Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 4, 2020

EVD²BSCAN: ENHANCED VORONOI DIAGRAM DENSITY BASED CLUSTERING ALGORITHM FOR DATA CLUSTERING

T. KAVIPRIYA[1], Dr. M. SENGALIAPPAN[2]

[1]Assistant Professor,Department of Computer Technology,Hindusthan College of Arts and Science. kkavipriya09@gmail.com
[2]Dean of Computer Science,Kovai Kalaimagal College of Arts and Science. cmsengs@gmail.com

Received: 05.01.2020 *Revised: 12.02.2020* *Accepted: 22.03.2020*

ABSTRACT

Clustering concern is a systematic approach for partitioning of data objects into matching clusters and typically regarded as unsupervised learning problem. The real-time multiple constraints for objects are difficult to achieve for conventional algorithms. The appropriate cluster quantity from the data attained by the distance-based clustering algorithm occurs in seldom manner autonomously. During clustering process, missing data problem is a serious issue in many applications as it has significant effects the conclusion drawn from the data. The missing values and irrelevant data are resolved by computing the mean of other data in either subject wise manner or student wise manner. In addition this work presents a Voronoi Diagram Density Based Clustering Algorithm (VD2BSCAN) by fast search and determining density peaks. The input samples' Voronoi diagram creation is a key factor in retrieving the density information which is accomplished by the suggested VD2BSCAN. The corresponding parts of the instance space point density is influenced directly by the point cells volume. The merging of densest parts of the instance space into clusters is achieved by scanning over the input points and their Voronoi cells at a time. However, wide density variation inside the clusters is a serious concern to be considered in DBSCAN algorithm. This issue is mitigated by enhanced EVD2BSCAN algorithm by estimation of growing Density Mean (DM) for some core object, in which density of its ε-neighborhood is considered with respect to DM has its own significance. The data are collected manually from the college placement departments for the evaluation in R tool. The data are stored and fetched from Structured Query Language (SQL) database. The clustering performance metrics are employed for the analysis of the performance. The EVD2BSCAN algorithm improves the clustering accuracy significantly compared than the DBSCAN, Hierarchical Clustering (HC) analysis which is revealed by the evaluation results. The proposed algorithm outperforms well which can be utilized for various applications.

Keywords: Missing Data, Data Clustering, Enhanced Voronoi Diagram Density Based Clustering Algorithm (EVD²BSCAN), R Tool.

DOI: xxxxxxx

INTRODUCTION

The process involved in extraction of inherent, nontrivial, earlier unidentified and potentially valuable data like knowledge rules, constraints, regularities from data in databases [1-2] is termed as data mining. The transformation of processed data into valuable information as well as knowledge is greatly required due to the high demand of business managements, government administration, and scientific data analysis since there is rapid evolution of data and databases involved in those sectors.

The group formation of organizing identical unlabeled objects for making generalization through labeling is characterized as Clustering. Clustering is highly demanded in various applications since comprehensively greater technique of clustering instances is necessitated. This motivated to survey various unsupervised clustering algorithms (see e.g. [3]).

There exist high inter-cluster and low intra-cluster distance in clusters obtained from the worthy clustering technique [4]. The main motto is that each cluster data points similarity and dissimilarity across clusters is to be maximized [5]. Many researches are being carried out in machine learning, pattern recognition, data mining and statistics domains [6]. Apart from this, k-means [7], entropy soft k-means [8], inverse exponential k-means [9-10], Fuzzy C Means (FCM) [11-12], Self-Organizing Map (SOM) [13-14], Artificial Neural Networks (ANNs) [15] and Support Vector Machines (SVMs) [16-17] are utilized by the researchers for experimentation.

There are four major classification of data clustering algorithms .They are partitioning, hierarchical, density-based and grid-based which may come under more than single category. The recognition of dense and sparse regions is accomplished by clustering and for realization of overall distribution patterns. The greatest challenge involved in determining the clusters which are of various dimensions, shapes and densities involved in data along with noise and outliers. Though there exist several algorithms for detection of clusters with different sizes and shapes densities, only few available for clusters with dissimilar densities [18].

The significant factor of Density Based Clustering algorithm (DBSCAN) is that determination of clusters of diverse shapes and sizes. The capability of managing the local density variation that happen inside the cluster in high dimensional database is another difficult task by utmost of the density based clustering algorithms.

The DBSCAN has the advantage of managing noise-handling capabilities in which clusters are stated as typical densities regions alienated by low or no density regions.

