ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

Stock Prediction Using Clustering And Regression Techniques

[1] Shalini Lotlikar [2]Megha Ainapurkar
[1][2] Master of Engineering(IT)
Padre Conceicao College of Engineering
Margao, India

Abstract: **A stock market is the aggregation of buyers and sellers. A stock exchange is a place where, or an organization through which, individuals and organizations can trade stocks. In the financial market the decision on when buying or selling stocks is important in order to achieve profit. There are several techniques that can be used to help investors in order to make a decision for financial gain .The data set is taken from yahoo finance and preprocessing is performed on it. Clustering and regression are two techniques of data mining used here .For prediction of future stock price multiple regression technique is used which helps the buyers and sellers to choose their companies for stock.This paper has introduced the k means algorithm after which multiple regression is applied to predict the stock.**

***Keywords*—Data Mining, k-means, multiple regression, rule of thumb, centroid, clustering.**

I. INTRODUCTION

Stock markets are aggressive in nature. It is very difficult to predict the future stock price of the companies. The main categories of data mining are classification, clustering and regression. Clustering, as a generic tool for finding groups or clusters. Several algorithms have been proposed in the literature for clustering. .In the present work the techniques used are clustering and regression. The k-means clustering algorithm is the most commonly used because of its simplicity.Clustering is the process of creation of clusters of similar objects which is an unsupervised technique since it does not make use of class labels. K-Means can solve the well-known clustering problem.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. Regression is a supervised technique, which has a predefined target. Multiple regression in regression technique is used for this work.

Regression is used for predicting an outcome based on given input.The simplest regression technique is linear regression and advanced regression technique is multiple regression. If a single descriptive variable is used then it is known as simple linear regression and if more than one descriptive variable is used then the technique is multiple regression.

II. BACKGROUND KNOWLEDGE

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data.

Before clustering and regression techniques data pre-processing should be carried to make sure that there is no noise or errors present.

In k-means the main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

III. DIFFERENT APPROACHES TO SELECTING THE RIGHT NUMBER OF CLUSTERS IN K-MEANS CLUSTERING

There have been a number of different proposals in the literature for choosing the right K after multiple runs of K-Means, among them we focus on following approaches

A. By rule of thumb
B. Elbow method
C. Information Criterion Approach
D. An Information Theoretic Approach
E. Choosing k Using the Silhouette
F. Cross-validation

The best method that can be used is rule of thumb for selecting the right number of cluster.

This method can be used for any type of dataset

$K \approx \sqrt{n/2}$

Where n is the number of objects

A. Pre-Processing of stock data

The data is taken from yahoo finance and the weekly historical data is downloaded and saved in terms of csv files. The pre-processing is carried on close values of the data files.

If there are any empty values then it will be take 10 previous value which is summed up and divide by 10.

B. K-means Algorithm:

The algorithm starts its working by assigning different objects to different dataset randomly and then at each iteration it reallocates the data objects to another partition. Each partition is represented as a centroid were it is an average of all data objects in a partition. Here every partition must contain leastways a single data object, and each data objects must contain exactly one cluster

K-Means is a widely known unsupervised techniques. It assigns a given set of data objects into „k" clusters, where k shows the number of clusters and it should be mentioned in advance. Each cluster should have a centroid. Based on the distance of each data point to centroid. Data point is assigned to cluster having closest centroid. Then at each iteration the data points are reassigned to different clusters by recalculating the distance. This process continues until there is no further change in centroid location

Algorithm:

Input: The number of clusters created: k the number of objects assigned: x

Output: Based on the given similarity function „k" clusters are obtained.

Steps:

i)select „k" data entity randomly and assign it as the first cluster centroids;

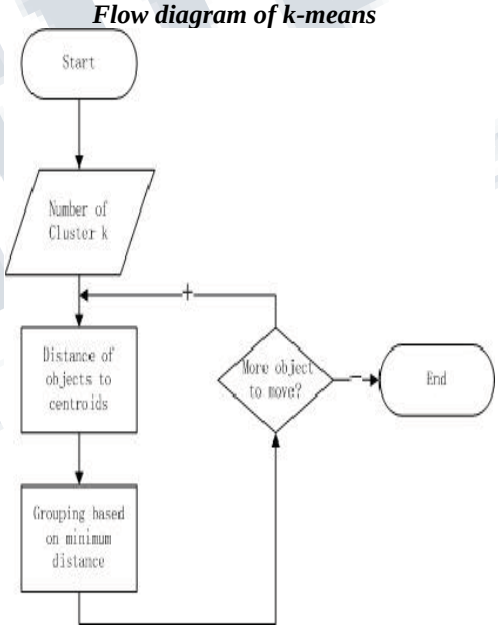
ii) Continue,

a .With respect to similarity function set the remaining data points to each cluster.

b .Revise the centroid for each cluster by taking the cluster mean

iii) Until no further change occurs.

Flow diagram of k-means



C. Categorization:

After data has been clustered we categories the companies with respect to its company department .for example Honda motors is an automobile company ,cisco is an IT company like wise we have categories the 67 companies into its department .The idea behind this is to put companies in similar industries together for comparison purpose.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

DATASET

We have taken historical data downloaded from Yahoo! Server of one seventy five companies from a time period of January 2016 to October 2016.

IV. CONCLUSION

Our data is collected from yahoo finance about seventy five companies and several data mining algorithms are applied on that dataset. In our proposed model k-means clustering method was performed on data where all the close value falls in one cluster which will predict the best companies' .Then we will use multiple regression which will predict the stock of the companies

REFERENCES

1. R. C. Dubes and A. K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.

2. Andrew Ng, Clustering with the K-Means Algorithm, Machine Learning, 2012

3. Trupti M. Kodinariya1 Dr. Prashant R. Makwana ,Review on determining number of Cluster in K-Means Clustering

4. Hailong Chen ,Chunli Liu Research and Application of Cluster Analysis Algorithm,2013 2nd International Conference on Measurement,

5. Zhigang Xiong ,Assco. Prof. Zhongneng Zhang,A Data Preprocessing Method Applied to Cluster Analysis on Stock Data by Kmea

6. Bini B.Sa*,Tessy Mathewb,clustering and regression techniques for stock prediction

7. Saeid Fallahpour 1, Mohammad Hendijani Zadeh & Eisa Norouzian Lakvan, Use of Clustering Approach For Portfolio Management, Inter- national SAMANM Journal of Finance and Accounting ISSN 2308-2356 January 2014, Vol. 2, No. 1