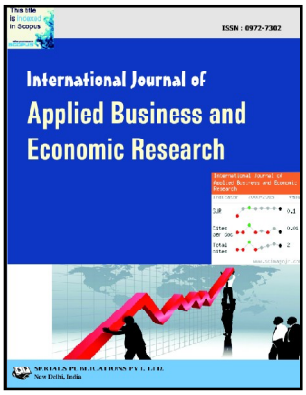

International Journal of Applied Business and Economic Research

ISSN : 0972-7302

available at http: www.serialsjournals.com

Volume 15 • Number 19 (Part-II) • 2017

Evaluation of Mutual Influence and Correlation Performance of Global Stock Indices Using Unsupervised Learning

Dilip Singh Sisodia[1] and Sagar Jadhav[2]

[1,2] National Institute of Technology Raipur

Abstract: In this study, the mutual influence of global stock indices on Indian stock index is evaluated using unsupervised learning methods. The unsupervised learning techniques are applied to the publicly accessible data set collected from past stock indices performance records. A historical data of 17 global stocks indices are used in this study for mutual influence and correlation analysis according to the market capitalization. The raw stock indices data set is pre-processed by removing inconsistencies and computing features. The model applies three independent partition based unsupervised learning techniques including K-Means, K-Medoids, and clustering large application (CLARA) on this dataset. The optimal number of clusters is decided empirically by using a popular internal clustering validity measures including Davies Bouldin (DB), Dunn's validity (DV), and Silhouette index. The result suggests that K-Means perform better than other two on the used dataset. The result generated by this model is then analyzed for knowledge discovery, and the stock indices with maximum impact on Indian Stock market are found. The indices which have a greater number of instances in the cluster containing highest instances of the Indian stock market are considered to have a high impact or high correlation with Indian stock market.

Keywords: clustering, cluster validity measures, CLARA, K-means, K-medoids, stock market analysis, unsupervised learning.

1. INTRODUCTION

A standout amongst the most significant and broad money related wonder in the late twentieth century and the forepart of this century is the dangerous development in stock exchanges, and capital flows among different equity markets in developed and developing nations. The wonder in the global fund is not just a consequence of the liberalization of financial markets in the advanced and developing countries. The expanding assortment and many-sided quality of money related instruments, additionally an aftereffect of the expanding relativity of the developing and developed economies as developing countries get to be more coordinated in universal streams of exchange and installments. There has been a focus on the degree

of interdependence for returns among international stock markets and find that the first moments of equity returns among the equity markets exhibit a high level of interaction. Many researchers and economists have studied the correlation between different global stock indices of the various countries. Various theories have been proposed [1], and many techniques have been applied to establish a relationship between the global indices. The impact of international indices on a specific stock index has been a topic of growing interest in recent times [2]. One of the reasons for mutual inductive effect between stock indices may be the more flow of capital across national borders. The other important reason assumed behind this phenomena is the rapid transmission of information across the globe with the advancement in technology and information dissemination techniques [3]. There has also been an interest in finding out the influence of developed nations stock indices on the indices of developing countries. There has also been an investigation on the interdependence of the second moments, i.e., volatility [4] transmissions, among international markets. More opportunity in the moving of capital streams enhances the assignment of capital all around, permitting assets to move to territories with higher rates of return. Oppositely, endeavors to limit capital streams lead to mutilations of capital structure that are for the most part expensive to the economies forcing the controls. Along these lines, the help in worldwide capital streams and monetary exchange is an in progress and, to some extent, irreversible procedure. The goal of this paper is to build a system capable of performing the following tasks:

a) Creating a dataset by raw data downloaded from the internet
b) Calculate the optimal number of clusters needed for the model
c) Knowledge discovery in the result generated by the model to find out the stock indices with maximum influence on Indian stock market.

2. RELATED WORK

The short term and long term relationship between the Indian Stock indices and other global indices including USA, UK, and Japan are studied using the autoregressive moving average (ARMA). The results suggest that there exist a correlation between these markets and also indicate that the mature markets have a high correlation with the Indian Stock Market [5]. In [6] author has applied clustering methods such as k-Means, Self-organizing maps (SOM), Fuzzy C-means to create an efficient portfolio among the selected stocks from Bombay stock exchange. The different indexes and Intra-class inertia are used to decide to decide an optimal number of fixed clusters to help investors for maintaining effective and diversified the portfolio and finding the stock market trends. A three-phase techniques pre-clustering, splitting and merging are applied on the stock data of Kuala Lumpur stock exchange (KLSE), Malaysia to categorize different companies according to their similarity/dissimilarity score [7]. The effect on one stock market can be transmitted to other global markets as shown in [8] and [9]. The correlation between markets can also be found using switching ARCH (SWARCH) model [10]. There has been a high correlation between the stock markets of developed countries and developing or newly emerging economies after the 1987 stock market crash [11] and the Asian financial crises of 1997 as described in [1]. The effect of short-run and long-run relationship between markets of South Asia is analyzed in [3] using high-dimensional dataset for testing. The issue of dimensionality reduction is addressed in[12] using principal component analysis (PCA). The PCA results suggest that Expectation-Maximization (EM) algorithm does not perform well as compared to K-Means algorithm on high-dimensional data. Clustering can be carried out on both labeled and unlabelled data. The performance of labeled data clustering is measured by external measures as discussed in[13][14].

