. K-means is utilized as a part of the primary methodology of our as DDC. By taking the mean estimation of age and the groups will be encircled. Be that as it may, the groups are just mean based arrangement so we propose another methodology after K-Means based manufactured FFM. Taken after by that arbitrary markers set to get the achievable consequences of classification furthermore recurrence of appearance regarding allocated irregular scope of interesting qualities to every last existing mix tuple. To accomplish the wanted arrangement of activity we propose another methodology recognized bunching and attainable recurrence with extraordinary result for each procedure. These successions are trailed by non-standard pre-handling like self-cleaning of information and stemming. The pre-handling is finished by semi fluffy system. To accomplish these things of procedure we propose another calculation called DDC (unmistakable relocation for grouping) and UTOF (Unique recurrence result in expandable information. This procedure is absolutely on extensive therapeutic information which is constantly expandable with different new infections. The above calculation takes after a specific new system called doable fluffy mining (FFM) strategy

Keywords: K-Means, DC, UFOE, FFM, fuzzy clustering, frequency.

I. INTRODUCTION

These days in medicinal fields the expansive information examination is dependably a major assignment with step by step developing of patients and their malady. To arrange that new manmade brainpower of information analyzer which synchronizes to convey the information to an appropriate justifiable configuration and simplicity method for presentation. For this vast information to be cleaned and stemmed before applying the information analyzer strategies. In the first place our work is separated into 2 sections one is bunching the information in light of mean quality in view of age which is K-Means and groups encircled and organized information will be populated taking into account the groups range. Second one is to produce the bunches in light of fake recurrence surrounding and exactness ward figuring. This knowledge procedure facilitates to apply recurrence estimation levels took after by brilliant grouping component. The unstructured information which will prompt unmistakable arrangement in organized and keen grouped component. The bunched information can be utilized for recurrence level order furthermore future investigation like expire. Chain of command method which is future upgrade for these proposed

calculations. Our work is completely tested and executed on thyroid information. The principle idea of DDC taking into account K-Means is to group the information in extents as for "AGE" quality of expired patient. Once the grouped result is on structure, this information can be examined utilizing FFM to get the doable recurrence level yield on a combinatorial characteristics i.e. age and ailment. The bunches are encircled with our system on DDC is completely taking into account most extreme age esteem. DDC is casings bunches yet information can't be ordered on different properties. So we recognize information as recurrence as for age and malady yet with fluffy rationale utilizing UFOE and FFM to get the one of a kind attainable results as for task of reach qualities to all combinational tuples of information.

