How to Cite:
Rashmi, M., & Varshney, M. (2022). Outlier analysis for microarray gene. *International Journal of Health Sciences*, *6*(S1), 4839–4849. https://doi.org/10.53730/ijhs.v6nS1.5925

Outlier analysis for microarray gene

Rashmi M
Research Scholar, MUIT, Lucknow.
Email: rashmimadan.11@gmail.com

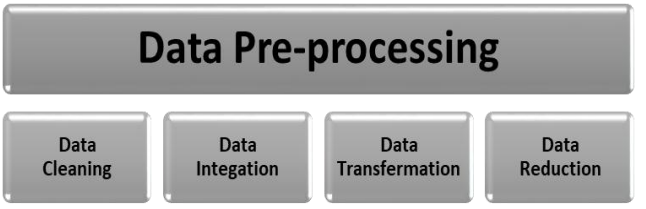
Dr Manish Varshney
Professor, Maharishi School of Engineering & Technology, MUIT, Lucknow.
Email: itsmanishvarshney@gmail.com

***Abstract*---**Pre-processing data is a critical component of data mining, as it comprises anomaly identification, outlier analysis, and dimensionality reduction utilising a distance-based technique. This research study demonstrates that in order to cope with scarcity difficulties in high dimensional spaces, computations should be limited to such data. A distance-based technique is seen more appropriate for the microarray-quality articulation of information delivered across many time zones.

***Keywords*---**Data Reduction, Outlier Analysis, Distance-based detection, Principle Component Analysis.

Introduction

Data pre-processing is a Data Mining technique that comprises transforming raw data into an understandable format. Real-world data is usually insufficient, inconsistent, and/or absent of certain behaviours or patterns, in addition to including other mistakes. This may result in poor data gathering and, as a result, poor models based on the data. Pre-processing data is one way to address these issues.



data
Figure 1.0 Data Pre-processing Groups

It is possible to transform the infeasible into the feasible through data pre-processing. Data pre-processing entails identifying, data reduction strategies, and reducing the complexity of the information, as well as removing noisy aspects. Pre-processing data may be accomplished in four distinct ways. Figure 1.0 shows the four categories are data cleansing/cleansing, data integration, data transformation, and data minimization. Data Cleaning: Data cleansing is the initial step in pre-processing data. The primary focus of this method is on removing outliers, eliminating duplication, dealing with missing data, noisy data, estimated biases, and detection within the data.

Data Integration: Data integration is a technique that aides in the collection of data from disparate data sources and the combining of that data to generate continuous data. After completing data cleansing, this ongoing information aids in analysis and data preparation [4].

Data transformation: We've already began modifying our data through data cleaning, but data transformation will begin the process of transforming the data into the correct format(s) for analysis and other downstream operations. This often occurs during one of the following processes: Aggregation, Normalization, Feature selection, Discretization, or the development of Concept hierarchies [5].
Data reduction: The more data you have, even after cleaning and changing it, the more difficult it will be to analyse. Depending on the nature of the activity at hand, you may find that you have more data than you require. Much of normal human speech, particularly when dealing with text analysis, is redundant or unrelated to the researcher's goals. Data reduction not only simplifies and improves the accuracy of analysis, but also reduces the amount of data stored. Techniques used in data reduction are:

- *Numerosity Reduction:* Alternative and more compact data representations such as parametric models (which contain only the model parameters rather than the actual data, such as Regression and Log-Linear Models) or non-parametric techniques are used to replace or estimate data (e.g. Clustering, Sampling, and the use of histograms).
- *Dimensionality Reduction:* The majority of real-world datasets contain an enormous number of characteristics. Consider an image processing problem: it may involve dealing with hundreds of characteristics, often known as dimensions. As the name implies, dimensionality reduction aims to reduce the number of features, but not just by picking a subset of features from the feature set, which is referred to as feature selection. Three techniques for dimensionality reduction are as follows: Subset Selection of Attributes, Wavelet Transform, and Principal Component Analysis

The research article illustrates the ideas of outlier analysis and dimensionality reduction using a microarray. The structure is as follows: anomaly identification using outlier analysis and dimension reduction of human serum microarrays.