The performance of unlabelled data clustering is evaluated by internal validation measures described in [15], along with the comparison[16]. Reviewing the above research work suggests that clustering can be applied to find out the correlation between global indices. K- means has been developed and applied to various fields on different sets of data either supervised or unsupervised as described in [17]. Thus, K-means appears to be the suitable and valid choice among various clustering algorithms to use as the core algorithm for completion of the objective of this paper.

From the literature reviewed, it can be inferred that various approaches have been used for finding the correlation between stock indices. In this paper, clustering techniques have been used to find those stock indices which have a high correlation with the Indian stock market.

3. METHODOLOGY

In this paper, Partitioning based Clustering algorithms are used for comparative analysis. All of the clustering algorithms, namely, Centroid-based k-means [17], Representative object based k-medoids and CLARA are implemented by using various libraries in JAVA, and their performance is analyzed based on their clustering quality.

3.1. K-Means Clustering

K-Means clustering creates a partition with K clusters randomly initialize k data points as cluster centroids (cluster centers) [12] and iteratively minimized the squared error between the observed mean of a centroid and the data points in the cluster. Suppose $X = \{x_i\}$, where $i = 1, 2 \ldots\ldots n$ data points are clustered into a set of K clusters, where $C = \{\{c_k, k = 1 \ldots\ldots K\}$. Assume u_k is the mean of cluster c_k. The squared error between L_k and the points in the cluster c_k is defined as Eq. (1) and objective function as Eq. (2) [17].

$$J(c_k) = \Sigma_{x_i \in c_k} ||x_i - u_k||^2 \quad (1)$$

$$J(C) = \Sigma_{K=1}^{K} \Sigma_{x_i \in c_k} ||x_i - u_k||^2 \quad (2)$$

Pseudo Code for K-Means algorithm [17]

Input: *k: the number of clusters; D: a data set of n objects.*

Output: *A set of k clusters.*

Method

1. Randomly choose any k objects from DB as the initial centers of cluster
2. **Repeat**
3. Calculate cluster centroids
4. Update the assignment of each object to the cluster to which the object is closer to its centroid, based on the mean value of the objects in the cluster
5. Recalculate the centroid for each cluster.
6. If Cluster Membership is stabilized BREAK;
7. **Until**
8. END

3.2. K-Medoids Clustering

The K-means clustering is sensitive to outliers and significantly distort the distribution of data[18]. Therefore in K-Medoids clustering in place of the mean value of the data points as cluster centroids the centrally located points may be assumed as initial cluster centers which known as medoids [19], [20], [21].

Pseudo Code for standard version of K-Medoid algorithms[19]

Input: K: the number of clusters**;** D: a data set containing n objects

Output: A set of k clusters

Method

1. Arbitrary select k seed data points from D as medoids m_h
2. Repeat
3. Allocate each remaining data point to the cluster with the nearest medoid
4. For each medoid m_h
5. Arbitrarily pick a non-representative data point d_{random}
6. Compute the total cost S of swapping medoid m_h with d_{random}
7. if S < 0 then replace m_h with d_{random}
8. Until no change

3.3 Clustering Large Application (CLARA)

Clustering Large Application (CLARA) algorithm is an amalgamation of partition around medoid (PAM) with sampling procedure and used to cluster large data sets. CLARA takes a small sample of a large data set and uses PAM to generate k medoids from the sample and uses the k medoids to cluster the rest of objects by the rules $\{x \in S_i \text{ if } d(x, q_i) \le d(x, q_j) \wedge i \ne j\}$, where $1 \le i, j \le k$, d is a dissimilarity measure, q_j is the medoid of cluster j and S_i is cluster i. For each sample, CLARA only goes through the large data set once. Its computational complexity depends on the computational complexity of PAM which is decided by the size of the sample[22].

Pseudo Code for CLARA algorithm [23]

1. For $i = 1$ to 5, repeat the following steps
2. Draw a sample of 40+2k objects randomly from the dataset and find k medoids using standard PAM algorithm
3. Find the k medoid which is most similar to the object from the data set O_j
4. From the previous step, find the average dissimilarity of the cluster. If the value is less than the current minimum, set this value to the current minimum and retain k medoids found in Step (2) as the best medoid.
5. Return to Step 1

4. HYPOTHESIS AND PROPOSED MODEL

4.1. Hypothesis

The proposed hypothesis works on the principle of partitioning based algorithms. Initially, the data set is collected, and then features are calculated. The ten features act as ten coordinates in a multidimensional space. The instances act as points with ten coordinates in a multidimensional space. The principle of partitioning based algorithms suggests that the points closer to each other in multidimensional space will fall into the same cluster. Therefore, after analysis of clustering results if maximum instances of any stock indices fall in the cluster of Indian stock market then such stock indices are having a major influence on the Indian stock market. For example, by considering two features and following hypothetical feature values for various stock indices such as India (1, 1), China (1, 2), USA (10, 10) and Brazil (10, 11), the clusters of correlated stock indices are formed as shown in Figure 1.

