GENETIC ALGORITHM BASED HYBRID APPROACH FOR CLUSTERING TIME SERIES FINANCIAL DATA

Dr.Chandrika.J[1], Dr.B.Ramesh[2] , Dr.K.R.Ananda kumar[3] and Raina.D.Cunha[4]

[1]Dept of CS & E, M C E,Hassan, Karnataka
jc@mcehassan.ac.in
[2]Dept of CS & E, M C E, Hassan, Karnataka
br@mcehassan.ac.in
[3]Dept of CS & E, SJBIT, Bangalore, Karnataka
kra_megha_tn @hotmail.com
[4]Infosys Technology, Mysore
raina.Dcunha@gmail.com

ABSTRACT

Stock market data is a high dimensional time series financial data that poses unique computational challenges. Stock data is variable in terms of time, predicting the future trend of the prices is a challenging task. The factors that influence the predictability of stock data cannot be judged as the same factors may or may not influence the value of the stock all the time. We propose a data mining approach for the prediction of the movement of stock market. It includes using the genetic algorithm for pre processing and a hybrid clustering approach of Hierarchical clustering and Fuzzy C-Means for clustering. The genetic algorithm helps in dimensionality reduction and clustering helps to create feature vectors that help in prediction.

KEYWORDS

Time series data, Genetic Algorithm, Clustering, Stock market prediction, fuzzy C Means.

1. INTRODUCTION

Time series refers to a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, and communications engineering. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series data have a natural temporal ordering.

Stock market prediction [1] is the act of trying to determine the future value of a company stock or a financial instrument traded on a financial exchange. Accurate predictions can help stock holders to invest further so as to gain profits from their investment, or sell their shares if there is a fall of market value. If more people want to buy a stock (demand) than sell it (supply), then the

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 39–52, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4806

Price moves up. Conversely, if more people wanted to sell a stock than buy it, there would be greater supply than demand, and the price would fall. Hence Stock prices change every day because of market forces such as “supply and demand”. Therefore stock data is a time series data.

Stock market prediction is one of the most challenging tasks. Many approaches and studies have been undertaken to understand the attributes that influence the stock market[1] [2]. Accurate predictions can help stock holders to invest further so as to gain profits from their investment, or sell their shares if there is a fall of market value. Since the stock market has a random behavior, accuracy of the prediction matters most for the analysts. Although, there cannot be hundred Percent accurate predictions, one can get the knowledge about the rise and fall. Stock prices change every day because of market forces such as “supply and demand”. It is easier to predict the short term price movements than the other sectors of long term market [3]. All the factors that influence the stock price are not specifically known but, to some extent the market value of short term stocks is usually influenced by structured data(price, trading volumes, accounting items) and unstructured data(financial news from newspapers, articles or internet). It is believed that it is not easy to predict how stock prices change, while certain statistical techniques on historical stock data can help to determine whether to buy or sell stock. But, stocks are volatile and can change in price rapidly [4].

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

The Knowledge Discovery in Databases (KDD) process is commonly defined with the following stages:

(1) Selection

(2) Pre-processing

(3) Transformation

(4) Data Mining

(5) Interpretation/Evaluation.

Selection refers to the collection of data required for the problem at hand. Data collected may be categorical data, numerical data, spatial data or temporal data. The raw data collected may be of high dimension, redundant, irrelevant, and may be prone to noise.

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Transformation of data includes dimensional reduction techniques like feature selection and feature extraction. How many features and what kind of features should be used, can be a difficult problem. There is much noise and redundancy in most high dimensionality, complex patterns. Therefore, it is sometimes difficult even for experts to determine a minimum or optimum feature set. The objective of these approaches is to find a reduced subset among the original N features such that useful class discriminatory information is included and redundant class information and/or noise is excluded.

Feature Selection is the task of finding the "best" subset of features from the initial 'N' features in the data pattern space. Feature Extraction defines a transformation from pattern space to feature space such that the new feature set used gives both better separation of pattern classes and reduces dimensionality. Thus feature extraction is a kind of feature selection, but also includes a space transformation. Feature extraction is a superset of feature selection; feature selection is a special case of feature extraction (feature extraction with the identity transformation).

Data mining uses two main core tasks clustering and classification. Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach is an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments.

Classification is similar to clustering in that it also segments records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For example, classes can be defined to represent the likelihood that a customer defaults on a loan (Yes/No). It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified.

