TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

Implementation of Map Reduce using Fuzzy C-Medoids Clustering on Time-Series Stock Market Big Data for Prediction

T P Sameerapallavi[1] B. Manjula[2]

[1]Research Scholar, Department of Computer Science, Kakatiya University, Warangal, Telangana, India
[2]Assistant Professor, Department of Computer Science, Kakatiya University, Warangal, Telangana, India

Article Info
Volume 82
Page Number: 14496 - 14507
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract:
Stock market data prediction is a problem of time series clustering and prediction. Clustering time series data can extract important information as well as additional features which help in efficient data classification. Several algorithms have been presented by researchers over the years in this domain. Most of the efficient clustering algorithms suffer from the drawback of higher complexity thereby increase in the computational time. This paper presents a weighted fuzzy C – Medoids algorithm with Dynamic Time Warping (DTW) modified according to parallel map reduce algorithm. The traditional fuzzy C - Medoids clustering is a sequential algorithm which clusters the input data into groups. As the historical data available for the stock market is huge, the time taken for clustering such data for further processing is time consuming. Modifying the existing algorithms based on the parallel map reduce would improve the efficiency of the overall method. The novel algorithm presented in the paper divides the data to be clustered into chunks and processes each chunk in parallel. The experimental results present how the proposed method has a reduced complexity of *O(nk/m)*, which is *m* times less than the conventional weighted fuzzy C - Medoids. The experimental results prove that the accuracy of the proposed algorithm has increased compared to the existing approaches.

Keywords: *Stock market, Data prediction, Dynamic Time Warping, Fuzzy C – Medoids, Map Reduce.*

I. INTRODUCTION

Stock market prediction is a process that tries to define the upcoming price of a corporate stock to be able to assess the appreciation or depreciation of a stock. Predicting a stock's performance can be useful as this leads to increase in the profits for the traders. Any publicly known information about the company will have a direct effect on its stock price. The expectation on the stock price may have both positive and negative effect on the stock. Thus the stock value is deemed to be random and change instantaneously. Market hypothesis tries to provide predictions based on the previous data available, the world markets and financial condition of the company.

People use various methods and techniques to predict these fluctuations. These methodologies are generally classified into three types:

- Fundamental
- Technical
- Technological

Fundamental analysts focus on the company of which the stock belongs to. They study the past development of the company and its financial status. Many performance ratios like P/E ratio are calculated and a hypothesis is formulated on the basic of which the user tires estimate the future stock value. The primary goal of this type of analysis is to find out a stock's true value and compare it with the current trade value. This comparison will yield whether a stock is undervalued or not, there by predicting the future outcome.

Technical analysts do not concentrate on one single

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

stock but watch the entire market state. The analysis is done through numerical charts and their patterns. The future price is predicted by the current price, the past price over 10 years, the past price over 30 years and so on. To facilitate this, patterns such as moving average, exponential moving average, candle charts, bar graphs etc are used. The volatile nature of the stocks can be clearly viewed in this method.

Technological analysis uses the modern day computers to predict the stock market data. These system take the input from all sources available in the internet. Time series data forecasting is one such method where the outcome of a data can be predicted before the actual value is revealed. This paper presents a method for time series data prediction for stock market analysis using fuzzy c-mediods and map reduce [1].

Data objects are divided into homogeneous combinations using clustering to facilitate grouping of objects into similar groups based on their characteristics. Researchers have offered a huge number of clustering algorithms up to now [2-3]. Fuzzy clustering assigns multiple clusters to each object in the initial stage. Then, by mimicking the human brain, the algorithm decides the cluster in the final stage based on the cluster similarities. A very frequently used clustering technique is Fuzzy C-Medoids (FCMdd). FCMdd is similar to Fuzzy C Means (FCM) clustering in terms of the objective function minimization and the procedure to obtain the cluster centers (V). This method also produces the partition matrix (U) containing the cluster weights of each particle. The major difference is the way of producing the cluster centers. In FCM virtual objects are selected as initial centers where as the FCMdd selected the same objects form the dataset as initial centers. In comparison to FCM, FCMdd is noise resistant. The clustering results in FCMdd are also more accurate and the algorithm runs much faster than FCM.

Euclidean Distance (ED) measures the distance between two vectors. The value of ED is less when the vectors are similar to each other. If finds out the square root of squared difference between individual elements in a vector. This feature becomes unreliable when the vectors are too random. Therefore it may be observed that Euclidean distance doesn't remain the optimum selection aimed at time sequences clustering. As a result, Dynamic Time Warping (DTW) distance is chosen in this approach. DTW remains the utmost eminent procedure that remains utilized entirely aimed at computing comparison between two progressive series that might differ in speediness in time sequence investigation. This procedure computes a best counterpart between twofold time sequences and therefore may calculate the correspondence further precisely considering time periods.

DTW grounded fuzzy clustering aimed at time sequences information as well as anticipated three substitutions is deliberated by Izakian et al [4]. By means of extending or condensing sections of progressive information, their mechanisms demonstrate DTW which remains a desired select aimed at fuzzy clustering of time sequences. Nevertheless, in large-scale information handling, their investigation remains quite restricted. By means of the unceasing progress of science as well as technology, composed using continually growing of the scales of time sequences information, customary approaches uncover certain limitations:

- Time sequences information remains very large that it may not remain overloaded in memory by a instance in several circumstances,
- The information might constantly attain to facilitate, there exists no technique aimed at us towards acquiring the entire information by an instance.