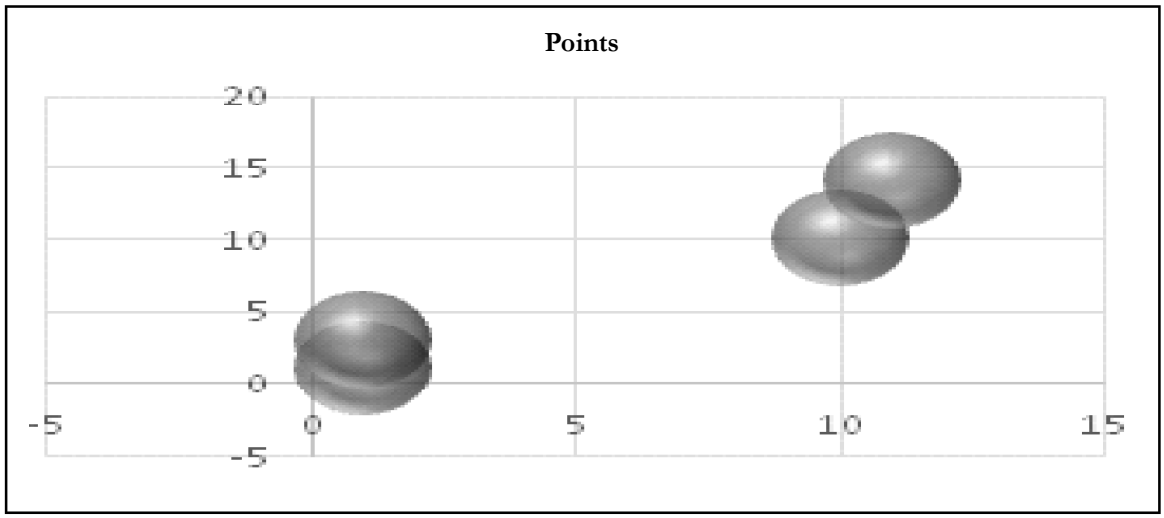

Figure 1: Hypothetical Clusters of Correlated Stock Indices

4.2. Model Description

The data set needed for the model needs to be generated. Hence, the software has been developed to produce the dataset as shown in Figure 2. It downloads the historical data of stock indices from various sources. If any inconsistencies exist, then it eliminates them and calculates the features required for the model proposed in this project.

As shown in Figure 3, initially the software converts the CSV file into data matrix form. The number of clusters is taken as input from the user, and the method is selected to among K-Means, CLARA, and K-Medoids. Error measures are calculated and analyzed to find out the best method and an optimal number of clusters.

After finding an optimal number of clusters, the best algorithm for this dataset is applied and the clusters formed are analyzed for knowledge discovery to find the instances of indices which are in the same cluster as that of the instances of Indian Stock market cluster as shown in Figure 4.

