IJCTA, 10(01), 2017, pp. 13-18

Performance Analysis of Dynamic Clustering Algorithm for Biological Dynamic Data

S. Angel Latha Mary* K. Anuradha and K.Uma Maheswari*****

Abstract : Most of the existing clustering algorithms give entirely different results when the data consists of diverse shapes, densities and sizes.This paper proposes a new density based dynamic data incremental clustering algorithm named as normal Dynamic DBSCAN. This normal Dynamic DBSCAN algorithm performs well and is capable of doing dynamic clustering. This algorithm is a dynamic, density based and clusters the incremental dynamic dataset and capable of adding/classifying only one data point at a time. This algorithm considers outliers, border objects and noisy objects and these objects added with new data. This normal Dynamic DBSCAN algorithm performance compared with the existing density based clustering algorithm DBSCAN. The performance of the algorithm evaluated using Generalized Dunn Index(GDI) as the cluster validation metric as well as time taken for clustering.

Keywords : Incremental Clustering, Dynamic Clustering Dynamic Dataset, DBSCAN Algorithm, Dynamic DBSCAN.

1. INTRODUCTION

Clustering in data mining is a process of grouping a set of data. Existing clustering algorithms are designed to find clusters using some static models. CURE and ROCK algorithms uses a static model to determine the most similar cluster [1]. CHAMELEON measures the similarity of two clusters based on a dynamic model. But this algorithm very much affected by the size of the dataset. All the existing clustering algorithms use static dataset. K-Means, CHAMELEON, CURE ROCK and BIRCH [9] [10] algorithms can give incorrect results if the choice of parameter is incorrect with respect to the data set being clustered, or the data consists of clusters are of diverse shapes, densities, and sizes [8]. Dynamic clustering [5][7] [13] is a mechanism to adopt and discover clusters in real time environments [6][12][14]. Most of the supervised classification algorithms performs well and will give ideal results with good accuracy metrics [2]in change over time[3].

In this work a density based dynamic data incremental clustering algorithm for clustering dynamic dataset is presented and compared to its performance with existing clustering algorithm DBSCAN[15] the real data sets ZOO Dataset and Wisconsin Breast Cancer Dataset which are downloaded from UCI Machine Learning Repository[11].

2. IMPLEMENTATION

The existing DBSCAN algorithm defines a cluster to be a maximum set of density-connected points. The two parameters Eps and MinPts used by existing DBSCAN algorithm. Every core point (points inside the cluster) in a cluster must have at least a minimum number of points (MinPts) within a given radius (Eps). DBSCAN can find arbitrary shape of clusters. It could retrieve all points that are density-reachable

* Professor / Department of CSE Karpagam College of Engineering Coimbatore, Tamilnadu, India *xavierangellatha@gmail.com*

** Associate Professor/Department of MCA Karpagam College of Engineering Coimbatore, Tamilnadu, India *K_anur@yahoo.com*

*** Assistant Professor/ Department of MCA Karpagam College of Engineering Coimbatore, Tamilnadu, India *umamaheswarikit@gmail.com*

from the given point using the correct parameters. DBSCAN uses global values for Eps and MinPts, tt. the same values for all clusters. This algorithm handles outliers well and not includes them in any cluster. This paper also addresses the problems of clustering an incremental dynamic dataset [4] in which the data set is increasing in size over time by adding more and more data.

This new algorithm uses DBSCAN algorithm for clustering existing data. This proposed Density based normal Dynamic DBSCAN algorithm dynamically changes the epsilon (Eps) value during each batch of insertion. Another most important variation is, during each step of batch insertion it considers the data points which are classified as noise(outliers) or border objects (border of the cluster) are removed and once again they marked as unclassified points or outliers and combined with the new data which is to be inserted. During insertion, it inserts only one data point at a time and then re-estimates the cluster IDs. This normal Dynamic DBSCAN algorithm performs well and is capable of doing dynamic clustering. To find a cluster DBSCAN starts with an arbitrary object p in D and retrieves all objects of D density-reachable from p with respect to Eps and MinPts. If p is a core object, this procedure yields a cluster with respect to Eps and MinPts. If p is a border object, no objects are density-reachable from p and p is assigned to the noise. Then DBSCAN visits the next object of the database D.

