Journal of Theoretical and Applied Information Technology

30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SCALABLE AND ROBUST CLUSTERING FRAMEWORK FOR HANDLING LARGE DATASETS

[1]Mrs. RAMANA LAKSHMI ADUSUMILLI, Prof. M. SHASHI[2]

[1]Research Scholar, Dept of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India,

[2]Professor, Dept of Computer Science and Systems Engineering, AU College of Engineering. Andhra University, Visakhapatnam, India,

E-Mail: aramanalakshmi@gmail.com, smogalla2000@yahoo.com

ABSTRACT

, Clustering large datasets often suffer from scalability issues as they involve pair-wise distance computations among all the instances of the dataset. Grid based algorithms achieve scalability as they circumvent the distance estimation step but are vulnerable to quality and coverage issues. In this paper, the authors proposed a new framework for cluster formation with high scalability while maintaining the coverage and quality. The Scalable and Robust Clustering (SRC) framework has three modules. The first module involves PCA to convert the multi-dimensional dataset into a low dimensional grid space and based on the joinable boundaries, the dense grid cells are merged to form macro-clusters representing dense regions. The second module involves density based clustering applied separately within the dense regions to form micro-clusters. Finally the third module involves statistical technique to find appropriate clusters for data objects located in non-dense grid cells. The first two modules of the framework handles scalability issues while the third module focuses on improving coverage without affecting the quality of clusters. The experimental results obtained on bench mark datasets show that the framework could achieve scalability, coverage and quality while handling large multi-dimensional datasets.

Keywords: *Hybrid Clustering, Grid Based Approach, Density Based Clustering, Post Processing*.

1. INTRODUCTION

One of the most dominating sub-domains in data mining is clustering. Clustering aims to discover patterns of similar objects from large datasets, wherein each object is represented as a data point in a multi-dimensional space. In clustering, data points are divided into groups or clusters such that the data points that are present in the same cluster are similar to each other and are different from the data points of the other clusters. Hence, each of these clusters represents a pattern obeyed by a significant group of objects and the patterns extracted from larger datasets are of higher quality with better generalization and robustness. At present many data collection tools are available to generate huge amounts of data with large number of attributes. Clustering being an unsupervised learning doesn't need the data points to be annotated or labelled. Hence, there is data available in abundance, but the existing clustering algorithms are not capable of processing large datasets due to their limited scalability. Scalability issues arise in clustering, when there is an increase in dataset cardinality or dataset dimensionality. Clustering is a time consuming process due to high time complexity of most of the clustering algorithms except for Grid based clustering approaches. Grid based clustering algorithms are found to be scalable when dealing with datasets of high cardinality. The time complexity of grid based clustering algorithms relies on the number of grid cells rather than the cardinality of the data set. However, the number of grid cells increases exponentially with the dimensionality of a dataset. Hence, the scalability of grid based clustering is curtailed, if the dimensionality of the dataset is also high unless the grid resolution is lowered at the cost of cluster quality.

The authors proposed a Hybridization of Grid and Density based Clustering (HGDBC) methodology[2] for scalable clustering of large datasets. The methodology starts by applying Principle Component Analysis (PCA) on a large multi-dimensional dataset to reduce the dimensionality of the dataset without losing the essential information related to variance of the data points. In the low dimensional data space, a fixed number of grid cells are created and among them the dense grid cells are identified based on a threshold on the number of data points in a grid

cell. Dense regions or dense grid neighborhood's are formed by merging the dense cell boundaries that are adjacent either horizontally, vertically or diagonally indicated by a difference in their corresponding co-ordinates limited to **1.** Finally HGDBC algorithm goes by divide and conquer strategy to achieve scalability as it applies density based clustering within each of the dense regions to discover density based clusters of high purity. The details of HGDBC methodology for scalable clustering are available in [2].

However, the limitation of the original HGDBC algorithm is low coverage; data points contained in the grid cells of the non-dense regions are ignored and left beyond the scope of the clustering process. This low coverage implies loss of useful information and hence the authors addressed the problem of improving the coverage of the proposed clustering methodology by a post processing step. In this paper the authors propose a post processing technique to handle the data points belonging to the grid cells of non-dense regions. In this paper, A Scalable and Robust Clustering Framework for handling large datasets is proposed to improve the coverage of data points in clusters and solve the scalability issue without effecting the accuracy and quality of the clusters.