Anomalies Detection

As previously discussed, genuine information will typically be fragmentary, noisy, and contradictory. As outliers (irregularities) might have a cumulative effect on

the quality of microarray organization can achieve information, they are given further consideration. The purpose of abnormality detection is to identify artifacts that are not identical to the majority of other items. Frequently, strange artefacts are referred to as anomalies. Oddity discovery is a key subfield in mining techniques, which is concerned with the disclosure of data that deviates significantly from conventional information designs. Exception detection and elimination is critical in data mining.

Anomaly identification, or exception mining, is a technique for identifying abnormalities in a collection of data. The method of anomaly identification has applications in finance, advertising, extortion detection, interrupt recognition, biological system circulations, general health, and treatment, as well as the assessment of high-quality articulation data. As a result, identifying and investigating exceptions is an enthralling and essential information mining endeavor. Figure 2.0 illustrates an information mining measure for recognising exceptions.

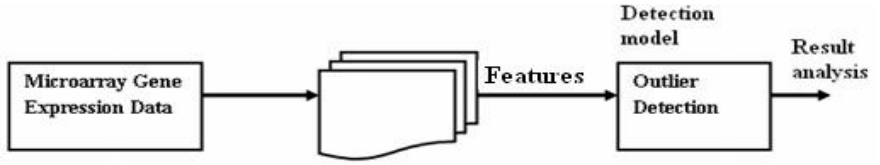


Figure 2.0 Outlier detection in Data Mining

Outlier Analysis

- Outliers are a necessary component of pre-processing data. An outlier is an observation point that deviates significantly from other observations. An outlier is significant because it indicates an experimental inaccuracy. Outliers are widely employed in a variety of applications, including identifying fraud and proposing possible new market trends. Typically, outliers are mistaken for noise[3]. Outliers, on the other hand, are distinct from noise input data in the number of ways:
- While noise is a random mistake, an outlier is an observation point which is geographically far from other observations.
- Noise should be reduced to aid in the discovery of outliers.

Importance to Handle Outliers in Data Mining

There are several reasons why outliers should be handled in Data Mining. Several of these justifications are stated below:

- Outliers have an effect on the database's results.
- Outliers frequently provide helpful or beneficial outcomes and conclusions, allowing for the identification of diverse trends or patterns.
- Outliers might be advantageous in the research area as well. They can be incredibly beneficial in some instances of discovery.
- Outliers are a critical component of data mining.

Applications of Outlier Detection in Data Mining

Outlier Detection is widely utilised in Data Mining. It is used in data mining to discover patterns or trends. Outlier Detection in Data Mining is used in the following applications: Intrusion Detection in Cyber Security, Fraud Detection, Medical Analysis, Telecommunication Fraud Detection, and Environment Monitoring for Cyclone, Tsunami, Floods, and Drought.

Approaches in Outlier Detection

Outlier detection employs primarily three methodologies. The statistical scientific methods are, the distance-based method, and the deviation-based approach.

Distance Based Approach for Outlier Detection on Human Serum Microarray

Among several ways for detecting exceptions, a distance-based strategy was chosen to discriminate anomalies in microarray datasets of human blood, Table 1.0. If the neighbourhood of an object o does not contain a sufficient number of other points, the object o is considered an outlier. This technique is deemed more appropriate for communicating microarray organization can achieve information across many time zones. The subsequent sections explore the implications of the exception detection process on four independent quality articulation samples as referenced before [1][2].

Another approach for detecting anomalies that was used in this investigation was a graphical strategy. Among several graphical tools, a box plot was created to illustrate the abnormalities in the first informative collection. Box plots are an excellent tool for conveying location and diversity data in informative sets. Box plots are defined by a vertical hub for response factors and a level pivot for the factor of concern. There is a helpful variation on the crate plot that distinguishes anomalies even more explicitly.