The following pseudo code explains the algorithm for Clustering Evolving Data over Time.

1. The existing data is assumed as already clustered.
2. So the existing data are clustered using normal standard DBSCAN algorithm and it uses the global values for Eps and MinPts, *i.e*. the same values for all clusters.
3. The outliers and border objects are removed for validation and the clusters are validated using GDI.
4. Add the incremental data and increases the Eps and MinPts based on existing and incremental data size.
5. Separate, the noise object (outliers and border objects) as unclassified data points.
6. If the object is unclassified, create a cluster for border and noise objects and Merge them with nearby clusters, if possible.
7. For incremental data find the neighborhood of each data based on new Eps value
8. If no nearby points, then this data is a noise object.
9. If some of the object are unclassified it merges all points and assign a new class ID
10. In case of absorption in existing cluster assign cluster-id based on its neighborhood
11. If Some of the object are unclassified merge all clusters and assign a common class ID
12. Remove outliers for validation
13. Now validate the cluster accuracy using GDI.

This algorithm is a dynamic, density based and clusters the incremental dynamic dataset. But this algorithm is capable of adding/classifying only one data point at a time. The results of the proposed Dynamic DBSCAN algorithm performed significantly well in terms of speed. The algorithm is capable of creating, modifying and inserting clusters over time. The performance of clustering is evaluated using some of the real data sets Zoo Dataset and Wisconsin Breast Cancer Dataset.

3. RESULTS AND DISCUSSION

The static data set is simulated as incremental dynamic dataset and initially 50% of the data is randomly selected from the given dataset and it clusters the data using DBSCAN algorithm. This data is considered as existing data. Then every time 10% of the incremental data added and this new data is updated in the existing cluster or it creates new cluster based on the Eps and Minpts values. The outlier data separated and marked as unclassified objects. These unclassified objects once again added with new incremental data. This algorithm adds one object at a time. The performance in terms of time for the cluster using the real datasets (ZOO Dataset and Wisconsin Breast Cancer Dataset) are given in table 1.

Table 1

Performance of normal Dynamic DBSCAN algorithm in terms of GDI

Dataset	*% of data used*						*AVG*
	50	*60*	*70*	*80*	*90*	*100*	
ZOO dataset	0.71	0.69	0.88	0.8	0.8	0.88	0.80
Wisconsin Breast Cancer	0.99	0.98	0.75	0.78	0.82	0.84	0.85

Larger values of GDI correspond to good clusters and the number of clusters that maximizes GDI is taken as the optimal number of clusters.

Table 2

Performance in terms of Time using normal Dynamic DBSCAN algorithm

Dataset	*% of Data used*					
	50	*60*	*70*	*80*	*90*	*100*
Zoo	0.03	0.04	0.03	0.02	0.03	0.03
Wisconsin Breast Cancer	0.03	0.06	0.08	0.08	0.09	0.09

The following section compares the performance of normal Dynamic DBSCAN algorithm and the existing DBSCAN algorithm with the real datasets in terms cluster accuracy using GDI(figure 1 and figure 2) and speed (figure 3,4 and 5). The performance of the normal Dynamic DBSCAN algorithm is best in terms of accuracy and cluster speed.