Hence, clustering in place of large-scale time sequences information requires an increasing procedure, whose aim remains, specified a series of time sequences towards constructing a combination of decent dividers commencing the information flow by means of an insignificant quantity of memory as well as time.

Single-Pass FCM (spFCM) as well as Online FCM (oFCM) remains the twofold increasing fuzzy

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

clustering procedures anticipated by Hore et al [5]. Single-Pass strategies as well as online strategy are the twofold procedures signify dual application approaches aimed at increasing clustering correspondingly. Huge information is sort out piece by piece, and the earlier piece remains signified using the aforementioned centroids that would remain incorporated by the freshly approaching piece aimed at the subsequent turn of clustering in the previous approach. Every single piece remains categorized exclusively as well as characterized using its centroids, and at that time the entire centroids produced will remain clustered another time in the concluding approach. In handling large-scale information several investigations [6-7] have presented that together of the approaches remain operative.

Traditional incremental procedures meant for fuzzy co-clustering of co-occurrence environments are extended by Honda et al [6] and utilizing categorical multivariate information (FCCM) in addition to fuzzy CoDoK, Single-Pass or online schemes are used in such fuzzy clustering procedures. Mei et al [8] proposed twofold incremental clustering procedures to handle huge data groups that may not be suitable for complete memory. One of the methods is obtained by modifying the present FCM-built incremental clustering, whereas Single-Pass or Online technique using subjective fuzzy co-clustering is the additional incremental clustering technique. Single-Pass fuzzy c-means (spFCM-IB) along with online fuzzy c-means (oFCM-IB) [8] are the two incremental procedures on the basis of data blockage proposed in 2016 that would transform the conventional procedures in assuming numerous loads for each centroid.

Hadoop environment builds Map Reduce which is a parallel programming model [8]. The Mapper along with Reducer tasks may be modified to the user's necessity in order to the process of the information is included in Map Reduce. The data is forwarded towards the participant node for processing and within smaller chunks of information exchanging amid 16MB to 64MB, the input information is split. The processing in the Hadoop would take place in this manner. Using the Mapper function, the data is processed while the worker nodes would receive the portions of information through which the intermediate file i.e., List (key, value) is attained by processing the input file which consists of (Key, Value) pairs. The data can be reduced into desired outcomes using the Reducer function that has the previous intermediate result as the input by the detailed instructions provided by the planner. [9]

In market place inquiry, pattern detection, as well as image processing, has been analyzed regarding the clustering which is employed in the data extraction. While the information is independent of outliers as well as noise, one of the better performing and simple clustering procedures is K-Means. There would be a compromised result if the K-Means is influenced by the outliers that are included in the data. While measuring the performance, another better choice i.e., K-Medoids can be utilized which will not match with K-Means and is less affected by outliers and noise. Whenever the information set is ranging from small to medium, the performance of K-Medoids will get improved. The complication of computing is high whenever the size of the data set is large and becomes unsuitable. To discover the medoids [2], K-Medoids in addition to sampling method uses CLARA (Clustering Large Applications). However, depending on sampling, the medoids can be attained completely using this method and sometimes optimal clustering may be attained [9].

To obtain improved performance, K- Medoids might be implemented correspondingly which could be a far better resolution. When performing serially, the complication of K-Medoids is provided as $O(k*(n-k)^2)$where k represents the numerous clusters as well as n represents several data points and the procedure has been turned out to be expensive in computing, if the values k & n are considerable [4]. The complication may be decreased in the Hadoop by Map Reduce whenever the algorithm execution is achieved correspondingly in contrast [9].

As stated by A. Mohamed Ashik et al, K.

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

Senthamarai Kannan et al. [10] the Nifty 50 closing stock market prices were computed and predicted the trend of stock market fluctuations with the help of time series modeling methods such as exponential smoothing as well as autoregressive integrated moving average. The forecasted values of Nifty 50 closing stock price are computed for both models separately and also compared the error rates. From the results, the autoregressive integrated moving average model performed effectively compared to another models.

Alexander Vlasenko et al , Olena Vynokurova et al, Nataliia Vlasenko et al, Marta Peleshko et al [11] proposed a hybrid five-layer neuro-fuzzy pattern in addition to an equivalent learning procedure within stock market time-series estimation works using the application. In order to attain improved computational performance as well as representative capabilities during the reprocessing of the extremely non-linear variable information, Multi-dimensional Gaussian operations are employed rather than polynomial functions. As a result of a combination of the ability as well as strength in approximating and the computing efficiency as well as capability of acquiring the ANNs, Neurofuzzy models had become most popular.

According to Nakagawa et al, K., Imamura et al, M., & Yoshida, K. [12], the stock price fluctuation's model has employed the patterns which were never used in terms of an input feature of prediction within the financial market. This technique extracts the patterns of representative price fluctuations by the Clustering of k-Medoids using the procedure of Indexing Dynamic Time Warping. The application of k-medoids clustering is done for the dissimilarity matrix with the help of Indexing DTW (IDTW) for extracting the stock price fluctuation's representing models and to measure the distance of DTW (Dynamic Time Warping) within the indexed stock price fluctuations. The k-medoids clustering is proposed by this method using IDTW wherein the distance of DTW within the fluctuations of the indexed price is measured for implementing the k-medoids clustering in the dissimilarity matrix.

