International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Ensemble Hybrid K- Means and DBSCAN Clustering Algorithm – HDKA for Cancer Dataset

Sangeetha.M, Kousalya.R

Abstract*— Data Mining is the foremost vital space of analysis and is pragmatically utilized in totally different domains, It becomes a highly demanding field because huge amounts of data have been collected in various applications. The database can be clustered in more number of ways depending on the clustering algorithm used, parameter settings and other factors. Multiple clustering algorithms can be combined to get the final partitioning of data which provides better clustering results. In this paper, Ensemble hybrid KMeans and DBSCAN (HDKA) algorithm has been proposed to overcome the drawbacks of DBSCAN and KMeans clustering algorithms. The performance of the proposed algorithm improves the selection of centroid points through the centroid selection strategy.For experimental results we have used two dataset Colon and Leukemia from UCI machine learning repository.

Keywords : K-Means, DBSCAN, HDKA, Colon, Lukemia

I. INTRODUCTION

Data Mining is the foremost vital space of analysis and is pragmatically utilized in totally different domains like finance, education, clinical analysis, healthcare, agriculture etc. within the aim of discovering helpful info from great amount of dataset. This analysis uses totally different data processing techniques to cluster medical Data.

Data processing tasks in Data Mining is categorized in to Two types: Supervised tasks and unsupervised tasks. Supervised tasks have datasets that contain each the instructive variables, dependent variables. the target is to get the associations between the instructive and dependent variables. On the opposite hand, unsupervised tasks have datasets that contain solely the instructive variables with the target to explore and generate postulates regarding the hidden structures of the data.

Manuscript published on November 30, 2019.
* Correspondence Author
M.Sangeetha, Assistant Professor of Computer Applications , PSGR Krishnammal College for Womrn ,Coimbatore,
Dr. R. Kousalya, Head of Department/Professor of Computer Application, Dr. N.G.P Arts and Science College, Coimbatore,

Clustering is taken into account as a vital technique in data mining and is a lively analysis topic for the researchers.

The objective of cluster is to partition a group of objects into clusters such objects within a bunch are more similar to each other than patterns in several clusters.

These algorithms are divided into many classes. The basic three distinguished classes are partitioning, hierarchical and density-based. The challenges in the clustering issues treating

Vast quantity of information in massive databases can be attempted by solving these three classes. However, none of them are the foremost effective. DBSCAN (Density based mostly spatial clustering of Applications with Noise) [5] may be a typical density-based clustering algorithmic rule.

The K-Means Algorithm

The K-Means algorithm attempts on partitioning the dataset into ‘k’ subsets such that all records, from current subsets are noted as points, in an exceedingly set which "belong" to a similar center. Conjointly the points in an exceedingly set area unit nearer be a center than to the other center. The algorithm analysis the centroids of the subsets, and takes the straightforward iterations.

Problems with KMeans clustering algorithm:-

Some of the weaknesses of k-means are

- Once the numbers of data don't seem to be numerous, initial grouping can confirm the cluster considerably
- The result is circular cluster form because based on distance.
- The amount of cluster, K, should be determined beforehand. Choice of import of K is itself a problem and generally its hard to predict beforehand the amount of clusters that might be there in data.
- We never understand the important cluster, using identical data, if it is inputted in a completely different order could manufacture different cluster if the amount of data is few.
- Sensitive to initial condition, different initial condition could manufacture different results of cluster. The algorithm could also be trapped within the native optimum.
- It is difficult to understand the attribute contribution. A lot of grouping method assume that every attribute has identical weight.
- The arithmetic mean value isnot sturdy to outliers. Very far data from the center of mass could pull the center of mass far from the data.

DBSCAN ALGORITHM

Density-Based Spatial Clustering and Application with Noise (DBSCAN) was a clustering algorithm based on density. It did clustering through growing high density area, and it can find any shape of clustering (Rong et al., 2004). The Density Reachability, Density Connectivity and Density reachability is the first stage in DBSCAN. It defines whether two distance close points belong to the same cluster.

II. LITERATURE SURVEY

Vikas Chaurasia, Saurabh Pal1 and BB Tiwari[1] applied three prediction models for breast cancer survivability on two parameters: benign and malignant cancer patients. The parameters are measured on the most popular datamining methods likely - Naive Bayes, RBF Network, and J48 Decision Tree. The Authors used dataset from UCI Machine learning repository.