The remaining sections of the paper are as follows. Section 2 presents recent developments towards clustering large datasets. Section 3 provides an overview of the existing scalable clustering algorithm, HGDBC[2] developed by the authors through hybridization of grid and density based clustering algorithms in order to gain scalability. Section 4 discusses the limitations of HGDBC algorithm and provides a solution to overcome the limitations by developing a Scalable and Robust Clustering(SRC) Framework for handling large data sets to discover clusters with high scalability, quality and coverage. The experimental results are presented in the Section 5 followed by the Section 6 with conclusions.

2. LITERTURE SURVEY:

Dorit S. Hochbaum et al.,[1] proposed a method of sparse computation that generates a sparse similarity matrix which contains only relevant similarities without computing all pair wise similarities. This method adopts an efficient algorithm that provides an "approximate Principal Component Analysis" and grid based structure is used to identify groups of identical objects. This approach is effective for large data sets.Ramana Lakshmi Adusumilli et al.,[2] proposed a highly scalable methodology by Hybridization of Grid and Density Based Clustering (HGDBC) techniques for clustering large datasets. They developed a two step method to reduce the number of pair wise distance computations instead of the normal step involving the time consuming distance estimation among all possible pairs of instances of the data set. In the first step, the large dataset is divided into partitions by applying grid based approach into multiple smaller dense regions called macro level clusters each containing the set of candidate co-members of clusters at micro level. In the second step, it employs density based clustering algorithm within the individual dense regions for micro level clustering, leading to the elimination of pair-wise distance computations between instances belonging to different dense regions.D. Cheng et al., [3] proposed a novel MST-based clustering algorithm called LDP-MST. It first uses local density peaks to construct MST and then repeatedly cuts the longest edge until a given number of clusters are found..Chunhu Ren et al., [4] proposed density peaks clustering (DPC) algorithm which is used to analyze the complex data such as the datasets that shows uneven density distribution and number of peaks in same cluster. To solve these issues the author developed an improved density peaks clustering algorithm based on the layered k-nearest neighbours and sub-cluster merging (LKSM_DPC).Zheng Qin et al.,[5] introduced an efficient graph-based partitioning algorithm for extended target tracking. To reduce the computational load and the interference of clutter on the measurement set partition, a measurement set pre-processing method based on density-based clustering algorithm is presented. Zhi liu et al., [6] proposed a novel method for clustering with local peaks in the symmetric neighbourhood.A graph-based scheme is adopted here to merge similar clusters based on their similarity in the symmetric neighbourhood graph, followed by assigning each outlier to the closest cluster. Chunrong WU et al., [7] proposed a robust clustering method which establishes a symmetric neighbourhood graph over all data points, based on the k-nearest neighbours and reverse k-nearest neighbours of each point.Initial centres for clusters are estimated over the peaks and similar clusters are aggregated on the symmetric neighbourhood graph.Shen Xinglin et al.,[8] proposed a fast density peak-based clustering partitioning algorithm is applied to the measurement set partitioning. The simulation results show that the proposed algorithm can get the most informative partition and obviously reduce

computational burden without losing track of performance

3. RELATED CONCEPTS

Clustering is a process of finding similar objects from a dataset and grouping them into valid Clusters. The aim of the clustering algorithms is to maximize the inter cluster distance and minimize the intra cluster distance. Distance calculation between pairs of data points is the basic operation which determines the efficiency of clustering algorithms. Specifically the number of distance calculations required in Density based and Hierarchical Clustering algorithms increases quadratic with the number of data objects and makes the process of distance estimation prohibitively expensive while clustering very large datasets. An algorithm for Hybridization of Grid and Density based Clustering (HGDBC)[2] is developed for clustering large datasets. When compared to DBSCAN, the HGDBC methodology proposed by the authors is highly scalable as it reduces the number of distance estimations required for cluster formation. The methodology employs a Two Phase of clustering approach, the first phase named Macro-Clustering phase identifies dense regions of data points using Grid based approach and the Second Phase named Micro-Clustering groups the objects within dense regions into Clusters of High Quality. PCA is applied to reduce the dimensionality of the dataset and then Grid based Clustering is used in low Dimensional space to partition the large dataset into smaller dense regions. In the second phase, the instances of different dense regions need not be considered as co-members of a micro-cluster and hence the inter-region pair-wise distances need not be estimated. Thus the number of distance computations in full feature space will be reduced to a great extent in the second phase cluster formation leading to sparse computation of distance matrix that doesn't compromise on the essential information required for density based clustering within the individual dense regions to generate high quality micro-clusters. Thus the HGDBC methodology achieves two fold advantage to overcome the scalability limitation; grid based clustering is applied on dataset represented in low dimensional space that limits the number of grid cells irrespective of the original dimensionality of the dataset to achieve scalability while identifying dense regions in the first fold and then distances are estimated only among the co-members of the dense regions to form accurate micro-clusters. Thus HGDBC methodology effectively overcomes the scalability issue generally suffered by the density based clustering algorithms. However, it was observed experimentally [2] that almost 30% of the data objects are not covered by the dense clusters discovered using HGDBC algorithm.

