**MOTIF DISCOVERY IN NON LINEAR TIME SERIES**

**A PROJECT REPORT *submitted by***

**CB.EN.U4EIE10018** **ERPINA SATYA RAVITEJA**

**CB.EN.U4EIE10042** **NIKHIL NAYANAR**

**CB.EN.U4EIE10032** **SRITEJA KOTTAMASU**

**CB.EN.U4EIE10075** **VIGNESWARI**

***under the guidance of***

**Mr BINOY B NAIR**

***in partial fulfilment of the requirements for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**IN**

**ELECTRONICS AND INSTRUMENTATION ENGINEERING**



**AMRITA SCHOOL OF ENGINEERING**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE 641 112**

**MAY 2014**

**AMRITA VISHWA VIDYAPEETHAM**

**AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112**



**BONAFIDE CERTIFICATE**

This is to certify that the project report entitled **MOTIF DISCOVERY IN NON LINEAR TIME SERIES**

submitted by

**CB.EN.U4EIE10018** **ERPINA SATYA RAVITEJA**

**CB.EN.U4EIE10042** **NIKHIL NAYANAR**

**CB.EN.U4EIE10032** **SRITEJA KOTTAMASU**

**CB.EN.U4EIE10075** **VIGNESWARI**

in partial fulfillment of the requirements for the award of the **Degree of Bachelor of** **Technology** in **ELECTRONICS AND INSTRUMENTATION ENGINEERING** is a

bonafide record of the work carried out under my guidance and supervision at Amrita

School of Engineering, Coimbatore.

|  |  |
| --- | --- |
| Project Advisor | Chairman, ECE |
| Mr Binoy B Nair | Dr Jaya Kumar M |
| Assistant Professor |  |
|  |  |

The project was evaluated by us on:

Internal Examiner External Examiner

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE** |
| **NO.** |
|  |  |
|  | **ABSTRACT** | i |
|  | **LIST OF FIGURES** | ii |
|  | **LIST OF ABBREVIATIONS** | iii |
|  | **LIST OF TABLES** | iv |
| **1** | **INTRODUCTION** | 2 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **2** | **LITERARY SURVEY** | 5 |  |

**3 TREND PREDICTION**

**3.1 Time Series**

**3.2 SAX Algorithm**

3.2.1 Normalization of the time series

**3.3 Symbolic representation**

**3.4 Motif discovery**

**3.5 Genetic algorithm**

**3.6 HP filter**

**4 CONCLUSION**

**4.1 Conclusion**

**4.2 Future scope of work**

**5 REFERENCES**

**ABSTRACT**

The project deals with the design of an effective system which can predict trends in non linear financial time series using Motif discovery algorithm and Genetic algorithm

The non linear financial data is first symbolically represented using SAX algorithm and then

motif discovery is used to predict trends. Genetic algorithm is used to optimize the parameters of

SAX algorithm. Results have been compared with other optimization algorithm ABC .

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE NO.** |

**LIST OF ABBREVIATIONS**

PAA Piecewise Aggregate Approximation

SAX Symbolic Aggregate Approximation

USD/CAD U.S. Dollar / Canadian Dollar

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE NO.** | **TITLE** | **PAGE** |
|  |  | **NO** |

**CHAPTER 1**

**INTRODUCTION**

**1.INTRODUCTION:**

Since the project aims at pattern recognition in stock values, it is necessary to have an understanding of what stocks are and why it would be useful to be able to its future value.

STOCK MARKET:

Basics and its importance:

A stock represents a small share of a company. The stock market is a term used to describe the physical location where the buying and selling as well as overall market activity takes place. Companies issue stocks to acquire capital while investors buy them to own a portion of the company. Investors buy stocks with the belief that the company will grow continuously to raise the value of their shares. Even a small share in the company will give the investors a say in how the company runs. Only companies listed in public stock exchanges like NYSE, BSE, LSE etc. are capable of stock trading. Generally, investors can hire a ‘broker’ to make the transactions for him/her. The transactions here refer to selling or buying of stocks. So for any sale (through a transaction), the broker receives a commission.