This research concentrates on improving the renowned algorithm DBSCAN in a manner that recognition of clusters from uneven datasets which contains homogenous densities clusters regions. The Voronoi Diagram is greatly utilized for achieving the scalability of the proposed EVD²BSCAN algorithm by means of initial partition of the dataset followed by use of improved VD²BSCAN on each partition. The merging procedure is utilized for obtaining the actual natural number of clusters in the underlying dataset.

LITERATURE REVIEW

Spatial data clustering is particular among the significant methodologies in data mining, concerning the derivation of information through gatheringhuge quantity of spatial data from several applications, especially GIS, Remote Sensing, Computer Cartography, Environmental Assessment And Planning, etc.

In past few decades, a number ofbeneficial and standard algorithms have been introduced in spatial data clustering method, including the strategy called DBSCAN. This strategy enables the identification of all arbitrary-shaped clusters and efficient handling of noise points, whereas demands huge memory support, as it runs on overall database.

Borah and Bhattacharyya [19] introduced the enhanced version of DBSCAN, namely sampling-based DBSCAN, for efficient clustering of extensive spatial databases. The suggested sampling-based DBSCAN surpasses DBSCAN and other similar approaches, concerning the process time and sustaining the clustering efficiency, which has been proved by the provided empirical findings.

Cao et al [20] confers a novel strategy to recognize the clusters from progressive data stream, namely Den Stream. In order to abridge the clusters according to arbitrary shape, the "dense" micro-cluster (known as core-micro-cluster) has been introduced, whereas the structures of potential core-micro-cluster and outlier micro-cluster are suggested to manage and differentiate the potential clusters and outliers.

On the basis of these notions, a modern pruning system has developed to ensure the accurate weights of the micro-clusters through less memory. The value and competency of DenStream method represented through its implementation study on numerous real and synthetic datasets.

Tsai and Liu [21] bestows an innovative approach focused on the notion of IDBSCAN, that applies K-means to recognize center points of high density and enlarge the clusters from these points. Besides, it lessens the process time through choosing representative points in seeds. The simulation demonstrates the capability of suggested KIDBSCAN to deliver the clustering outcomes with optimum accuracy, whereas surpasses the methods of DBSCAN and IDBSCAN.Moreover, this approach diminishes the spending on I/O.

Chen and Tu [22]suggests a clustering stream data structure, called D-Stream that includes density-based method. This algorithm employs an online component, and official component. Every source record of data is mapped into a grid by online component, whereas the official component calculates the density of the grid and clustering the grids according to their density.

The algorithm involves the method of density decomposition, in order to grasp the dynamic variations of data stream. The suggested algorithm assures the real-time generation and modification of clusters in an efficient way, by utilizing the complicated correlations within data density, decay factor, and structure of cluster. Besides, a speculatively strong approach is designed, which ensures the identification and eradication of mapped sporadic grids by outliers, concerning the radical enhancement of the efficiency of system's space and time.

This approach accelerates the data stream clustering, simultaneously maintains the standard of clustering. The empirical findings depict that the suggested algorithm possesses the optimum quality and efficacy, besides it has the potential to identify the arbitrary-shaped clusters, and properly observe the progressing activities of real-time data streams.

Liu et al [23] presents an innovative algorithm to assess the datasets with varied-density, namely VDBSCAN that considered a few approaches prior to implying conventional DBSCAN algorithm, in order choose many parameter EPS values for varied densities corresponding to k-dist plot VDBSCAN. The utilization of various EPS values enables the recognition of the clusters that accompanying varied densities simultaneity.

DBSCAN algorithm is adopted to ensure the clustering process of overall clusters with their respective density for each EPS value. For further process, the instance of tagging both of the denser areas and sparser areas as single cluster have evaded by skipping the clustered points. Ultimately, the trials depict that the VDBSCAN approach is proficient to handle the clustering of inconsistent datasets by using a synthetic database accompanying two-dimension data.

Borah and Bhattacharyya [24] expand the algorithm of DBSCAN, concerning the identification of the clusters with density variations. In fact, the local densities are homogeneous inside a cluster, if there is any major variation in densities then the neighboring areas will be divided into individual clusters.

Hence, the algorithm tries to identify the natural clusters on the basis of density, as it is expected to be undivided by any sparse area. The algorithm contains the complicacy in calculation, which is O (n log n). The further benefit is known to be the reactivity of input parameter ϱ, which diminishes the crucial drawback of DBSCAN.