4. METHODOLOGY

Most of the existing clustering algorithms that aim to form high quality clusters like DBSCAN are not suitable for handling large datasets due to scalability issues. The HGDBC methodology proposed by the authors solves the scalability problem as discussed briefly in the previous section. Using HGDBC Scalability issues are handled effectively and accuracy and quality of the clusters are maintained on par with the other non-scalable micro-clustering algorithms. However, the limitation of the HGDBC methodology is that it can cover only the data points that fall into the dense regions of the low dimensional feature space, thereby leading to ignoring almost 30% of the data points during the clustering process. A framework is developed to extend the utility and effectiveness of the proposed HGDBC methodology to handle large datasets without any data loss.

Figure 1 shows the proposed Scalable and Robust Clustering(SRC) Framework for handling of large datasets. The first few modules of the framework implements the original HGDBC methodology to create micro-clusters within each of the dense regions formed in the low dimensional space. It may be observed that locating data points on a low dimensional grid and formation of macro

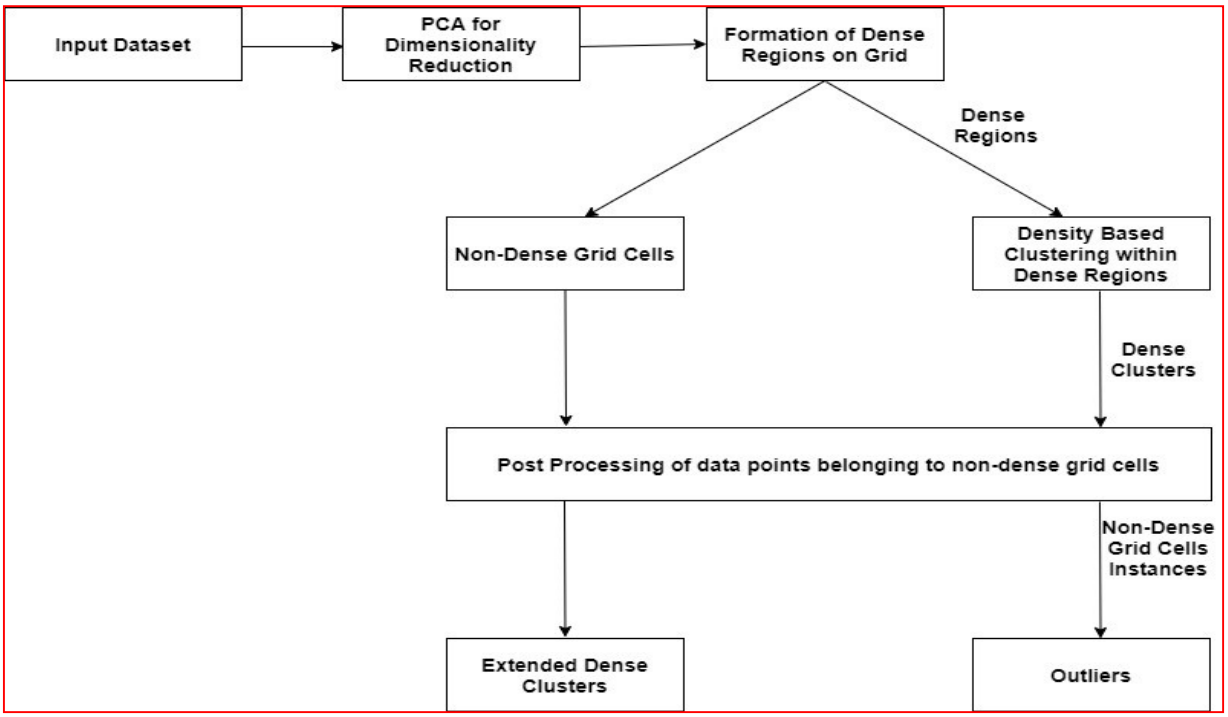

Figure 1: Architectural diagram of the proposed Scalable and Robust Clustering Framework