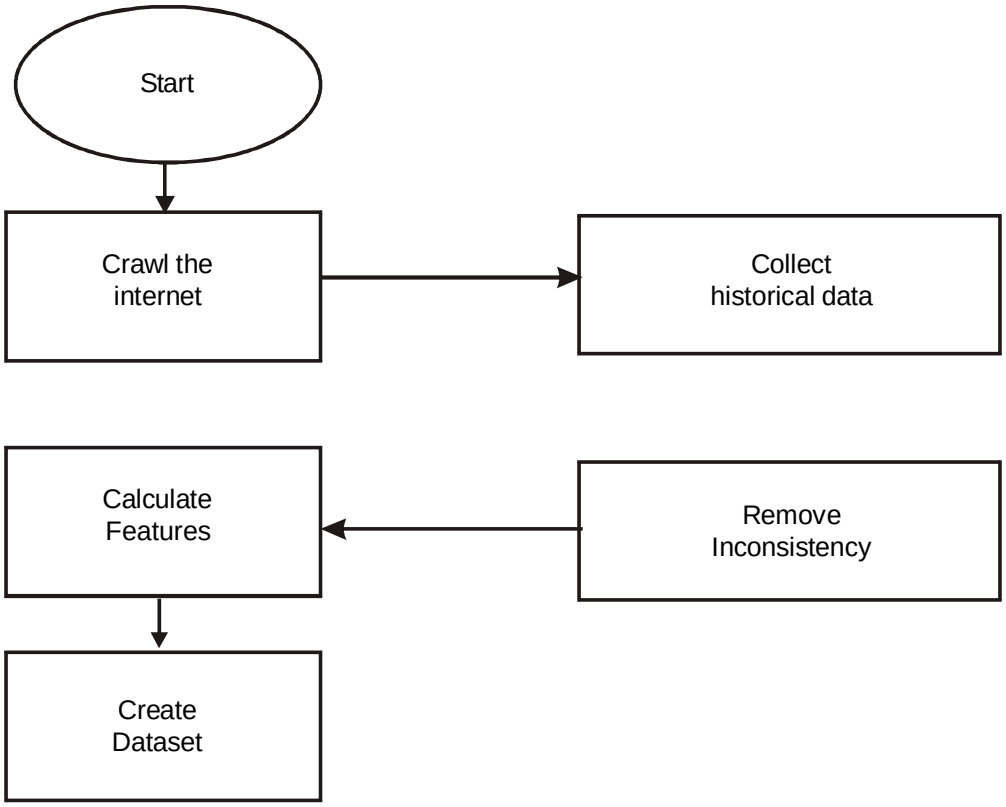

Figure 2: Process Flow for Preparing the Dataset

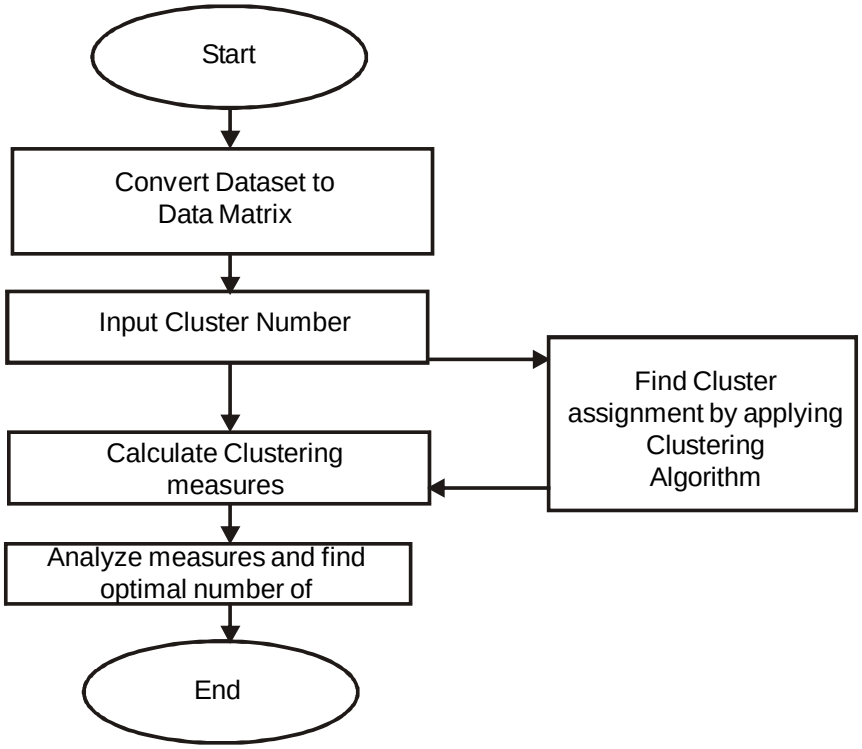

Figure 3: Flowchart for Determining the Optimal Number of Clusters

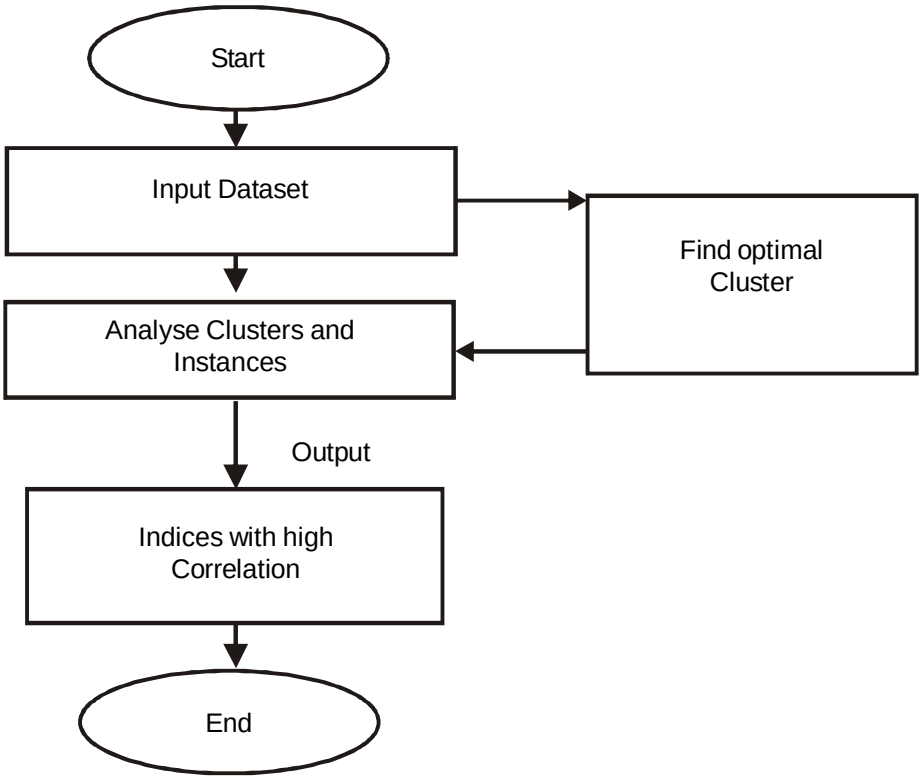

Figure 4: Flowchart for Finding the Indices with High Correlation

4.3. Validation Measures

The number of optimal clusters and generated cluster quality is evaluated with external and internal validity measures[16][24]. The clusters generated by unsupervised learning are assessed only with internal validation measures. In this study, three performance validity measures are used to find the optimal number of clusters[25].