Ram et al [25] suggest an advanced DBSCAN algorithm that continuously monitors the changes in local density inside the cluster, besides it estimates the variance of density for each core object, pertaining to its ε-neighborhood. The core object has been enabled to expand, if its value is lower than or similar to threshold value, whereas fulfilling the homogeneity index according to it'sε-neighborhood. The empirical findings indicates that the suggested algorithm of clustering provides enhanced outcomes.

Fahim et al [26] intends to improve the renowned algorithm of DBSCAN, through which it can be scaled and determine the clusters from inconsistent datasets, where the clusters are the regions of homogenous densities.

Eventually, the outcomes denote that the proposed algorithm has obtained the scalability, in which the k-means algorithm has employed to acquire primary partition of the dataset that includes the advanced DBSCAN over each partition, besides the amalgamation process, have performed to extract the authentic natural number of clusters from an underlying dataset. Therefore, the suggested algorithm comprises of three phases. Empirical findings derived by employing synthetic dataset, which signifies that the suggested algorithm of clustering is more quick and scalable when compared to other advanced identical DBSCAN.

Fawzy et al [27] suggests a modern algorithm, which is an extended version of DBSCAN which enable the discovery of the cluster based on density, namely DBCLUM.

Though the DBSCAN recognizes the arbitrary-shaped clusters, it still lacks to identify the clusters with varied density.DBCLUM has designed to surpass these issues. DBCLUM identifies the individual clusters and incorporates them if they have identical density and linked. Thus, the clusters with varied densities and nearby clusters have discovered by DBCLUM, which considered as a major advantage of this method. Implementations exhibit the aforementioned ability of DBCLUM, besides it depicted that the DBCLUM is 11% to 52% quicker than DBSCAN.

PROPOSED METHODOLOGY

In this section, the proposed model of data clustering for grouping the similar data and then clustering algorithm details are explained. The proposed model consists of two major processes: 1) Irrelevant data processing and 2) Missing value data processing.3)EVD²BSCAN algorithm

The dataset considered for the processing is the student database collected manually from the colleges for the job placement process. This database contains the students' information including name, father's name, gender, date of birth, roll number, register number, name of schools studied SSLC & HSC, name of college, medium of study, degree, the marks and percentage obtained in each course with individual subjects and marks for soft skills and achievements. The above details are collected for 6324 students completing their college under graduate (UG) degree from Bharathiar University in the 2017 academic year.

The students belong to different specializations namely Computer Science, Information Technology, computer applications, commerce and others in the UG. Thus the dataset consists of 75 columns and 6324 rows of data which are processed in R.

In this context, the irrelevant data is considered as the marks obtained by students that are mistakenly marked above 100. These data are needed to be removed and replaced by a new concept. The removal of the irrelevant data also results in missing data. Similar to irrelevant data, the missing data are also caused due to mistakes like breaks in

collection, non-inclusion of data or removal of false data without replacement.

Missing Value Data Processing

In this proposed model, two methods are utilized for filling the missing data and also for replacing the irrelevant data. The two models are subject wise and student wise. The first model is by computing the mean of the particular subject across SSLC, HSC & US and using it to fill the missing values.

The second model is by taking mean of the marks obtained by the particular student from in the course where the data is missing and uses it for replacing the data. These two methods of filling the irrelevant data is analyzed using correlation and regression analysis [28]. This can be illustrated by the sample example. Randomly consider a subset of data from the whole dataset.

DBSCAN

Density-Based Spatial Clustering of Application with Noise (DBSCAN) [29] is an approach that generatedensity-based clusters in high-dimensional database, which is either spatial or non-spatial, in context of noise and outlier.The subsequent definitions are the fundamental of this work, where the comprehension of DBSCAN included in [29]:

Definition.1: Nε(p) indicates the ε-neighborhood of an object p, which is determined as the sum of objects existing in the radius, that is expressed by Nε (p) = {qεD | dist (p, q) ≤ }.

Definition 2: Where, | $N\varepsilon\ (p)$| ≥ μ (least objects), an object p is assumed to be Core object.

Definition 3: Where, $p \in N\varepsilon(q)$ and q is a Core object, then object p is considered as directly density reachable from an object q as to μ.

Definition 4: In context of the chain of objects $p_1 \ldots p_n$, p_1= q, p_n= p, an object p considered as density-reachable from an object q, hence p_{i+1} is direct density-reachable from p_i as forε and μ.

Definition 5: If object ooccurs, then object p implied as density-connected to an object q according to $\varepsilon$and μ, hence p and q are known to be density reachable from o as for ε and μ.

Definition 6: An object that lies at the boundary is not a Core object but a member of a cluster, whereas an object that exists in none of the clusters has considered as an object of noise.

