IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Framework of Selective Clustering Ensemble Based On K-Nearest Neighbour Approach

Miss.R. J. Wadnare
SGBAU, Amravati
India.

Dr. S. S. Sherekar
SGBAU, Amravati
India.

Dr.V. M. Thakare
SGBAU, Amravati
Indi

ABSTRACT

Grouping of data according to their similarity is called clustering; multiple clustering algorithms are invented in previous decades. As the time passed, the diversity, shape, size and ambiguity present in the data is increased. This paper examines five different techniques such as selective clustering ensemble, DBMAC algorithm, fast D-peak algorithm, multi-density clustering algorithm, ADA-DBSCAN algorithm. These approaches having limitation like dependency of parameter, time complexity, less scalability of algorithm. To improve these limitation this paper proposed a new method called "Fast selective clustering ensemble with k-NN". This method reduces the time complexity of selective clustering ensemble by using fast k-NN approach with it.

***Keywords*—nearest neighbour, Density peak clustering,ADB-DBSCAN, k-NN.**

I. INTRODUCTION

Clustering is widely used data mining technique, various clustering algorithm are available in literature like k-means, DBSCAN, K- spectral clustering. Different clustering algorithm has different objectives so no one clustering algorithm is accurate results on categorical data. To more accurate, stable clustering result cluster ensemble is a method to combine the results of different base clustering results.[1]Clustering is widely used technique in data mining to discover pattern or extract knowledgeable information from data. Most of the available clustering algorithm work well when data contain in arbitrary shape or with low noise. But in actual it is not possible to always data contain high percentage of data with low noise because intrinsic randomness or error in measurement.[2]Clustering is the process indata mining to organized data according to the groups for knowledge extraction. Many clustering algorithm are available like hierarchical clustering algorithm, partition based clustering algorithm, density based clustering algorithm. Density based algorithm able to find cluster accurately in data with different size and shape and also robust to noise. Most of the density based clustering algorithm fail to identify cluster properly in massive data because of computational complexity.[3]In data mining clustering aims to group data according to their objectives. Different clustering algorithms are available but still clustering algorithm does not gives proper cluster if data contain noise or prior knowledge of data set is not available. It is also sensible to parameter and cannot identify cluster with different size and densities. Density based clustering algorithm can identify cluster with different size and without prior knowledge of dataset. But it fails to identify the cluster with different densities.[4]**:** As the data is increasing day by day in many applications like business management, cloud computing, social media clustering algorithm is having more in demand. DBSCAN is density based clustering algorithm widely used because it can identify cluster with varying shape and size and also robust to noise. But DBSCAN suffer many problems in cluster identification. DBSCAN performance dependent on parameters selection which cause the DBSCAN suffer for identifying cluster in heterogeneous density data sets [5] Clustering is a widely used data mining technique. The target of clustering is to cluster information as indicated by their likeness, so that it can be used in machine learning, pattern reorganization, etc. Different clustering algorithms are available such as hierarchical clustering, partition-based clustering, density-based clustering. In density-based clustering, the density depends upon the cut-off distance which may lead