Interpretation and Evaluation refers to the tasks of validating the results obtained through the data mining tasks.

In this paper we propose the use of both data mining techniques clustering and classification for predicting the rise or fall of stock data. The rest of the paper is organized as follows. Section 2 outlines the related work in this area. Section 3 gives a detailed account of proposed algorithm. Section 4 depicts the experimental results. Section 5 concludes the paper with direction for future enhancements.

2. RELATED WORK

Brown et. Al and Jennings et. Al [16] [17] proposed two types of stock market analysis. First, the fundamental analysis derives stock price movements from financial ratios, earnings, and management effectiveness. Second, the technical analysis identifies the trends of stock prices and trading volumes based on historical prices and volumes.

There are no specific ratios that contribute to the prediction. Various ratios calculated on the stock data are present that is derived by each analyst according to his or her observation. Hence using only limited ratios may omit an important factor that implies the most on the prediction. In the project, the historical data along with the earnings and trading volumes is considered.

Lin et al. [18] have proposed a method based on structured data such as price, trading volume and accounting items for stock market prediction. However, it is much more difficult to predict stock price movements based on unstructured textual data such as financial news published on the newspapers or Internet.

The textual data in the form of financial news and stock quotes require very high effort of fetching for some terms that are likely to be used by the companies in their documentation. After the searching is done the frequency is to be maintained and the contribution level of each word to the outcome is to be considered. This is very tedious and hard to implement. Hence historical structured data which is primarily made up of categorical data is considered.

Schumaker et.al [19] used news articles to predict stock prices. Another kind of unstructured textual data is gathered from financial reports, which contain not only textual data but also numerical data. The numerical data provides quantitative information and the textual data contains a large amount of qualitative information related to the company performance and future financial movements.

The research of Kogan et al. [20] explains that using the quantitative and qualitative information can improve the prediction accuracy. The words that are to be searched in the document are not standard and are left to the user or analyzer's choice and observations.

Though the combined effect of numerical and textual information provides more accuracy of prediction, the difficulty of implementation is observed and not included.

Genetic algorithm is used for feature selection and classification by Pie et.al [5]. Two approaches are elaborated, where Genetic algorithm is combined with the KNearest - Neighbor decision rule (GA/KNN) and a production decision rule (GA/RULE). The computational cost of the GA/KNN method was very high and required parallel or distributed processing to be attractive for large, high-dimensionality problems. As an Improvement, GA/RULE was used. The objective of the GA/RULE approach was to find an optimal transformation that yields both the lowest error and smallest feature set. The results of experiments proved that GA/RULE required substantially fewer computation cycles to achieve answers of similar quality as that of GA/KNN. The test results obtained showed that GA/RULE outperformed the standard KNN method in every case, and its performance approached the hybrid GA/KNN method in most cases. The problems faced where that the sample (training) dataset needs to be representative, and it must also be large enough to allow for effective training. Otherwise, it will allow 'false' rules to be induced.

Since stock data is of high dimension, the GA/RULE method can be used for feature selection and extraction. Since the GA/RULE works best on binary data, an approach to convert the categorical data to binary is to be considered.

Many stock prediction methods based on SVM have been proposed in [21]. The SVM-based predictive models are developed with different feature selection methods from ten years of annual reports. The results showed that document frequency threshold is efficient in reducing feature space while maintaining the same classification accuracy compared with other feature selection methods. Furthermore, the results showed the feasibility of using text classification on current year's annual reports to predict next year's company financial performance, namely the return on equity ratio.

A clustering approach for stock market prediction [6] was experimented with a hybrid approach using Hierarchical Agglomerative clustering and the K-means algorithm which was named as HRK. Both numerical and textual information from historical financial news and stock quotes were considered. The proposed method consists of three phases. First, each financial report was

converted into a feature vector and the hierarchical agglomerative clustering method was used to divide the converted feature vectors into clusters. Second, for each cluster, K-means clustering method was applied recursively to partition each cluster into sub-clusters so that most feature vectors in each sub cluster belong to the same class. Then, for each sub-cluster, the centroid was chosen as the representative feature vector. Finally, the representative feature vectors were employed to predict the stock price movements.The experimental results showed that it outperformed SVM in terms of accuracy and average profits. The total average profit of 10 industry sectors of HRK is 3.95%, while the total average profit of SVM is 1.46%. The HRK outperformed SVM to produce 73% accuracy.