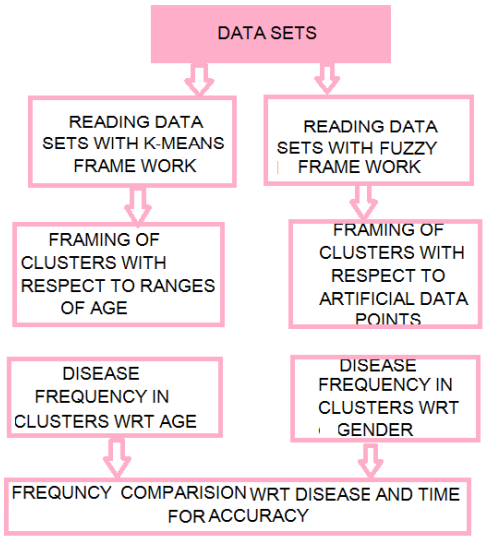

Fig: 1 Flow chart

II. PROPOSED WORK

Our work is completely tested and actualized on thyroid information. The fundamental idea of DDC in view of K-means is to bunch the information in extents as for "AGE" characteristic of perished patient. Once the bunched result is on structure, this information can be examined utilizing FFM to get the doable recurrence level yield on a combinatorial characteristics ie age and sickness. The groups are surrounded with our structure on DDC is completely in view of greatest age esteem. DDC is casings bunches yet information can't be arranged on different characteristics. So we recognize information as recurrence as for age and sickness yet with fluffy rationale utilizing UFOE and FFM to get the one of a kind attainable results regarding task of reach qualities to all combinational tuples of information. By taking the figure 2 every one of the information will be gathered as crude information before preprocessing and our piece of work is to clean the information before applying to handling of grouping. In this level the vast majority of unstructured information will be organized with our casing work before procedure to individual bunching approachs. Every one of the information will be considered as bunches with deference age openings (1 - 10 , 11 – 20 … and so forth). Once these bunches surrounded K –Means will take greatest filled group (without considering GENDER as male or female) and concentrate the number of malady and discovers its exactness. Yet, this is altered for non-developing information. So by giving the one of a kind reach information focuses (for the most part dynamic and one of a kind for every procedure) with fluffy every one of the groups as for those extents encircled with information arrangement as Gender with male and female. So here taking most extreme recurrence group and get the precision of expire in that for thought. In both cases handling time utilization is practically close to coordinating however precision is absolutely more in fluffy.

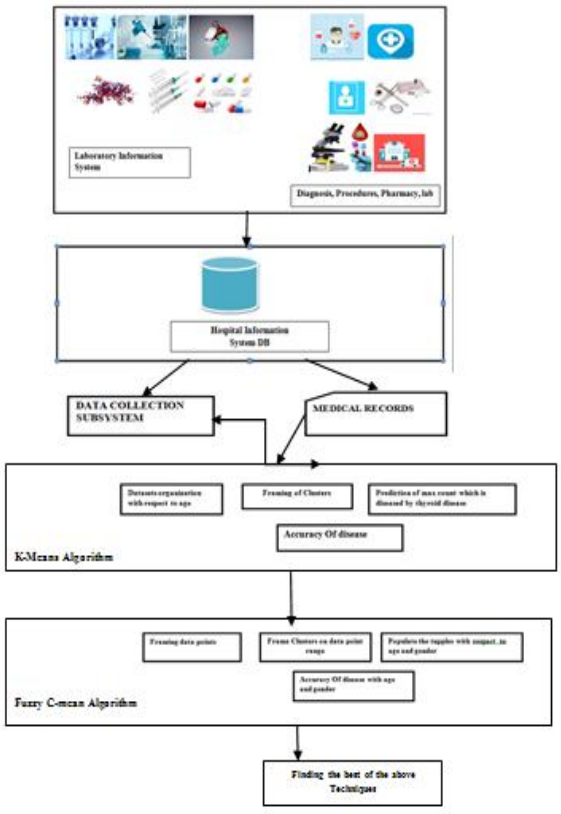

Fig 2: Architecture

III. MINING TECHNIQUES

The contribution for this technique can be either a video document or a video stream from the camera upheld by OpenCV camera interface. The code recognizes Space-Time Interest Points (STIPs) and figures comparing neighborhood space-time descriptors. This rearrangements seems to deliver comparative (or better) brings about applications (e.g. activity acknowledgment) while bringing about an impressive rate up and near video-rate run time. The code distinguishes extraction focuses utilizing edge discovery system and from the extraction focuses order the activity. By the characterization gainn the information from the recordings and perceive the what kind of activity it is.

K-Means:

k means accuracy architecture is shown. Normally datasets will be read with our own buffer readers and maintains the attributed buffers. So if n attributes then our framework maintains n individual buffers to relate the data in reading and representing after mining followed by clustering. Here after framing clusters according to third box of figure K-means framework predicts maximum count of cluster and calculate the accuracy of that particular disease with respect to age and gender considering the attribute “GENDER”.