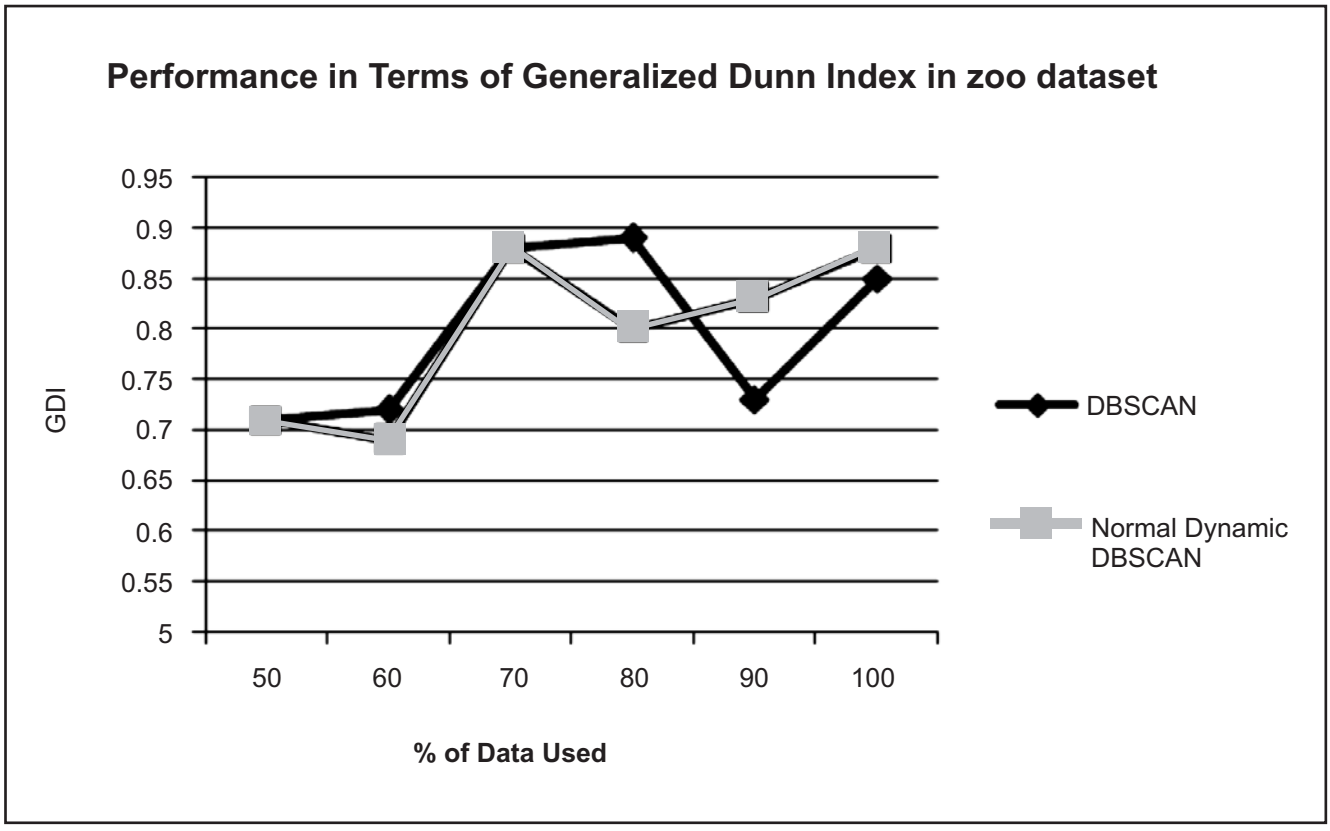

Figure 1: Performance in Terms of Generalized Dunn Index in zoo dataset

4. CONCLUSION

The proposed density based Dynamic DBSCAN clustering algorithm is successfully implemented using MATLAB platform and evaluated. This algorithm is able to insert data objects one by one and then re-estimate the cluster IDs during every point**,** which is inserted. The performance of the algorithm is evaluated using DBSCAN algorithm with cluster accuracy and cluster speed as two parameters.