Definition 7: A cluster X is known to be non-empty subset of database for every p, q, as regards $\varepsilon$and μ: then,$q \in X$ and p is density-connected to q, if $p \in X$, q is density reachable from p.

DBSCAN [29] recognizes the density-connected clusters through identifying specific p among its core object. Besides, it estimates the entire density-reachable objects from p, and this estimation process has iterated in order to operate the set of density-reachable objects.

The ε-neighborhood of every object p has examined by DBSCAN in the database, where, N (p) of an object p involves at least μ objects, that is to say, a new cluster X has generated, that comprising overall objects of $N\varepsilon\ (p)$, if p is a Core object.

Afterwards, it has examined the neighborhood of overall objects that is yet to be processed. In further process, the neighbors of object q included to X, which are not designated to X earlier and their ε neighborhood has been inspected, where object q also considered as a core object. The repetition of this process continued till the current cluster X could not be included with any further object. In the DBSCAN algorithms the representation of the data points becomes very difficult task.

Voronoi Diagram Density Based Clustering Algorithm (VD²BSCAN)

To solve the DBSCAN density information representation problem, proposed work accesses density information by constructing a Voronoi diagram for the input sample. The volumes of the point cells directly reflect the point density in the respective parts of the instance space.

A multidimensional Voronoi diagram [31] is a partitioning of the instance space $\mathbb{R}^d$ in regions R_j with the properties: Each center c_j lies exactly in one region R_j , which consists of all points x 2 Rd that are closer to c_j than any other center c_k, $j \neq k$:

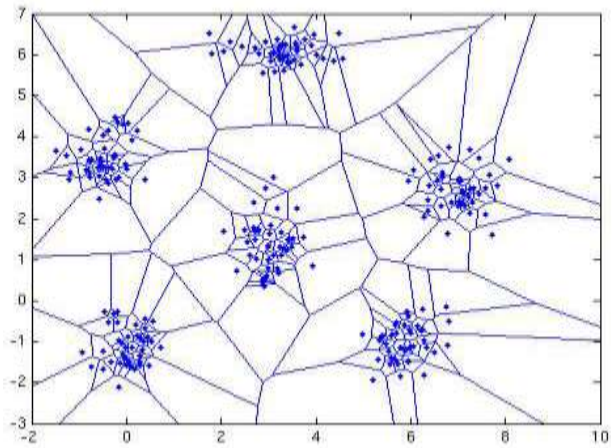

Fig. 1: A Voronoi Diagram for Data Coming from Six Clusters

Voronoi points indicates the centers c_j. In unsupervised leaning environment, the sample points si apparently specifies Voronoi points, where sample $S = \{s_1, .., s_n\}$ is expressed. Let us call the region allied with a Voronoi point c_j its cell, as the overall points of a cell are nearby to its respective center instead of some other center, besides the borders of the cell properly rest in between of two centers.Figure 1 illustrates the Voronoi diagram for dataset properly accompanied by six separate clusters. The Delaunay tessellation (also known as Delaunay triangulation) of Voronoi diagram for a point set is known to be its dual.In the Voronoi diagram, the vertices represent the centers of Voronoi cells (the initial data points) and edges that unites any two vertices have mutual boundary. The correlation of any two-point p and q acquires the same set of edges, which accomplished by a ball B which already existed there to progress through p and q, whereas its interior or boundary does not include any other points of S. The triangulation of a point set S known to be a planar graph along vertices S, and the highest set of straight-line edges, thus it was termed as Delaunay triangulation. If any additional straight-line edge is included,will exceed to other edges. A subset of edges of a triangulation a tessellation of S if it contains the edges of the convex hull and if each point of S has at least two adjacent edges.

Expressed a data set of points $S = \{s_1, .., s_n\} \subseteq$ to be clustered initiate through the Voronoi diagram construction for S. Generally, (n log n) has viewed as the calculation complicacy of constructing the Voronoi diagram [31][32], whereas solely needs the linear time in a few limited cases. Therefore, the two-dimensional plane has taken to account, concerning the simplification of illustration. The algorithm is provided as an input parameter a threshold value max, which represents that extreme volume has authorized to a cell and yet able to be merged into a progressing cluster. Then, the space density of local instance has to be estimated using cell volume, as it is necessitated to ensure the sufficient high density in order to process the further incorporation into a cluster.So, the distinct clusters expected to be enlarged to any size until the adequate local density obtained. After all cells of at most volume max have been taken care of have comparatively large cells remaining.