Clustering Algorithm for Healthcare Dataset using Silhouette Score Value[2] The objective of this method is to analyze the performance of two clustering algorithms K-Means and DBSCAN.The experimental results indicates that the K-means algorithm gives better results in terms of clustering accuracy and execution time when compared to DBSCAN Algorithm

Bharat Singh1, Kundan Kumar, et al.[3]proposed a hybrid approach, using bits from k-mean techniques and called it Feature Co-association Ensemble (CFE). For experimental analysis, they have used seven benchmark high dimensional datasets from UCI repository and validation proves its worthiness among the exiting approaches of feature selection methods.

K. Mumtaz and Dr. K. Duraiswamy [5] proposed clustering and outlier detection system that has been implemented using Matlab and tested with the data synthetically created by Gaussian distribution function. The authors have formed data circular or spherical clusters in space. They proposed Dbkmeans algorithm which perform very well than DBSCAN and k-means clustering in term of quality of classification measured by Rand index. One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data.

Mengxing Huang, Qili Bao et al .[6] proposed a new hybrid an algorithm based on Optimization of Initial Points and Variable-Parameter Density-Based Spatial Clustering of Applications with Noise (OVDBCSAN) and support vector regression (SVR). At the initial point of optimization, ε and MinPts, which are global parameters in DBSCAN, mainly deal with datasets of different densities. According to different densities, appropriate parameters are selected for clustering through optimization.

S. Sharma, A. K. Sharma et al. [7]proposes a methodology for the improvement in DB-SCAN algorithm to improve clustering accuracy. They proposed that the improvement is based on back propagation algorithm to calculate Euclidean distance in the dynamic manner. Also they obtained results of implemented proposed and existing methods and it compares the results in terms of its execution time and accuracy.

D. Jain, M. Singh et al. [8] have used the DBSCAN algorithm and applied to compute the EPS value and Euclidian distance on the basis of similarity or dissimilarity of the input data. Also back propagation algorithm is applied to calculate Euclidian distance dynamically and simulation study is conducted that shows improvement to increase accuracy and reduce execution time.

M. S. Premkumar and S. H. Ganesh [9] have proposed a work on novel median based initial centroids have been generated and imposed onto an experimental dataset to analyze the performance of the proposed work. The results have shown that the proposed work, improved the accuracy of clustering with reduced number of iterations.

G. Jagatheeshkumar and S. S. Brunda [10] find a problem using many applications in Market Analysis, web mining and indexing. In this analysis covers of clustering methods similarity measures based on distance. To discover related work this cluster technique find a new proposal for our further work in text documents, similarity meaning data mining.

III. PROPOSED ALGORITHM

In this section, the proposed Algorithm is described in brief. The paper combines existing clustering algorithms to mine, form a model and predict relevant medical data. The publicly available data sets have been pre-processed and this is done by calling the function Preprocess in Matlab function , that code pre-processing step can be applied to Colon and Leukemia data, The dataset are downloaded from UCI Machine Learning Repository.

First this article finds the limitations of both DBSCAN and K-Means algorithms. Second the algorithms are improved by creating a cluster ensemble. This Cluster Ensemble involves the formation of Type I, Type II and Type III cluster ensemble from the mixed datasets. Here, Type I or first cluster ensemble takes into account the categorical attribute value to solve the issues in forming the cluster ensembles. The categorical attribute value or label is now considered as a cluster with data points D of N sets, where D = {d1, d2… dN}.

Thus categorical attribute set, A = {a1, a2,…,aM} and set of M partitions as π = { π 1, π 2,…, π M}. The generation of π I (partition) for categorical attribute values are defined as π i= {C1

i, C2 ,i,…,Ci k} and Ukj j=1 = ki, where, aM represents total number of attribute values. Then the categorical data is transformed to cluster ensemble, however, accuracy is not affected. The generation of high quality clustering is promised using this ensemble technique through the Transformation of categorical data.

Type II Full-space ensemble follows the results of base clustering that is ascertained through clustering algorithm over the numerical and categorical dataset. Here, categorical dataset is taken into consideration. The k-mode study is used for the creation of base clustering after the selection of cluster center. To avoid instability during k-mode procedure over categorical dataset, the definite clusters for base clusters are selected using two methods: Fixed-k method where k = $\sqrt{N}$and random-k with k $\sqrt{N}${2,…,[($\sqrt{N}$)]}.