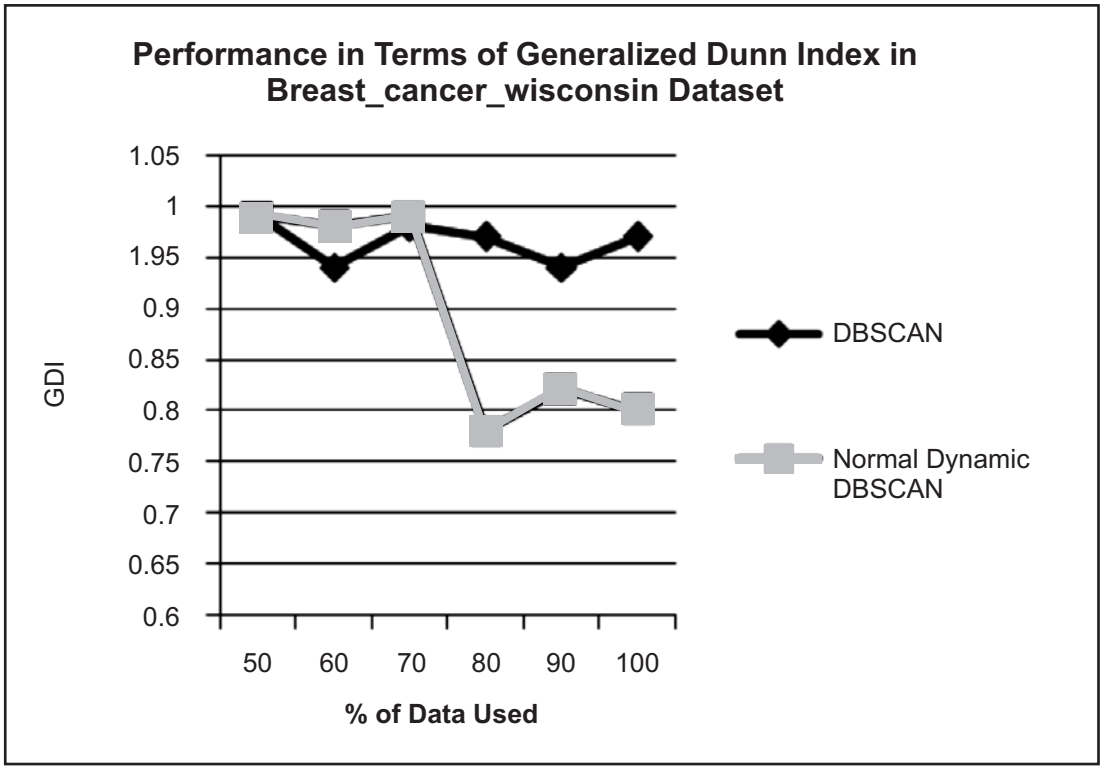

Figure 2: Performance in Terms of Generalized Dunn Index in Breast_cancer_wisconsin Dataset