From the above survey conducted, we drove to the conclusion of using only the categorical data from the historical stock data. A model is to be developed where the production rule based system (GA/RULE) can be used for finding the best rules to predict the stock price movements. The rules generated can be further combined with genetic algorithm for dimensionality reduction of the historical data considered. After the dimensional reduction, the data is given to a hybrid clustering approach that uses hierarchical agglomerative clustering and fuzzy C-Means or Hierarchical Agglomerative Clustering and K-Medoids clustering. Two hybrid approaches are used for comparative analysis. And the end result would be the accuracy of the prediction.

3. PROPOSED METHOD

The proposed work works in three phases – Data collection, Preprocessing, Dimensionality reduction and Clustering. A detailed outline of these phases is given below.

3.1 Data Collection

A company's financial data of five years is considered.[7] The five years data is divided into training data set which is of three years and testing data set which is of two years. The data should have a predefined class label for each row. '1' indicates that there is a rise and '0' indicates fall of stock value.

3.2 Pre-Processing

The training data set is pre processed to remove the row and column heading that is the date and the attribute names. Columns having empty value for each row are removed. The copy of this data is saved for future use. The mean obtained by each column is subtracted with each row of column entries. If the value obtained after subtraction is between the range +1 and -1 a value '1' is assigned if not '0' is assigned for that field. The fields are converted to 1's and 0's specifically because the data is given as an input to the genetic algorithm. The genetic algorithm works best for binary attributes. [5]

3.3 Dimensionality reduction

Genetic Algorithm is used both for classification and Feature Extraction. The production decision rule based approach is used. [5].

