Time Series Representation for Identification of Extremes

Rajesh Kumar
943A/28, Bharat Colony, Rohtak, Haryana

ABSTRACT
Extracting information from the huge time series is a challenging task. Databases are prepared by keeping in mind the type of information required. Indexing a time series is a difficult task, where shape may not be exact. Finding the minima and maxima of the time series is another difficult job dependent on the expert's subjectivity. In this information age algorithmic trading is the buzz word. For the identification of various patterns, a machine can be made intelligent by embedding some good algorithm in the trading module to identify the various patterns. In this paper an attempt has been made to identify the extremes, where profit probability is maximized.

Keywords
SAX, time sequence, DFT, DWT

1. INTRODUCTION
A time sequence is a sequence of values produced by a process. All these values obtained from a particular process have a time stamp. Sampling frequency is decided on the basis of change in the values of time stamp. Time series are ordered by time hence all the values obtained by a process must be preserved on the time basis. Time series databases are prepared in keeping the mind whether the query on the time series will be of shape based or value based .Objective of shape query is to find the shapes from the time series whose shape is similar to the training data set shape. Mammoth size time series makes it difficult to extract information from the data; hence important features are extracted so that information is not lost. Feature extracted from the time series is dependent on the types query to be made on the time series. Searching may be approximate. Discrete fourier transformation is the common method in which signal is viewed as a frequency domain instead of values by keeping the few frequency components where as noise and irrelevant attributes are dropped. In features extraction using times derivatives, the difference of signal values with the point and the next point values, this value is then mapped to a symbol from an alphabet. This alphabet is called a shape description alphabet. A signal described as a symbol string will have an envelope and all signals within that same symbol string [1].

Table1. Example of a shape description alphabet [1].

Symbol	Meaning	Definition
A	High increasing transition	d/dt >5
B	Slightly increasing transition	2< d/dt < 5
C	Stable transition	-2 < d/dt <2
D	Slightly decreasing transition	-5 < d/dt < -2
E	Highly decreasing transition	d/dt <-5

Derivatives are the efficient means for the noise shaping [3]. First derivative of any function can be interpreted as the rate of change of signal where as the second derivative is the rate of change of slope or curvature [2]. Principal component analysis is a good candidate of essential features extraction methods. All components having Eigen values greater than one are considered essential and rest are dropped without losing the essential information [2]. In features extraction using polynomial, A sequence is broken into numerous segments and each segment is approximated with a signal. In first degree of approximation start trend, end trend, slope and constant are identified to form the equation y= mx +b. [4]. Researcher's interest remains in the time series data mining task of indexing, clustering, classification, summarization and anomaly detection. Indexing is finding the most similar time series using some similar or dissimilar measure. Clustering a time series is grouping on some similarity distance measure [5]. Classification is assigning a time series to a predefined class [5]. Summarization is the keeping the essential features without destroying the information [5].Various methods of features extraction have been discussed above. Anomaly detection is the finding of all surprising, unexpected behavior in the time series which deviates from the normal behavior [5].

Rest of the paper is organized as follows

Section2 covers symbolic time series representation.

Section3 covers discrete fourier transformation.

Section4 covers proposed model.

Section5 covers data analysis.

Section6 covers conclusion.

Section7 covers references.

2. SYMBOLIC TIME SERIES REPRESENTATION [5]
A Symbolic aggregate approximation is one of the method of symbolic time series representation, ideally suited for the time series where lower bounding of the distance is required. SAX representation is taking a real valued signal and that signal is distributed among the various sections as displayed in fig. 1. By substituting each section with its mean a reduced dimensionality, piece wise constant approximation is obtained. For a long time series, A single SAX word is not obtained but a shorter window or feature window is

obtained[8].Sax representation is a discrete time series representation and chaos game bitmap representation can be used for visualizing discrete sequence. SAX has the potential to build word of any size but it has been found out that cardinality of four is best [5]. Initial ordering of four SAX symbols are obtained as given in Figure 2. Frequencies of the subsequences are counted from the SAX words. Streams are not mixed in order to count the frequency of sub words. Once the raw count of sub word of desired length have been obtained and recorded in the pixels, then normalization is done by dividing the every number in cells by the largest number. Colors are identified from Zero to one and values in the cells are mapped with a color and the time series takes the form of a bitmap. Anomaly detection is done by the creation of two windows lag window and the lead window. Lead window looks ahead in the time series for anomalous pattern. Lag window size is determined by the fact that the how much memory of the past is required. Each window bitmap is generated using the SAX representation. Distance between two bitmap is measured and as reported.

3. DISCRTE FOURIER TRANSFORMATION.

Discrete fourier transformation is the projection of a time signal from time domain into frequency domain and can be given by equation 1.

$$C_f = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} f(t) exp^{\frac{-2\pi i f t}{n}} \qquad \text{Eq. 1}$$

Where f= 1 ,2… n, and i=$\sqrt{-1}$

C_f are complex numbers and represent the amplitudes and shifts of a decomposition of signal into sinusoidal functions. It measures global frequencies and assumes to be periodic. Fast fourier algorithm has the time complexity O(n log n) [9]. If the problem is to index the collection of smaller sequences, discrete transformation is applied to each sequence and two to four components are selected to represent the entire sequence. These components are our feature factor. Since very little information is extracted hence compression will be easier and the pattern will be matched easily. For finding the sequences from the large sequences a sliding window of length n is selected. Initially window is placed over the first n values of the time series. The feature vector is extracted from the n values using DFT. This process is repeated for the rest of the time series [4]. Discrete wave transform measures frequency at different time. Signal is projected into time frequency plane. Basis function is given in equation 2.

$$\varphi_{j,k}(t) = 2^{\frac{j}{2}} \varphi(2^j t - k) \qquad \text{eq. 2.}$$

Where φ is the mother wavelet function. Any square integrable real function f(t) can be represented in terms of this bases as[9].

$$f(t) = \sum_{j,k} c_{j,k} \varphi_{j,k}(t) \qquad \text{eq. 3.}$$

Commonly used wavelet is the HAAR wavelet with mother function.

$$\varphi Haar(t) = \begin{cases} 0 \; if \; t \; greater \; than \; 0 \; , less \; than \; .5 \\ -1 \; if \; t \; greater \; than \; .5, less \; than \; 1 \\ 0 \; otherwise \end{cases}$$

eq4.