4.3.1. Dunn's Validity (DV) Index

Dunn's validity (*DV*) index (Eq. 3) is the ratio of the smallest distance between data points in different cluster to the largest intra-cluster distance [26]. The value of Dunn's index lies between zero and infinity and maximum value of this index suggest the optimal cluster number.

$$\mathcal{DV}_{index} = \min_{i=1,2..k}\left\{\min_{j=i+1,..k}\left(\frac{d(c_i,c_j)}{\max_{l=1,2,..k} diam(c_l)}\right)\right\} \quad (3)$$

Where, $d(c_i,c_j) = \min_{\mathcal{P}_{i\in\mathcal{C}_i},\mathcal{P}_{j\in\mathcal{C}_j}} d(\mathcal{P}_i,\mathcal{P}_j)$ and $diam(c_l) = \max_{\mathcal{P}_i,\mathcal{P}_j\in c_l} d(\mathcal{P}_i,\mathcal{P}_j)$

4.3.2. Davies-Bouldin (DB) Index

The Davies-Bouldin (*DB*) index (Eq. 4) is based on a ratio of inter-cluster and intra-cluster distances. The minimum value of this index suggests the optimal cluster number [27].

$$\mathcal{DB}_{index} = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1,2,..k,j\neq i}\left(\frac{diam(c_i)+diam(c_j)}{d(c_i,c_j)}\right) \quad (4)$$

Where, $d(\mathcal{P}_i, \mathcal{P}_j)$ is the distance between two sessions in cluster $\mathcal{C}_i$ and $\mathcal{C}_j$, $d(c_i, c_j)$ is the dissimilarity function between two cluster $\mathcal{C}_i$ and $\mathcal{C}_j$ and $diam(c_\ell)$ diameter of cluster $\mathcal{C}_\ell$.

4.3.3. Silhouette index (S)

The Silhouette (*S*) index (Eq. 4) for each data point p is defined as the difference between *b*(*i*) B and *a*(*i*) divided by the maximum of the two max {*a*(*i*), *b*(*i*)}. The optimal cluster number is determined by maximizing the value of this index [28].

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (5)$$

Where *a*(*i*) is cohesion- the average distance between p and all other points in the same cluster and *b*(*i*) is separation- the average distance between p and all points in the nearest clusters.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

The historical data of 17 global stock indices have been used for experimentation which includes 200 instances of daily trading values of every stock index. The software has been developed for experimentation which applies K-means, K-Medoids, and CLARA on the dataset to generate results. The results have been analyzed by the software, to display the final indices which have a high correlation with the Indian stock market.

The experiments are performed using K-means, K-Medoids, and CLARA clustering algorithms on the preprocessed dataset and results are reported in Table 2. The value of three internal clustering measures including DBI, Dunn's index, and Silhouette index are evaluated. The maximum value of DBI and a minimum value of Dunn's index is considered as optimal while the value of Silhouette index varies from range {-1, 1} and values nearer to 1 is considered as optimal. It can be observed from results that DBI and Silhouette index are producing optimal value for two numbers of clusters for all clustering methods while Dunn's index is suggesting two clusters only for k-means algorithms and a different number of optimal clusters for other algorithms.

The analysis of experimental result suggests that different internal cluster validation measures more or less satisfying the condition of optimality for two number of cluster for all algorithms. The graphical comparison of all the 3 clustering algorithms has been shown in Figure 5, Figure 6, and Figure 7 for DBI, Dunn's, and Silhouette index. The analysis of these graphs infers that K-Means may be considered a better algorithm.

The DBI index plot for all the 10 clusters is shown in Figure 5. The observations show that DBI index is low for Cluster = 2. Hence, the optimal number of clusters considering the DBI index is 2.

The Dunn index plot for all the 10 clusters is shown in Figure 6. The observations show that Dunn index is maximum for two clusters. Therefore, the optimal number of clusters considering the Dunn index is two.