Algorithm 1:- Centroid Selection Algorithm

Choose users (k) as centroids from the entire datasets
Input : Total Users in training set, u and total clusters, k
Output: Total centroids, k i.e given by $\{l_1,l_2....l_k\}$
Step 1: Define the needed clusters, k
Step 2: Compute the mean of all datasets
Step 3: Sort mean with Euclidean Distance from average mean-pairwise distance between the data points
Step 4: While C: C= {1, 2 ,…. k} do
Step5: The cluster center is chosen for centroid

$S_c = X_{1+(c-1)*(M/k)}$

Step6: Repeat step 5 until centroids(k) are found
Step7: Return $\{l_1,l_2....l_k\}$
Step8: End While
Step9: The objects that remains in cluster space is assigned to the nearest cluster using nearest neighbour distance measurement based on cluster centers $\{l_1,l_2....l_k\}$. This process is associated with data points of numerical attributes and categorical attributes using k-mode clustering.
Step10: If cluster = balanced&& cluater=non-cluster data points.
Step11: Then find the nearest neighbour in balanced clusters using K-mode clustering for numerical data sets
Step12: For object x_i and x_j distance is defined as :

$d(x_i .x_{j)} = \sqrt{\sum_{a=1}^{c} (x^{a_i}.x^{a_j}, d^{a}(.)}$ is the distance of a^{th} attribute possessing numerical data

$d(x^{a}_i,x^{a}_j) = \begin{cases} 0 & x^{a_i} = x^{a_j} \\ 1 & x^{a_i} \# x^{a_j} \end{cases}$ is the distance of a^{th} attribute and categorical distance is

$d(x^{a}_i,x^{a}_j) = \sum_{a=1}^{c} \delta (x^{a_i}.x^{a_j})$

Step13: End For
Step14: Check cluster ≠non-cluster data points
Step15: If yes End If

The Algorithm 1 generated the centroids selected based on the instants and the random selection of clusters that degradessystem performance. Hence, the performances of the selection of centroids points are improved through the proposed centroids selection strategy.

Algorithm 2:- Hybrid Model HDKA

Step1: Run the enhanced K-Means algorithm and identify the clusters.
Step2: Execute step 2a for each data point in each and every cluster
a. Calculate the number of points in each cluster point. If this number is less than the minimum point then return the centre as noise data or return the centre point as the core data.
Step3: Execute the enhanced DBSCAN for the core points identified. The non- core points are considered as noise and are not included in the clustering.
Step4: Resolve the multiple cluster IDs and let the resultantant number of clusters be m
Step5: For each m clusters {
find the cluster centers C_m by taking the mean
and find the total number of points in each clusters }
Step6: if m>k (from enhanced K) then
{Select two clusters that have similar ε,minimum points, and merge them
Find the new cluster center. Repeat until achieving k clusters. }
else
{ l=k-m
if (m>=l) {
Select the cluster whose density and No.of points is equal to ε and minimum poimts
Use the enhanced HDKA clustering algorithm
Repeat it until k clusters are achieved. }
End if

After selecting the Centroids it is applied for the K-Means Algorithm. It calculates the number of points in each cluster. The DBSCAN algorithm is then executed for the core points identified by the K-Means. In this process the Non-Core points are considered as noise and are not included in the clustering. Resolve the multiple cluster IDs and set the resultant number of clusters. Select two clusters form the algorithms that have similar Ɛ, minimum points and merge them. The steps can be repeated until all the new clusters have been achieved. The above enhanced algorithm thus gives the higher performance in terms of Classification and Average Accuracy. The noise in the dataset has also been removed.

IV. RESULTS AND DISCUSSIONS

a. DataSet:

The two input set of dataset colon and leukemia are used for analysis. The dataset are downloaded from UCI Machine Learning Repository. The publicly available data have been pre-processed and this is done by calling the function PreProcess in Matlab function , that code pre-processing step and that can be applied to *Leukemia data*, *Colon data*.

b. Metrics used for Evaluation

To measure the performance of the experimental results we have used the technique of Rand index or Rand measure for finding the similarity between two data clusters. We Proposed an algorithm HDKA (Hybrid DBSCAN and K-Means Algorithm) a method to combine the K-Means and DBSCAN algorithms using Ensemble methods.
The Rand index has a value between 0 and 1 where the point 0 indicates two data clusters which do not agree on any pair of points and point 1 indicaties the similarity between the dataset.

c. Performance in terms of Accuracy.