Stocks are advantageous over savings investments in the fact that they actually give you a right to participate in making company decisions. This partial ownership also allows the stock-holders to benefit from the profits the company makes which are distributed among the share-holders. When the company makes profits, the values of stocks go up and when the company suffers so does its stocks. So compared to savings investment, stocks have a higher chances of making profits as well as losses.

Stock markets are closely related to the economic health/prosperity of a country. They play a crucial role in development of commerce and industry which have direct implications on the development of the country. Whenever a company wants to raise money for business it can approach the banks or it can issue shares of its company through the stock market. So the primary aim of a stcok market is to facilitate the selling and buying of stocks of listed companies, thereby supporting the economic growth of the country. They make it possible to sell the stocks at any point of time and get back the investment with the loss/profit correspondingly made. This liquid nature is what attracts investors to the stock market.

To verify, we have to apply the time-series forecasting to markets that are both old and long established and relatively new. The newer market we have chosen is the Bombay stock exchange and the established one is the London Stock Exchange.

BSE:

Bombay Stock Exchange (BSE) was set up in 1875 and has since then facilitated the growth of Indian corporate sector, acting as an efficient capital-raising platform.BSE provides an efficient and transparent market for trading in equity, debt instruments, derivatives, and mutual funds.BSE Ltd is world’s fifth most active exchange in terms of number of transactions handled through its electronic trading system. BSE systems and processes are designed to safeguard market integrity, drive the growth of Indian capital market and stimulate innovation and competition across all market segments. BSE also provides a host of other services to capital market participants including risk management, market data services and education. BSE’s popular equity index – the S&P BSE SENSEX is India’s most widely tracked stock market benchmark index.

LSE:

The London Stock Exchange was founded in 1801 and its current premises are in the City of London. There are currently 2938 companies from over 60 countries listed on the London Stock Exchange. The LSE today offers trading in more emerging markets exchange traded funds than any other exchange in the world. The LSE supply its participants with real time prices and trading data creating the transparency and liquidity. LSE allows companies to raise money, increase their profile and obtain a market valuation through a variety of routes. The London Stock Exchange runs several markets for listing, giving an opportunity for different sized companies to list. International companies can list a number of products in LSE including shares, depository receipts and debt, offering different and cost-effective ways to raise capital.

STOCK MARKET FORECASTING:

Hence, if the stocks prices can be previously determined, it will help in trading favourably such that the investors can maximize their earnings by selling/retaining the stocks. This project aims at devising an algorithm to predict the values of stock so that only favourable trading is done. Broadly, the existing forecasting methods can be classified into the following:

1. Fundamental Analysis: These methods are a sort of self-evaluation by the concerned company itself; based on the previous performances of the company stocks and the validity of their accounts. It is built on the fundamental hypothesis that to progress humans need capital and if a company is prospering it should be given further capital resulting in an increase in stock market. This underlying principle makes it very reasonable and objective so that it is the most widely used by fund managers. An alternative method of fundamental analysis is beyond bottom to up analysis. It involves first analyzing the global economy, followed by country analysis and then sector analysis and finally company-level analysis. Many performance ratios are created that aid the fundamental analysis with assessing the validating of a stock.
2. Technical Analysis: This method aims at forecasting the direction of prices through the study of past market data. Models and trading rules from prices and volume transformations, moving averages, regressions or through recognition of chart patterns. An underlying principle of technical analysis is that market price reflects all relevant information .So price action tends to repeat as investors tend to move toward a patterned behaviour.

Stock Market direction prediction is a key area of interests for professionals and most investors. Soft-computing techniques cover a set of techniques that employ methods of reasoning that are approximate rather than exact. The difference between hard computing and soft computing is the importance given to factors like precision and rigor. These methods are already playing a key role in areas like securities and foreign exchange prediction, trading, automatic credit storing, business failure prediction.