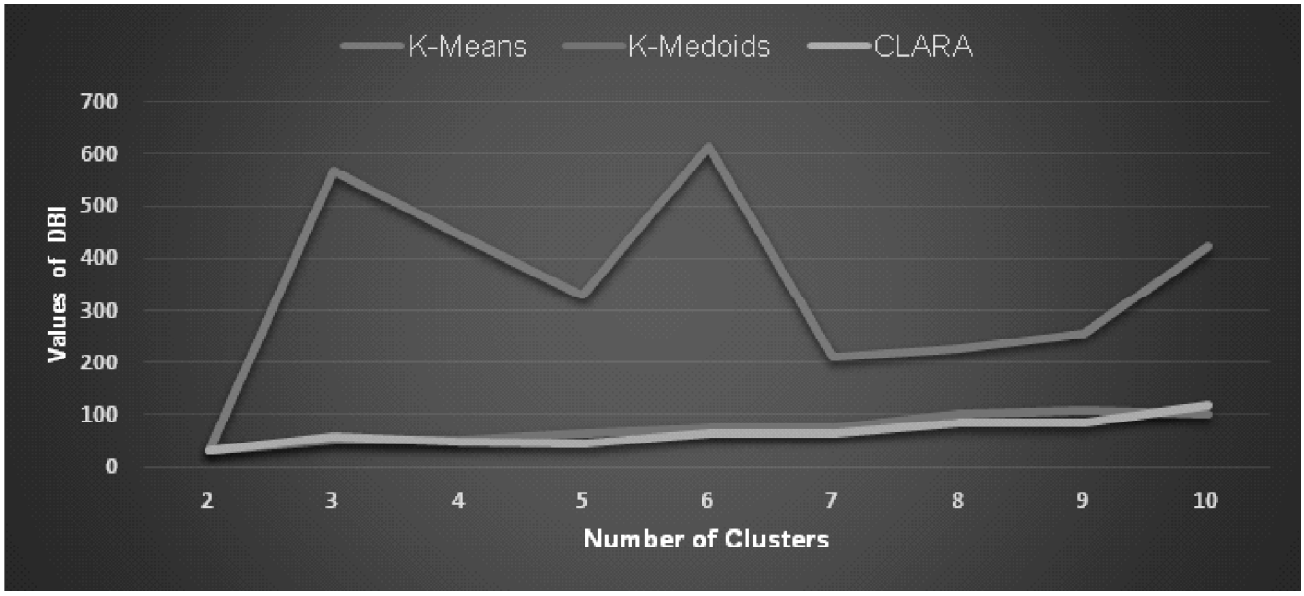

Figure 5: DB index vs. Number of Clusters

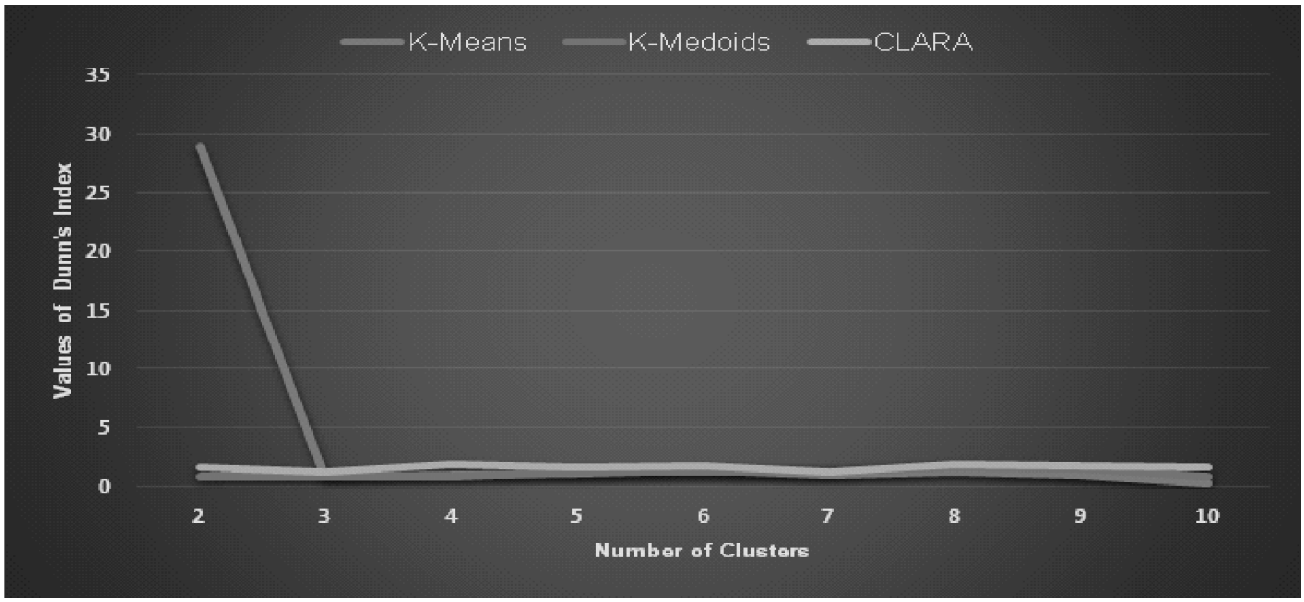

Figure 6: Dunn index vs. Number of Clusters

Figure 7 shows the graphical plot for values of Silhouette Index. The graph clearly shows that the values at cluster number two are closer to 1 than others. Therefore, the optimal number of clusters considering the silhouette index is two.

Table 1 shows the number of instances of each Index in both the clusters. It can be inferred from the table that Australia, Brazil, China, Malaysia, and the USA have a maximum number of instances in the cluster 1 which contains all the instances of the Indian cluster.