Pier Francesco Procacci et al and Tomaso Aste et al. [13] proposed a new technique for defining, analyzing as well as forecasting the market conditions. The reference sparse precision matrix in addition to the values of vector expectation is used to identify the conditions of the market. Every single multivariate observation and presented market condition are connected together in order to minimize the distance of the penalized Mahalanobis in this approach. The forecasting of the off-sample future market condition using essential prediction accurateness is done in this approach. The single observations are considered by this approach for proposing comparable Covariance centered Clustering.

Wang, W. et al & Mishra, K. K. et al [14] proposed a new recommendation system as well as stock trading prediction which are easily understandable methods. Initially, the raw time-series is transformed within the expressive as well as suggestive particles and the periods upon the basis of the information granule are achieved in the proposed system. Later, the fuzzy groups are defined and the historical information is fuzzified. Thirdly, fuzzy relations are constructed and weights are assigned. Lastly, the implementation of the prediction and recommendation is done.

Aboul Ella Hassanien et al, Mohamed F. Tolba et al, Khaled Shaalan et al and Ahmad Taher Azar et al [15] proposed an approach depending upon the usage of constructive features of the interval study in strengthening a FPN with a feasible pattern. Evaluating the behaviour of a thermoelectric cooler that is vulnerable for the two magnetic field effects in addition to the air convection vibrating stream at various input values of electrical current with the help of fuzzy logic device as well as the comparison of the investigational outcomes is the primary objective of this method.

All the existing techniques described in the literature are limited by the execution time when the input data is huge. The inherent nature of the C – Medoids algorithm allows for parallel execution of each cluster. Combining this feature with a parallel execution technique like mapreduce could improve

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

the performance of the proposed algorithm by dividing the steps and executing them in parallel. The following sections describes the proposed method where weighted fuzzy C Medoids is modified as per mapreduce algorithm to parallelize the weighted fuzzy C Medoids algorithm.

II. FCMDD AND WEIGHTED FCMDD

Please use automatic hyphenation and check your spelling. Additionally, FCMdd remains one of the demonstrative procedures of fuzzy clustering as stated in the previous section. The objective function is minimized using the formulae presented in equation (1):

$$J_{FCMdd} = \sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci}^{m} \, d(x_i, v_c) \quad (1)$$

Where $x_i$is the i-th object, d (x_i, v_c) remains the *Euclidean* distance between x_i as well as v_c, and m (m>=1) is the fuzzifier parameter.

FCMdd has to be transformed into a weighted procedure in increasing clustering efficiency that investigates weighted datasets having medoids and other objects. The objective function of weighted FCMdd (WFCMdd) to be minimized is given below:

$$J_{WFCMdd} = \sum_{c=1}^{C} \sum_{i=1}^{N} w_i u_{ci}^{m} \, d(x_i, v_c) \quad (2)$$

Where $w_i$is an optimistic real worth, connecting by every objective x_i. In the restriction situation$\sum_{c=1}^{c} u_{ci} = 1$, the value of u_{ci}can be computed as follows:

$$u_{ci} = [\sum_{l=1}^{c} (\frac{d(x_i, v_c)}{d(x_i, v_l)})^{\frac{1}{m-1}}]^{-1} \quad (3)$$

It remains critical for selecting the optimum objective as the medoid in FCMdd. Dependent on their participation towards the cluster; the mutual methodology remains to choose the objective that decreases its expanse by entire entities in the data groups [9]. Nevertheless, the time complication remains great. A linearization procedure that studies merely the q points, which enhance the participation to every cluster by means of medoid candidates, which is anticipated by Nasraoui et al [16]. Hence the medoid v_c of the cluster c remains deliberated as below:

$$v_c = min_{x \in \xi} \sum_{i=1}^{N} w_i u_{ci}^{m} d(x_i, x) \quad (4)$$

Where, ξ is the set of q medoid candidates.

Since they frequently reserve greatly additional data, dissimilar from usual objects, medoids would remain allocated complex loads in FCMdd. This weighted procedure remains broadly employed in the increasing clustering procedure as well as support to increase the presentation of our procedure in this broadsheet. The parameter description of the terms used in this section are listed in table 1.

TABLE 1
Parameter description

Notation	Description
C,N	Numbers of clusters, objects
uci	Fuzzy object partitioning membership
vc	Centroid/Medoid of the c-th cluster
m	FCM user-defined parameters
w	Weights of centroids/medoids and objects

III. DTW BASED INCREMENTAL FUZZY C MEDOIDS CLUSTERING WFCMDD-DTW

The objective utility of WFCMdd-DTW is:

$$J_{WFCMdd-DTW} = \sum_{c=1}^{C} \sum_{i=1}^{N} w_i u_{ci}^{m} \, dtw(x_i, v_c) \quad (5)$$

Where the function $dtw(x_i, v_c)$ remains the DTW distance amid time series x_i as well as the medoid v_c.

The significance of *dtw* ***(a, b)*** remains computed by the DTW procedure, assumed two time sequences **a** as well as **b**, using length S as well as T correspondingly. Every topic in a remains related by some topic in bin this procedure. Therefore, the identical shape lets commencing a as well as b will remain observed, though they might exist in dissimilar time intervals [17]. The pseudo-code aimed at computing DTW distance amid a as well as b is depicted in figure 1.