We evaluated two dataset Colon and Leukemia in terms of performance on classification and accuracy.
The Attributes of two Dataset are:
The Number of Clusters: 8
K-Means Iteration: 40
Minimum Distance: 0.8325
Number of Cluster Labels: 7

d. Performance in terms of Classification Accuracy and Average Accuracy for Colon Dataset:-

S.No	Performance in terms of classification and Average accuracy	Name of the Algorithm		
		K-Means	DBSCAN	HDKA

1	Classification Accuracy Measured by Rand Index (Found Using Cluster Centers)	68.087	68.172	68.229
2	Average Accuracy Measured by Rand Index (Found Using Cluster Centers)	95.7719	95.09.3	95.6419

Table1:Performance Comparison of Algorithms

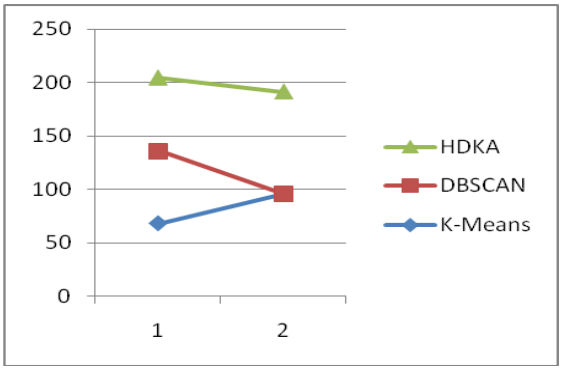

Fig 1: Performance Comparison for Colon Data

e. Performance in terms of Classification Accuracy and Average Accuracy for Lukemia Dataset:-

S.No	Performance in terms of classification and Average accuracy	Name of the Algorithm		
		K-Means	DBSCAN	HDKA
1	Classification Accuracy Measured by Rand Index (Found Using Cluster Centers)	68.087	68.172	68.229
2	Average Accuracy Measured by Rand Index (Found Using Cluster Centers)	95.7719	95.09.3	95.642

Table2:Performance Comparison of Algorithms

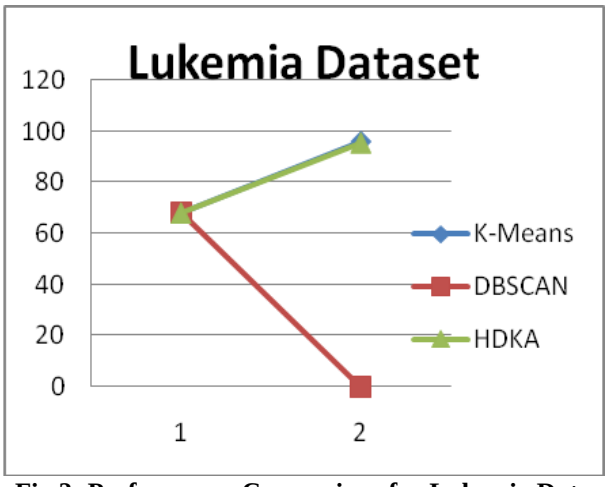

Fig 2: Performance Comparison for Lukemia Data

f. The clustering results for HDKA for Colon and Lukemia Dataset:-

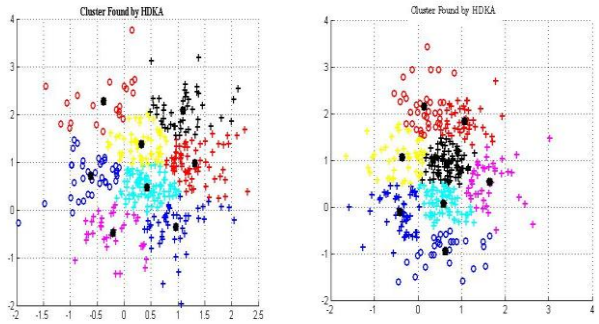

Fig-3:- Cluster Points by HDKA for Colon and Lukemia

From the Plotted results it is noted that HDKA Algorithm has a better Classification Accuracy in Lukemia Dataset than Colon Dataset and also gives an better Average Accuracy in both the data sets. It could be found that the hybrid ensemble model HDKS Algorithm improves the Speed and make it robust in terms of Classification and Accuracy.

V. CONCLUSION

In this paper, we proposed a new ensemble hybrid algorithm HDKA for Cancer Dataset using Matlab and tested with the two dataset – Colon and Lukemia. HDKA algorithm performed very well in term of classification accuracy and average accuracy measured by Rand index This provides direction for effective clustering of datasets of different densities. The extraction of knowledge in medical domain is considered as a biggest problem for medical data diagnosis The proposed HDKA clustering algorithm provides a better pavement for the medical data mining. The proposed HDKA algorithm can be further amended by considering a number of efficient clustering methods such as hierarchical method.