Initially, the smallest Voroni cell of the point is selected, as it does not include any known neighboring cells. Hence, it simply designate this cell with class label 1. The further consideration by the cell might be a neighborhood of former cell's object p, which would be identified by the two sharing a corner point q. Consequently, if current cellhas lower value than threshold value μ, they will be united together and adopt the class label of first cell, if not, the second cell possibly will not become a ε-neighborhood of an object p of the former one, that is to say, it acquires self-class label in this instance. Nevertheless, yet there is possibility to merge the two points p, q from the same cluster and later cluster into the same cell, by reason of the mutual ε-neighborhood of an object p cell. Generally, numerous neighboring

cells accompanying various class labels have comprised in a cell and a few of them have yet to be labelled. According to the magnitude the known neighbors are performed. The considered cell has amalgamated with its neighbor, which has least class label, besides its labeled neighbors consider the same class. The cells remain integrated until their volume keeps below the value of μ. It is not required to process through the remaining neighbor cells, once obtained the highest value. Conversely, the center points accompanying a cell with higher volume than threshold μ have merged to the nearby neighboring class.

Algorithm 1: Voronoi Clustering(S, ε, μ)

1. Create the Voronoi diagram for the sample S = {s_1,..,s_n}.
2. Estimate the Voronoi cell volumes and sequence the points correspondingly. Without loss of generality, let the acquired order be S = {s_1,..,s_n}
3. For i = 1 to n do

Getε-neighborhood of an object p
If the cell volume $R_i$relatedto s_i is at most μ
Then merge R_i with an adjacent cluster with the smallest class number if one exists, otherwise assign a new class number to the cell
Else
Designate R_i to the closest $N\varepsilon(p)$neighboring cluster

EVD²BSCAN Algorithm

Nevertheless, since VD2BSCAN struggles with the issue of density disparity inside the clusters, the algorithm of EVD²BSCAN has been suggested in this segment to surpass this issue, which focus on the notion of creating the clusters by choosing core object. Afterwards, the Density Mean (DM) has estimated for developing cluster prior to enabling the enlargement of a core object p, which is yet to process. Subsequently, it calculates the Density Mean (DM) accompanying the ε -neighborhood of the unprocessed core object p. If the Density Variance of the developing cluster (according to DM) is lower than the defined value of threshold μ and variation within lowest and highest objects existing in the ε -objects' neighborhood, which are known to be developing cluster's objects, besides the unprocessed core object p is authenticated for enlargement, but only if its neighborhood objects are lower than the defined value of threshold μ, or else, the object just included into the cluster.

Density Mean (DM): It is indicated by DM(C). The subsequent expression defines the DM of a growing cluster:

$$DM(C) = \frac{\sum_{o \in C} N_\varepsilon(O)}{|C|} \quad (1)$$

Where the $N_\varepsilon(O)$ is the density of the object O existed in the $\varepsilon-$ neighborhood.

PERFORMANCE EVALUATION

Herein, two types of implementation have assessed in R, which is subject-based calculation and subject-based calculation. R is known as a language as well as an environment, specifically made for computational statistics and graphics, which is a GNU project and the product of Bell Laboratories, which is identical to the S language and environment. It can be said that R is another version of S, as it operates on more of the code, which has written for S and it stays unchanged, whereas R has a few unique features. R offers wide-ranging approaches in the domains of Statistics (e.g. Clustering, Classification, Classical Statistical Tests, Linear and Nonlinear modeling, etc.) and Graphics, besides it is extremely expandable. Frequently the S language is the vehicle of choice, concerning the analysis of statisticalapproach, while R delivers an Open Source for involvement in that operation.

The identification of effective pre-processing approach has accomplished through determining the performance, concerning the analysis of correlation and regression. The correlation is particular among most familiar as well as highly beneficial statistics, besides it defines the value of relationship within two variable by a solitary number. The correlation analysis process denotes the observation of strength regarding the accessible data's relationship, whereas the regression analysis deals with the identification of the relationship amid the dependent variable and independent variable/s. The relationship system has conjectured and evaluated values of the parameters have employed for developing an estimated regression equation.

Concerning the measurement of the precision of clustering, cluster association of the pairs, which contains four categories: Pairs that are mutual in both clustering (entirely shaded segments), Pairs that do not appear in both clustering, Pairs only appear in the first clustering, and Pairs appear only in the second clustering Table 1.