clusters representing dense regions [2] doesn't require distance estimation. Thus HGDBC methodology gains time as it reduces the effort required for distance estimation between all pairs of data points. The dense clusters discovered separately from every dense region involve data points located in dense-grid cells only. This SRC framework extends the functionality by adding a post-processing module to handle the data points located in non-dense grid cells so that they would be inserted into nearby dense clusters with minimal distance estimates. The authors devised a statistical technique for finding appropriate clusters for the data objects located in non-dense grid cells.

This statistical technique calculates the clusters centroids to represent the micro-clusters already discovered by HGDBC and finds the distances from every non-dense cell data object to the cluster centroids. The closest cluster is found for every non-dense cell data point. Assuming that the distance from a data object to it closest cluster are Gaussian distributed, the mean, and standard deviation which are the parameters that define the distribution are estimated. Accordingly the upper threshold on the distance of a data object from its closest cluster is fixed. Based on the threshold, the data points of non dense grid cells are added to their nearest clusters. The details are further elaborated in the following sub-section.

4.1 Scalable and Robust Clustering (SRC) Framework:

Step 1: PCA is applied on the large datasets for dimensionality reduction so that data points are located on a 2 or 3 dimensional grid irrespective of the original dimensionality of the dataset.

Step 2: Simplified grid based clustering methodology devised by the authors [2] is used to partition the large datasets into smaller dense regions so that the objects within a dense region are similar and objects located in different dense regions are dissimilar. Hence, the distance computations in full feature space will be reduced to a great extent. Density based clustering is called on dense regions.

Step 3: Some of the instances do not belong to any dense clusters. It was observed that upto 30% of the data points were not assigned to any of the clusters identified by HGDBC methodology. At this stage, the following post processing technique is proposed to assign appropriate clusters to data objects that belong to non-dense grid cells.

4.2. Post processing Technique:

Post processing technique aims to assign cluster membership to the non-dense grid cell data points. Let the number of data points to be considered for cluster membership are N and the total number of micro-clusters discovered are M.

- Estimate the cluster centroids for each of the micro-clusters identified in all dense regions. These M cluster centroids represent the M micro-clusters identified by HGDBC algorithm.

- Calculate the distances from each non-dense grid cell data object to the cluster centroids and a Distance matrix(D) is formed to hold the distances between every non-dense cell data object and various cluster centroids at appropriate rows and columns respectively. The order of the matrix D is N x M

- Calculate the minimum distance in each row and assuming that these N row_mins are Gaussian distributed, estimate the mean (μ) and standard deviation (σ) as the parameters of the distribution.

- Upper bound on distance is taken as the threshold value (T) for cluster membership. Threshold is defined as given below for different settings of C varying in the range of 0.5 to 3 for experimentation.

$T = \mu + C \times \sigma$

- Closest cluster is identified for each non-grid cell data point indicated by the column number corresponding to the row_min in the distance matrix, D.

- Compare the row_min distance in each row with threshold, T, and if the distance is less than the threshold,T, the corresponding data point is added to the closest cluster otherwise the data point is declared as outlier. In other words, based on the indices of the row_min value, the row number identifies the data point and column number identifies the cluster, provided that the row_min is less than the threshold.

Using the proposed Scalable and Robust Clustering Framework most of the data objects belonged to non-dense grid cells are inserted into their nearest clusters. Thus the coverage of the data points is significantly extended beyond the dense regions compared to the originally proposed HGDBC. Experimentation is conducted to investigate the effect of the post processing technique on the Accuracy and Quality of clusters at different threshold settings.

5. EXPERIMENTAL RESULTS

The SRC framework is implemented using Python Programming language with I7 processor, Ram with 32 GB and 2TB hard disk. Bench mark datasets such as forest cov type are used for investigating the performance of the proposed scalable and robust clustering framework.

