JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Cluster-Then-Predict and Predictive Algorithms (Logistic Regression)

A. Bhattacharjee[1*], J. Kharade[2]

[1*] Bharti Vidyapeeth's Institute of Management and Information Technology, Mumbai University, Navi Mumbai, India
[2]Bharti Vidyapeeth's Institute of Management and Information Technology, Mumbai University, Navi Mumbai, India

[*]*Corresponding Author:* anikb48@gmail.com , *Tel.: +91-91673-87781*

Available online at: www.ijcseonline.org

Received: 26/Jan//2018, Revised: 06/Feb2018, Accepted: 21/Feb/2018, Published: 28/Feb/2018

Abstract— Stock market is playing a vital role as investments option and investors make short-term investments as well as long-term investments. But here the main question arises "Where to invest?" and "when to invest?" even if an investor is aware about where to invest, it is still unpredictable whether or not stocks will have good future returns over time. To eliminate this dilemma predictive algorithms were introduced that will help investors in making investments by predicting which stocks will have positive expected returns. However, predicting stock returns with predictive algorithms alone is not enough. Clustering algorithms are widely used to cluster the stocks that have related returns over time. Using Cluster-Then-Predict approach we are going to prove that it provides more accurate results than the original predictive (Logistic Regression) model.

Keywords— Predictive Algorithms, Regression, Classification, Clustering, Logistic Regression, Stock Returns, Cluster-Then-Predict

I. Introduction

Machine learning is described as the data which is obtained by knowledge extraction. Machines do not require to be programmed directly instead it's trained to make decisions driven by the data. Rather than writing a code for every specific problem, data is provided to the algorithms and logic is developed on the basis of that data. When the machine improves itself depending on its previous experiences it can be concluded that machine has truly learned on its own [1].

Different databases such as RDBMS, multimedia databases, ORDBMS etc use Data mining. Data mining is used on wide applications like stock prediction, market analysis, stock forecasting etc. Item sets that are frequently used in data mining have an important role to find out the correlations between the database fields.

The objectives of the research paper are:

1) To study Predictive Algorithms
2) To study Cluster-Then-Predict
3) To compare Predictive Algorithms with Cluster-Then-Predict

The purpose of carrying out this study is to spread the awareness about the methodologies that can help in solving real-world problems and can derive better results than the traditional approaches. In this study, Section I contains the introduction of the study which has been discussed already, Section II contain the related work of the study that describes the previous research works, Section III discusses thea two different methodologies in detail. Section IV contains the comparison of results between the two approaches. Section V contains the conclusion and the future scope.

II. Related Work

In paper [2] presents analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER public datasets has been used in this project. The preprocessed dataset consists of 151, 886 records which have 16 fields from the SEER. Three data mining techniques like Naïve Bayes, back propagated neural network and C4.5 decision tree algorithm are investigated and compared with the achieved prediction performance. it is It was concluded that C4.5 algorithm has a much better performance than other two techniques.

In paper [3] they try to help investors with the better period for the selling and buying stocks within the stock market on the knowledge of past historical experiences. It is analysis that past investigations help to predict future in data analysis. In this paper they used decision tree algorithm which is one of the best data mining techniques. They also explained that forecasting stock return is one the important topic to be learnt for prediction of data.

In paper [4] they explained the main content of the study is to predicting the changes of balance of the users of balance of Yuebao. The methods of prediction adopts first clustering, then predicting. First of all, according to the user's balance of account information, the user's basic information and operating characteristics of the user, this paper made the classification for users. And then the amount of the user's balance of each class are predicted, so that authors can greatly guarantee the loss of information in the forecast process, this can greatly to ensure the accuracy of data prediction, and the empirical data analysis of the results is also proved that the forecasting model can well describe and forecast the change of the balance data, which can get more excellent results than the direct forecasting.

Algorithm	Usage	Pros	Cons
Linear Regression	Predicts a continuous outcome (price, salary, etc)	1) Simple, well recognized 2) Can work on small and large datasets	Consider a linear relationship
Logistic Regression	Predicting a categorical outcome (Yes/No, True/False, etc)	1) Calculates probabilities that is used to asses confidence of the prediction	consider a linear relationship
CART	It can Predict a categorical outcome or continuous outcome	1) Handles datasets has no linear relationship 2) Easy to explain and interpret	1) Does not work well on small datasets
Random Forests	Predicting a categorical outcome or continuous outcome	1) Has better accuracy over CART	1) Parameter tunings are needed 2) Cannot easily interpret unlike CART

Table 1. Comparison of predictive algorithms

III. Methodology

3.1) Predictive Algorithms (Logistic Regression)

Logistic Regression is commonly applied to a group of independent variables to predict an outcome which is binary. This binary outcome can be represented in either in 0 or 1/ True or false. Dummy variables often represent this binary outcome. Mostly, it is used for solving Classification problems. Logistic regression is considered as an exceptional case of linear regression in which the target variable or the dependent variable is categorical

For the dependent variable it uses the logit function, it fits the data that consists the log of odds using the possibility of the occurrence of the event.