Table 1
Sample input dataset for human serum

Gene Index	15 MIN	30 MIN	1 HR	2 HR	4 HR	6 HR	8 HR	12 HR	16 HR	20 HR	24 HR	28 HR
361771	-0.47	-3.32	-0.81	0.11	-0.6	-1.36	-1.03	-1.84	-1	-0.6	-0.94	-0.84
120386	-0.45	1.62	1.83	0.03	0.33	0.25	-0.07	0.23	-0.4	-0.1	-0.36	-0.32
26474	1.42	3.03	**3.67**	0.58	0.66	0.78	0.3	-0.38	0.19	-0.01	-0.17	0.11
162772	0.56	2.05	2.43	0	1.36	0.06	-0.58	-0.04	-0.76	0.16	0.21	0.07
254436	0.01	2.24	3.41	1.58	1.86	0.69	0.08	-0.22	0.74	0.61	-0.32	-0.23
510136	-0.07	-0.14	0.01	0.1	2.8	1.34	0.56	0.55	0.48	0.18	0.33	-0.3
23464	-0.54	-0.27	-1.06	0.43	1.66	1.7	1.52	0.64	0.21	0.2	-0.12	0.23
364959	0.07	0.5	-0.09	0.01	1.57	1.71	1.54	0.86	-0.09	-0.49	-0.64	0.71
108837	0.25	0.82	0.78	0.61	2.26	2.61	1.77	1.17	0.66	-0.18	-0.29	1.14
328692	1.42	1.27	1.91	2.63	**5.28**	**6.44**	**4.68**	**3.89**	2.75	1.44	1.28	0.53

The data collection contains 517 objects with 12 characteristics each, where the items are genes and the properties are the relevant gene's expression values at various time intervals. In this dataset, outliers are eliminated using a distance-based outlier identification approach. This approach was used to screen the original data set and resulted in the identification of a maximum of twenty-eight outliers out from 6204 values, as shown in Tables 2 and 3.

Table 2
Outliers detected on human serum dataset

3.3400	3.3800	3.4000	3.4100	3.4400	3.4600	3.4700	3.4900	3.5200	3.5400
3.5500	3.5700	3.6300	3.6500	3.6700	3.7800	3.8900	3.9300	4.0400	4.0500
4.1200	4.1700	4.3300	4.4000	4.6800	4.8200	5.2800	6.4400		

Table 3
Summed up anomalies recognized from human serum dataset

Dataset	Human Serum
Observations	6204
Outliers detected	28

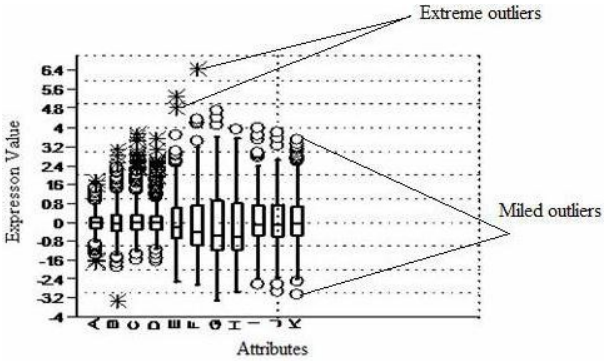


Figure 3. Box plot outliers on human serum

Figure 3. depicts a crate plot for the supplied human serum dataset, with circles representing exceptions. This plot has 28 anomalies out of around 6204 perceptions. While this approach for detecting anomalies is good for datasets with a small number of items and attributes, it is ineffective for datasets with a large number of articles and high-dimensional data. Due to the fact that the upsides of things are not visible due to the covering of foci, the plot's size is limited, as seen in the picture. Additionally, a fraction of the points indicated as anomalous by the case plot are not truly so, for example, the worth - 3.4. (lower outrageous exception). As a result, it is not recommended to use graphical approaches for identifying exceptions in high-dimensional data, as the results may be solid.

As illustrated in Figure 4.0, a dispersion plot is created for objects (qualities) versus articulation esteems transmitted at distinct time points. The x hub is concerned with characteristics, whereas the y hub is concerned with articulation esteems. It demonstrates simply that a percentage of the characteristics over 3.340 are dissimilar to other qualities, and therefore may be considered exceptions.