to stoical error, and the density-based grouping isn't appropriate for multi-scale information [6]. Clustering is coordinated information with the end goal that comparative articles are in a similar cluster and disparate in another cluster. Clustering techniques are classified into partitioning, hierarchical, model, density, and grid-based approaches. A density-based clustering algorithm can identify clusters properly in arbitrary shape and automatically identify a number of clusters, and it is powerful to ambiguity.[7] The data which is more similar are put into one cluster and data with dissimilarity is put into another cluster. Different clustering techniques are available like k-means, DBSCAN but their performance depends upon the parameter selection like K-means algorithm accuracy depends on initial parameter selection and number of clusters. [8] The finding cluster with descending density. This approach needs to characterize various clusters ahead of time. It likewise needs a manual determination of cluster center in the decision graph which may lead to data points may assign to the wrong cluster or multiple clusters.[9] In clustering, a density-based clustering algorithm is more popular because it can identify clusters with various sizes, shapes. In a density-based clustering algorithm data points that have high density and relatively longer to each other are selected as a cluster center. The points with low density are assigning to the cluster using assignment procedure. . DPC and DBSCAN are parameter dependent algorithm need to giver input from users like dc in DPC which is cut off distance and €and MinPts in DBSCAN which is the radius of the neighbourhood for an object, and a minimum number of points in a neighbourhood respectively.[10] Clustering is an exceptionally viable solo learning approach and it is broadly utilized in different fields including information mining design acknowledgment and picture investigation. Due to the evidence importance of clustering various, clustering algorithms have been proposed in the previous many years, which include partitioning methods, density based clustering and hierarchical clustering. The partition based clustering is simple and effective but more fragile to noise. A density based algorithm- DBSCAN which is robust to noise, shape and densities but it's performance totally depends upon the parameter cut off distance, which it need to select in advanced. In Density based clustering framework (DCF)Density partition needs to be repeated for large data sets and cluster with more overlap so it may be time consuming[11].Recently Density and Distance based clustering algorithm is invented in which time utilization is high yet center thought is novel.. Inspired by their work another KMDD technique for enormous spatial informational collections and high dimensional datasets is imagined, .AS the k-means are request delicate to the data, result of KMDD are not stable. In some case core and noise are difficult to identify [12] .Density Peak Clustering With Symmetric Neighbourhood Graph(DPC-SNR) is suitable for larger data sets. Clustering of data sets can be done correctly. Cluster center can successfully identify cluster center regardless of their distributions and dimensionality. The efficiency of algorithm is depends upon the selection of parameter k[13]. ICCK-K-a means novel method for initial cluster center selection which automatically determines the number of clusters. It is vigorous, more steady, and gives excellent outcomes yet the age of a distance framework causes a lot of time utilization [14]. Clustering algorithm based on message passing.(MPSC) effectively deal with multi-scale data sets, however, it devours a great deal of time when managing huge scope information, very much like conventional otherworldly bunching [15]. Clustering is a broadly utilized unaided information mining procedure. In clustering, the main aim is to put similar data objects in one cluster and dissimilar in another cluster**. The k-implies is the most famous clustering algorithm because of its effortlessness.** But the execution of the k-means clustering algorithm depends upon the variable selection. **Parameter selection like number of cluster and initial cluster center are key of k-means algorithm. Distance augmentation method, density method quadratic clustering method is for the most part utilized to initial cluster selection.** [16] Clustering is widely used for information extraction. In Natural Language Processing (NLP) extracting information from text sources is an important task. Some language technology required text information for better performance. Topic modelling is important for some applications likes (NLP) and information retrieval. It is an unsupervised methodology where a pre-determined number of themes is separated from a specific arrangement of reports on measurable ideas.[17]For convenient use of social media site, the user uses personalize tags and familiar words according to their own understanding. Tag is a keyword that gives more information about the object. Many developers take the advantage of tag information to build personalize tag recommendation systems for users. But there are many problems in tagging system because of its free nature and lack of explicit meaning in the social tag. Different clustering techniques are used in tag development such as K-means and it's improve version, hierarchical clustering, latent semantic analysis (LSA) with clustering. But this technique doesn't use semantic relation between the tags hence less accurate and real clusters are found [18]. The actionable knowledge extraction from text documents is a complex process and required a lot of

expertise. In-text mining needs to find previously unknown and implicit data from text documents which include a grouping of data with similar content, topic modelling and detection, clarification model, document summarizations, and document querying. It is a multi-step process that required multiple algorithm implementation and parameters set by the user. It has high computation cost and time consuming because it needs the best joint analysis selection of techniques.[19]Day to day data available on crime is increased. It is not feasible to study that data manually to solve crime related queries. Thusly natural language preparing strategies are most broadly utilized for handling and taking care of such unstructured information for criminal examination. Past strategies utilized in natura language handling are administered procedures and required a ton of human oversight from the criminal business.[20]

This paper study different clustering methods like selective clustering ensemble, DBMAC algorithm, fast D-peak algorithm,multi-density clustering algorithm,ADA-DBSCANalgorithm and proposed improved approach.