DDC (Distinct displacement for clustering using kmeans):

In K-Means we utilize limit mean worth is 10. On the off chance that greatest age in our information is 98 then the groups size is as per the following:

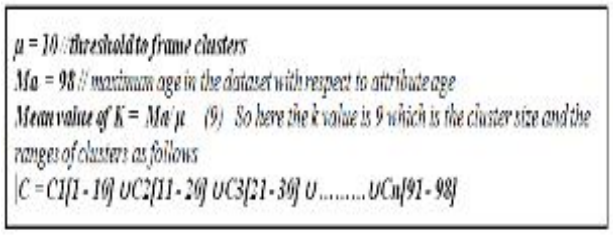

Fig 2: algorithm for K-Means

This essential use is to bunch the substantial information on premise of AGE property. This is to recognize the substantial tuples into sorted out groups. On the off chance that the age is 95(maximum) in our accessible datasets (which can be expandable) then by utilizing the edge of bunching is dependably 10 so accessible groups will be 10. Taking after is the scientific representation to outline the groups. Every one of the bunches will be confined with unmistakable AGE however with relative influenced with malady. So every single group will be isolated with removal of 10.

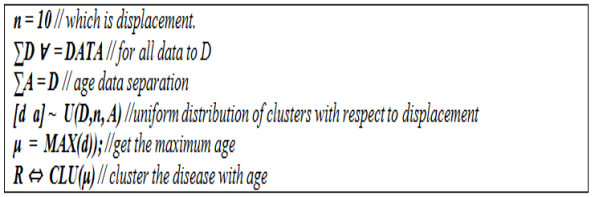

Fig 3:K-Means Algorithm

Demerits of DDC:

- ✓ Only AGE and DISEASE is countable
- ✓ Clustering is totally based on AGE
- ✓ Other attributes grouping costs much with thisapproach
- ✓ Time complexity is much with stemming onlyfor grouping if this technique expands andadding new attributes.
- ✓ SEO presentation of data is weak.
- ✓ No frequency of data tuples.

So by taking the conclusion and demerits our work is migrated to present this thyroid data in unique way which is FFM to overcome them.

Fuzzy C-mean:

The FFM exactness is appeared. In view of datasets every one of the information will be relegated with one of a kind buoy recognizable proof took after by 4 relocations focuses i.e. information focuses which is surrounded in the second box in the above picture. Taken after at this point the groups confined in light of most extreme of age the R contains the bunches with scope of AGE with particular relocation of 10(n) and sickness will be available in groups.

Example:

Input tuple is: 29(age), PITUITARY MALFUNCTION(disease). So the group reach is (30 - 40) and this Malady will fall under that specific cluster.That every one of the information will be populated in important clusters.Based on greatest recurrence of group size dioceseaccuracy will be anticipated furthermore the resultant vector isbased on "Sex" and "AGE" variable. (*NOTE:* Based on both K-means and FMM, prediction will be on thyroid's accuracy for "GENDER" and "AGE" attributes).

FFM(Feasible fuzzy mining on k means):

This is absolutely extraordinary and every one of the information is countable andmainly work is done on the premise of doable presentationof the relative information. Consistent procedure of sequencegives the practical results. Furthermore, recurrence will be assignedto tuple in a split second by own framework utilizing fluffy. Thefrequency tuning is done misleadingly and taking number ofall required traits.

Steps of FFM with K-Means work (non- mathematical):

1. Filter the datasets and allot remarkable buoy values (for uniqueness and doable results incrementally).
2. The scope of this buoy qualities is 0.0 to 4.99.
3. Utilizing fluffy create 4 one of a kind buoy values and the extent is 0.0 to 4.99. The fluffy will examine this extraordinary qualities for each tuple. So every single tuple will be its own worth in view of the extent.
4. The past 2 stages yields created by UFOE(part of fluffy system) and gets remarkable qualities in fancied scope of recurrence.
5. These 4 interesting focuses named as information focuses and 4 singular scope of recognized groups encircled.
6. In the event that the information focuses are 4.566, 2.9085, 3.678, 1.907 and these will be sorted by fluffy and yield will be 4.566, 2.9085, 3.678, 1.907.
7. So now 4 recognized groups depend on these sorted information focuses are(0.1 – 1.907) , (1.907 – 2.9085) , (2.9085 – 3.678) , (3.678 – 4.566)
8. All the tuples(whole information) ie allocated as of now with exceptional buoy esteem and by contrasting the quality with every single range and classified and put into that group range by taking GENDER additionally as combinational property.