Dataset Description Forest cover dataset is a huge dataset which consists of 5,81,012 instances and 54 attributes. This is used for research work. With this huge dataset various issues such as scalability issues occur. The experiments are conducted by using different sizes of dataset.

The performance of the proposed Scalable and Robust Clustering (SRC) Framework for handling large datasets is measured in terms of coverage of data points and purity value at various threshold values where Threshold=Mean+C*Std, where C varies in the range of 0.5,1,1.5,…3.0. It was observed from the experimental results that as the threshold value is increased, the coverage of data points in clusters is also increased. Thus the data lost in HGDBC algorithm is avoided using the proposed SRC framework.

The results are shown in the cross table 1. Each row in the cross table presents the performance metrics obtained for different size samples of forest cover dataset. Specifically % coverage and purity at different threshold values (C varied from 0.5 to 3.0) for each sample dataset are presented in the table.

The Graph shown in Figure-2 is drawn between increasing values of C on X axis (varies in steps of 0.5) and percentage of coverage on Y axis in red coloured line. The graph also shows the variation of purity values with the increasing values of C in green line. Thus the graph simultaneously shows the % coverage of data points and purity values at various threshold settings for upper bound on the distance to the closest cluster, where, Threshold = Mean + C * Standard Deviation (C varies in steps of 0.5). It may be observed from the experimental results that the coverage of data points is increased from 79% to 99% with the value of C increased from 0.5 to 3. There is no significant decrease in the purity of clusters at higher threshold setting upto C=3.0. Hence, the proposed technique is applied with C=3.0 setting for best results. Compared to the original HGDBC methodology as standalone, the holistic SRC framework is performing better as the data loss is reduced and the quality of the cluster measured in terms of purity is almost maintained. The SRC framework is scalable as it involves the highly scalable HGDBC methodology for formation of dense clusters which are improvised for the purpose of increasing its coverage in the final module. Thus the authors advocate that the proposed SRC framework is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

scalable and robust as well based on the experimental results.

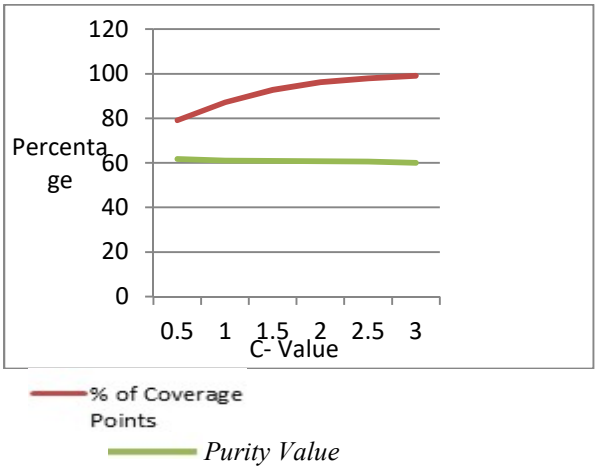

Figure-2: Coverage of Data Points and purity of a Scalable Clustering Framework

6. CONCLUSION

Existing density based clustering algorithms are not suitable for handling large data sets .The HGDBC algorithm developed by the authors is scalable and hence, handles large data. But at the same time HGDBC algorithm doesn't consider almost 30% of the data points for clustering and those data points may contain useful information. This limitation is rectified in our proposed Scalable and Robust Clustering(SRC) framework. The proposed SRC framework synergies the strengths of PCA at dimensionality reduction, grid based clustering for scalability to handle large data sets, density based clustering towards formation of quality clusters and finally statistical approach for post processing of data objects located in non-dense grid cells for improving the coverage. The framework is developed by extending the HGDBC methodology proposed by the authors using the newly proposed statistical technique for post processing the results to achieve coverage while maintaining quality of clusters. Thus the data points which are hitherto outside the dense regions are now included in appropriate clusters. The coverage of the data points is significantly improved from 70% to 99% in SRC framework while maintaining the quality of the cluster.

REFERENCES

[1]. Hochbaum, Dorit S., and Philipp Baumann. (2016) Sparse computation for large-scale data mining. IEEE Transactions on Big Data 2.2: 151-174.

[2]. Ramana Lakshmi Adusumilli, M. Shashi. (2020). Hybridization of Grid and Density

[3]. Based Clustering for Formation of Macro and Micro Level Clusters. International Journal of

[4]. Advanced Science and Technology, 29(06), 3050 - 3060.