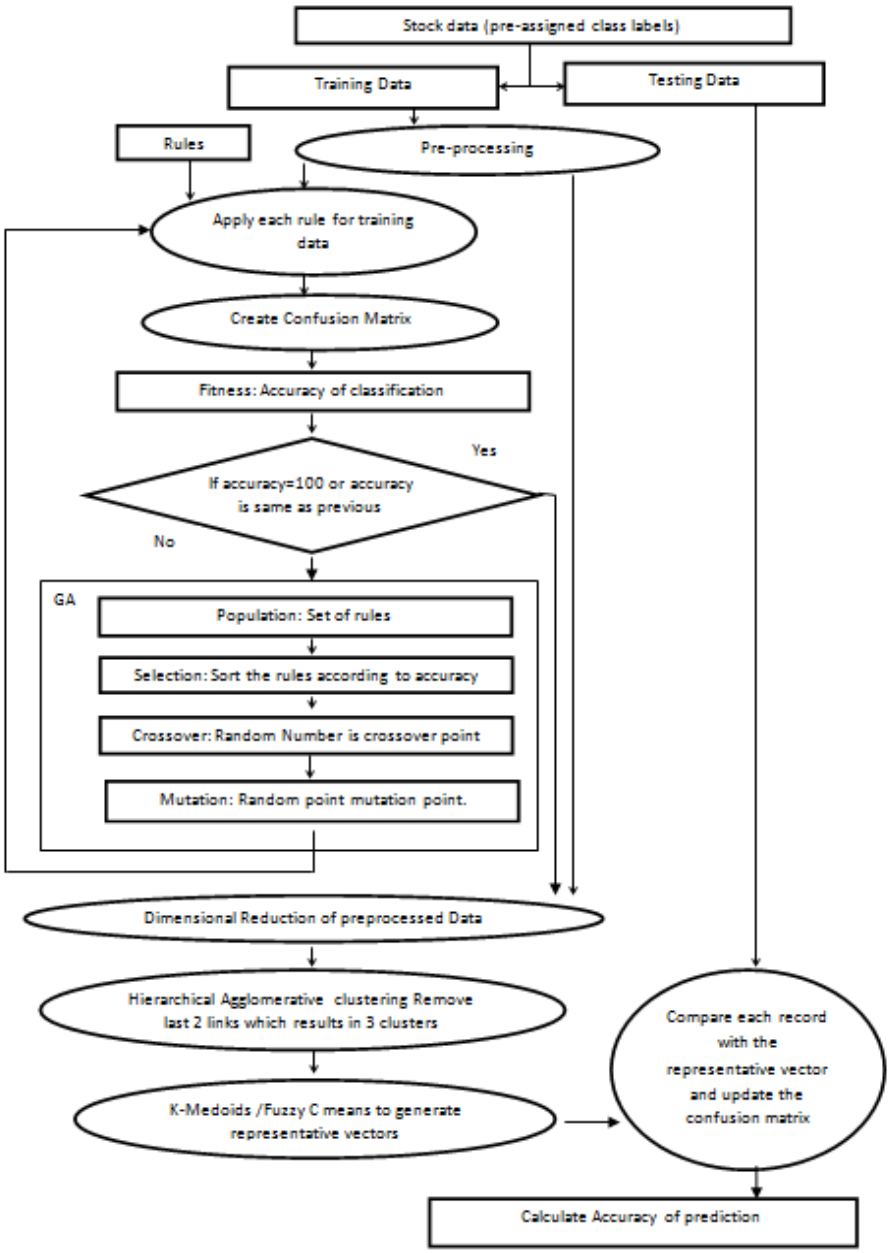

Figure 1. Outline of the model.

Genetic Algorithm (GA): Genetic algorithms imitate the evolution of the living beings, described by Charles Darwin [8]. GA is a part of the group of Evolutionary Algorithms (EA). Genetic Algorithm works with a set of individuals, representing possible solutions of the task. The selection principle is applied by using a criterion, giving an evaluation for the individual with respect to the desired solution. The best-suited individuals create the next generation. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures as to preserve critical information. Genetic algorithms although are randomized, use historical information to find an optimal solution within the search space. The genetic algorithm is as below[9]:

1. **[Start]** Generate random population of n chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness f(x) of each chromosome x in the population
3. **[New population]** Create a new population by repeating following steps until the new population is complete
    a. **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
    b. **[Crossover]** With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
    c. **[Mutation]** With a mutation probability mutate new offspring at each locus (position in chromosome).
    d. **[Accepting]** Place new offspring in the new population
4. **[Replace]** Use new generated population for a further run of the algorithm
5. **[Test]** if the end conditions are satisfied, stop, and return the best solution in current population
6. [Loop] **Go to step 2**

A solution generated by genetic algorithm is called a chromosome, while collection of chromosome is referred as a population [10]. The implementation of a genetic algorithm has random chromosomes called population as input. The fitness function is applied on the chromosomes to measure the suitability of solution, more suitability gives more reproductive opportunities. Some chromosomes in population will mate through process called crossover thus producing new chromosomes named offspring which its genes composition are the combination of their parent. Mutation means random change of the value of a gene in the population. In a generation, a few chromosomes will also undergo mutation in their gene. The number of chromosomes which will undergo crossover and mutation is controlled by crossover rate and mutation rate value. Chromosomes for the next generation will be selected based on Darwinian evolution rule [11], the chromosome that achieve a better solution to the problem are given more chances to reproduce than those which are poorer. Therefore the solution is typically based on the current population. After several generations, the chromosome value will converges to a certain value which is the best solution for the problem.

Applying Genetic Algorithm - The inputs for this phase are the set of initial rules and the preprocessed data in binary. The rules have values '0' or '1' or '2' for each field and the last column of the rule indicates the class label for which the rule is defined. Then the genetic algorithm is executed with the following attributes.

Initial population: The set of initial rules represented as bit vectors.

Fitness Function: Accuracy of classification for each rule. It is based on the number of test records predicted correctly by the classification model. The frequency of incorrect and correct predictions is recorded in a table called confusion matrix.

If f_{ij} indicates the number of records of class 'i' predicted as the records of class 'j', then f_{11} and f_{00} are the number of correct predictions while f_{10} and f_{01} are the number of incorrect predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$Accuracy = \frac{f11 + f00}{f00 + f11 + f01 + f10}$$

A classification model should always try to attain higher accuracy.

Selection: The genetic algorithm arranges the rules based on this fitness function i.e. the rule having the highest accuracy is at the top. Then the genetic algorithm selects the top two rules to perform crossover and mutation.

Crossover: Crossover takes place on the selected rules at a randomly generated point.

Mutation: The mutation rate considered is 2%. The mutation occurs at a randomly generated point.

Population: The rules generated after crossover and mutation are the population for the next iteration of the genetic algorithm.

Stopping Condition: The algorithm is executed until, the accuracy reaches 100 or the genetic algorithm executes for 1000 generations.