So now the categories of clusters are 2 types with respect to GENDER, so 8 clusters will be generated.

All the data exactly placed on the basis of ranges and easy to pop out data for SEO and to present the desired output based on the queries.

```
FFM algorithm:
Initialization:
∑D = 0 //whole data set initialization
∑A = 0 // all ages
∑G = 0 // all genders
DP = {4.566 , 2.9085 ,3.678 ,1.907} // generated data points
FP = 0 //all unique float values assigned to tuples
T = {place , adhar , ,,,,}
m = 'm' // male code
f = 'f' //female code
GET(D);
A = D ~ [A , T ] // assign only ages to vector A
G = D ~ [G , T] // assign only gender to vector G
t = 0 //iterator
MT <= {r1 r2 r3 r4] // all male ranges ie clusters
FT <= [r1 r2 r3 r4] // all female ranges ie clusters
ct = 0 // total generated cluster sizes with values
//getting random float values to each and every tuple
for each t in D
Start:
 FP ∨Rand(t{0.0 4.99]) End;
t = 0 //neutralization
for each t in FP
temp = COMP(FP(t) , DP) //compare the all float assigned values to the above range
ff temp ∃ MT
MT(r) = temp
end if;
Else temp ∄ MT
FT(r) = temp
end
ct = D ~ t //extract exact number of final values in clusters with ranges.
_____

Final Result for SEO on K means with FFM:
temp1 = MAX(MT & FT) // get the maximum cluster size among all
SEO(r) = EXTRACT(D(ct) D['disease'],temp1) // get all diseases with respect to max cluster's age
PRESENT(SEO(r);
```

Fig : 4 Fuzzy C-mean Algorithm

IV. RESULTS AND OBSERVATIONS

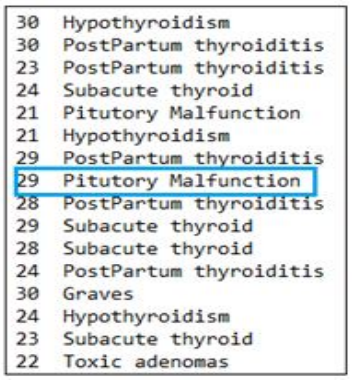

Fig : 5 Result for K-mean

All the ages are in the range of 20 to 30 and this cluster number is 3. Highlighted disease is with age of 29 and effected disease is Pituitary malfunction.

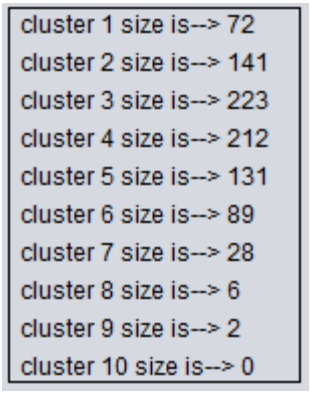

Fig :6 Clusters Conclusion

Here total number of clusters generated with DDC is 10 and maximum number of infected with thyroid is with the age between 30 to 40 age and 223 people fall under this cluster.

Fig: 7 displaying of decease and elapsed time

In this figure total time calculation is counted as the process time starts to the end.

AGE	SEX
65	F
18	F
46	M
56	F
19	M
20	F
8	M
7	M
9	M
3	M
1	M
2	F
3	F

Fig: 8 Fuzzy Result

Using the two attributes i.e. age and gender Fuzzy algorithm executed d to display the disease.

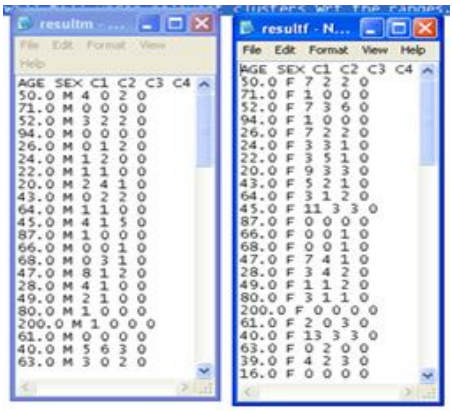

Fig : 8 Result with respect to gender.