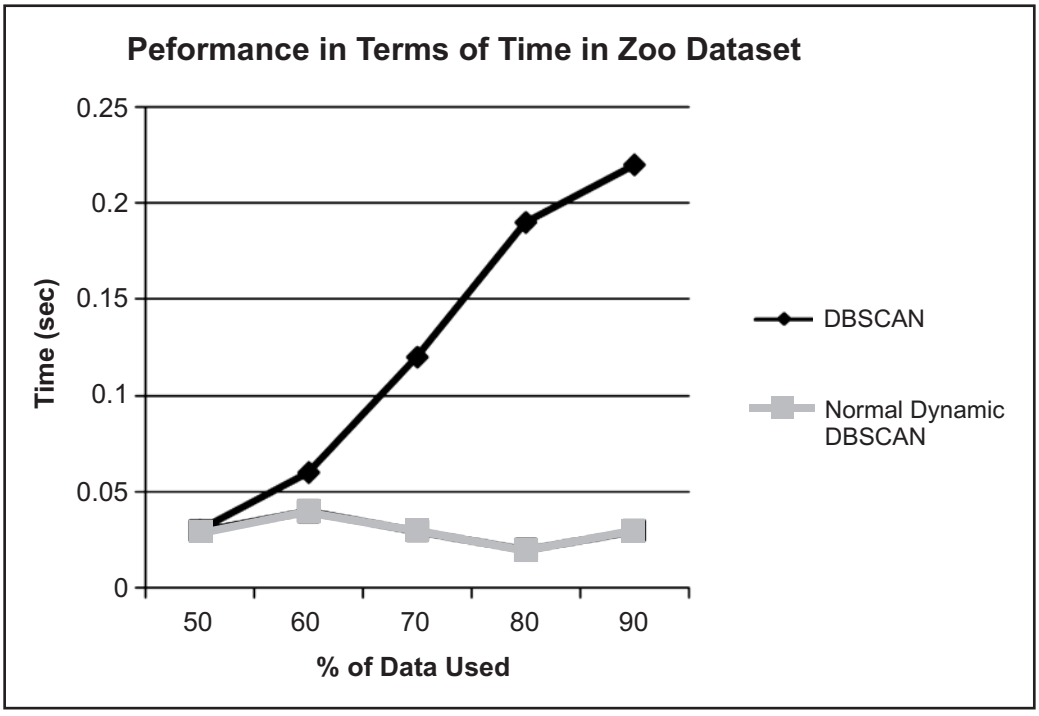

Figure 3: Performance in Terms of Time in Zoo Dataset

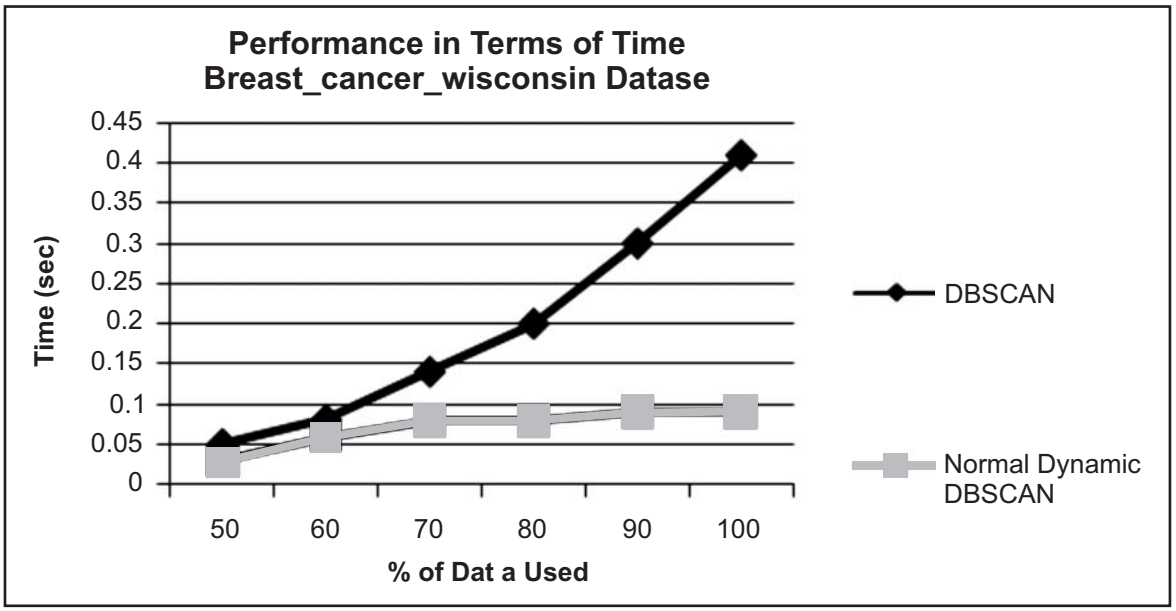

Figure 4: Performance in Terms of Time in Breast_cancer_wisconsin Dataset

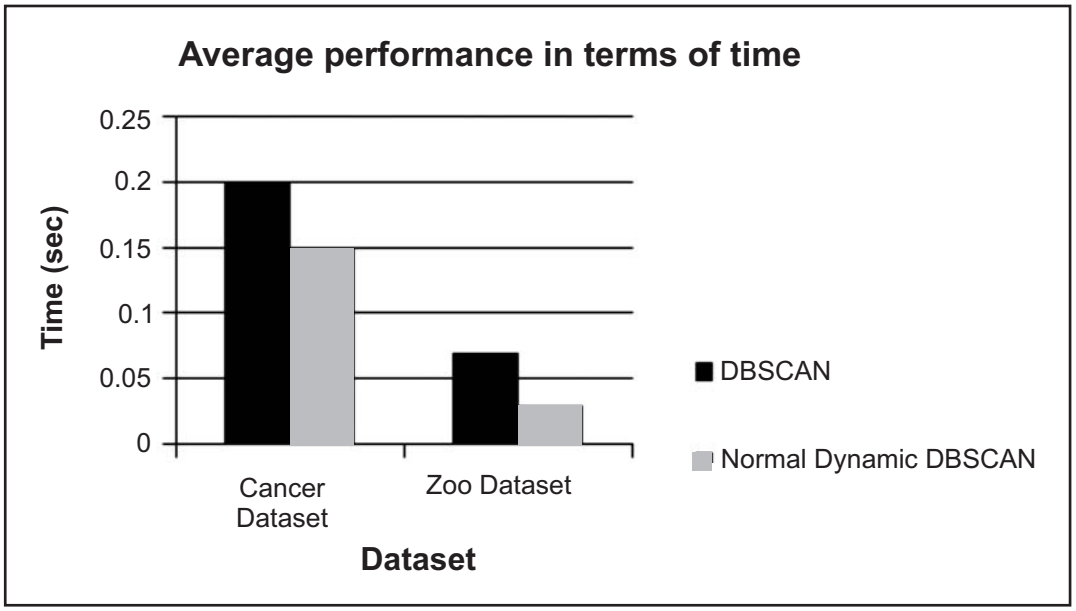

Figure 5: Average performance in Terms of Time

5. REFRENCES

1. Karypis, G., Eui-Hong (Sam) Han and VipinKumar, "Chameleon: Hierarchical Clustering Using Dynamic Modeling", IEEE Computers, (1999): 68-75.
2. Seret, Alex, Bart Baesens, and Jan Vanthienen. "A dynamic understanding of customer behavior processes based on clustering and sequence mining." Expert Systems with Applications, (2014): 4648-4657.
3. Bai, Liang, Jiye Liang, Chuangyin Dang, and Fuyuan Cao. "A novel fuzzy clustering algorithm with between-cluster information for categorical data." Fuzzy Sets and Systems, (2013): 55-73.
4. Weber, Richard. "From Operations Research to Dynamic Data Mining and Beyond." In Zukunftsperspektiven des Operations Research. Springer Fachmedien Wiesbaden, (2014) :343-356
5. Campbell, Trevor, Miao Liu, Brian Kulis, Jonathan P. How, and Lawrence Carin. "Dynamic clustering via asymptotics of the Dependent dirichlet process mixture." In Advances in Neural Information Processing Systems, (2013): 449-457.

6. Rakthanmanon, Thanawin, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. “Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping.” ACM Transactions on Knowledge Discovery from Data (TKDD) 7, no. 3 ,(2013): 3047-3051.