II. BACKGROUND

Author proposed a new selective clustering ensemble algorithm MMSCE based on multi-modal metrics. To obtain diverse base partitions it uses alternately k-means and hierarchical clustering algorithm combine with random projection method. To improve the quality of clustering it proposed new selection strategy for number of cluster in k-means algorithm. Multi-model matrix is used to evaluate quality and diversity of clustering partition. Clustering results on benchmark data set shows that it improves the both quality and diversity of clusters.[1]

Author proposed Density-Based Multi-scale Analysis for Clustering (DBMAC)-II which is the advance version of DBMAC. DBMAC has an assumption that all clusters are homogenous and con not work well when data having varying density cluster. (DBMAC)-II remove the limitation of DBMAC it can find cluster accurately without having any assumption by performing multi-scale analysis iteratively. [2]

Author proposed Fast density peak clustering for large scale data based on k-NN. It computed density by using fast KNN algorithm such as cover tree which improve the speed of density computation. This density is called KNN density. These densities are classified as local density peaks and non-local density peak. This paper also proposed algorithm to calculate density in two ways to reduce time complexity. An experimental result shows effectiveness and superiority of fast d-peak than other density peak algorithm.[3]

DENSS algorithm to identify cluster with different data size, density and noise is proposed by author. It does not require prior knowledge of dataset and it is insensitive to parameter also. It identify cluster based on the similarity of neighbourhood distribution and number of same neighbour of two objects. If the densities of the data points has difference then it group into different clusters. This algorithm is compare with the seven clustering algorithm on real and synthetic data sets. The result shows it is superior to other clustering algorithm. **[4]**

Authorproposed a novel Adaptive Density-Based Spatial Clustering for Massive Data Analysis (ADA-DBSCAN). To address the issue of linear connection this paper proposed the concept of data splitter and data block merger which can find local to global cluster. This paper also proposed simple parameter adaptation technique to find cluster with varying density data and remove the parameter dependency of algorithm. Experimental result shows that (Ada-DBSCAN) can perform better than DBSCAN and identify cluster accurately.[5]

This paper focused on five different techniques such as selective clustering ensemble, DBMAC algorithm, fast D-peak algorithm, multi-density clustering algorithm, ADA-DBSCAN.

This paper is organized as follows

Segment I,gives brief about Introduction. Part IIexplains the Background. Segment III, analysis the previous works. Segment IV give view about existing methodologies. Segment V discusses attributes and parameters and how these are affected by clustering techniques, Segment VIproposed method Section VII for outcome result section VIII Conclude this. Finally, Segment IX comment on future scope.

III. PREVIOUS WORK DONE

Hongling Wang et al [2018][1] had proposed Two-level-oriented selective clustering ensemble based on hybrid multi-modal metrics. It is novel method which gets robust and accurate cluster result according to the quality and diversity and accessed by hybrid multi-model metric from two levels: clustering labels level and data structure level.[1]

Tian-Tian Zhang et al [2018][2] had proposed Density-Based Multi-scale Analysis for Clustering in Strong Noise Settings with Varying Densities Adaptive. This is an improved version of DBMAC which is invented for identifying cluster with

strong noise, arbitrary shape, and different size data. But it cannot find cluster with varying density having assumption cluster are homogenous in nature. To remove the limitation it used multi-scale analysis for finding cluster iteratively. It also developed self stopping multi-scale test method and uses different radius rather than fixed one. It is superior to strong noise robust clustering technique such as Skinny-dip.[2]