The widely used soft-computational techniques are artificial neural networks (ANN) , linear and multi-linear regression (LR,MLR), ARMA and ARIMA models, genetic algorithms (GAs) , random walk (RW) , buy and hold (B & H) strategy.

This paper aims at evaluating motif-discovery as a soft-computation technique and comparing it to other methods. Time-series motifs are highly related to clusters. Basically they are over-represented strings in discrete strings, for example, in musical or DNA sequences (Reinert et al., 2000).Before explaining what motifs is, it is essential to know what trivial matches are. They are simply the closest matching sub-sequences in a time series including the possibility of sharing elements. In this context, motifs are those patterns with the highest number of non-trivial matches. By ensuring that these matches are a minimum distance apart, we eliminate any possibility of them sharing elements and hence are independent at the same time matching as well. It is assumed that motifs do exist in non-linear time series like the stock market values. Extracting these patterns can hence help develop trading rules.

In order to extract such information, it is essential that the data itself is represented in a suitable form. The SAX (Symbolic Aggregate Approximation) is one such method. Here, the entire data set is represented as a string of alphabets. The data series is firstly piece-wise approximated (PAA) for easier computation without losing significant data. Cut points are defined based on the amplitude such that all data-points lying between two such cut points are represented by a particular alphabet. As the name suggests, this methods approximates the series to symbols. From this string of alphabets, motifs are hence extracted.

So, definitely questions will arise as to how the cut points are arbitrarily fixed at a value. This calls for the optimisation algorithm GA (Genetic Algorithm). By selecting sufficiently large data as a training set, the most suitable cut points are selected for the testing set, where the technique is tested.

The performance of this technique is compared to that of association rule mining (ARM) and optimisation technique artificial bee colony algorithm (ABC), association rule mining (ARM) and optimisation technique genetic algorithm (GA), motif discovery and optimisation technique artificial bee colony algorithm (ABC).

**CHAPTER 2**

**LITERATURE SURVEY**

**2.LITERATURE SURVEY**

“Extended SAX: Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation” [1]

Financial time series has few critical points. In Sax algorithm dimensionality reduction can be achieved by using PAA (Piece wise Aggregate Approximation) .The main disadvantage of using SAX is that some important points in the time series can be ignored which results in inaccuracy. This paper introduces a new approach called Extended SAX where in each segment (time series is divided into segments of equal length) including average of the segment two extra points i.e., min and max points are also considered so that accuracy of data representation can be improved. To illustrate this they have taken 20 data sets with an average length of 5000 points each which are both computer generated and taken from internet and subsequence search is carried out by taking a random query and using Brute Force algorithm and the average number of matching results using both SAX and Extended SAX are tabulated. It is shown that Extended SAX gives better results. But the problem is that more memory is required for Extended SAX.

“An Empirical Study of Similarity Search in Stock Data” [2]

The main objective of this paper is to compare numerical and symbolic representation with respect to similarity search in stock data. For this purpose they have taken stock data of Malaysian exchange from Dec 31, 2001 to Dec 31, 2006 with a total of 1233 days. Six parameters of stock data have been considered. They include: a)Opening value b)Closing value c) Highest d)Lowest values e)trading volume f) Price changes( Closing value –opening value).For numerical representation of stock data they used Euclidean distance for similarity search. For symbolic representation they have used three matching symbols (up, down, same) .The conclusion is that numerical representation of stock data gives consistent and intuitive results compared to symbolic representation. They suggested opening, closing, highest and lowest valued are best parameters to be used when performing similarity search.

“Combining Clustering and a Decision Tree Classifier in a Forecasting Task” [3]

This paper deals with new product demand forecast for next one year based on historical data. They constructed demand images by historical data clustering and classification is done using inductive decision trees. K means algorithm is used for clustering and only those clusters with minimum mean clustering error are taken for forecasting. They suggested that Clustering using k means algorithm has many shortcomings and new techniques should be adopted. They tested their methods on very little data. Model has to be tested on different volumes and types of initial data for more accurate evaluation.