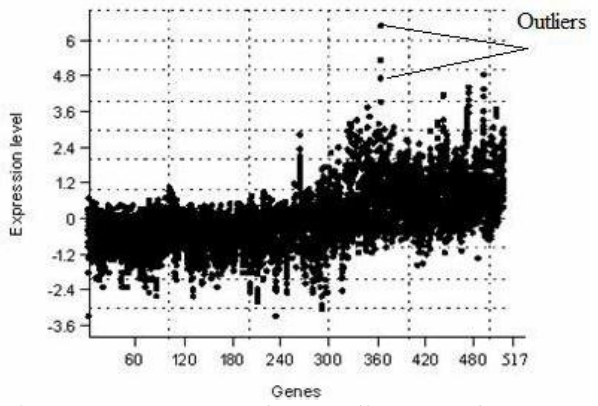


Figure 4. Scatter plot outliers on human

Principal Component Analysis (PCA)

Assume we have a data collection that has to be analysed and it contains tuples with n properties. Principal component analysis generates k unique tuples with n features that may be used to define the data collection. This way, the original data may be projected into a considerably smaller space, resulting in dimensionality reduction. Principal component analysis can be used to analyse data that is sparse or skewed. Head Component Analysis (PCA) is a time-honoured technique for reducing the dimensionality of a large data set comprised of several elements. This is performed by rearranging the factors (Principal Components) in such a way that the initial few retain the bulk of the variances found in the full of the original factors. Thus, dimensionality reduction by PCA can result in typically low-dimensional data, making it possible to use techniques that do not perform well with high-dimensional data, such as microarray quality articulation data [6][7].

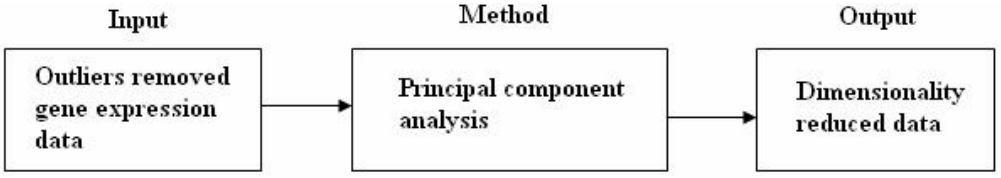


Figure 5. Framework of dimensionality reduction

Dimensionality reduction technique is applied on outliers removed gene expression dataset. Outliers present in dataset has already been detected and removed in the previous section. Figure 5. show the process of performing

dimensionality reduction, it is demonstrated in the following sections how dimensions of dataset is considerably reduced. An algorithm of PCA is as follows:

Algorithm: Principal Component Analysis (PCA)

Input: Multidimensional data in a data matrix in which the rows are genes and columns are conditions.
Output: Dimensions reduced dataset
Step 1: Calculate mean value of each object in data matrix.
Step 2: Calculate the covariance of data matrix.
Step 3: Calculate the eigenvectors and eigenvalues of the covariance matrix.
Step 4: Find the number of Components and forming a feature vector.
Step 5: Transpose the feature vector and mean adjusted data matrix.
Step 6: Infer the new dataset with diminished measurements by duplicating the component vector and mean changed information.

This section examines the ramifications of doing Principal Component Analysis (PCA) on a microarray dataset. Four unique microarray quality articulation data sets with measurement scopes ranging from two to twenty-five are investigated. After doing PCA on these datasets, it is observed that the measurements are greatly reduced, and it is recognised that components with a lower number would have produced identical results as unique measurements. This study examined four datasets: a) human serum, b) yeast, c) lung illness, and d) blood malignant growth. The preceding sub-segments discuss the representation of datasets and their results.

Human Serum Data: This dataset comprises 517 characteristics in the form of lines and 12 measurements in the form of segments. The measures must be reduced to making the dataset suitable for cluster analysis. As previously stated, a large number of measurements may result in mistake and may violate the criteria for several grouping strategies.

PCA will determine the eigenvectors and eigenvalues associated with the data by applying a covariance network as described in Table 4.0. Eigenvectors can be thought of as specific directions of an informative index, or as fundamental designs in the information. PCA may be used to two types of profiles: PCA on qualities and PCA on conditions; in this investigation, PCA on quality is used exclusively. In the context of PCA on characteristics, an eigenvector may be thought of as an articulation profile that is usually representative of the data, whilst eigenvalues can be thought of as a quantitative evaluation of how well a segment addresses the data. The more information agent, the greater the eigenvalues of a segment [7].

Eigenvalues can also be used to visualise the degree of clarified variation as a measure of total difference. Eigenvalues are not instructional in and of themselves. The percentage of variation clarified is contingent upon how well each of the parts summarises the data. In principle, the total number of segments clarifies 100 percent of the information's volatility.