7. Angel Latha Mary .S, Shankar Kumar.K.R “Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset” Journal of Computer Science 8 (5): 656-664, 2012 ISSN 1549-3636.

8. Sim, Kelvin, Vivekanand Gopalkrishnan, Arthur Zimek, and Gao Cong. “A survey on enhanced subspace clustering.” Data mining and knowledge discovery 26, no. 2, (2013): 332-397.

9. Dong, Jianqiang, Fei Wang, and Bo Yuan. “Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism.” In Intelligent Data Engineering and Automated Learning–IDEAL 2013, Springer Berlin Heidelberg, (2013):409-416.

10. Angel Latha Mary. S, Shankar Kumar.K.R “Evaluation of Clustering Algorithm with Cluster Validation Metrics” in European Journal of Scientific Research Vol.69 No.1 (2012), pp.61-72 ISSN 1450-216X.

11. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.

12. S. Angel Latha Mary, A. N. Sivagami and M. Usha Rani “Cluster Validity Measures Dynamic Clustering Algorithms“, ARPN Journal of Engineering and Applied Sciences Vol:10, Issue:9, Pages 4009-4012,2015

13. S. Angel Latha Mary, D. Sivaganesan and R. Vinothkumar; “An Empirical Research of Dynamic Clustering Algorithms”ARPN Journal of Engineering and Applied Sciences Vol:10, Issue:9,Pages 4002-4004. (Annexure-II) 2015

14. Srinivas, Chintakindi, and C. V. GuruRao. “Clustering Text Data Streams–A Tree based Approach with Ternary Function and Ternary Feature Vector.” Procedia Computer Science 31 ,(2014): 976-984.

15. Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland , (1996) :226-231