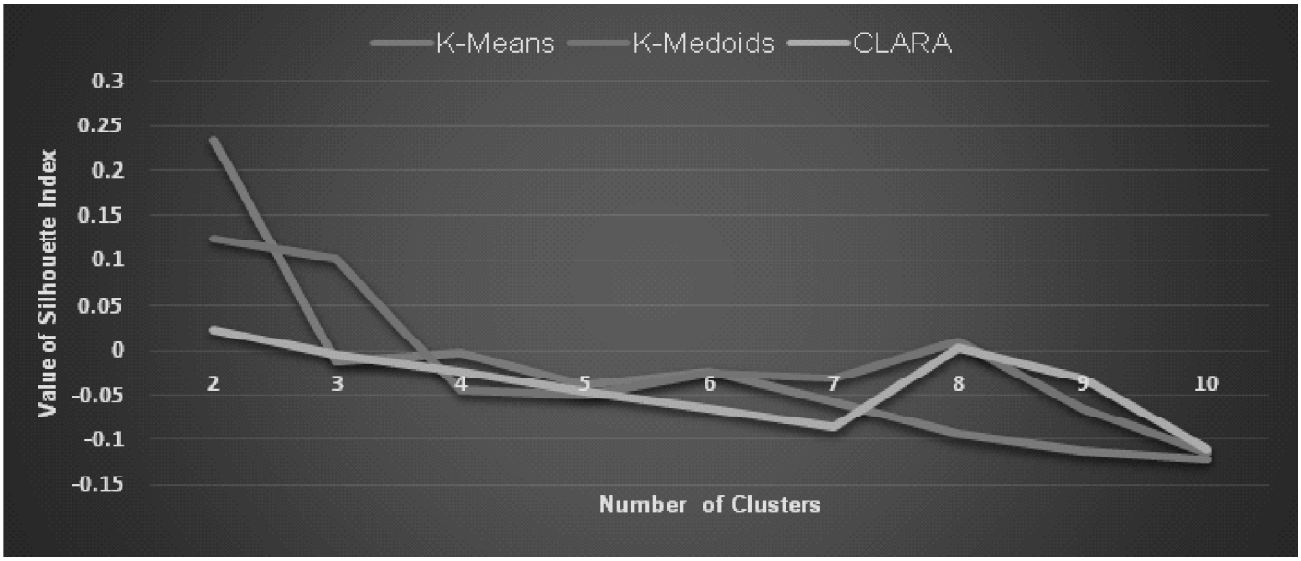

Figure 7: Silhouette index vs. Number of Clusters

Table 1
Instances of Indices in different Clusters

Index Country	*Instances in Cluster #1*	*Instances in Cluster #2*
Australia	200	0
Brazil	200	0
Hong Kong, China	178	22
Germany	0	200
India	200	0
Israel	82	118
Jakarta	65	135
Japan	77	123
South Korea	0	200
London, UK	97	103
Malaysia	165	35
Mexico	0	200
Paris, France	70	130
Switzerland	90	110
Toronto, Canada	87	113
USA	190	10
Vienna, Austria	112	88

The threshold values have been assumed to be 150, but may vary according to the distribution of the instances. The number of instances of each of these high impact indices has been showed in Table 2.

Table 2
Indices having high correlation with Indian Index

Indices with High Correlation with Indian Stock Market	*Instance in Cluster containing Indian Index (out of 200)*
Australia	200
Brazil	200
China	178
Malaysia	165
USA	190

The scatter plot for the clusters formed from experimentation is shown in Figure 8.It includes all the instances of all 17 stock indices considered in this paper for experimentation.

