A Novel Approach for Improvement of Denclue in Clustering for Incomplete Systems

Y. Vijay Bhaskhar Reddy[1], Dr L.S.S Reddy[2], Dr.S.S.N. Reddy[3]
[1]*Research Scholar, Rayalaseema University, Kurnool, AP.*
[2]*Vice Chancellor, KLEF, Vaddeswaram.*
[3]*Principal, Vardaman College of Engineering, Hyderabad, TS.*

Abstract- One of the challenge in data world is grouping of the data aplenty of types of data is available for example stream data, data from social network videos, images, tent type … etc. The clustering methods revise this data into various groups. Many issues are identify such as computation time, cannot handle high dimensional data. A point is said to be one it within its surrounding radius of € contains minimize number of points. The number of minimum points is decided based on the problem. This algorithm gives efficient results when compared to K- Means and other clustering algorithms. All the noise points are remains in DENCLUE Algorithm. Sometimes two points are within the region it is not possible to establish direct connection between these two points and also the density parameter and the noise threshold need to be selected carefully as it significantly affects the quality of results we proposed a novel approach for improvement of denclue in clustering data world. In which one to one corresponding relation is maintained between noise points to its nearest neighbor node and unwanted links with in the dense region are remains to proposed approach leads to better results. When compared to DENCLUE algorithms no data loss for Noise points also the implementation cost is decreased by removing unwanted links.

Keywords- Cluster, DENCLUE, K Means, Optimal, dense mining.

I. INTRODUCTION

From the past few years multiple sources produce numerous amounts of data. The sources of data generation are social networks, mobile networks, air traffic data...etc. A well-defined procedure is required to organize this heterogeneous data. In the area of data mining address various methods for organization of this heterogeneous data.

In this context, the classification is done using data mining techniques [1, 2, 3]. The raw data is grouped basing on the similarity. These are further processed and apply variour compression techniques. The operation applied on intra class is different from operation applied on inter class.

The supervised clustering is done into phases they are learning phase and validation phase. In learning phase the learning model is constructed for the given data and is validation phase the testing is performed with the given constraints. The learning phase will take more time when compared to testing phase by using validation algorithm (4). These are many methods for supervised clustering, such as KNN [5], NN, FL and SVM.

The unsupervised classification is called as clustering, in which the number of clusters is decided without any prior knowledge of data. All these clustering algorithms are divided into five types they are Model based clustering, Grid based clustering, and Density based Clustering, Partitioning Clustering and hierarchical clustering.

The paper is designed as follows; Section-II will give the related work in clusters Section-III will give the proposed algorithm section-IV the proposed algorithm will be evaluated based on the results, in section V conclusion of the proposed Novel clustering ALGORITHM.

II. RELATED WORK

DBSCAN and OPTICS are two of the most understood density based clustering algorithms. An intriguing property of density based clustering is that these algorithms don't accept groups to have a specific shape. Moreover, the calculations permit "noise" questions that don't have a place with any of the groups. K-means for illustrations parcels the information space in Voronoi cells (a few people assert it produces circular groups - that is off base). For the genuine state of K-means clusters and an illustration that cannot be grouped by K-means. Inside measures for group assessment likewise generally expect the clusters to be all around isolated circles (and don't permit noise/outlier v questions) - of course, as we tend to try different things with manufactured information produced by various Gaussian disseminations.

Model-based clustering:

Model-based cluster could be noteworthy thanks to subsume grouping investigation. cluster algorithms may be developed supported likelihood models, like the finite mixture model for likelihood densities. The model represents the sort of constraints and geometric properties of the variance matrices. Within the family of model-based cluster algorithms, one uses sure models for clusters and tries to optimize the work between the information and therefore the models. Within the model-based cluster approach, the information ar viewed as coming back from a combination of likelihood distributions, every of that represents a distinct cluster. In different words, in model-based cluster, it's assumed that the information ar generated by a combination of likelihood distributions within which every part represents a distinct cluster. therefore a specific cluster methodology will be expected to figure well once the information adapt to the model.

Grid-based clustering:

Grid-based approaches are fashionable for mining clusters during a giant two-dimensional house whereby clusters are thought to be denser regions than their surroundings. The complexness of most cluster strategies is a minimum of linearly proportional to the scale of the information set. the nice advantage of grid-based cluster is its important reduction of the process complexness, particularly for cluster terribly giant information sets. The grid-based cluster approach differs from the standard cluster algorithms therein it's involved not with the information points however with the worth house that surrounds the information points. In general, a typical grid-based cluster algorithmic rule consists of the subsequent 5 basic steps (Grabusts and Borisov, 2002):

1. Making the grid structure
2. Hard the cell density for every cell.
3. Sorting of the cells consistent with their densities.
4. Distinguishing cluster centers.
5. Traversal of neighbour cells.