“An Effective Algorithm for Clustering Time-Series Data by Fuzzy Logic” [4]

The main objective of this paper is to perform clustering using LCCS (Longest Common Subsequence), K means algorithm and Fuzzy logic. The data set is taken from Centre for Atmospheric studies that have two attributes humidity and temperature with 4600 data points. Euclidean distance measure has many shortcomings as it is not able to manage time axis gap. So LCCS technique is chosen. There is no clarity about how fuzzy logic is being used. Only one data set is taken into consideration for training. The results show the clusters with their centers.

“TIME SERIES MOTIFS STATISTICAL SIGNIFICANCE” [5]

Researchers in the field of data mining mainly developed algorithms on extraction of motifs in a dataset and evaluated them either by using their expertise or by probability based techniques(motifs are highest occurrence are considered most significant).This papers says that many frequent patterns are spurious. Their approach is to estimate time series statistical significance. To achieve this they first used SAX algorithm to get symbolic representation. Then motifs are extracted and probability of occurrence of each motif is evaluated using Marchov Chain Models. A new parameter called p value which gives difference between observed count and estimated count. Estimated count is obtained by algorithm published by Robin and Schbath in 2007 [10].52 datasets are several sources like brain activity (EEG), motion capture (Mocop) etc. Motifs are ranked according to their statistical significance.

“CLUSTERING OF TIME SERIES SUBSEQUENCES IS MEANINGLESS: IMPLICATIONS FOR PREVIOUS AND FUTURE RESEARCH” [6]

The paper aims to prove that it is using certain restrictions that clusters are obtained from time series such that it is very improbable that these restrictions are unsatisfied by any series and so the clusters obtained are arbitrary. In order to prove the above statement, the authors have used the general clustering algorithms of k-means clustering, hierarchical clustering, EM, SOM. For the clustering techniques, the series used were S&P’s 500 Index closing values and random walk data. The idea is to take data sets as unrelated and random as possible and show that the clusters extracted from both are similar i.e. the clusters obtained are independent of the datasets. The reason for this anomaly is the presence of trivial matches which restrict viability of clusters to a small subset of datasets .In order to make clustering provide useful information the authors have stated consideration of the overtly presented subseries in a time series, called ‘motifs’. This definition of ‘motif’ as can be seen in it itself eliminates the problem due to trivial matches. Future work aims to extensively validate motifs and examine how the above mentioned problems in clustering will affect streaming clustering of time series.

“Making Subsequence Time Series Clustering Meaningful” [7]

Recently a surprising claim has been made that time series clustering is meaningless. The main problematic question that arises is how cluster centers are chosen. This claim invalidates the research done in this area. This paper tries to prove that time series clustering can be meaningful. They state that Euclidean distance taken as a measure of distance is the centre for all the problems associated with clustering and they developed a new algorithm for distance measure and tested on time invariant deterministic that produces time series that are cyclic. The results show that clustering is meaningful.

“Combining SAX and Piecewise Linear Approximation to Improve Similarity Search on Financial Time Series” [8]

The traditional SAX algorithm is based on Piece Wise Aggregate Approximation (PAA).SAX has many advantages as many advantages but main issue is PAA used in SAX can miss many important points in each subsequence. This paper proposes a new technique where PLA (piece wise linear approximation) is used where in each a linear line represents a subsequence and then a shape hierarchy decision tree is build which shows that has branches that categorize linear lines fitted in each segment as sharp up, steady up and sharp down. The subsequence in each branch will be similar and by finding distance between two slopes the similar subsequence can be found. They tested this technique on data set downloaded from internet with 5000 points each and with every 1000 points an average of 4.52 resultant patterns are matched.

Clustering Time Series Data Stream – A Literature

Survey [9]

Time series clustering is an important but difficult task. A lot of research has gone into data mining using clustering techniques. This paper studies various clustering techniques and summarized the merits and demerits of them. Similarity and dissimilarity methods are covered in clustering which helps to understand the need and purpose of clustering .It has been stated that many clustering techniques are not able differentiate random and real patterns. Their future work is to develop an effective clustering algorithm for time series.