Yewang Chen et al [2019][3] had proposed Fast density peak clustering for large scale data based on KNN. This paper replaces density with KNN density and used fast KNN based algorithm to calculate density. This paper also proposed algorithm to calculate the density in two different strategies with complexity O (n). An experimental result on data shows effectiveness and speed of density computation for massive data.[3]

Xingxing Zhou et al [2019][4] had proposed a multi-density clustering algorithm based on similarity for dataset with density variation. This algorithm can identify cluster with different density, size and shape. It is robust to outlier and noise. It identify cluster based on similarity diversion and shared neighbours. The data in same cluster having homogeneous density[4]

ZihaoCai et al [2020][5] had proposed Adaptive Density-Based Spatial Clustering for Massive Data Analysis. This paper proposed the concept of data splitter and block merging to identify cluster from local to globally. It developed the adaptive parameter selection technique which improves cluster quality in heterogeneous density data and reduced parameter dependency of algorithm which is suitable for massive data.[5]

IV. EXISTING METHODOLOGIES

Many clustering algorithm for multi- value density data have been implemented over the last several decades. There are different methodologies that are implemented for different heterogeneous density data i.e. selective clustering ensemble, DBMAC algorithm, fast D-peak algorithm, multi-density clustering algorithm, ADA-DBSCANalgorithm

A. Selective clustering ensemble:

This paper proposed a new selective clustering ensemble algorithm MMSCE based on multi-modal metrics. For generating clustering base partition it uses different clustering algorithm. Not only diversity but also quality is used as objectives to generate the base cluster. So base cluster generated by proposed method is full of diversity and quality. For selection of sub-cluster it proposed new selective method based on multi-model matrix which takes advantages from both label information and structure information. The following figure represent model of selective clustering ensemble algorithm.[1]

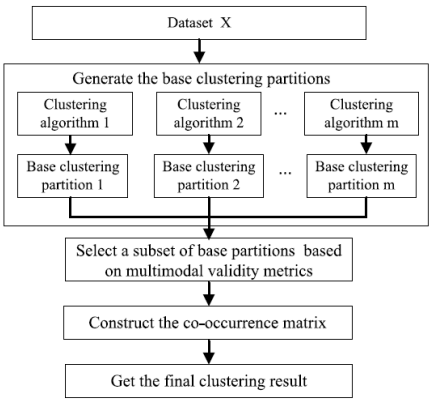

Fig1: Framework of proposed selective clustering ensemble.

B. DBMAC algorithm:

The main objective of this paper is to develop a method which can identify cluster in strong noise data of arbitrary shape and different densities. This paper develops density based filtering criteria for noise data and then applied density based clustering method to find cluster. It applies multi-scale analysis iteratively and find cluster with homogenous density in each iteration. Then it apply standard density based algorithm to find the partial results. A statistical test is adopted to stop the multi-scan iterations. Data is uniform when $E[T] = \frac{2M1M2}{L}$ where T is the X-Y join count in MST with where C is the number of edge pairs in the MST sharing a common node and L=M1+M2. $M1$ given data points after gap filling, and the $M2$ points labelled Y are uniformly generated. [2]

C. Fast D-peak algorithm:

This paper proposed fast D-peak algorithm with natural dimension clustering capability and able to deal with massive data within expected time. To improve the speed of density computation it converts the D-peak density into KNN density which can be computed by fast KNN algorithm. It distributed density as local density and non-local density and computer the distance (δ) for both of them by using their KNN nearest neighbors. Construct the hierarchical tree. Last it select data points as clusters centers having largest value of KNN density and distance and find the final cluster.The local density ρi of data point i is defined as:$\rho i - \sum_j X(di, j - dc)$. Where, χ(x) = 1 if x <0 else χ(x) = 0, and dc is a cut off distance which is predefined by user.δi is measured by computing the minimum distance betweenthe point i and any other point with higher density: $\delta i = \min i, j$ where $\rho i > \rho j$ [3]

