ANALYSING THE STOCK MARKET USING CLUSTERING IN TIME SERIES DATA MINING

R. Saravanan kumar¹, R. Renuka²,
M.Phil Research scholar¹, Assistant Professor²,
Department of Computer Science,
Sri Ganesh College of Arts and Science, Salem.
Email: rsaravanakumar333@gmail.com

Received 22 Oct 2016; Accepted 11 Nov 2016

Abstract

A new framework for analysing time series data called Time Series Data Mining (TSDM) is introduced. This framework adapts and innovate data mining concepts to analysing time series data. It creates a set of methods with the growing deployment of a large Number of sensors, telemetry devices and that reveals hidden temporal patterns that are characteristic and predictive of time series events. The TSDM methods overcome limitations of traditional time series analysis techniques by adapting data mining concepts for analysing time series. Data Mining is the analysis of data with the goal of uncovering hidden patterns. Data Mining encompasses a set of methods that automate the scientific discovery process. In general, applied the clustering algorithm to stock market data and showed the results based on the stock market index (high return, low return, high risk and low risk). Therefore in this paper applied K-Means Clustering algorithm is applied for Stock Market. The goal of stock market data clustering is to predict the period of highly returned investments. The clustering algorithms were successfully applied to cluster stock market data comprising into two distinct clusters based on the similarity of stock market index profiles and prior share market knowledge.

Key Words: Time Series Data Mining (TSDM), k-means clustering, Data mining, Stock market, hidden temporal patterns.

1. Introduction

The Time Series Data Mining (TSDM) framework, introduced by this dissertation, is a fundamental contribution to the fields of time series analysis and data mining. Methods based on the TSDM [1] framework are able to successfully characterize and predict complex, nonperiodic, irregular, and chaotic time series. The TSDM methods overcome limitations (including stationarity and linearity requirements) of traditional time series analysis techniques by adapting data mining concepts for analyzing time series. This chapter reviews the definition of a time series, introduces the key TSDM concepts of events and hidden temporal patterns,
and provides examples of problems the TSDM framework addresses. A time series $X$ is a sequence of observed data, usually ordered in time

$X(t) = x_t, t=1, ..., N$

Where $t$ is a time index, and $N$ is the number of observations. Time series analysis is fundamental to engineering, scientific, and business endeavors. Researchers study systems as they evolve through time, hoping to discern their underlying principles and develop models useful for predicting or controlling them. Time series analysis may be applied to the prediction of data which are observed over time.

The TSDM framework focuses on predicting events, which are important occurrences. It is commonly assumed that the ARIMA [4] time series models developed with past data will apply to future prediction. This is stationary assumption that models will not need to very through time. ARIMA model can be defined by linear difference equations, but system generating the time series are not necessary linear or stationary. In contrast, the TSDM framework and methods built upon it can handle nonlinear and non-stationary time series [5]. The TSDM framework innovate data mining concepts for analysing time series data. Therefore in this paper applying clustering is using k-mean algorithm.

Remaining of the paper organized as follows section 2 contains related work. The proposed work is discussed in section 3. Result and discussion of the proposed method discussed in section 4. Finally section 5 contains conclusion.

2. Related Work

Tsang, P.M., Kwok, P [6] made an empirical study on building a stock buying/selling alert system using back propagation neural networks (BPNN), their NN was codenamed NN5. The system was trained and tested with past price data from Hong Kong and Shanghai Banking Corporation Holdings over the period from January 2004 to December 2005. The empirical results showed that the implemented system was able to predict short-term price movement directions with accuracy about 74%.

Wang, J.L. [7] applied the concept of serial topology and designed a new decision system, namely the two layer bias decision tree, for stock price prediction. The methodology developed by the authors differs from other studies in two respects; first, to reduce the classification error, the decision model was modified into a bias decision model. Second, a two-layer bias decision tree is used to improve purchasing accuracy. The empirical results indicated that the presented decision model produced excellent purchasing accuracy, and it significantly outperformed than random purchase.

Enke, D., Thawornwong [8] presented an approach that used data mining
methods and neural networks for forecasting stock market returns. An attempt has been made in this study to investigate the predictive power of financial and economic variables by adopting the variable relevance analysis technique in machine learning for data mining. The authors examined the effectiveness of the neural network models used for level estimation and classification. The results showed that the trading strategies guided by the neural network classification models generate higher profits under the same risk exposure than those suggested by other strategies.

Anil Rajput [9] has proposed a rule based classification method which generates rules with preprocessed data. It focuses on each class separately and increases the probability of exact classification. PRISM algorithm is applied to generate rules by examining the training data and identifying suitable rules that covers majority of class. Also it measures the success of each rule. It serves as a suitable tool for generating rules for buy and sell.