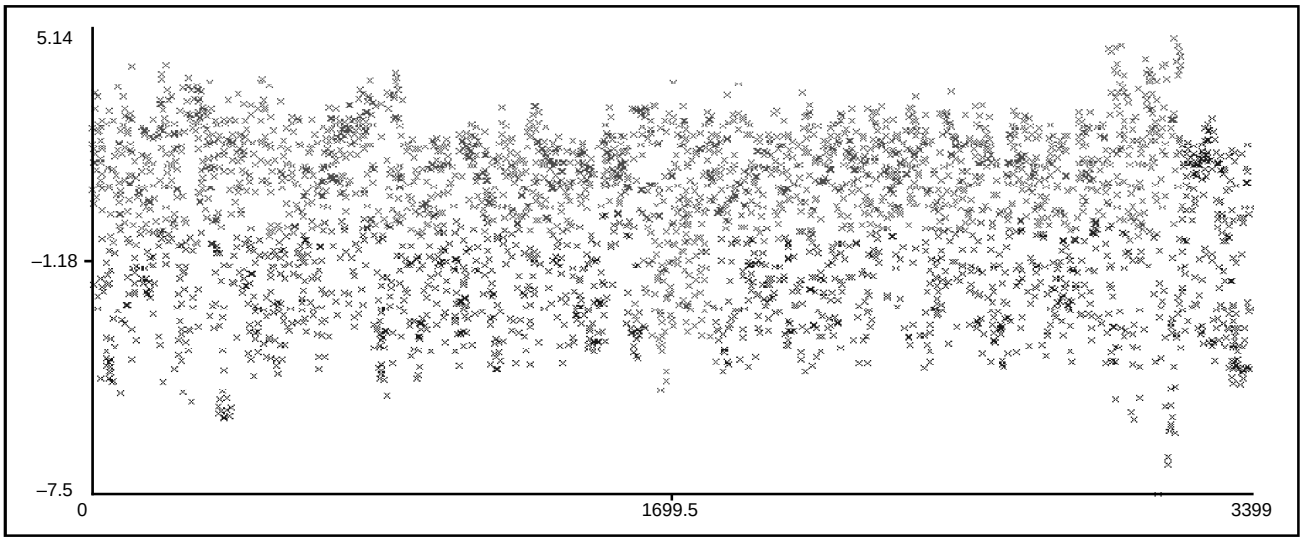

Figure 8: Scatter plot of clusters

6. CONCLUSION AND FUTURE WORK

Analyzing the stock market has always been an area of vast research and interest. In this paper, the impact of global stock indices on Indian stock market was analyzed using clustering techniques, and knowledge discovery on the result set received. Initially, the dataset is created with a specific number of instances of 17 major stock indices around the globe and K-Means has been applied on the dataset to find the optimal number of clusters. Once the optimal numbers of clusters are found which is found to be 2, then knowledge discovery or analysis is done on the result to know the stock indices which have a major impact on the Indian stock market. Among the 17 stock markets, five equity markets are considered to have a high correlation with the Indian stock market, which includes the markets of Australia, Brazil, China, Malaysia, and the USA. The hypothesis proposed in this study can be improved or extended to other dataset and also can be applied on other stock indices which are not included in this study. The model can also be used to analyze the impact of other sectors/stocks in a stock market on a specified sector/stock. Thus, this study has much scope in the future and can be extended to explore many more areas

REFERENCES

K. J. Forbes and R. Rigobon (2016), "No Contagion , Only Interdependence/ : Measuring Stock Market Comovements," *The Journal of Finance*, vol. 57, no. 5, pp. 2223–2261.

C. Sinha and N. C. Pradhan (2008), "Regional Financial Integration in Asia: Present and Future," *BIS Papers*, vol. 42.

P. Narayan, R. Smyth, and M. Nandha (2004), "Interdependence and dynamic linkages between the emerging stock markets of South Asia," *Accounting and Finance*, vol. 44, no. 3, pp. 419–439.

I. Y. Chuang, J. R. Lu, and K. Tswei (2007), "Interdependence of international equity variances: Evidence from East Asian markets," *Emerging Markets Review*, vol. 8, no. 4, pp. 311–327.

W. Wong, J. Penm, R. D. Terell, and K. Y. Ching (2004), "The Relationship Between Stock Markets Of Major Developed Countries And Asian Emerging Markets," *Journal of Applied Mathematics & Decision Sciences*, vol. 8, no. 4, pp. 201–218.

S. R. Nanda, B. Mahanty, and M. K. Tiwari (2010), "Clustering Indian stock market data for portfolio management," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8793–8798.

S. Aghabozorgi and Y. W. Teh (2014), "Stock market co-movement assessment using a three-phase clustering method," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1301–1314.

C. S. . Eun and S. Shim (2010), "International Transmission of Stock Market Movements," *Journal of financial and quantitative Analysis*, vol. 24, no. 2, pp. 241–256.

B. Solnik, C. Boucrelle, and Y. Le Fur (1996), "International market correlation and volatility," *Financial Analysts Journal*, vol. 52, no. 5, pp. 17–34.

L. Ramchand and R. Susmel (1998), "Volatility and cross correlation across major stock markets," *Journal of Empirical Finance*, vol. 5, no. 4, pp. 397–416.

B. Arshanapalli and J. Doukas (1993), "International stock market linkages: Evidence from the pre- and post-October 1987 period," *Journal of Banking and Finance*, vol. 17, no. 1, pp. 193–208.

N. Alldrin, A. Smith, and D. Turnbull (2003), "Clustering with EM and K-means, Tech Report, University of San Diego, California,".

O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona (2013), "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256.

M. Halkidi, Y. Batistakis, and M. Vazirgiannis (2002), "Cluster Validity M e t h o d s/ : Part I," vol. 31, no. 2, pp. 40–45.

Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu (2010), "Understanding of Internal Clustering Validation Measures," in *IEEE International Conference on Data Mining (ICDM)*, pp. 911–916.

E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz (2011), "Internal versus External cluster validation indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27-34.

A. K. Jain (2010), "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666.

T. Velmurugan and T. Santhanam (2010), "Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points," *Journal of Computer Science*, vol. 6, no. 3, pp. 363–368.

M. D. Boomija (2008), "Comparison of Partitioning Based Clustering Algorithm," *Journal of Computer Applications*, vol. 1, no. 4, pp. 18–21.

H. S. Park and C. H. Jun (2009), "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341.

Z. Huang (1998), "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304.

Kaufman, Leonard, and Peter J. Rousseeuw (2009), *Finding groups in data: an introduction to cluster analysis*.

Kaufman, Leonard, and Peter J. Rousseeuw (2008), "Clustering large applications (Program CLARA).," pp. 126–163.

D. S. Sisodia, S. Verma, and O. P. Vyas (2016), "A Discounted Fuzzy Relational Clustering of Web Users' Using Intuitive Augmented Sessions Dissimilarity Metric," *IEEE Access*, vol. 4, no. 99, pp. 6883–6893.

D. S. Sisodia, S. Verma, and O. P. Vyas (2016), "Augmented Intuitive Dissimilarity Metric for Clustering of Web User Sessions," *Journal of Information Science, Vol. 43, No.4, DOI: 10.1177/0165551516648259*, pp. 1–12.

J. C. Dunn (1974), "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104.

D. L. Davies and D. W. Bouldin (1979), "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, vol. 1, no. 2, pp. 224–227.

P. J. Rousseeuw (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65.