D. Multi-density clustering algorithm:

This paper proposed a new multi-density clustering algorithm. It uses the concept of diversion similarity and neighbour correction to identify cluster with multiple density data set. For neighbour correction it uses SNN algorithm with contain intersection of KNN of two objects such that it is symmetric. Then k-nearest-neighbour based on divergence (KNND) is calculated based on the criteria diversion similarity of point is greater than another point. Data points that satisfy both can group into clusters. The diversion value of a point is computed as the average distance of any two points in set with its neighbours Data point in same cluster having the same density. The proposed algorithm is universal can identify cluster with different data size, density and shape, outlier of any feature. **:** Similarity between Point p and Cluster C= {p1, p2, p3….pm} (SPC) is obtained by using following function:

$$SPC(C,P) = \frac{\sqrt{\min(P.div, C.div)}}{\sqrt{\max(P.div, C.div)}} * \frac{\sqrt{\min((dis(p,C), C.div)}}{\sqrt{\max(P.div, C.div)}}$$

Where dis (p; C) refers to the smallest distance between p and the points in cluster C. [4]

E. ADA-DBSCAN Algorithm:

This paper proposed Adaptive Density-based Spatial Clustering of Applications with Noise (ADA-DBSCAN) to solve the problem of linear connection and improve parameter selection and efficiency of algorithm is improve when clustering massive data. By applying data splitter it divide the data uniformly. . The uniform distribution means data block have same density data which is independent density of data block and containing linear connection. Local density cluster is found for each data block for finding the cluster globally it merges data obtained cluster. First this paper adopts top-down approach for splitting data into blocks and bottom up approach for clustering. The framework for adaptive DBSCAN as follows [5]

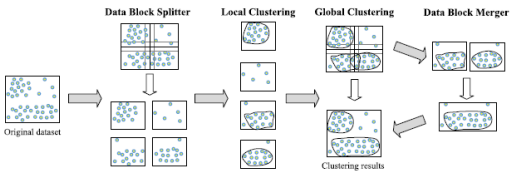

Fig2: framework for adaptive DBSCAN

V. ANALYSIS AND DISCUSSION

The experiments conducted on five UCI data set using MMSCE, FCEKM, FSCE, RCEKM, RSCE and SCEKM algorithms used to ensemble clustering partition and compare by CA and NMI. The result shows that MMSCE algorithm gives highest accuracy of classification compare to other algorithm. It is superior in quality on iris, wine, breast cancer datasets than other algorithm. The proposed strategy to ensemble partition cluster is outperformed in diversity maintenance than other.[1]Compare to other clustering algorithm DBMAC-II show superior robustness against dimensionality, Numbers of clusters, noise percentage, and density levels of DBSCAN and OPTICS performance is largely dominated by relative density of data. DBMAC-II can identify noisy object correctly. [2]The experiment on synthetic data is done and comparison with EDDPC, PDPC, D-Peak is performed. From experiments it observed that the distance computations of Fast-D-Peak are far less than EDDPC. The cost of Fast-D-Peak is lies in computation of nearest neighbour and it reduce the cost of determining δ. [3]On comparing DENSS with OPTICS, DBSCAN, K-means, SNN-DPC on multi-density synthetic data only DENSS can identify cluster accurately and boundaries of cluster identified precisely as well. It evaluate usability and effectiveness of algorithm it take four feature from iris data sets randomly where it outperforms other algorithms.[4]**:** Comparison ofAda-DBSCAN has done with different density based algorithm such as DBSCAN, OPTICS, HDBSCAN, Any-DBC, SDC. The time cost of Ada-DBSCAN is similar DBSCAN to small data set but need less time when apply on large data set because of data splitting and merging technique. The efficiency of Ada-DBSCAN is superior to all density-based clustering algorithm.[5].