Density-based clustering:

Density based mostly cluster algorithmic rule has contend a vital role find non linear shapes structure supported the density. Density-Based spacial cluster of Applications with Noise (DBSCAN) is most generally used density based mostly algorithmic rule. It uses the idea of density reach ability and density property.

Density Reach ability 'a' is claimed to be density accessible from purpose 'b' if point 'a' is inside ε distance from point 'b' and 'a' has decent range of points in its neighbours that are inside distance ε.

Density property - some extent 'a' and 'b' are aforesaid to be density connected if there exist some extent 'c' that has decent range of points in its neighbours and each the points 'a' and 'b' ar inside the ε distance. this can be chaining method. So, if 'b' is neighbor of 'c', 'c' is neighbor of 'd', 'd' is neighbor of 'e' that successively is neighbor of 'a' implies that 'b' is neighbor of 'a'.

III. NEW DENSITY-BASED ALGORITHM

We present below our algorithm, called A Novel Approach Algorithm

Input: The dataset length d and attributes t_1, t_2, t_3..... t_n are initialized.

Step 1: Pre-processing of dataset.

Step 2: Initialize the two class p and e for the mushroom datasets.

Step 3: Here for every attribute of mushroom dataset initialize with the cluster (as shown in the table-1).

Step 4: In every attribute positive and negative classes are divided for every attribute in dataset and this is consider as cluster.

Step 5: Initialise attributes values as shown in table-1.

Step-6: Initialize the ArrayList<HashMap<Character, HashMap<Character, Integer>>> counts;

Step-7:

```
public final void initCounts() {
    counts = new ArrayList<HashMap<Character, HashMap<Character, Integer>>>();
      Attribute clsAtr = dataSet.classAttribute();
      for (int i = 0; i < dataSet.numAttributes(); i++) {
        HashMap<Character, HashMap<Character, Integer>> attrCounts = new
HashMap<Character, HashMap<Character, Integer>>();
        Attribute atr = dataSet.attribute(i);
        for (int j = 0; j < atr.numValues(); j++) {
          HashMap<Character, Integer> valCounts = new HashMap<Character,
Integer>();
          for (int k = 0; k < clsAtr.numValues(); k++) {
            valCounts.put(clsAtr.value(k).charAt(0), 0);
          }
          attrCounts.put(atr.value(j).charAt(0), valCounts);
        }
        counts.add(attrCounts);
      }
    }
```

Step 8: Call the method initCounts().

Step 9: Clusters based on attributes.

Output: Display the poisonous and edible results for the given dataset.

IV. EXPERIMENTAL RESULTS

This paper is implemented in R-Programming Language. Based on the total number of attributes and total number of tuples the result is displayed. The aim of the paper is to calculate the dataset mushrooms based on the attributes present in the dataset. Here two classes are very important that are p (poisonous), e (edible). The description of dataset is described as follows.

A. Explanation of Datasets:

To evaluate our approaches, we have used mushroom datasets. In this section, we will explain about the dataset used is mushroom dataset. Which is used to find out the poison and edible with features of the mushrooms?

{'x', 'b', 's', 'f', 'k', 'c'}, // cap-shape
{'s', 'y', 'f', 'g'}, // cap-surface
{'n', 'y', 'w', 'g', 'e', 'p', 'b', 'u', 'c', 'r'}, // cap-color
{'t', 'f'}, // bruises
{'p', 'a', 'l', 'n', 'f', 'c', 'y', 's', 'm'}, // odor
{'f', 'a', 'd', 'n'}, // gill-attachment
{'c', 'w', 'd'}, // gill-spacing
{'n', 'b'}, // gill-size
{'k', 'n', 'g', 'p', 'w', 'h', 'u', 'e', 'b', 'r', 'y', 'o'}, // gill-color
{'e', 't'}, // stalk-shape
{'e', 'c', 'b', 'r', 'u', 'z', '?'}, // stalk-root
{'s', 'f', 'k', 'y'}, // stalk-surace-above-ring
{'s', 'f', 'y', 'k'}, // stalk-surface-below-ring
{'w', 'g', 'p', 'n', 'b', 'e', 'o', 'c', 'y'}, // stalk-color-above-ring
{'w', 'p', 'g', 'b', 'n', 'e', 'y', 'o', 'c'}, // stalk-color-below-ring
{'p', 'u'}, // veil-type
{'w', 'n', 'o', 'y'}, // veil-color
{'o', 't', 'n'}, // ring-number
{'p', 'e', 'l', 'f', 'n', 'c', 's', 'z'}, // ring-type