In this example the value of 'y' can be 0 to 1 it is represented by the equation [5] odds = probability of event occurrence divided by the probability of the event.

$$Odds = \frac{P}{1-P}. \quad (1)$$

$$\ln(Odds) = \ln\left(\frac{P}{1-P}\right). \quad (2)$$

$$\log(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + (\beta_1 \times 1\,) + \cdots + (\beta_k \times k). \quad (3)$$

Where, P is the probability of interested characteristic presence. As there is a binomial distribution of dependent variable implemented, there has to be a link function which will be best fitted for the distribution, which is the logit function. The equation in the above selects the parameters to maximize the odds of receiving the sample values instead of reducing the sum of squared errors (as seen in the ordinary regression). [6]

Points to be considered before implementing logistic regression:

- It handles non-linear relationships efficiently between the target variable and predictors. It can take various relationships types because it uses a log transformation which is non-linear for predicting the ratio of odds.
- To remove overfitting as well as underfitting, all significant variables should be included. A better method to assure this practice is by using a step-by-step method to compute the logistic regression.
- It needs samples which are large in size. Since, maximum likelihood that calculations are less accurate at low sample sizes in comparison to the simple least square.
- It does not have multi-collinearity i.e. independent variables need not be inter-related with each other. However, there are still options to consider interaction impacts of categorical variables during modeling and computation.
- It will be known as Ordinal logistic regression when the values of dependent variable are ordinal. [7]

3.2) Cluster-Then-Predict

Current section explores the method of clustering. Clustering is an classification method which is part of unsupervised learning that focuses on creating groups of clusters, in a way that entities which belong to the same cluster are very identical and entities in other clusters are quite distinct [8]. Analysis of clusters is one of the oldest topics in the data mining field. It is the initial step in the direction of exciting knowledge discovery. Clustering is a procedure of grouping data entities into a group of different classes, called clusters. Now entities within a class have high similarity to one another whereas entities in separate classes are more different.

Analysis of clusters has been used majorly in wide areas, that includes data analysis, bioinformatics, pattern recognition machine learning and text mining. In Industries, clustering can help entrepreneurs discover interests of their customers based on purchase patterns and characterize groups of the customers. In geology, the expert can apply clustering to identify regions of houses that are in a city and lands. Clustering can also used in classification of documents that are on Web for the discovery of information.

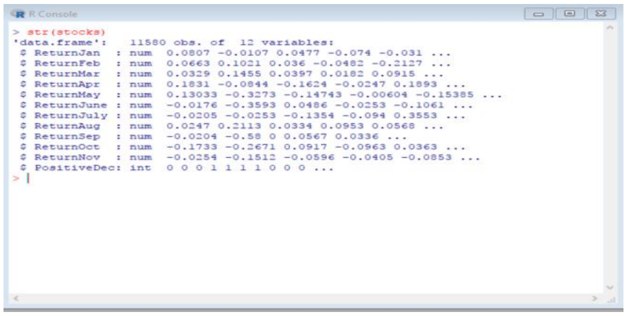
In K-Means clustering similar kind of data are clustered for data prediction. In K-Means clustering, each data point is assigns itself to its nearest cluster and using Euclidian distance formula the data points are clustered. Using this, it will it improves the clusters and improve the Euclidian distance formula. This improvement is based on normalization. Two new features are added because of this improvement. Based on normalization, the first feature will calculate normal distance metrics. Because of the majority voting the second feature will cluster the data points. The proposed technique will be implemented in R.