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

Permitting towards the procedure of WFCMdd, we can obtain the standards of u_{ci} and v_c as:

$$u_{ci} = [\sum_{l=1}^{c} (\frac{dtw\,(x_i, v_c)}{dtw\,(x_i, v_l)})^{\frac{1}{m-1}}]^{-1} \quad (6)$$

$$v_c = arg\ min_x \sum_{i=1}^{N} w_i u_{ci}^m d(x_i, x) \quad (7)$$

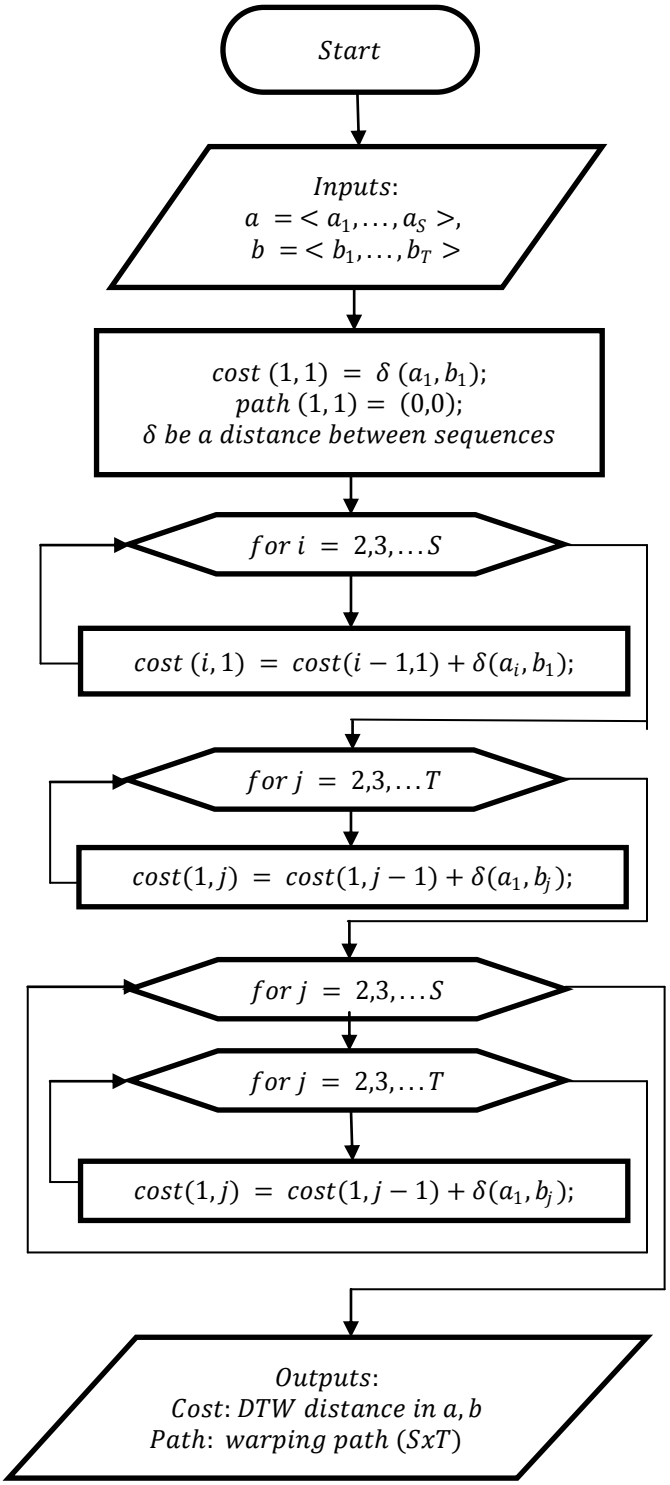

Fig 1: DTW flowchart

IV. MapReduce

MapReduce is an algorithm introduced to process large data sets in a parallel fashion. A large task is divided into programmable sub tasks where the algorithm can run each task in parallel with higher efficiency and reliability [18]. The algorithm is based on the concepts of functional programming where the technique of split-apply-combine. The best performance is obtained when the algorithm is implemented on multicore multiprocessor system.

MapReduce can be implemented on data present in the file system as well as large datasets. The algorithm, as the name implies, has two stages, a map stage and a reduce stage. The Map stage performs operations of filtering and sorting. The Reduce stage summarizes the results of the Map stage and produces the final result. This process is explained with the following functions.

- Input Preparation: The first step is to analyse the processors of the stem and assign keys to each of the processor. The required input is given in this stage.
- Map Function: A set of key pairs are given as input in this stage and output key pairs are produced as the output. For example, in an application of word counter, the input sentence is divided into words and each word is assigned a key pair.
- Shuffling Function: This function takes the key values and assigns them to different nodes to perform the next stage.
- Reduce Function: The reduction operation is performed on the data to get the final output

To following figure 2 summarizes the flow of Map reduce algorithm:

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

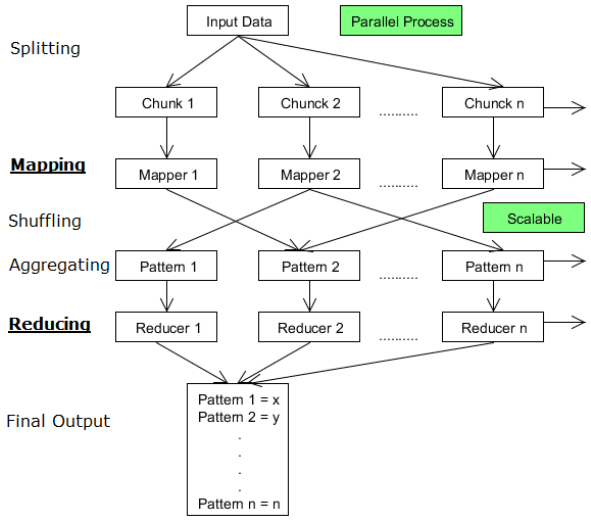

Fig 2: Map Reduce algorithm flow

In the above map reduce flow:

Algorithm 2
Step 1: The first step is to divide the data into n chunks based on the size of the input data.
Step 2: These chunks are passed to the mapping functions simultaneously.
Step 3: The data is then shuffled to aggregate related patterns.
Step 4: The reducers merge the results to get the combined output in the end logically.