[5]. Cheng, D., Zhu, Q., Huang, J., Wu, Q., & Yang, L. (2019). Clustering with local density peaks-based minimum spanning tree. IEEE Transactions on Knowledge and Data Engineering, 33(2), 374-387.

[6]. Ren, C., Sun, L., Yu, Y., & Wu, Q. (2020). Effective Density Peaks Clustering Algorithm Based on the Layered K-Nearest Neighbors and Subcluster Merging. IEEE Access, 8, 123449-123468.

[7]. Qin, Zheng, Thia Kirubarajan, and Yangang Liang. (2020) Application of an Efficient Graph-Based Partitioning Algorithm for Extended Target Tracking Using GM-PHD Filter." IEEE Transactions on Aerospace and Electronic Systems 56.6: 4451-4466.

[8]. Liu, Zhi, et al. (2019) Local Peaks-Based Clustering Algorithm in Symmetric Neighbourhood Graph." IEEE Access 8: 1600-1612.

[9]. Wu, C., Lee, J., Isokawa, T., Yao, J., & Xia, Y. (2019). Efficient clustering method based on density peaks with symmetric neighborhood relationship. IEEE Access, 7, 60684-60696.

[10]. SXinglin, S. H. E. N., Zhiyong, S. O. N. G., Hongqi, F. A. N., & Qiang, F. U. (2019). Fast density peak-based clustering algorithm for multiple extended target tracking. Journal of Systems Engineering and Electronics, 30(3), 435-447.

[11]. Datasets available at UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php

Table-1: Performance Analysis Of The Proposed Scalable Clustering Framework

Forest Covtype Samples	Threshold Mean + C*Std C varied from 0.5 to 3.0 in steps of 0.5	Number of clustered points				Purity of Proposed SCA Algorithm	Purity of HGDBC Algorithm
		HGDBC Algorithm		Proposed SCA Algorithm			
		Number of covered points	% of covered points	Number of covered points	% of covered points		
50K	Mean + 0.5 Std	27007	54	37175	74.35	0.666	0.754
	Mean + Std			42551	85.10	0.636	
	Mean + 1.5 Std			45630	91.26	0.626	
	Mean + 2Std			47706	95.41	0.620	
	Mean + 2.5 Std			48888	97.7	0.619	
	Mean + 3 Std			49417	98.8	0.619	
1Lakh	Mean + 0.5 Std	64659	64	79794	79.79	0.691	0.696
	Mean + Std			88435	88.43	0.688	
	Mean + 1.5 Std			93703	93.70	0.687	
	Mean + 2Std			96560	96.56	0.686	
	Mean + 2.5 Std			98167	98	0.684	
	Mean + 3 Std			99943	99.9	0.683	
2Lakh	Mean + 0.5 Std	130756	65.37	156130	78.06	0.664	0.673
	Mean + Std			171143	85.57	0.665	
	Mean + 1.5 Std			182736	91.36	0.669	
	Mean + 2Std			190718	95.35	0.672	
	Mean + 2.5 Std			195359	97.6	0.670	
	Mean + 3 Std			197680	98.8	0.669	
3Lakh	Mean + 0.5 Std	205018	68.33	240939	80.31	0.599	0.617
	Mean + Std			263451	87.81	0.595	
	Mean + 1.5Std			280769	93.59	0.594	
	Mean + 2Std			289918	96.64	0.593	
	Mean + 2.5 Std			295076	98.3	0.592	
	Mean + 3 Std			297535	99.1	0.592	
4Lakh	Mean + 0.5 Std	268085	67.02	316831	79.20	0.557	0.573
	Mean + Std			345367	86.34	0.554	
	Mean + 1.5Std			368537	92.13	0.552	
	Mean + 2Std			384031	96	0.551	
	Mean + 2.5 Std			392016	98	0.550	
	Mean + 3 Std			396008	99	0.550	
5Lakh	Mean + 0.5 Std	351688	70.33	413129	82.62	0.522	0.539
	Mean + Std			450722	90.14	0.522	
	Mean + 1.5Std			471756	94.35	0.523	
	Mean + 2Std			484682	96.93	0.523	
	Mean + 2.5 Std			492013	98.4	0.521	
	Mean + 3 Std			495984	99.1	0.520	