Steps for implementing K-Means clustering:

1. Define number of clusters k
2. Put each point to a cluster randomly
3. Calculate the cluster centroids
4. Again assigning each point to the nearest cluster centroid
5. Re-Calculate the cluster centroids
6. Repeat steps 4 and 5 until no improvement is made

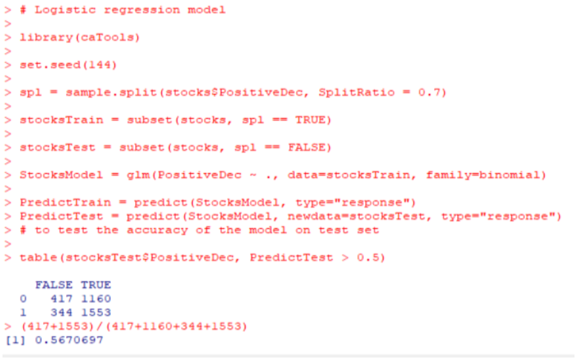
IV. COMPARISON OF PREDICTIVE ALGORITHMS WITH CLUSTER-THEN-PREDICT

The StockCluster.csv dataset contains monthwise stock returns that has been provided by the NASDAQ stock exchange. This dataset has been taken from infochimps which provides access to many datasets. The data of stock price in this dataset has 12 variables and 11580 observations as shown in Figure 1. The dataset is loaded in R.



```
> str(stocks)
'data.frame':   11580 obs. of  12 variables:
 $ ReturnJan   : num  0.0807 -0.0107 0.0477 -0.074 -0.031 ...
 $ ReturnFeb   : num  0.0663 0.1021 0.036 -0.0482 -0.2127 ...
 $ ReturnMar   : num  0.0329 0.1455 0.0397 0.0182 0.0915 ...
 $ ReturnApr   : num  0.1831 -0.0844 -0.1624 -0.0247 0.1893 ...
 $ ReturnMay   : num  0.13033 -0.3273 -0.14743 -0.00604 -0.15385 ...
 $ ReturnJune  : num  -0.0176 -0.3593 0.0486 -0.0253 -0.1061 ...
 $ ReturnJuly  : num  -0.0205 -0.0253 -0.1354 -0.094 0.3553 ...
 $ ReturnAug   : num  0.0247 0.2113 0.0334 0.0953 0.0568 ...
 $ ReturnSep   : num  -0.0204 -0.58 0 0.0567 0.0336 ...
 $ ReturnOct   : num  -0.1733 -0.2671 0.0917 -0.0963 0.0363 ...
 $ ReturnNov   : num  -0.0254 -0.1512 -0.0596 -0.0405 -0.0853 ...
 $ PositiveDec: int  0 0 0 1 1 1 1 0 0 0 ...
> |
```

Figure 1: List of variables in the Dataset

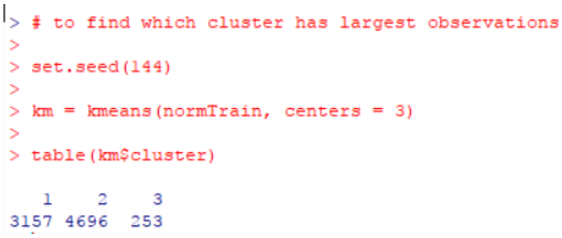


```
> # Logistic regression model
>
> library(caTools)
>
> set.seed(144)
>
> spl = sample.split(stocks$PositiveDec, SplitRatio = 0.7)
>
> stocksTrain = subset(stocks, spl == TRUE)
>
> stocksTest = subset(stocks, spl == FALSE)
>
> StocksModel = glm(PositiveDec ~ ., data=stocksTrain, family=binomial)
>
> PredictTrain = predict(StocksModel, type="response")
> PredictTest = predict(StocksModel, newdata=stocksTest, type="response")
> # to test the accuracy of the model on test set
>
> table(stocksTest$PositiveDec, PredictTest > 0.5)

    FALSE TRUE
  0   417 1160
  1   344 1553
> (417+1553)/(417+1160+344+1553)
[1] 0.5670697
```

Figure 2: Test set accuracy of Logistic Regression model

In Figure 2, using this dataset the Logistic Regression model is created. This model gives the accuracy of 57% on the test set over the threshold of 0.5 which obviously beats the baseline model.