Mobility scheme	Advantages	Limitations
selective clustering ensemble	It combines characteristic of dataset and different evaluation criteria in base cluster partitions. It can produce cluster with desire quality and diversity by adjusting parameters in k-means algorithm. It gives more accurate and robust cluster.	Time complexity is high.
DBMAC algorithm	It can identify cluster with strong noise data. Dimensionality, variation in density not affects the algorithm performance.	. Scalability of algorithm is less.
fast D-peak algorithm	It is suitable for large scale data. Reduce time for computation of density.	The value of K impact the performance of algorithm
multi-density clustering algorithm	It can identify cluster with different size, shape and density. It is robust to outlier and noise. It is not sensitive to parameter and not require prior knowledge of data set.. It can identify cluster with distinct diversion zone.	Similarity threshold may vary with different data set.
Ada-DBSCAN algorithm	Massive data analysis. It reduced the time complexity of data computation. Need not explain properly.	Data blocking and merging

TABLE 1: Comparisons between different clustering techniques

.

VI. PROPOSED METHODOLOGY

These paper proposed method called "Fast selective clustering ensemble with k-NN". A selective clustering ensemble is able to find more accurate result with respect to diversity and quality but its time consuming. In these proposed method , K-nearest neighbour is apply to fast compute the density of the data point by using KNN density concept. This approach build the KNN cover tree and divide the data point into local and non-local by calculating KNN density. The parents of the non-local data peak are calculated and build the cover tree. At the same time, the values of local density peak and k is compare and build cover tree .After finding the cover tree partition based clustering algorithm is applied and subset of based partition is selected based on multi-mode validity matrix. By constructing the concurrent matrix final cluster is found.

Basic steps of algorithm:

Step1.: Load the data set.

Step2: Build the KNN cover tree.

Step3: Differentiate the local and non-local density peaks.

Step4: For the non-local density peak calculate the parents and build the KNN tree.

Step5: For the local density peak compare the number with value of k.

Step6: If the number of local density peak is less than k then find the parent of local density peak.

Step7: If the number of local density peak is more than k then construct the KNN tree and find their parents

Step8: Build the cover tree.

Step9: apply partition Based algorithm.

Step10: Select the subset partition by using multi-mode validity matrix.

Step11: construct the concurrent matrix

Step12: find final cluster

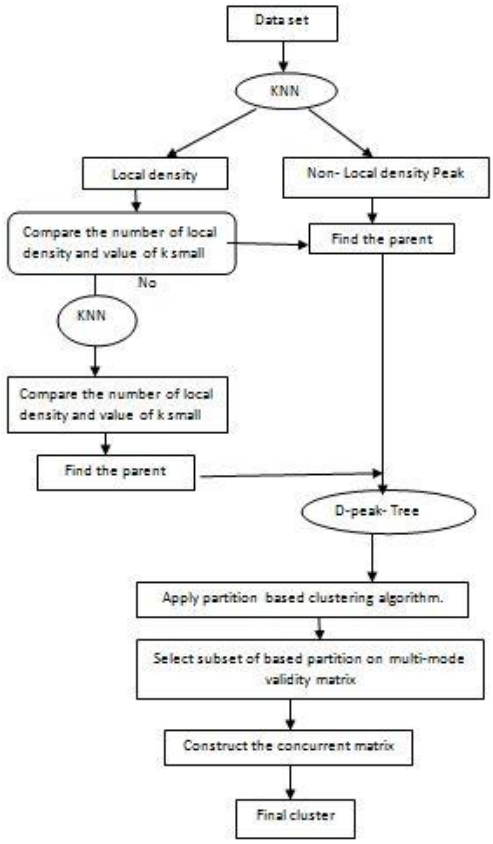

Fig3:Flowchart of "Fast selective clustering ensemble with k-NN"

VII) SIMULATION AND RESULT

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

Fig4: dataset loaded

In. fig. theGolf dataset is uploaded in the rapid miner tool.