Table 1: Categories in Pairing Clustered Objects

Clustering	Pairs in P	Pairs not in P
Pairs in Q	a:=Pairs in both	b:Pairs in Q
Pairs not in Q	c:=Pairs in P	d:Pairs in none

Table 2:Some Well Known Pair Counting Formulas

Precision	Recall	F-measure	Rand	Jaccard
a/(a+c)	a/(a+b)	2a/(2a+b+c)	a+d/(a+b+c+d)	a/(a+b+c)

Such categories are the working group of overall measure indices for pair counting. It is able to derive an indicator for agreement and disagreement of both clustering, through totaling the pairs in every category. Table 2 demonstrates the precision and recall, which has computed by familiar asymmetric approaches. For comparison of two clustering without a given gold standard, symmetric measures, especially the F-measure (combining precision and recall) or the Jaccard-index [32] are highly suggestable. The Rand index [33] is described as the ratio of the pairs in both as well as the pairs in no clustering to all pairs, represents the probability that two objects are considered as same in both clustering.

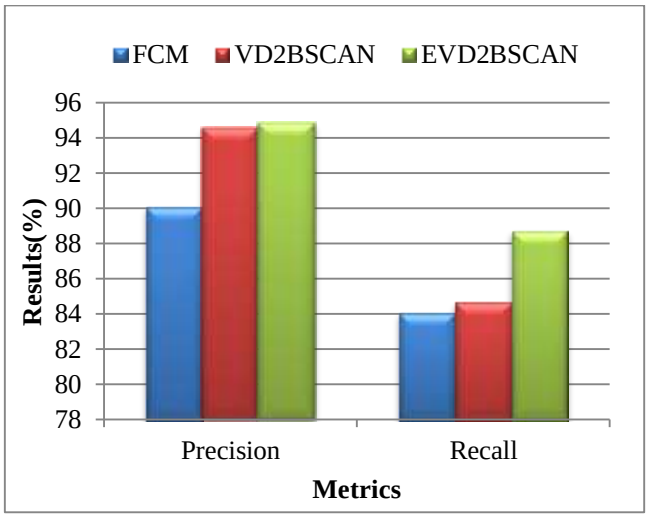

Fig. 2: Precision and Recall Metrics Comparison vs. Density based Clustering Algorithms

Fig. 2 compares the implementation outputs of three individual approaches of clustering, like FCM, VD²BSCAN, EVD²BSCAN, concerning Precision and recall metrics.

From the results it concludes that the proposed EVD²BSCAN produces precision results of 94.83% which is 0.24% and 3-4% higher when compared to VD²BSCAN and FCM clustering methods respectively. Similarly the proposed EVD²BSCAN clustering algorithm produces recall results of 88.71% which is 4.03% and 4.68% higher when compared to VD²BSCAN and FCM clustering methods respectively.

Table 3: Performance Comparison Metrics

Metrics	FCM	VD²BSCAN	EVD²BSCAN
Precision	90.09	94.59	94.83
Recall	84.03	84.68	88.71

F-measure	86.96	89.36	91.67
Rand	80.00	83.33	86.67
Jaccard	76.92	80.79	84.61
Error rate	20.00	16.67	13.33

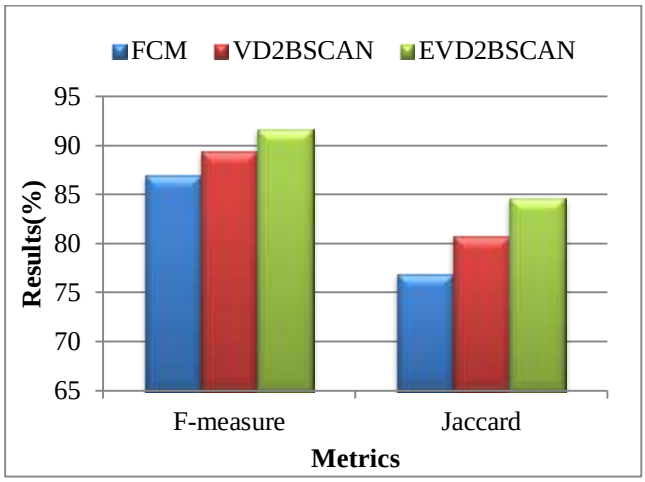

Fig. 3: F-Measure and Jaccard Metrics Comparison vs. Density based Clustering Algorithms

In Fig. 3, the chart compares the results derived from the performance of three individual clustering strategies, like FCM, VD²BSCAN, and EVD²BSCAN, concerning the F-measure and Jaccard metrics. The results conclude that the proposed EVD2BSCAN produces F-measure results of 91.67% which is 2.31% and 4.71% higher when compared to VD²BSCAN and FCM clustering methods respectively. Similarly the proposed EVD²BSCAN clustering algorithm produces Jaccard results of 84.61% which is 3.82% and 7.69% higher when compared to VD²BSCAN and FCM clustering methods respectively.