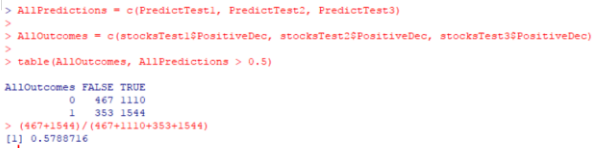


```
> # to find which cluster has largest observations
>
> set.seed(144)
>
> km = kmeans(normTrain, centers = 3)
>
> table(km$cluster)

   1    2    3
3157 4696  253
```

Figure 3: Assigning each observation to the cluster

In Figure 3, it can be observed that the K-Means clustering algorithm created 3 clusters and assigned each observation to the corresponding cluster based on the seed value that has been specified in the R console. The table command in R console showed the number of observations that each cluster has which helps the investigator, to find out the cluster that has the most number of observations.



```
> AllPredictions = c(PredictTest1, PredictTest2, PredictTest3)
>
> AllOutcomes = c(stocksTest1$PositiveDec, stocksTest2$PositiveDec, stocksTest3$PositiveDec)
>
> table(AllOutcomes, AllPredictions > 0.5)

AllOutcomes FALSE TRUE
          0   467 1110
          1   353 1544
> (467+1544)/(467+1110+353+1544)
[1] 0.5788716
```

Figure 4: Test set accuracy using Cluster-Then-Predict

As shown in Figure 4, there is a modest improvement on the accuracy of the test set over a threshold of 0.5. Using Cluster-Then-Predict approach we got the accuracy of 58% which is anytime better than the baseline model but also, it slightly gives better results than the traditional Logistic Regression model.

V. CONCLUSION AND FUTURE SCOPE

In this paper we came to know that Cluster-Then-Predict methodology can provide more accurate results than the simple implementation of the predictive algorithms. During the small experimentation conducted on the dataset, we witnessed that, there was not a lot of improvement over the implementation of the single predictive algorithm (Logistic Regression). The reason behind that is predicting stock returns is a ridiculously hard problem considering that fact we can see a good increase in accuracy. The main objective is to make investors feel more confident that they will have positive returns on the stocks that they have invested in. We can improve our overall accuracy on this dataset by implementing other predictive algorithms that are new in the field of machine learning (e.g XGBOOST, LightGBM, CatBoost etc) which can provide better accuracy than the traditional predictive algorithms. Furthermore, to improve the results even further it can be recommended to implement clustering before applying any of the Predictive Algorithms mentioned above.

REFERENCES

[1] W. Huang, Y. Nakamoria and S. Wang, " *Forecasting stock market movement direction with support vector machine*", Computers & Operations Research, Vol. 32, pp. 2513 – 2522.

[2] A. Bellaachia, E. Guven, "*Predicting Breast Cancer Survivability Using Data Mining Techniques*", In the Proceedings of the 2010 Department of Computer Science The George Washington University, Washington DC, pp. 20052, 2010

[3] S. Gour, "*Developing Decision Model by Mining Historical Prices Data of Infosys for Stock Market Prediction*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.10, pp. 92-97, 2016.

[4] Y. X. Lu, T. Zhao, "*Research on time series data prediction based on clustering algorithm*", In the Proceedings of the 2017 American Institute of Physics Conference, United States, pp. 1864-020152, 2017.

[5] S. S. Sathe, S. M. Purandare, P. D. Pujari and S. D. Sawant, "*Stock Market Prediction Using Artificial Neural Network",* International Education and Research Journal. Vol. 2, Issue 3, pp. 2254-9916, 2016

[6] M. Mittermaye, "*Forecasting Intraday Stock Price Trends with Text Mining Techniques*", In the Proceedings of the 2004 37th Hawaii International Conference on System Sciences, , pp. 0-7695-2056-1/04, 2004.

[7] K. S. Kannan, P. S. Sekar, M. M. Sathik and P. Arumugam "*Financial Stock Market Forecast using Data Mining Techniques*", International Multi Conference of Engineers and Computer Scientists, Vol 1, I,IMECS 2010, March 17-19,2010, Hong Kong. pp. 2078-0966.

[8] Swati Joshi, Farhat Ullah Khan and Narina Thakur, "*Contrasting and Evaluating Different Clustering Algorithms: A Literature Review*", International Journal of Computer Science and Engineering, Vol. 2, Issue.4, pp. 2347-2693, 2014.

Authors Profile

Mr. A. Bhattacharjee pursed Bachelor of Science (Information Technology) from University of Mumbai, Mumbai, India in 2015. He is currently pursuing Masters Of Computer Applications from University of Mumbai, Mumbai, India.

Dr Jyoti . Kharade, Bachelor of Science,Master of Computer Application from Shivaji University, M.Phil from Bharati Vidyappeth deemed University and Ph.D from SNDT University. She is currently working as Associate Professor in Bharati Vidyapeeth's Institute of Management and Information Technology, University of Mumbai, since 2004. She is a member of CSI. She has published more than 27 research papers in reputed international journals including conferences. Her main research work focuses on e-Governance, Data Mining. She has 16 years of teaching experience.