The genetic algorithm executes repeatedly to generate the best set of rules for predicting stock market data. The rules that are generated by the genetic algorithm are analyzed. If a column has a value 2 for each rule that column in the saved training data is deleted. Because a value 2 indicates that the value in that field doesn't contribute to the outcome. Hence we get a data set with reduced dimensionality.

3.4 . Clustering

The dimensional reduced data is given for the clustering process. We propose a hybrid clustering mechanism, which includes Hierarchical Agglomerative Clustering and Fuzzy-C-Means Clustering or Hierarchical Agglomerative Clustering and K-Medoids clustering. A comparative analysis is to be performed using Fuzzy C-Means and K-medoids.

1) *Hierarchical Agglomerative Clustering:* Hierarchical agglomerative clustering [12] or HAC is a bottom-up hierarchical clustering technique. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. An HAC clustering is typically visualized as a dendrogram. Each merge is represented by a horizontal line. The y-coordinate of the horizontal line is the similarity of the two clusters that were merged, where documents are viewed as singleton clusters. Hierarchical clustering does not require a pre-specified number of clusters.

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Step 3 can be done in different ways, single-linkage, complete-linkage and average-linkage clustering.

In single-linkage clustering (also called the connectedness or minimum method), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, we

consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

In complete-linkage clustering (also called the diameter or maximum method), we consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.

In average-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

Fuzzy C-Means*:* FCM algorithm is one of the most important fuzzy clustering methods, initially proposed by Dunn, and then generalized by Bezdek [13]. The Fuzzy C-means clustering algorithm is a variation of the K-means clustering algorithm, in which a degree of membership of clusters is incorporated for each data point. FCM algorithm is a technique of clustering which permits one piece of data to belong to two or more clusters [14]. The aim of the FCM algorithms is to assign the data points into clusters with varying degrees of membership values. Membership values lie between 0 and 1. This membership value reflects the degree to which the point is more representative of one cluster than the other. The centroids of the clusters are computed based on the degree of memberships as well as data points. The algorithm consists of the following steps:

1. Let us suppose that M-dimensional N data points represented by x_i (i = 1, 2. . . N) are to be clustered.
2. Assume the number of clusters to be made, that is, C, where $2 \leq C \leq N$.
3. Choose an appropriate level of cluster fuzziness f >1.
4. Initialize the N × C × M sized membership matrix U, at random, such that $U_{ijm} \in [0,1]$ and $\sum_{j=1}^{c} U_{ijm}$=1.0, for each i and a fixed value of m.
5. Determine the cluster centers CC_{jm}, for j cluster and its m dimension by using the expression given below:

$$CC_{\text{jm}}=\frac{\sum_{i=1}^{N} U_{ijm}^{f} x_{im}}{\sum_{i=1}^{N} U_{ijm}^{f}}$$

6. Calculate the Euclidean distance between i^{th} data point and j^{th} cluster center with respect to say m^{th} dimension like the following:

$$D_{ijm}=\left\|\left(x_{im}-CC_{jm}\right)\right\|$$

7. Update fuzzy membership matrix U according to D_{ijm}. If $D_{ijm} > 0$, then

$$CC_{ijm}=\frac{1}{\sum_{c=1}^{C}\left(\frac{D_{ijm}}{D_{icm}}\right)^{\frac{2}{f-1}}}$$

   If D_{ijm}= 0, then the data point coincides with the corresponding data point of j^{th} cluster center C_{jm} and it has the full membership value, that is, U_{ijm}= 1.0.
8. Repeat from Step 5 to Step 7 until the changes in the U ≤ ∈, where ∈ is a pre-specified termination criterion.

K-Medoids Clustering: The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoidshift algorithm [15]. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. It is more robust to

noise and outliers as compared to k-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances.

A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the cluster. The algorithm can be given as:

1. Initialize: randomly select (without replacement) k of the n data points as the medoids
2. Associate each data point to the closest medoid.
   ("closest" here is defined using any valid distance metric, most commonly Euclidean distance, Manhattan distance or Minkowski distance)
3. For each medoid m
4. For each non-medoid data point o
   a. Swap m and o and compute the total cost of the configuration
5. Select the configuration with the lowest cost.
6. Repeat steps 2 to 4 until there is no change in the medoid.