Table 4
Co-variance matrix for human serum data

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
0.160	0.146	0.138	0.115	0.083	0.092	0.082	0.075	0.052	0.053	0.056
0.146	0.443	0.385	0.221	0.291	0.322	0.307	0.249	0.087	0.091	0.074
0.138	0.385	0.545	0.350	0.423	0.436	0.414	0.309	0.132	0.122	0.111
0.115	0.221	0.350	0.460	0.440	0.462	0.468	0.348	0.184	0.162	0.142
0.083	0.291	0.423	0.440	1.096	1.199	1.193	0.963	0.512	0.477	0.396
0.092	0.322	0.436	0.462	1.199	1.517	1.574	1.304	0.739	0.691	0.588
0.082	0.307	0.414	0.468	1.193	1.574	1.796	1.547	0.895	0.862	0.734
0.075	0.249	0.309	0.348	0.963	1.304	1.547	1.549	0.977	0.962	0.859
0.052	0.087	0.132	0.184	0.512	0.739	0.895	0.977	0.897	0.899	0.876
0.053	0.091	0.122	0.162	0.477	0.691	0.862	0.962	0.899	1.039	1.042
0.056	0.074	0.111	0.142	0.396	0.588	0.734	0.859	0.876	1.042	1.146

The eigenvalues and aggregate level of variance for eigenvalues from head part examination of human serum information are shown in Table 5.0. Among these, the first three PCs with the most pronounced modifications are chosen for study. One should select the components (factors) that exhibit the greatest change in the most essential eigenvectors. The first four eigenvalues and their respective level of fluctuation, as listed in the table, correspond to the 93 level of total variation in the autonomous components. Essentially, the first three eigenvalues correspond to the 90th percentile of the total variation in the autonomous components. Seven eigenvalues have a smaller difference in their change, implying that these factors are connected to the other major components. To calculate the eigenvalues of head parts, co-fluctuations of eigenvectors are computed for all components.

Table 5
Eigenvalues and its percentage of variance of human serum data

Principal Componen t	Eigenvalue	Percentage of Variance	Cumulative Percentageof Variance
1	7.72484	66.199	66.199
2	1.9893	17.048	83.247
3	0.807397	6.9191	90.166
4	0.30341	2.6001	92.766
5	0.236446	2.0263	94.792
6	0.1978	1.6951	96.487
7	0.116029	0.99433	97.482
8	0.0880234	0.75433	98.236
9	0.0743113	0.63682	98.873
10	0.0647379	0.55478	99.248
11	0.0389406	0.33371	99.762
12	0.027799	0.23823	99.999

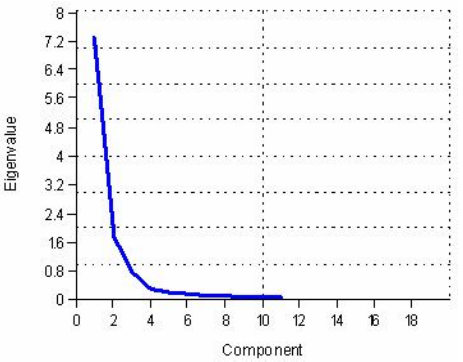


Figure 6. Scree plot for human serum data to determine number of components

Choosing the appropriate number of computers is also critical. On human serum data, there are 12 head sections from which only the segments with the biggest absolute difference should be chosen. With the aid of the scree plot, this should be doable. In Figure 6.0, when the bend is hazardous, the number of segments to be chosen is three. Thus, the first three portions may be used to get the total difference. The Scree diagram method of determining the number of PCs is extremely spontaneous and emotive.

As seen in Table 6.0, the simplified form of human serum data is utilised to cluster the dataset's patterns. The number of clusters generated by the Hybrid Clustering Technique following PCA is deemed optimal.