By considering the maximum cluster part of male and female with respect to age and get the effected disease are separated using fuzzy algorithm for cluster count.

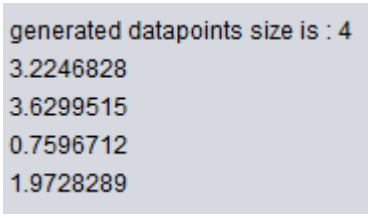

Fig: 9 Generating data points using Fuzzy

In Fuzzy algorithm this data points are generated by the float values which are assigned before the execution start time. For every new execution the data points are change due to dynamic.

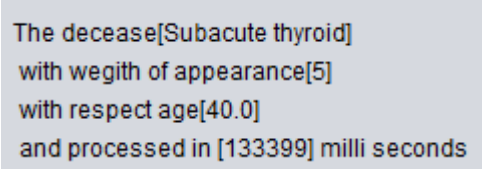

Fig: 10 Fuzzy total result

Based upon the frequency the highest rating of disease is displayed with respect to age and gender (male or female) .And the processed time is also calculated from the process execution start time to end time.

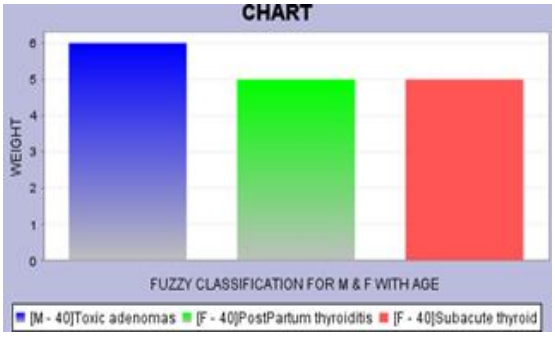

Fig: 11 Frequency graph of fuzzy

Frequency level is calculated by range of decease with respect to age and gender.

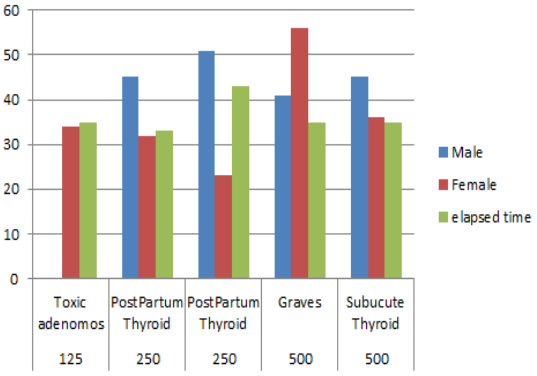

Fig: 12 Accuracy level by K-mean algorithm.

Using k-mean algorithm the accuracy o the decease is calculated by using the six thyroid types.

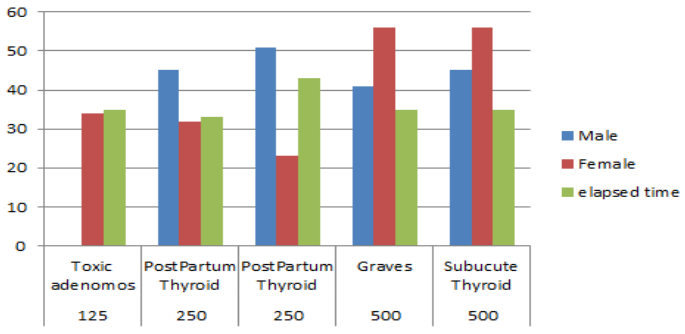

Fig: 13 Accuracy level by Fuzzy algorithm.

The accuracy of fuzzy is best when compared to k-mean algorithm .And the elapsed time of this algorithm is high due to the gender separation with respect to age.