V. Map Reduce with Weighted Fuzzy C Medoids Clustering

Please note that Map Reduce algorithm is used for processing large datasets in parallel. This algorithm becomes very useful when the programming can be executed on parallel programming platforms thus ensuring improved performance. Map Reduce relies on the concept of functional programming that would be making it possible to be extended to C-Medoids clustering by distributing the data on the different nodes then accumulate the result on to another node. This section describes how the two stages in Map Reduce, map and reduce, are modified according to weighted fuzzy C-medoid algorithm. The map stage calculates the global threshold value for clustering in one pass and then the data is passed to reduce stage after shuffling. The reduce stage processes the subsets by calculating the optimal local threshold value for clustering. The iteration between both these stages continuous till the global optimal threshold for clustering is discovered.

Map stage: As Map Reduce is designed for large datasets, Hadoop platform is used with HDFS which is relying on the key-value pair's sequence. Every entry is represented by <key, value> pairs. "key" denotes the offset from the commencement and the "value" represents the actual object value. In the map stage, the data is divided into different subsets. The global clustering result is optimized after every local clustering solution. The mapper function calculated the global clusters as follows:

A. Mapper Algorithm:

Input: N sets of objects in the input with k local medoids (n_1, n_2, n_3, . . . n_k)

Output: Global medoids (n_1', n_2', . . . n_k')

Step 1: Select m medoids form the input data randomly.

Step 2: Calculate the mean value of all the medoids n_1, n_2, n_3, . . . n_k.

Step 3: Find the medoids which are close to the mean value and output them as global medoids using DTW algorithm mentioned in section III.

Step 4: Repeat the steps 1 to 3 till the global mean value remains constant.

Reduce stage: In this phase, the local medoids are calculated by adding all the local data points. This generates new medoids which have been utilized as inputs in the subsequent iteration of map.

Reduce Algorithm:

Input: Local k medoids (n_1, n_2, … n_k) of i^{th} dataset portion.

Output: Recalculation of local k medoids (n_1', n_2', … n_k')

Step 1: In the i^{th} partition, minimize the WFCMdd objective function as stated in equation 2.

Step 2: Repeat this step for all the partitions obtained from map stage.

Step 3: Allotment of every object to the cluster that has the most comparable medoid.

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

VI. RESULTS

The reduce stage processes This section discusses the findings of implementing Incremental Map Reduce for weighted fuzzy C-Medoids Clustering of Big Time-Series Stock market data for prediction. The data set for the experimentation is chosen from quandl.com website.

Complexity Analysis

The traditional K-Means segmentation algorithm calculates the distances of each point form the centroid. However, the C-Medoids algorithm needs to find the similarity of data points with the medoids. Consider the number of data points as *'n'* and the number of clusters as *'k'*, then the complexity of the C-Medoids is given by *O(nk)*. In the proposed algorithm, the data points are divided into parallel chunks using Map Reduce algorithm. Considering *'m'* clusters in Map Reduce, the complexity of the proposed algorithm becomes *O(nk/m)*. This analysis proves that the proposed algorithm performs much faster the proposed technique. The performance analysis in the following section substantiates this theory with numerical analysis.

PERFORMANCE EVALUATION

In this section, we discuss about the performance evaluation of proposed approach i.e. fuzzy k-medoids based Map-reduce with traditionally developed fuzzy k-medoids in terms of analysis of stock market time series data with respect to accuracy in formation of attributes with similar relations using clustering based on k-means presentation. The environment used for the simulations are Intel core I5 with 8GB RAM, 1TB external with high speed configurations, and also the latest version of R-studio to process and evaluate stock market data with different notations.

The stock market time series data is downloaded from UCI repository. This repository is used to maintain different synthetic related data sets i.e. medical, sports and etc. The dataset contains different attributes like year and analysis of data with respect to closing values by different years. Here we discuss experimental results i.e. accuracy, sensitivity, specificity in formation of cluster with different attribute relations. Description of experimental results described as follows:

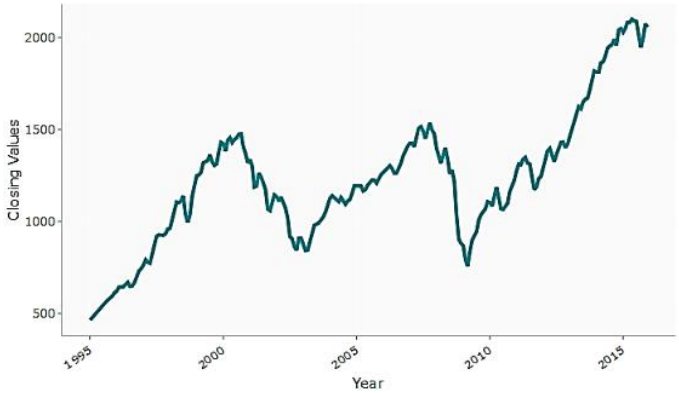

Fig 3: Plot of S&P 500 Time Series (1995 - 2015)

Figure 3 shows the analysis of time series in between years from 1995-2015 with different closing values. Closing values are represented in y-axis and year is represented in x-axis, from year 1995 closing values started from 500, if every year closing values are increased with different notations.