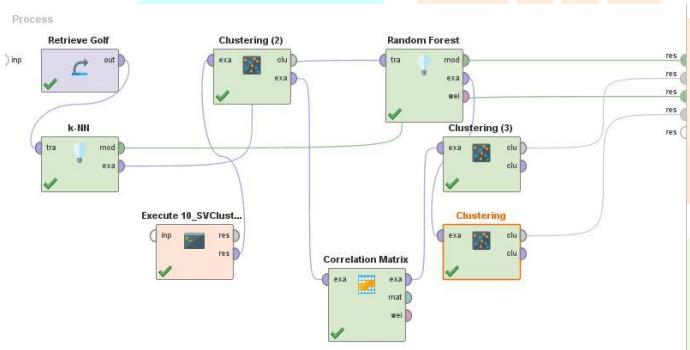

Fig5: Design of proposed Framework

Fig shows the design of proposed method constructed and executed.

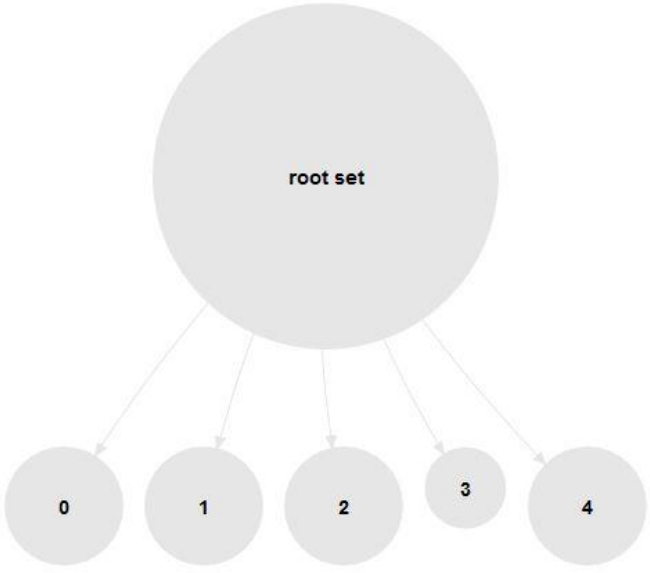

Fig6 : clustering result "Fast selective clustering ensemble with k-NN".

Theclustering result in the form of graph is shown in following fig.

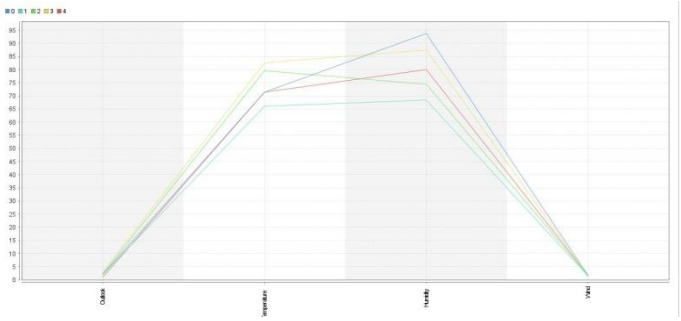

Fig7: result in graph format.

VIII) RESULT AND DICUSSION

It gives more accurate result in the form of quality and diversity without more prior knowledge about the data. It is fast approach as compare to selective clustering ensemble.

VII. CONCLUSION

This paper look into five paper which work on multi-value dataset, noise data set and data set with diversities namely selective clustering ensemble, DBMAC algorithm, fast D-peak algorithm, multi-density clustering algorithm, ADA-DBSCAN. A selective clustering ensemble can identify cluster with quality and diversity without more knowledge of data set but its consuming approach. This paper modifies this approach with concept of KNN density which reduces the time complexity of selective ensemble.

VIII. FUTURE SCOPE

Calculation complexity needs to reduce.