REFERENCES

1. Vikas Chaurasia, et.al"Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms &Computational Technology – 2018 DOI: 10.1177/1748301818756225
2. Godwin Ogbuabor1 and Ugwoke,F.N – "Clustering Algorithm for Healthcare Dataset using Silhouette Score Value", International Journal of computer science and Information Technology-Vol 1, No2, April 2018. DOI10.5121
3. Bharat Singh, Kundan Kumar,et.al."Ensemble of Clustering Approaches for Feature Selection of High Dimensional Data", 2nd INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING AND SOFTWARE ENGINEERING (ICACSE-2019)
4. K. Mumtaz1 and Dr. K. Duraiswamy," A Novel Density based improved k-means Clustering Algorithm – Dbkmeans" International Journal on Computer Science and Engineering ISSN : 0975-3397 213 Vol. 02, No. 02, 2010, 213-218
5. Mengxing Huang 1,2, Qili Bao et al." A Hybrid Algorithm for Forecasting Financial Time Series Data Based on DBSCAN and SVR" Information 2019, 10, 103; doi:10.3390/info10030103
6. S. Sharma, A. K. Sharma and D. Soni, "Enhancing DBSCAN algorithm for data mining," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, 2017, pp. 1634-1638. doi: 10.1109/ICECDS.2017.8389724

7. D. Jain, M. Singh and A. K. Sharma, "Performance enhancement of DBSCAN density based clustering algorithm in data mining," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, 2017, pp. 1559-1564. doi: 10.1109/ICECDS.2017.8389708
8. M. S. Premkumar and S. H. Ganesh, "A Median Based External Initial Centroid Selection Method for K-Means Clustering," *2017 World Congress on Computing and Communication Technologies (WCCCT)*, Tiruchirappalli, 2017, pp. 143-146. doi: 10.1109/WCCCT.2016.42
9. G. Jagatheeshkumar and S. S. Brunda, "An analysis of efficient clustering methods for estimates similarity measures," *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, 2017, pp. 1-3. doi: 10.1109/ICACCS.2017.8014710
10. U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Noida, 2017, pp. 527-530. doi: 10.1109/CONFLUENCE.2017.7943207
11. E. N. Benderskaya, "Cluster analysis problems and bio-inspired clustering methods," *2017 XX IEEE International Conference on Soft Computing and Measurements (SCM)*, St. Petersburg, 2017, pp. 162-164. doi: 10.1109/SCM.2017.7970526
12. G. Baruch, D. Shapira and S. T. Klein, "Compressed Hierarchical Clustering," *2018 Data Compression Conference*, Snowbird, UT, 2018, pp. 399-399. doi: 10.1109/DCC.2018.00052
13. E. B. Fatima and E. M. Abdelmajid, "Study of efficiency k-means clustering using Z-test proprieties," *2017 Intelligent Systems and Computer Vision (ISCV)*, Fez, 2017, pp. 1-5.doi: 10.1109/ISACV.2017.8054926
14. Batra, S. Verma and Kavita, "Performance Analysis of Data Mining Techniques in IoT," *2018 4th International Conference on Computing Sciences (ICCS)*, Jalandhar, 2018, pp. 194-199.doi: 10.1109/ICCS.2018.00039
15. Rashmi Amardeep1" The MATLAB Data Mining Software – Study" International JoUrnal of Recent Innovation in Engineering and Research Scientific JoUrnal Impact Factor - 3.605 by SJIF e- ISSN: 2456 – 208

AUTHORS PROFILE

M.Sangeetha, received the M.Sc. Information Science degree from SNS Rajalakshmi College of Arts and Acience, India in the year 2004 and M.Phil degree in Computer Science from Bharathiyar University, India in the year 2010 respectively. Currently, she is an Assistant Professor of Computer Applications , PSGR Krishnammal College for Womrn ,Coimbatore, affiliated to Bharathiyar University. Shehas a total experience of over 10 years. She has published 3papers in Journals. Her area of interest is Data Mining

Dr. R. Kousalya, received the B.Sc. degree in Physics from P.S.G.R. Krishnammal College for Women, India in the year 1997, the MCA degree in Computer Applications from Bharathiar University in the year 2000, the M.Phil degree in Computer Science from Manonmaniam Sundaranar University in the year 2003 and the Ph.D degree in Computer Applications from Manonmaniam Sundaranar University in the year 2016. Currently, she is a Head of Department/Professor of Computer Application, Dr. N.G.P Arts and Science College, Coimbatore, affiliated to Bharathiyar University. She has a total experience of over 19 years. She has published 33 papers in Conference and 11 papers in Journal. Her area of interest includes data mining and web mining.