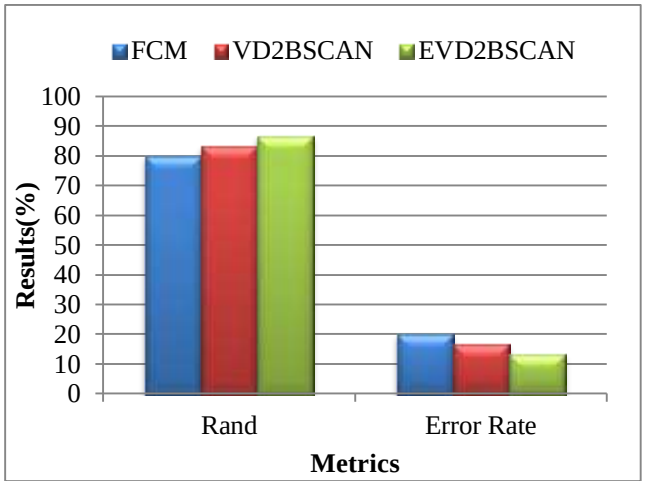

Fig. 5:Rand and Error Rate(ER) Metrics Comparison vs. Density based Clustering Algorithms

In Fig. 5, the chart compares the implementation outputs of three individual approaches of clustering, specifically FCM, VD²BSCAN, EVD²BSCAN, concerning Error Rate (ER) and Rand metrics. The results conclude that the proposed EVD2BSCAN produces Rand results of 86.67% which is 3.34% greater than VD²BSCAN and 6.67% greater than FCM clustering methods. Similarly the proposed EVD²BSCAN clustering algorithm produces Error Rate(ER) results of 13.33% which is 3.34% and 6.67% higher when compared to VD²BSCAN and FCM clustering methods respectively.

CONCLUSION AND FUTURE WORK

In addition this work presents anEnhanced Voronoi Diagram Density Based Clustering Algorithm (EVD²BSCAN) by quick search and discover of density peaks. In the proposed ED²BSCAN algorithm which approaches the details of density through formulating a Voronoi diagram for the source model. Point cell volumes unswervingly influence the density of point within the corresponding portions of the instance space.The proposed clustering algorithm can find clusters that indicatecomparativelyidentical regions without being separated by sparse regions. EVD²BSCAN algorithm is enhanced by calculating the growing Density Mean (DM) for some core object, in which density of its ε-neighborhood is considered with respect to DM has its own significance.From the results it concludes that the proposed EVD²BSCAN produces Rand results of 86.67% which is 3.34% greater than VD²BSCAN and 6.67% greater than FCM clustering methods. Similarly the proposed EVD²BSCAN clustering algorithm produces Error Rate(ER) results of 13.33% which is 3.34% greater than VD²BSCAN, and 6.67% greater than FCM clustering methods.The empirical findings reveal that the suggested algorithm of EVD²BSCAN possesses optimum standard and effectiveness, which enables the identification of arbitrary-shape clusters, and proper observation of the progressing activities of student database. Further research work could be concentrated on the reduction of time complication of the algorithm, in future.

REFERENCES

1. Düsing, R., 2006. Knowledge Discovery in Databases. Analytische Informations system, *Berlin, Heidelberg: Springer*, pp.241-262.
2. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* 39(11), pp.27-34.
3. Aggarwal, C.C. and Reddy, C.K. eds., 2013. Data clustering: algorithms and applications. *CRC press.*
4. Kumar, P., Bapi, R.S. and Krishna, P.R., 2007. SeqPAM: a sequence clustering algorithm for Web personalization. *International Journal of Data Warehousing and Mining,* 3(1), 29-53.
5. K. Santhisree, A. Damodaram & SV Appaji. "An Enhanced DBSCAN Algorithm to Cluster Web usage Data using Rough Sets and Upper Approximations". *International Journal of Computer Science & Communication*. Vol.1, No. 1, 2010, pp. 263-265.
6. Md. Rafiul Hassan, Baikunth Nath and Michael Kirley. "A Data Clustering Algorithm Based On Single Hidden Markov Model". *Proceedings of the International Multiconference on Computer Science and Information Technology,* pp. 57-66, 2006.
7. Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters,* 31(8), pp.651-666.
8. Bai, X., Luo, S. and Zhao, Y., 2008, "Entropy based soft K-means clustering", *IEEE International Conference on Granular Computing,* 2008(GrC 2008), pp. 107-110.
9. W. Barbakh. The family of inverse exponential k-means algorithms, Computing *and Information Systems,* 11(1):1-10, 2007.
10. W. Barbakh and C. Fyfe. Inverse weighted clustering algorithm, *Computing and Information Systems,* 11(2)10-18, 2007.
11. Havens, T.C., Bezdek, J.C., Leckie, C., Hall, L.O. and Palaniswami, M., 2012. Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems,* 20(6), pp.1130-1146.
12. Lu, Y., Ma, T., Yin, C., Xie, X., Tian, W. and Zhong, S., 2013. Implementation of the fuzzy c-means clustering algorithm in meteorological data. *International Journal of Database Theory and Application,* 6(6), pp.1-18.
13. Sayad, S., 2010. Self-Organizing Maps (SOM). University of Toronto.
14. Kohonen, T., 2013. Essentials of the self-organizing map. *Neural networks,* 37, pp.52-65.