V. LITERATURE SURVEY

In information mining the procedure and systems t. cap will be utilized by points of the exploration are be fined, with a specific end goal to acquire the outcome to the issue. The conceivable assignments of a calculation of example extraction can be anticipated and graphic exercises.

The two primary sorts of errands for forecast are the arrangement and the relapse. The characterization comprises of the forecast of a downright variable, that is, to find an action that will delineate arrangement of registers in an arrangement of predefined variables called classes. This action can be connected to new registers, to anticipate the class in which these registers fit. A few calculations are connected in the arrangement assignments, yet those that seem most are Neural Networks, Back-Propagation, Bayesian Classifiers and Genetic Algorithms.

In the relapse, there is a quest for straight capacities or not, and the variable that will be anticipated comprises of a numerical property (constant) present in databases with genuine qualities. Keeping in mind the end goal to execute the relapse errand, the techniques for measurements and Neural Networks are utilized.

The searing undertaking is utilized to isolate the registers of databases into subsets or bunches, in a manner that the components of a group offer basic properties that serve to recognize the components in different groups, going for boosting intra-bunch similitudes and minimize bury group likenesses. Not at all like the order undertakings in which the variables are predefined, the searing needs, to distinguish consequently, the information gatherings, to which the analyst ought to property the variables. The most utilized calculations as a part of this assignment are the K-Means, K Modes, KPrototypes, K-Medoids, Kohonen, among others.

VI. FUTURE WORK

Our present work can be relocated in two distinctive ways .one is to reach out from typical grouping regarding AGE and other one is to move FFM (k implies) concerning all other conceivable properties in datasets. As for time of SEO presentation execution opening discarding component can be effectively supportive in the present design: Changing and amplifying information sets in juggling can be material as characteristics.

REFERENCES

[1]. Cardoso ONP, Machado RTM. Gestão do conhecimento usando data mining: estudo de casona Universidad Federal de Lavras. Rev Adm Pública. 2008;42(3):495-528. [Links]

[2]. Goldschmidt R, Passos E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elsevier; 2005.

[3]. Marcano Aular YJ, Talavera Pereira R. Minería de datos comosoporte a la toma de decisionesempresariales. Opcion. 2007;23(52):104-18. [Links]

[4]. AraujoJúnior RH, Tarapanoff K. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. Ci Inf. 2006;35(3):236-47.[Links]

[5]. Steiner MTA, Soma NY, Shimizu T, Nievola JC, Steiner Neto PJ.Abordagem de um problemamédicopormeio do processo de KDD com ênfase à análiseexploratória dos dados. Gest Prod. 2006;13(2):325-37.[Links]

[6]. Costa Lda F. Bioinformatics: perspectives for the future. Genet Mol Res. 2004;3(4):564-74. [Links]

[7]. Quoniam L, Tarapanoff K, AraújoJúnior RH, Alvarez L. Inteligênciaobtidapelaaplicação de data mining em base de tesesfrancesassobre o Brasil. Ci Inf. 2001;30(2):20-8. [Links]

[8]. Matos G, Chalmeta R, Coltell O. Metodologíapara la extracciónde lconocimientoempresarial a partir de los datos. InfTecnol. 2006;17(2):81-8. [Links]

[9]. Naãs IA, Queiroz MPG, Moura DJ, Brunassi LA. Estimativa de estroemvacasleiteirasutilizandométodos quantitativos preditivos.Ciênc Rural. 2008;38(8):2383-7. [Links]

[10]. Febles Rodríguez JP, González Pérez A. Aplicación de la minería de datos en la bioinformática. ACIMED. 2002;10(2):69-76. [Links]

[11]. Jones PBC. The commercialization of bioinformatics.Electron J Biotechnol. 2000;3(2):33-4.[Links] Fayyad UM, Shapiro GP, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. Menlo Park, Calif.: AAAI Press: MIT Press; c1996; 611p. [Links]