**CHAPTER 3**

**TREND PREDICTION**

**3.TREND PREDICTION**



Figure 3.1.block diagram

3.1 TIME SERIES:

A time series is a sequence of values that is obtained at uniform intervals of time. The set of values obtained is called as a data. Data can be monthly, annual, quarterly etc. Time series is of two types basically linear and non linear. Financial data are generally non linear. Usually financial time series data are usually continuous, large and unbound. There are many technical methods for financial time series data to predict the market behaviour. Time series data always exist in large data size and high dimensionality.SAX is used convert the time series into symbols. Determining recurring pattern from a time series and based on the probability of occurrence future values can be predicted. The extraction of frequent patterns from a time series is known as motif discovery.

3.2 SAX ALGORITHM:

Financial time series are usually large.SAX can handle large datasets and produce faster results than any other algorithm. SAX is the first symbolic representation that allows lower bounding, dimensionality reduction and numerosity reduction. SAX allows motif discovery. SAX requires less space.SAX converts the time series into symbolic representation whereas PAA ( piecewise aggregate approximation) gives only numerical results. The steps in SAX is converting the time series into PAA representation and then converting them into symbols. As financial time series are large datasets it is necessary to break down the data into smaller ones, to find motifs. The first step in PAA is dividing the time series into equal parts and the arithmetic mean of the points within each part represents a segment. Thus, the time series is now represented as a smaller set of numbers. This PAA representation is then converted into symbols to form a string. The next section presents SAX in detail.

3.2.1 NORMALIZATION OF THE TIME SERIES:

The time series is broken down into small time series windows and these windows are compared with each other. Inorder to apply windowing techniques they should have same magnitudes and base line, so the data has to be normalised. Normalization does not change the original shape of the time series. After normalization, dimension reduction is a problem so we use PAA to overcome this. In PAA , the time series is divided into equal sized segments. The points within the each part is added and arithmetic mean of each part is calculated , representing a segment. Thus the time series is represented as a set of numbers, each representing mean of

points within it.

Figure 3.2: PAA representation of non linear time series[6]

3.3 SYMBOLIC REPRESENTATION:

The time series is now set of numbers after PAA. Inorder to obtain a symbolic representation of the time series, the amplitude of the time series is divided into intervals. Each interval is given a symbol. Normal distribution curve is applied across the vertical axis and equiprobable intervals are produced. Breakpoints are calculated such that equal area is produced. After break points has been fixed, PAA levels are fixed and each segment is assigned with a symbol. Each segment is assigned with a symbol. Thus the time series is represented as symbols.

Figure 3.3 SAX representation of the non linear time series[2]

SAX represents time series in the form of a string which helps in the detection of frequent patterns.

3.4 MOTIF DISCOVERY:

Motifs are basically recurring patterns. The process of extraction of unknown recurring patterns from a time series is known as motif discovery. Motif discovery is useful in areas such as telecommunications, medicine, web, motion-capture and sensor-networks. The first step in motif discovery is to extract frequent patterns from the large time series. As we are interested in symbolic motifs, SAX is used to obtain symbolic representation of the time series. Sliding window is used to extract the subsequence and it is converted into symbolic representation using SAX. No of times the sub sequences occur in the time series are calculated. According to the probability, the next day’s value is forecasted.

Extraction of each subsequence is done using sliding window and converted to symbols using SAX. Total no of times each subsequence occurring is determined. According to the probability of the sub sequences occurring for the whole time series predicts the next day’s value. For example , if the symbolic representation of the given time series is aabbacabadacabacacaaac then using sliding window the first subsequence is extracted , which is aa in this case. The occurrence of aa in the whole time series is calculated.

The occurrence is 2. Likewise the occurrence of each subsequence is determined and tabulated in TABLE 3.1.

From the table, we can observe that the occurrence of ac is higher than that of the other sub sequences. This shows that the occurrence of ac has higher probability than other sub sequences. Thus, tomorrow’s value can be forecasted using today’s value.