15. Ermini, L., Catani, F. and Casagli, N., 2005. Artificial neural networks applied to landslide susceptibility assessment. *Geomorphology,* 66(1), pp.327-343.
16. Wang, L. ed., 2005. Support vector machines: theory and applications, *Springer Science & Business Media,* Vol. 177.
17. Mountrakis, G., Im, J. and Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing,* 66(3), pp.247-259.
18. Fahim, A.M., Saake, G., Salem, A.M., Torkey, F.A. and Ramadan, M.A., 2009. Improved DBSCAN for spatial databases with noise and different densities. *Georgian Electronic Scientific Journal: Computer Science and Telecommunications,* 3, pp.53-60.
19. B. Borah, and D.K. Bhattacharyya, "An Improved Sampling-Based DBSCAN for Large Spatial Databases" *proceedings of International Conference on Intelligent Sensing and Information,* 2004, pp. 92-96.
20. Cao, F., Estert, M., Qian, W. and Zhou, A., 2006, Density-based clustering over an evolving data stream with noise. In Proceedings of the 2006 SIAM international conference on data mining , *Society for Industrial and Applied Mathematics*, pp. 328-339.
21. C-F. Tsai, and C-W. Liu, "KIDBSCAN: A New Efficient Data Clustering Algorithm.", *ICAISC*, 2006, PP. 702- 711.
22. Chen, Y. and Tu, L., 2007, Density-based clustering for real-time stream data. *In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 133-142). ACM.
23. P. Liu, D. Zhou, and N. Wu, "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise", *International Conference on Service Systems and Service Management (ICSSSM),* 2007, pp.1-4
24. B. Borah, and D.K. Bhattacharyya, "DDSC: A Density Differentiated Spatial Clustering Technique", *Journal of Computers,* Vol. 3, No. 2, 2008, pp. 72-79.
25. Ram, A., Sharma, A., Jalal, A.S., Singh, R., agrawal, A. 2009. An Enhanced Density Based Spatial Clustering of Application with Noise. *In proceedings of IEEE International Advance Computing Conference.* pp.1475- 1478
26. Fahim, A., Salem, A.E., Torkey, F., Ramadan, M. and Saake, G., 2010. Scalable varied density clustering algorithm for large datasets. *Journal of Software Engineering and Applications,* 3(06), 593.
27. M. Fawzy, A. Badr, M. Reda, and I. Farag, " DBCLUM: Density-based Clustering and Merging Algorithm.", *International Journal of Computer Applications,*79, 2013, pp. 1-6.
28. Crawford, S. L. (2006). Correlation andregression. *Circulation*, *114*(19), 2083-2088.
29. Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996, "A density-based algorithm for discovering clusters in large spatial databases with noise", *In KDD* , Vol. 96, No. 34, pp. 226-231.
30. Aurenhammer, F. and Klein, R., 2000. Voronoi diagrams. *Handbook of computational geometry,* 5, pp.201-290.
31. Shirakawa, T., Adamatzky, A., Gunji, Y.P. and Miyake, Y., 2009. On simultaneous construction of Voronoi diagram and Delaunay triangulation by Physarum polycephalum. International *Journal of Bifurcation and Chaos,* 19(09), pp.3109-3117.
32. Bank, J. and Cole, B., 2008. Calculating the jaccard similarity coefficient with map reduce for entity pairs in wikipedia. Wikipedia Similarity Team, pp.1-18.
33. Hullermeier, E. and Rifqi, M., 2009, July. A fuzzy variant of the Rand index for comparing clustering structures. In *Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, IFSA-EUSFLAT 2009* (pp. 1294-1298).