{'k', 'n', 'u', 'h', 'w', 'r', 'o', 'y', 'b'}, // spore-print-color
{'s', 'n', 'a', 'v', 'y', 'c'}, // population

{'u', 'g', 'm', 'd', 'p', 'w', 'l'}, // habitat
{'p', 'e'} // class

TABLE I: Description of datasets used in the experiments.

Attribute Value	Attribute Sub Names(Cap Shape)
B	Bell
C	Convex
F	Flat
K	knobbed
S	sunken

Table 2: Attribute values

Attribute Value	Attribute Sub Names(Cap Surface)
F	Fibrous
G	Grooves
Y	Scaly
S	Smooth

Table 3: Attribute values

Attribute Value	Attribute Sub Names(Cap Surface)
N	brown
B	buff
C	cinnamon
G	gray
R	green
P	pink
U	purple
E	red
W	white
Y	yellow

B. Validity Metrics

For effectiveness of the clustering methods and improving the quality of the formation of the cluster and reducing the computation time.

Dunn Index (DI): This file evaluates the division degree between people of a similar cluster, i.e., the intra-group likeness. A high esteem shows a superior cluster.

Accuracy of Clusters (CA): CA calculates the level of effectively arranged protests in a group, in light of the pre-characterized class names. This record is utilized just with marked database, a high esteem shows a superior grouping.

Davies-Bouldin Index (DBI): This record, as DI, assesses additionally the division degree between groups (between bunch dislikeness), the littlest esteem demonstrates the better grouping.

Algorithms	Edible	Poisonous	Time	Accuracy
DENCLUE	4010	3810	240.12112	0.95411
NOVEL APPROACH	4208	3916	57.101636705 seconds	0.9958148695224027
ALGORITHM				

Table 4, Results of the existing and proposed algorithms.

C. Computation Time

The Computation time (in ms) of the algorithm as is shown table-4.

D. Comparison

Calculate the mean of each index for the mushroom dataset.

• Compare with existing algorithm novel approach algorithm minimize the execution time

• In terms of clustering performance Novel Approach Algorithm maximizes the DENCLUE

V. CONCLUSION

In this paper, a novel approach algorithm is utilized to improve the formation of clusters and performance of the clustering with time. The proposed system proves its performance by reducing the computation time, with acceptable clustering and better results. The issue found in two points are within the region it is not possible to establish a direct connection between these two points and also the density parameter and the noise threshold need to be selected carefully as it significantly affects the quality of results.

VI. REFERENCES

[1]. C. C. Aggarwal and C. Zhai, Mining text data. Springer Science & Business Media, 2012.
[2]. P. Berkhin, "A survey of clustering data mining tech-niques," in Grouping multidimensional data, pp. 25–71, Springer, 2006.
[3]. A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264–323, 1999.
[4]. P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in Encyclopedia of database systems, pp. 532–538, Springer, 2009.
[5]. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information The-ory, vol. 13, no. 1, pp. 21–27, 1967.
[6]. W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943.
[7]. V. Vapnik, The nature of statistical learning theory. Springer Science & Business Media, 2013.
[8]. R. Xu, D. Wunsch, et al., "Survey of clustering algo-rithms," Neural Networks, IEEE Transactions on, vol. 16, no. 3, pp. 645–678, 2005.
[9]. A. K. Jain, A. Topchy, M. H. Law, and J. M. Buh-mann, "Landscape of clustering algorithms," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 1, pp. 260–263, IEEE, 2004.
[10]. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," Emerging Topics in Computing, IEEE Transactions on, vol. 2, no. 3, pp. 267–279, 2014.
[11]. G. H. Shah, C. Bhensdadia, and A. P. Ganatra, "An em-pirical evaluation of density-based clustering techniques," International Journal of Soft Computing and Engineering (IJSCE) ISSN, pp. 2231–2307, 2012.

[12]. C. Ding and X. He, "Cluster merging and splitting in hierarchical clustering algorithms," in IEEE International Conference on Data Mining (ICDM'02), pp. 139–146, IEEE, 2002.