The dimensional reduced data is given to the hierarchical clustering algorithm. The last two links are removed to generate three clusters of the data. The Euclidean distance is used as a proximity measure. It is applied to two data points which lie in 1, 2, 3 or higher dimensional space. It is calculated as

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Where, n is the number of dimensions, x_k is the value of the data object x for dimension k, y_k is the value of the data object y for dimension k.

Each cluster is then given to the Fuzzy C-Means or the K-medoids algorithm to perform sub clustering. The Euclidean distance is used as a proximity measure. The clustering is performed until each cluster has a purity of 1. The centroid of each cluster along with the class label is transformed to a representative vector.

3.5 Prediction

The testing data is compared to each representative vector to find to which centroid the testing record is the closest. Then the class label of the representative vector is used as the predicted value for the testing data. This is done for each testing record. Since the testing data has a pre assigned class label, the accuracy of prediction is calculated.

4. EXPERIMENTATION

The proposed algorithm is implemented in MATLAB R2009b version. It provides plenty of user interface controls for creating a dynamic User Interface. It has several inbuilt functions that aids in the efficient implementation of proposed algorithm. The training data set, the set of rules and the testing data set are all stored as files. The software provides the flexibility of reading and displaying the file with numerical data or strings (rules).The training data is stored as a matrix format. The Rules are stored in a list and each rule is considered as a string. The software provides various internal functions for data type conversions and storing the data. For pre processing the column headers and the first column is deleted just by specifying the row and column index. To check for empty values an inbuilt function ‘isnan’ is used. The ‘fcm’ and

'kmeans' functions are already inbuilt that correspond to Fuzzy C-Means and K-Means (if the attributes are greater than 2 it works as K-Medoids) algorithms. The representative vectors and the testing data are stored in matrix format. The accuracy obtained by the genetic algorithm for feature extraction is 87%. The purity of each sub cluster obtained is 1. The accuracy of prediction obtained is 88%. Hence the model developed is better than the models developed previously.

Figure 2 shows the percentage of feature reduction obtained through the use of Genetic algorithm. The reduction rate of feature extraction obtained is 40%.

Since the accuracy of prediction obtained by both the algorithms is same, it is essential to observe the execution time taken by the hybrid algorithm using both the clustering approaches that is K-Medoids and Fuzzy C-Means. Figure 3 shows that Fuzzy C-Means takes more time to execute than K-Mediods as the number of clusters increase.

The accuracy obtained by our model is compared to the accuracy obtained from the existing model. From Fig 5 and Fig 4 we see that the HRK approach has achieved an accuracy of 73% while HAC with K-Medoids or HAC with Fuzzy C-Means has achieved 88% accuracy.