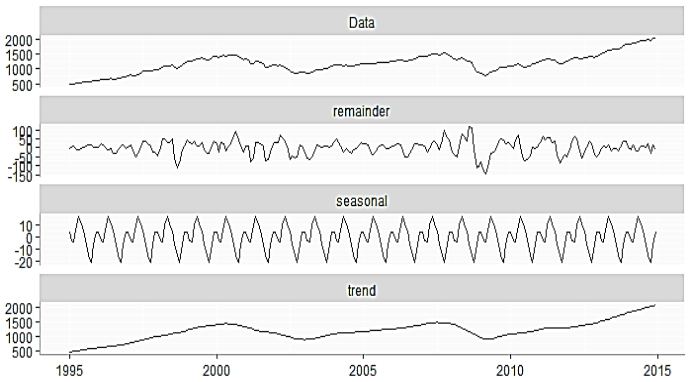

Fig 4: Decomposition plot S&P 500

Figure 4 describes the analysis of data when we use Map Reduce concept in processing attributes with respect to different notations i.e. data which describes stock market data with different attributes and evaluate the relation of remainder whenever if any closing values goes to negative value, also maintain the relation with seasonal attribute relation by trending data by year by year in parallel processing of all the attributes at a time.

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

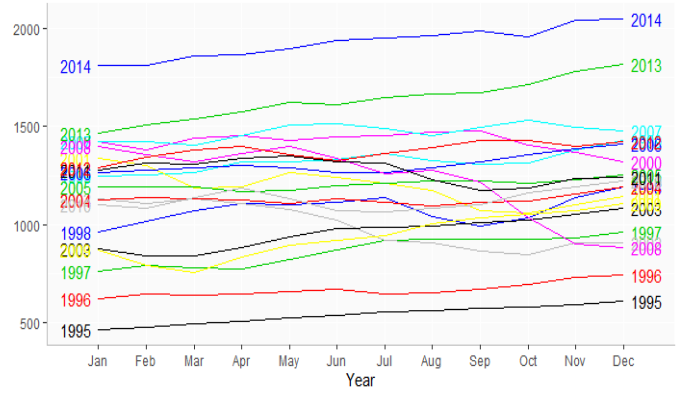

Fig 5: Seasonal plot of S&P 500

Figure 5 describes the seasonal plot which represents the relation between different years seasonal report by every month in every year, it represents from 1995-2000, it describes low closing values at middle stage i.e. from 2000-2012, it represents average seasonal rate and after 2014 it raises to high closing values with different attribute relations.

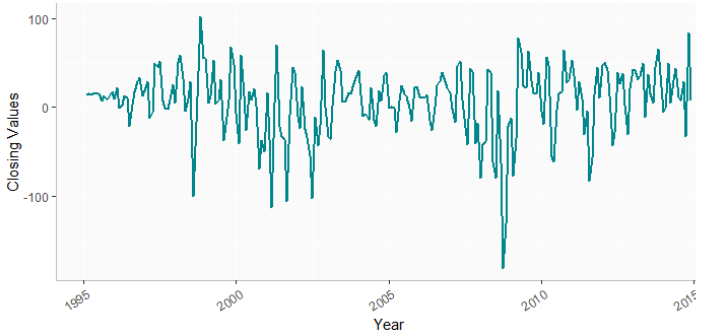

Fig 6: Plot of first difference time series (1995-2014)

Figure 6 represents the fluctuations appeared in between time series analysis with different years. Based on these attributes. The following graphs and figures show the calculation of accuracy, sensitivity, and specificity in processing attributes with different similarities in evaluation of relational time series data.

TABLE 2:
Accuracy

Data sets with different attributes by different years	Proposed Approach	Fuzzy k-medoids	naive	SVM based k-medoids
Dataset 1	92.1	81	71	70
Dataset 2	97.5	79	69	61
Dataset 3	93.4	75	60	74
Dataset 4	98.6	84	84	79
Dataset 5	98.4	94	87	68

Based on above accuracy values appeared from results, we plot the performance of proposed approach with respect to traditional approaches in processing attributes with analysis of stock market data in figure 7. Accuracy of proposed approach is better than existing methods in processing of attributes and formulate attributes by comparison of different attributes.

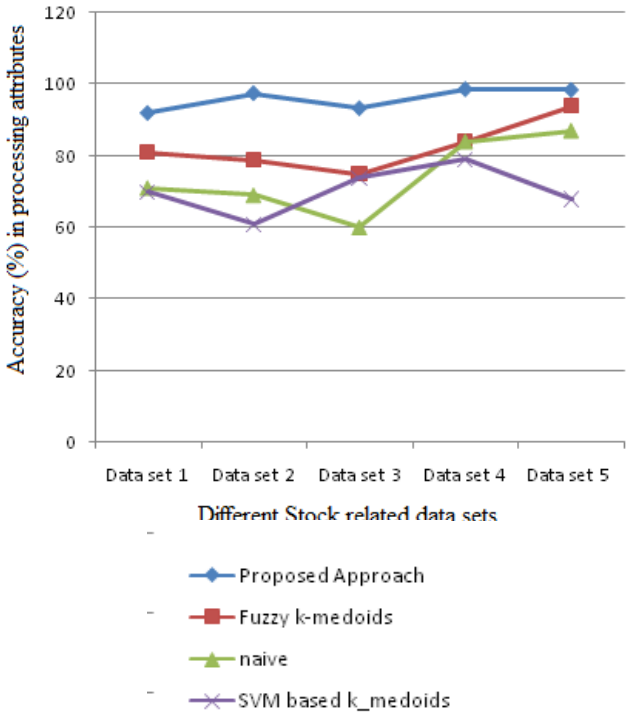

Fig 7: Different Accuracy values of proposed approach with different approaches.