The user can decide whether to trade or not using this algorithm. If yesterday’s value is a and he want to predict today’s value then using this algorithm he would have found that today’s value will be c and he can trade. If the probability of occurrence of aa is higher, then he will not go for trading. Thus forecasting of the time series is done using motif discovery.

In order to get efficient results, optimization algorithms can be used. Optimization algorithm used is Genetic Algorithm. Genetic algorithm is used to optimize the cut points and alphabet range. Optimization algorithm is used to reduce the error percentage and produce effective results.

|  |  |
| --- | --- |
| MOTIFS | PROBABILITY OF OCCURRENCE |
| aa | 2/10 |
| ab | 2/10 |
| ac | 5/10 |
| ad | 1/10 |

Table 3.1: probability of occurrence of motifs

3.5 GENETIC ALGORITHM:

Genetic algorithm is a mathematical optimization algorithm which generates good solutions using a process called pseudo-darwinian process for real-world problems.GA begins with the populational unit of analysis and each member of the population generates a potential solution to the targeted problem.GA uses pseudo-natural selection process and pseudo-genetic operators which provides variation in the population to produce next generation. These are the basic steps in GA.

The main objective or target is to maximize the profit which is done by optimizing the cutpoints using GA. The problem can be viewed as a black box with few parameters as input and the output of the black box is a value which represents effectiveness combination of the parameters to solve the optimization problem. In GA, the interactions between the parameters are considered to maximize the output. The interaction between the variables is known as epistasis.

The main steps in GA are

1. Initialisation
2. Selection
3. Crossover
4. Mutation

GA can be characterised as,

x[t + 1] = r(v(s(x[t])))….3.1

Where x[t] is the population of encodings at iteration t

v (.) is the random variation operator (crossover and mutation)

s (.) is the selection for mating operator and

r (.) is the replacement selection operator.

Key steps in GA:

1. Construct initial population
2. Decode each string into solution and determine the fitness of each solution.
3. Choose a selection process and select the parent chromosomes from the population which has better fitness.
4. Select a crossover process, to generate two new child solutions from the parent chromosomes.
5. Apply a mutation process to selected bit of the child solutions.
6. Store the new solutions to the new population.
7. Continue steps 3-7 until n new solutions are formed and replace the old population with new population.
8. Go to step 2 and repeat the process until the desired output is reached.

The selection process used is stochastic line selection to the best half of the population and then selects the parent chromosomes. Crossover and mutation techniques are used to generate off springs.

Crossover specifies how the two parent chromosomes combine to form child for next generation. Scattered crossover technique is used where the ones from first parent is combined with the 0’s of second parent and forms the child. Default crossover selection rate is 0.8, default mutation technique is Gaussian. Population size used is 40 and two generation has been obtained. The main objective of using GA is to optimize the parameters. Best chromosomes received from the application are tested to evaluate the results. At the end of time series, new population is generated with best individuals and new offspring’s. The stop criteria are the fixed number of generations.

Generate n population

Decode each string to solution and Determine fitness

Selection technique

Apply crossover and mutation techniques

Select two parent chromosomes

Replace existing population with new population

Figure 3.4 flowchart of genetic algorithm

3.6 HP FILTER:

Hodrick Prescott FILTER is an algorithm used to smoothen the time series. Hodrick Prescott FILTER decomposes the time series into trend and cyclic component. Cyclic component is the difference between the original time series and the trend component. This explained in the below equation,

yt = τt + ct

where τt is constructed to minimize:

(Yt—𝞽t)2 +𝞴[(𝞽t+1-𝞽t)-(𝞽t-𝞽t-1)]2

The first term is the sum of the squared deviations of yt from the trend. The larger the value of the positive parameter λ, the smoother the resulting trend will be.

If, e.g., λ = 0, then τt = yt , t = 1,…,t.

If λ→∞, then τt is the linear trend obtained by fitting yt to a linear trend model by OLS.