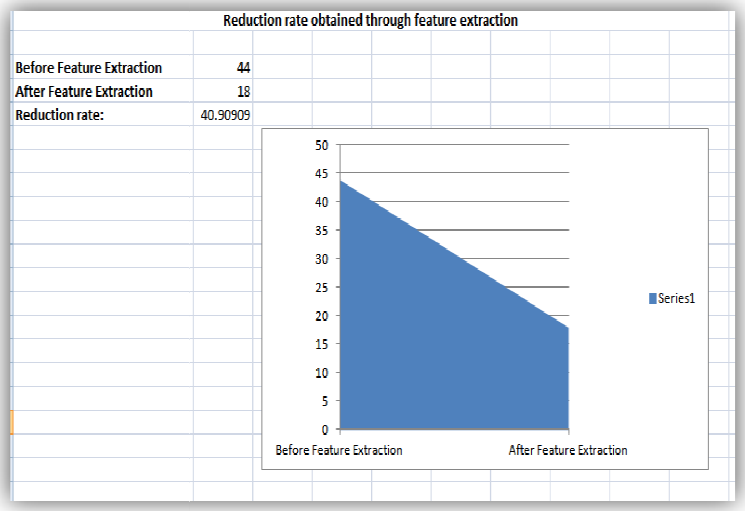

Figure 2. Reduction rate through Feature Extraction

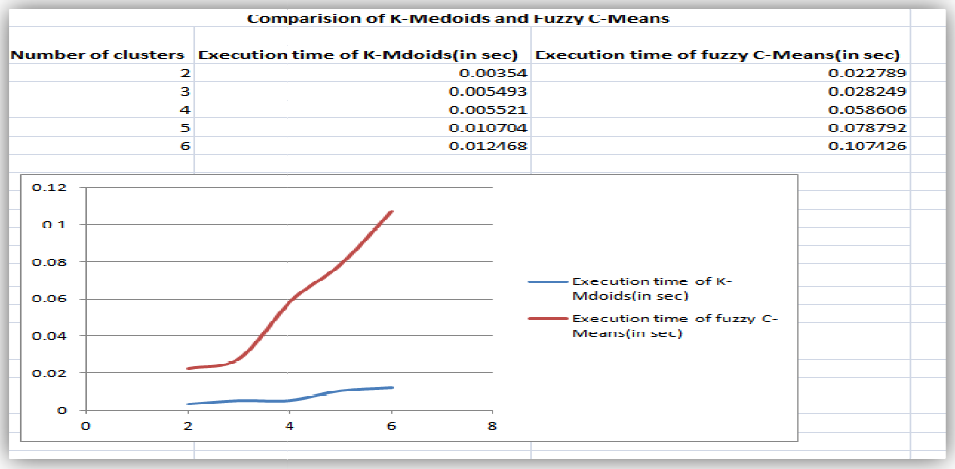

Figure 3. Comparison of fuzzy C-Means and K-Medoids

Figure 4. Accuracy achieved by existing Algorithms

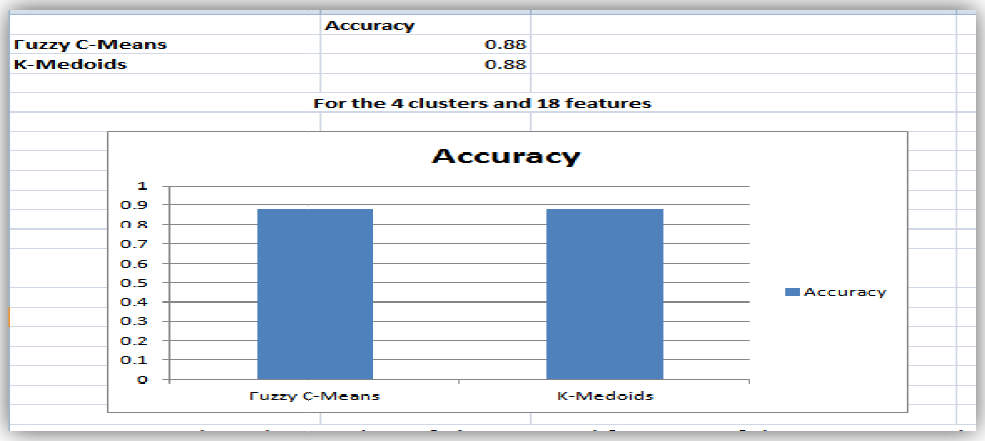

Figure 5 Accuracy obtained by proposed algorithms

5. CONCLUSION AND FUTURE ENHANCEMENTS

Data mining is found to be a useful domain that can be used to predict the stock price value and to build a structural model to predict accurate results. An efficient prediction can help the investors to gain a huge amount of profit. It is experimented that combining the methods of preprocessing, classification and clustering gives more accurate results than the other methods proposed to date. The model developed is highly dependent on the initial rules used. Even though the genetic algorithm is used for efficient processing of the rules, if the initial rules do not even provide half of the accuracy of classification, the genetic algorithm will evolve generating inefficient algorithms for feature extraction. From the experimental results conducted we found that fuzzy C-Means and K-Medoids both achieved an accuracy of 88% when using the same data set, rule set and the same number of clusters. But the execution time Fuzzy C-Means is greater than that of K-Medoids. Hence we conclude that K-Medoids is a better clustering algorithm for the model developed.

The future enhancements to this work would be considering categorical data along with numerical data. Instead of building a model for a particular company, it would be efficient to construct a general model so that even long-term company price movements can be predicted.

REFERENCES

[1] S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S., Jimoh, R. G (July 2011), "Stock Trend Prediction Using Regression Analysis –A Data Mining Approach", ISSN 2222-9833 ARPN Journal of Systems and Software.

[2] Clive W.J. G-anger, (1992),"Forecasting stock market prices: Lessons for forecasters", International Journal of Forecasting .

[3] http://www.indianstocktimes.com/study-zone.php

[4] Hongxing He,Jie Chen,Huidong Jin,Shuheng Chen(2006),"Stock Trend Analysis and Trading Strategy", ,jcis ,Taiwan 2006, DOI: 10.2991/jcis.2006.