Yin Song [10] proposed a method of analyzing coupled behavior of stocks in the market. CHMM (Coupled hidden markov model has been considered for coupled relations in hidden state space. He proposed a new graph based framework to find out abnormalities that might occur in coupled behaviors. He also applies the results to real world stock data and proved that their method beats the basic CHMM markov model in both technical and business measures.

3. Methodology

The first step in applying the TSDM method [11-12] is to define the TSDM goal, which specific to each application. The goal is to find hidden temporal patterns that are characteristics of events in time series where events are specified in context of the TSDM goal. Given a TSDM goal, an observed time series to be characterized, and a testing time series to be predicted, the steps in the TSDM methods are

3.1 Training Stage (Batch Process)
1. Frame the TSDM goal in term of event characterization function, objective function, and optimization formulation
   a. Define the events characterization function $g$
   b. Define the objective function $f$.
   c. Define the optimization formulation including independent variable over objective function.
2. Determine $Q$ i.e. the dimension of the phase space and the length of temporal patterns.
3. Transform the observed time series into the phase space using the time delayed embedding process.
4. Associates with each time index in phase space & form the augmented phase space.
5. in the augmented phase space, search for optimal temporal pattern cluster.
6. Evaluate training stage result repeat training stage necessary.

3.1 Testing Stage (Real Time or Batch Process)
1. Embed the testing time series into the phase space.
2. Use the optimal temporal patterns cluster for predicting events.
3. Evaluate testing stage results.

**K-Means Algorithm**

K-Means clustering [13-14] 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. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function: figure 1 shows k-means clustering algorithm.

![Algorithm 1](image)

### 4. Result and discussion

This work examines the daily change of the closing values of NSE-NIFTY and BSE-NIFTY based on the following predictors: Open price, High price, Low price and Close price. NSE-NIFTY and BSE-NIFTY values are obtained from the NSE and BSE websites respectively, for the period from Jan’2011 in April 2014 with a sample of 850 trading days. The data are divided into two sub-samples in the split up of 80:20 where the in-sample or training data spans from Jan’ 2011 to Sep’ 2013 with 680 trading days and the data for the remaining period from Oct 2013 to April 2014 with 170 trading days are used for out of sample or test data. Clustering Stock data is the process of grouping the similar Index values into the same cluster based on return value by applying clustering techniques. Index in the same cluster has been often similar returns. The K-Means techniques are implemented and tested against our Stock Market dataset.
Where, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

\[ \text{PSNR} = \log \left( \frac{(2^n-1)^2}{MSE} \right) \]

\[ \text{Sensitivity} = \frac{TP}{(TP+FN)} \]

\[ \text{Specificity} = \frac{TN}{(TN+FP)} \]

\[ \text{accuracy} = \frac{t}{N} \times 100 \]

Where, \( t \) is no of sample currently classified and \( N \) is total no of samples.

The following table 1 and figure 2 shows that the PSNR, sensitivity, specificity and accuracy value of the K-Means clustering.

Table 1 performance analysis of k-means

<table>
<thead>
<tr>
<th>Metrix</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>88.78</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>89.23</td>
</tr>
<tr>
<td>Specificity</td>
<td>87.47</td>
</tr>
<tr>
<td>Accuracy</td>
<td>89.82</td>
</tr>
</tbody>
</table>

From the figure, it is easy to justify that the k-means produces the best result which is described by using the PSNR, sensitivity, accuracy and specificity. The following table 2 and figure 3 shows the error rate of the k-means clustering.

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \]

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| \]

Where \( \hat{y} \) and \( f_i \) are the vector of n classified prediction. \( y_i \) is vector of observed value corresponding to the input.

Table 2 Error rate of k-means

<table>
<thead>
<tr>
<th>Metrix</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Square Error (MSR)</td>
<td>17</td>
</tr>
<tr>
<td>Mean Absolute Error (MAE)</td>
<td>20</td>
</tr>
<tr>
<td>Relative Absolute Error (RAE)</td>
<td>22</td>
</tr>
</tbody>
</table>

Figure 2 performance analyses of k-means

Figure 3 Error rates of k-means
From the analysis report of figure 3 describes the k-means has a minimum MSE and RAE which is used to produce high accuracy.

5. Conclusion

The clustering problem is a real important problem and has attracted much attention of many researchers. The Clustering algorithm K-Means algorithm is applied for Stock Market Dataset. This study is a first attempt from the different perspective of personalization and to best of our cognition. This attempt is a new direction in the field of Stock Market Prediction. The results obtained are very encouraging, proving the practical applicability of the stock market. The k-means has a minimum MSE and RAE which is used to produce high accuracy.

Reference