TABLE 3:
Sensitivity

Data sets with different attributes by different years	Proposed Approach	Fuzzy k-medoids	naive	SVM based k-medoids
Dataset 1	0.56	0.352	0.45	0.37
Dataset 2	0.52	0.321	0.43	0.42
Dataset 3	0.572	0.40	0.31	0.38
Dataset 4	0.545	0.43	0.392	0.37
Dataset 5	0.64	0.353	0.34	0.35

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

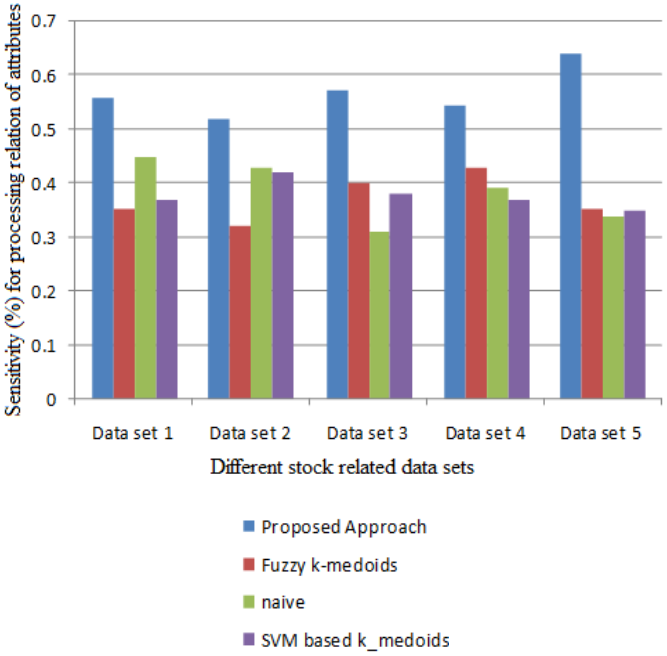

Fig 8: Sensitivity values of different approaches with different attribute relations

Performance evaluation of proposed approach with respect to different approaches in terms of sensitivity as shown in figure 8. High sensitivity implies that the performance of the approach has low false positive value in formation of attributes.

TABLE 4
Specificity

Data sets with different attributes by different years	Proposed Approach	Fuzzy k-medoids	naive	SVM based k_medoids
Dataset 1	0.51	0.541	0.61	0.71
Dataset 2	0.54	0.48	0.50	0.65
Dataset 3	0.62	0.51	0.46	0.69
Dataset 4	0.41	0.57	0.42	0.71
Dataset 5	0.39	0.42	0.41	0.72

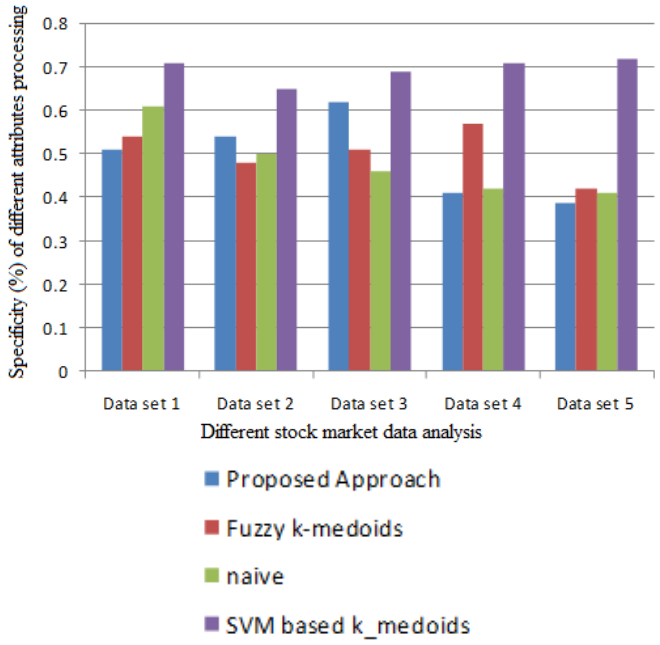

Fig 9: Different values of specificity in processing and evaluation of all the attributes based on different data sets.

Performance evaluation of specificity with respect to processing of different attributes and also describes the relation of each attribute, specificity of proposed approach is low because if specificity is low then performance of approach with parallel processing of all the attributes as shown in figure 9. Execution time for above mentioned methods described as follows:

TABLE 5
Time Comparison

Data sets with different attributes by different years	Proposed Approach	Fuzzy k-medoids	naive	SVM based k_medoids
Dataset 1	48	56	62	74
Dataset 2	54	48	57	65
Dataset 3	62	78	66	69
Dataset 4	41	57	67	73
Dataset 5	46	54	64	74

Different time comparison values with different approaches is shown in figure 10.