[5] Min Pei, Erik D. Goodman, William F. Punch III and Ying Ding, (1995),"Genetic Algorithms For Classification and Feature Extraction", Michigan State University, Genetic Algorithms Research and Applications Group (GARAGe),CSNA-95.

[6] M.Suresh Babu, Dr. N.Geethanjali, Prof B.Satyanarayana,, (2012), "Clustering Approach to Stock Market Prediction", Int. J. Advanced Networking and Applications Volume: 03, Issue: 04, Pages:1281-1291.

[7] http://finance.yahoo.com/q/hp?s=YHOO

[8] http://www.ro.feri.unimb.si/predmeti/int_reg/Predavanja/Eng/3.Genetic%20algorithm/_25.html

[9] http://en.wikipedia.org/wiki/Genetic_algorithm.

[10] Tom V. Mathew, "Genetic Algorithm", Indian Institute of Technology Bombay, Mumbai, available at http://www.civil.iitb.ac.in/tvm/2701_dga/2701-ga-notes/gadoc.pdf

[11] Ganesh Bonde, Rasheed Khaled ,(2012),"Stock price prediction using genetic algorithms and evolution Strategies", ,Intl. conference on artificial intelligence(ICAI2012), Las vegas.

[12] http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html

[13] N.R, Pal K, Keller J.M. and Bezdek J.C, (2005),"A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530.

[14] A.vathy-Fogarassy, B.Feil, J.Abonyi (2005),"Minimal Spanning Tree based Fuzzy clustering", Proceedings of World academy of Sc., Eng & Technology, vol-8.

[15] http://en.wikipedia.org/wiki/K-medoids

[16] David.P.Brown, Robert.H.Jennings (1989),"On Technical Analysis, in Review of financial studies", vol.2, issue 4,pp 527-557.

[17] Jeffery.S.Ababanell and Brain.J Bushee, (1998),"Abnormal returns to fundamental analysis strategy", The Accounting Review, Vol.73.

[18] ",L.Lin,Ren.R.E,D.Sornette,"Consistent model of explosive financial bubbles with mean reversing residuals", arXiv:0905.0128

[19] Robert.P.schumaker,Hisinchun chen,(2009), "Textual analysis of stock market prediction using breaking financial news: The AZFin text system, ACM transactions on information systems,vol. 27, issue 2.

[20] Shimon Kogan,Dimitry Levin,Bryan R.Routledge, Jacob.S.Sagi,Noah.A.Smith,(2009), "Predicting risks from financial reports with regression", Available at http://svmlight.joachims.org.

[21] TAY, Francis E. H. and Lijuan CAO, (2001),"Application of support vector machines in financial time series forecasting", Omega: The International Journal of Management Science, Volume 29, Issue 4, Pages 309-317

[22] Yuling LIN, Haixiang GUO and Jinglu HU, (2013),"An SVM-based Approach for Stock Market Trend Prediction", Proceedings of International Joint Conference on Neural Networks, Dallas, Texas, USA.

AUTHORS

Dr. J.Chandrika holds doctoral degree in computer science and engineering. Currently works as Associate professor in the department of computer science and engineering at Malnad college of Engineering, Hassan,Karnataka .Her areas of interest include, data stream mining ,Artificial intelligence and medical data mining She has six international conference publications and four international journal publications and one national conference publication to her credit.

Dr.B.Ramesh holds a doctoral degree in computer science and engineering. Currently works as Professor and Head of the department in the department of computer science and engineering ,MCE Hassan.He has a vast teaching experience of about 21 years.His research interest includes mobile adhoc networks, Computer networks and Data mining. He is currently guiding five research scholars.

Dr. K.R.Ananda kumar holds a doctoral degree in computer science and engineering. Currently works as Professor and Head of the department in the department of computer science and engineering ,SJBIT Bangalore.He has a vast teaching experience of about 25 years.His research interest includes medical data mining, data stream mining ,Artificial intelligence, Intelligent agents and web mining. .He has successfully guided two doctoral candidates. He has many publications in nternational and national journals to his credit.

Raina.D.cunha is a B.E student at Malnad College of Engineering Hassanand will be graduating in the year 2014. Presently she is working as trainee systems engineer at Infosys technology Mysore. She has undertaken many Data Mining projects during her UG course.