REFERENCES

I. **I.** Hongling Wang &Gang Liu"Two-level-oriented selective clustering ensemble based on hybrid multi-modal metrics." IEEE Acess2018

II. **II.** Tian-Tian Zhang And Bo Yuan**"**Density-Based Multi-scale Analysis for Clustering in Strong Noise Settings with Varying Densities Adaptive"

III. IEEE Access2018

IV. **III.**Yewang Chen Xiaoliang Hu, Wentao Fan, LianlianShen, ZhengZhang,Xin Liu, Jixiang Du, Haibo Li, Yi Chen, HailinLi**"**Fast density peak clustering for large scale data based on KNN" Knowledge-Based SystemsJune 2019

V. **IV**.Xingxing Zhou, Haiping Zhang, GenlinJi ,GuoanTang.**"**A Multi-Density Clustering Algorithm Based on Similarity for Dataset With Density Variation**"**IEEE Access.December 2019

VI. **V.**ZihaoCai ,Jian Wang , And KejingHe**"**Adaptive Density-Based Spatial Clustering for Massive Data Analysis"IEEE Access 2020**.**

VII. Dongdong Cheng Qingshenge Zhu, Jinlong Huang, Lijjun Yang" Natural Neighbor-based clustering

algorithmwith density peak", International joint conference on Neural network, 2016.

VIII. Avory Bryant and Krzysztof Cios. "RNN-DBSCAN: A Density-based Clustering Algorithm using Reverse nearest Density Estimates" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2017

IX. Jianhua Jiang, Yujun Chen ,Xianqiu Meng, Limin Wang, Keqin Li "A novel density peaks clustering algorithm based on k nearest neighbors for improving assignment process" Science Direct2019

X. Tong Liu, Hangyu Li, and Xudong Zhao "Clustering by Search in Descending Order and Automatic Find of Density"

XI. IEEE Access 2019

XII. Zhi Liu, Chunrong Wu, Qinglan Peng , Jia Lee And Yunni Xia "Local Peaks-Based Clustering Algorithm in Symmetric Neighborhood Graph" IEEE Access2020

XIII. Jianyun Lu, And Qingsheng Zhu, "An Effective Algorithm Based on Density Clustering Framework" IEEE Access vol.5 February 9, 2017.

XIV. Jiang Wang , Cheng Zhu, Yun Zhou , Xianqiang Zhu, Yilin Wang, And Weiming Zhang" From Partition-Based Clustering to Density-BasedClustering: Fast Find Clusters With DiverseShapes and Densities in Spatial Databases" IEEE Acess on SPECIAL SECTION ON ADVANCED DATA ANALYTICS FOR LARGE-SCALE COMPLEXDATA ENVIRONMENTS vol 6 September 4, 2017,

XV. .Chunrong Wu 1, Jia Lee TeijiroIsokawa, Jun Yao1,And Yunni Xia "Efficient Clustering Method Based on Density Peaks With Symmetric Neighbourhood Relationship" vol 7 April 4, 2019,

XVI. Yating Li, Jianghui Cai , Haifeng Yang , Jifu Zhang , And Xujun"A Novel Algorithm for Initial Cluster Center Selection" IEEE Acess vol.7 April 30, 2019,

XVII. Lijuan Wang, Shifei Ding, And Hongjie Jia "An Improvement of Spectral Clusteringvia Message Passing and Density

XVIII. Sensitive Similarity" IEEE Acess vol 7June 28, 2019,

XIX. Caiquan Xiong Zhen Hua KeLv Xuan Li"An Improved K-means text clustering algorithm By Optimizing initial cluster centers" International Conference on Cloud Computing and Big Data2016

XX. M.Alhawarat And M. Hegazi"Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents" IEEE Access2017

XXI. Jing Yang and Jun Wang

XXII. "Tag clustering algorithm LMMSK: improved K-means algorithm based on latent semantic "Journal of Systems Engineering and ElectronicsApril2017

XXIII. Tania Cerquitelli Evelina Di Corso Francesco Ventura Silvia Chiusano"Data miners' little helper: Data transformation activity cues for cluster analysis on document collections"ACM Reference format June 2017

XXIV. P. DAS,A. K. DAS, J. NAYAK, D. PELUSI, W. DING. "A Graph based Clustering Approach for Relation Extraction from Crime Data"