Table 6
Reduced representations of human serum data

15 MIN	30 MIN	1 HR	2 HR	4 HR	6 HR	8 HR	12 HR	16 HR	20 HR	24 HR	28 HR
0.160	0.146	0.138	0.116	0.083	0.092	0.082	0.076	0.052	0.053	0.057	0.081
0.146	0.444	0.385	0.221	0.291	0.322	0.307	0.249	0.087	0.091	0.074	0.094
0.138	0.385	0.545	0.350	0.423	0.436	0.414	0.309	0.132	0.122	0.111	0.129
0.115	0.221	0.350	0.460	0.440	0.462	0.468	0.348	0.185	0.162	0.142	0.178
0.083	0.291	0.423	0.440	1.097	1.199	1.193	0.963	0.512	0.477	0.396	0.341
0.092	0.322	0.436	0.462	1.199	1.517	1.574	1.304	0.739	0.691	0.588	0.535
0.082	0.307	0.414	0.468	1.193	1.574	1.796	1.547	0.895	0.862	0.735	0.680
0.076	0.250	0.310	0.349	0.963	1.304	1.547	1.550	0.978	0.962	0.860	0.734
0.052	0.087	0.132	0.184	0.512	0.740	0.895	0.978	0.898	0.899	0.876	0.7381
0.053	0.091	0.122	0.162	0.477	0.691	0.862	0.963	0.899	1.039	1.042	0.850
0.056	0.074	0.111	0.142	0.396	0.588	0.734	0.859	0.876	1.042	1.146	0.910
0.081	0.094	0.134	0.178	0.341	0.535	0.680	0.734	0.738	0.850	0.910	1.047

After completing PCA, the number of clusters formed by Hybrid Clustering Technique is regarded optimal. As indicated above, after lowering the dimensions of the original dataset, the efficacy of a new clustering algorithm is boosted. The percentage decrease in the dimensions of the human serum dataset is shown in Table 7. The decrease ranges from 40 to 66 percent.

Table 7
Original dimensions versus reduced dimensions

Dataset	Human Serum
Original Dimensions	12
Reduced Dimensions	8
Percentage of Reduction	66%

Conclusion

Prior to bunching, conflicting information such as exceptions should be deleted to improve the quality. It is discovered that an algorithmic method called distance-based exception detection technique is more trustworthy for obtaining high-quality articulation information, while a graphical strategy called box plot is more appropriate for a small dataset. Principal Component Analysis is used to reduce the dimensions of four datasets, and the competency of the new cross breed grouping computation is completely improved as a result of the large reduction in dimensions, as reviewed in the succeeding section.

Reference

[1] Bay, SD & Schwabacher, M 2003, 'Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule', Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).

[2] Yogeesh N. "Graphical Representation of Mathematical Equations Using Open Source Software." Journal of Advances and Scholarly Researches in Allied Education, vol. 16, no. 5, 2019, pp. 2204-2209 (6), www.ignited.in/p/304820.

[3] Edwin M Knorr, Raymond T Ng & Vladimir Tucakov 2000, 'Distance-based outliers: algorithms and applications', The International Journal on Very Large Data Bases, vol. 8, no. 3-4, pp. 237-253.

[4] Yogeesh N. "Mathematical Maxima Program to Show Corona (COVID-19) Disease Spread Over a Period." TUMBE Group of International Journals, vol. 3, no. 1, 2020, pp. 14-16.

[5] Hawkins, S, He, H, Williams, GJ & Baxter, RA 2002, 'Outlier detection using replicator neural networks', In: Kambayashi, Y, Winiwarter, W, Arikawa, M. (eds.) DaWaK 2002. LNCS, Springer, Heidelberg, vol. 2454, pp. 170-180.

[6] Girija D.K & M. S. Shashidhara, 'Data mining techniques used for uterus fibroid diagnosis and prognosis' In Kottayam, India, 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), INSPEC Accession Number: 13567035.

[7] Girija D.K & M. S. Shashidhara, 'Data mining approach for prediction of fibroid disease using neural networks', Bangalore, India, 2013 International

Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA), INSPEC Accession Number: 14130830.

[8] Ka Yee Yeung and Ruzzo W.L. (2000), 'An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data', Technical Report, Dept. of Computer Science and Engineering, University of Washington.

[9] Yogeesh N. "Study on Clustering Method Based on K-Means Algorithm." Journal of Advances and Scholarly Researches in Allied Education (JASRAE), vol. 17, no. 1, 2020, pp. 485-489(5), www.ignited.in//I/a/305304.

[10] Jolliffe I.T. (1986), 'Principal Component Analysis', Springer-Verlag, Second edition, New York.

[11] Yogeesh N. "Mathematical Approach to Representation of Locations Using K-Means Clustering Algorithm." International Journal of Mathematics And its Applications (IJMAA),