Hodrick and Prescott suggest that λ = 1600 is a reasonable choice for quarterly data

**CHAPTER 4**

**CONCLUSION**

**Conclusion:**

An effective algorithm has been developed that uses SAX and Genetic algorithm in an efficient way to predict patterns in financial time series and suggest the possible days of trade. The algorithm has been able to predict whether trading should take place on the next day based on today’s closing value.

**Future Work:**

Replace the genetic algorithm with other optimization algorithms and compare the results.

**CHAPTER 5**

**REFERENCES**

**REFERENCES:**

[1] Lkhagva, Battuguldur, Yu Suzuki, and Kyoji Kawagoe. “Extended sax: Extension of symbolic aggregate approximation for financial time series data representation." In *Proceedings of Data Engineering Workshop,* pp. 4A-i8, 2006.

[2] Soon, Lay-Ki, and Sang Ho Lee. "An empirical study of similarity search in stock data," *Australian Computer Society, Second International Workshop on Integrating AI and Data Mining, AIDM*. 2007.

[3] Kirshners, A. K., S. V. Parshutin, and A. N. Borisov. "Combining clustering and a decision tree classifier in a forecasting task," *Automatic Control and Computer Sciences* 44.3, pp.124-132. 2010.

[4] Sadiq Hussain and Prof.G.C. Hazarika. “An Effective Algorithm for Clustering Time-Series Data by Fuzzy Logic,”International Journal of Computer Science and Technology. Vol. 2, Issue 4, pp.283-288.Oct. - Dec. 2011

[5] NunoCastro and Paulo J. Azevedo. “ TIME SERIES MOTIFS STATISTICAL SIGNIFICANCE,” Semana da Escola de Engenharia October 24 - 27, 2011.

[6] Keogh, Eamonn, and Jessica Lin. "Clustering of time-series subsequence is meaningless: implications for previous and future research," *Knowledge and information systems,* pp.154-177, 2005

[7] Jason R. Chen. “Making Subsequence Time Series Clustering Meaningful,” Proceedings of the Fifth IEEE International Conference on Data mining .ICDM.2005

[8]Nguyen Quoc, Viet Hung and Duong Tuan Anh. “Combining SAX and Piecewise Linear Approximation to Improve Similarity Search on Financial Time Series,” International Symposium on Information Technology Convergence.pp.58-63.2007

[9] Warren Liao, T. "Clustering of time series data—a survey," *Pattern Recognition*38.11, pp.1857-1874, 2005.

[10] Robin, S. and Schbath, S. and Vandewalle, V. 2007. “Statistical tests to compare motif count exceptionalities”. *BMC bioinformatics*, 8(1):84.

[11] [António Canelas](http://www.sciencedirect.com/science/article/pii/S0957417412010561) , “ pattern discovery – a SAX-GA Based Investment Strategy**”,** [*Expert Systems with Applications*](http://www.sciencedirect.com/science/journal/09574174), [Volume 40, Issue 5](http://www.sciencedirect.com/science/journal/09574174/40/5), Pages 1579–1590, April 2013.

[12] Lin J, Li W, Keogh E, Lonardi S, “Experiencing SAX: a novel symbolic representation of time series**”,** [*Data Mining and Knowledge Discovery*](http://link.springer.com/journal/10618), Volume 15, Issue 2, pp 107-144, October 2007.

[13] Chiu B, Keogh E, Lonardi S, “Probabilistic discovery of time series motifs”. In: Proceedings of

the 9th ACM SIGKDD International conference on knowledge discovery and data mining,

Washington DC, USA. August 24–27, pp 493–498.

[14] Barbazon A, O’Neill M, Biologically inspired algorithms for financial modelling.springer,2006.

[15] Lkhagva, Battuguldur, Yu Suzuki, and Kyoji Kawagoe. "Extended sax: extension of symbolic aggregate approximation for financial time series data representation." In *Proceedings of Data Engineering Workshop*, pp. 4A-i8, 2006.