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

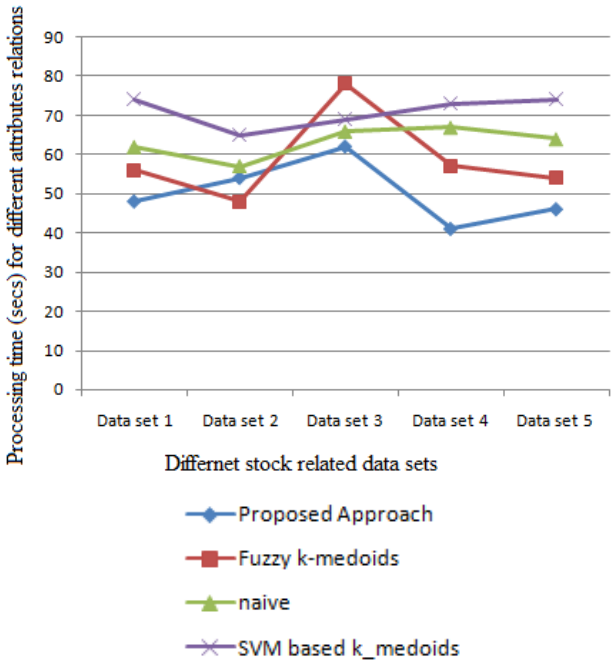

Fig 10: Performance evaluation of time in overall processing parallel attributes with different notations.

So finally based on above results, fuzzy k-medoids based Map Reduce approach gives better performance when compare to traditional approaches.

VII. Conclusion

Time series data clustering is a complex task as the input data to be clustered is very large. A conventional method used for this application is fuzzy C-Medoids. A major drawback of this method in the context is the time complexity. This paper presents an implementation of weighted fuzzy C-Medoids with MapReduce algorithm to cluster the time series data. The algorithm has proven to be faster than the conventional method. The experiments are performed on different dataset sizes to prove the efficiency of the algorithm. The proposed algorithm has operated as a speed times faster than the existing method.

VIII. References

[1] Poterba, James M., and Lawrence H. Summers. 1986. "The Persistence of Volatility and Stock Market Fluctuations." The American Economic Review, vol. 76, no. 5 (December):1142-51

[2] Liu Y, Wu S, Liu Z, Chao H (2017) A fuzzy co-clustering algorithm for biomedical data. Plos One 12.

[3] Hammouda KM, Kamel MS (2004) Efficient phrase-based document indexing for web document clustering. Ieee Transactions on Knowledge and Data Engineering 16: 1279-1296.

[4] Izakian H, Pedrycz W, Jamal I (2015) Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence 39: 235-244.

[5] Hore P, Hall LO, Goldgof DB, Gu Y, Maudsley AA, et al. (2009) A Scalable Framework For Segmenting Magnetic Resonance Images. Journal of Signal Processing Systems for Signal Image and Video Technology 54: 183-203.

[6] Honda K, Tanaka D, Notsu A. Incremental .algorithms for fuzzy co-clustering of very large co-occurrence matrix. IEEE International Conference on Fuzzy Systems; 2014; Beijing, China. Institute of Electrical and Electronics Engineers Inc. pp. 2494-2499.

[7] Mei J-P, Wang Y, Chen L, Miao C. Incremental fuzzy clustering for document categorization. IEEE International Conference on Fuzzy Systems; 2014; Beijing, China. Institute of Electrical and Electronics Engineers Inc. pp. 1518-1525.

[8] Liu Y, Wan X (2016) Information bottleneck based incremental fuzzy clustering for large biomedical data. Journal of Biomedical Informatics 62: 48±58. https://doi.org/10.1016/j.jbi.2016.05.009 PMID:27260783.

[9] Labroche N. New incremental fuzzy C medoids clustering algorithms. Annual Conference of the North American Fuzzy Information Processing Society ĐNAFIPS; 2010; Toronto, ON, Canada. Institute of Electrical and Electronics Engineers Inc. pp. Ryerson University; The Institute of

January - February 2020
ISSN: 0193 - 4120 Page No. 14496 - 14507

Electrical and Electronic Engineers (IEEE); University of Waterloo.

[10]Mohamed Ashik A, Senthamarai Kannan K. Time Series Model for National Stock Price Prediction. Research & Reviews: Journal of Statistics. 2018; 7(1): 85s–90sp.

[11]A. Vlasenko, O. Vynokurova, N. Vlasenko, M. Peleshko, A Hybrid Neuro-Fuzzy Model for Stock Market Time-Series Prediction", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP).

[12]Nakagawa, K., Imamura, M., and Yoshida, K. 2019. Stock price prediction using k- medoids clustering with indexing dynamic time warping. Electronics and Communications in Japan, 102(2), 3-8.

[13]P. F. Procacci, T. Aste, "Forecasting market states", Quantitative Finance, vol. 19, no. 9, pp. 1491-1498, July 2019.

[14]Wang W, Mishra KK (2018) A novel stock trading prediction and recommendation system. Multimed Tools Appl 77(4):4203–4215.

[15]Alqaryouti O., Farouk T., Siyam N. (2019) Clustering Stock Markets for Balanced Portfolio Construction. In: Hassanien A., Tolba M., Shaalan K., Azar A. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018. AISI 2018. Advances in Intelligent Systems and Computing, vol 845. Springer, Cham.

[16]Nasraoui O, Krishnapuram R, Joshi A, Kamdar T (2002) Automatic Web User Profiling and Personalization Using Robust Fuzzy Relational Clustering: Physica-Verlag HD. 233-261 p.

[17]Y. Liu, J. Chen, S. Wu, Z. Liu, and H. Chao, "Incremental fuzzy C medoids clustering of time series data using dynamic time warping distance," PLoS One, vol. 13, no. 5, Article ID e0197499, 2018.

[18]J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.