4. PROPOSED MODEL .

In time series, order of values dependent on time is mandatory. An attempt has been made to identify the maxima or minima. In this model for identification of minima or maxima, Price and Volume weighted moving averages of ten days is used on weekly basis taken from the NSE tame software. In the previous works, it was found out that Fibonacci series can predict the time period when peak or trough will occur. Fibonacci series can be obtained recursively by $f(n) = f(n-1) + f(n-2)$.It can be interpreted as when price crosses the VWMA in uptrend then peak will occur either in 1,2,3,5,8,13…. time slots. It is difficult to predict the accurate time period because of infinite series. Similarly in down trend when prices goes down from the VWMA, trough will occur in 1,2,3,5,8,13….. time slots. Same dilemma exists in finding both minima and maxima because Fibonacci series is not restricted to one number. In this model all the prices of NIFTY are measured on weekly basis and trend reversal occurs when $p_t > v_t$ and in previous time slot $v_t > p_t$ was true. Where p_t is price at time t and v_t is the vwma at time t. Similarly another trend reversal occurs when $v_t > p_t$ and in previous time slot $p_t > V_t$ was true. Proposed algorithm for the finding minima, maxima, local minima and local maxima is given below.

Step 1.Δp is calculated as ($price_t - price_{t-1}$) , Δ v is calculated as ($vwma_t$ - $vwma_{t-1}$), $\Delta^2 p$ is calculated as $\Delta p_t - \Delta p_{t-1}$ and $\Delta^2 v$ is calculated as $\Delta v_t - \Delta v_{t-1}$.

step 2. $\Delta p / \Delta v$ is calculated, if ratio is not near to zero. Then go to step 4.else go to step 3.

Step 3. If $p_t > v_t$ then

Step 3.1.1 If $\Delta^2 p$ and $\Delta^2 v$ are of different signs then maxima has occurred.

Step 3.1.2 If $\Delta^2 p$ and $\Delta^2 v$ are both negative and $\Delta^2 p / \Delta^2 v$ is greater than 15 then maxima has occurred.

Step 3.2. If $\Delta^2 p$ and $\Delta^2 v$ are of same signs then local maxima has occurred.

Else if $p_t < v_t$ then

Step 3.3 If $\Delta^2 p$ and $\Delta^2 v$ are of different signs then minima has occurred.

Step 3.4. If $\Delta^2 p$ and $\Delta^2 v$ are of same signs then local minima has occurred.

Step 4. Continue with step 1 for next time slot iteration unless the trend reverses.

5. DATA ANALYSIS.

In the proposed model, model calculates $\Delta p / \Delta v$ at every time slot. Where ever $\Delta p / \Delta v$ is near zero is identified, it is further analyzed to find the maxima and minima. A comparison is made between the first maxima signal value and the actual maxima value. It has been observed that signal occurred just near the extremes and gave very few signals of extremes even on very large time dependent financial series which can be verified by the results given in table 2,table 3,table 4,table 5.

6. CONCLUSION

Time series produces huge amount of data and it is very difficult to save the complete time series data and to retrieve the information accurately due to its mammoth size, Representation schemes solves the problem without losing the

essential information. Symbolic time representation and discrete fourier transformation keeps the essential component to facilitate the indexing of time series. These representation schemes fails to identify the extremes of the time sequence, hence a derivative method is used in identification of extremes. It has been found out that extremes occur when $(p_t-p_{t-1})/(v_t-v_{t-1})$ is near zero and $(\Delta p_t-\Delta p_{t-1})$ and $(\Delta v_t-\Delta V_{t-1})$ are of different sign. Maxima may encounter soon even if $(\Delta p_t-\Delta p_{t-1})$ and $(\Delta v_t-\Delta V_{t-1})$ are of same negative sign provided ratio is very high.

Table2. Signal of maxima and local maxima.

Time series	**Purchase price at reversal**	**Price at signal**	**$\Delta p/\Delta v$**	**$\Delta^2 p$**	**$\Delta^2 v$**	**Prediction of model**	**Actual maxima**
1.	1418	1622	.25	-124	-5	Maxima	1756
		1599	0	21	-1	Maxima	
2.	1467	1471	-0.16	4	8	Local maxima	1517
		1509	-0.7	-54	-3	Maxima	
3.	1242	1312	-0.14	-38	-1	Maxima	1313
4.	1263	1378	.53	-33	2	Maxima	1405
5.	976	1067	.33	-15	7	Maxima	1112
6.	1123	1163	.57	-32	2	Maxima	1187
		1187	.75	-6	-1	Local maxima	
7.	990	1079	-0.7	-24	-4	Local max.	1089
		1089	-0.75	-28	2	Maxima	
		1086	.66	15	2	Local Maxima.	
8.	973	1138	.5	-12	6	Maxima	1138
9.	1553	1633	.08	-30	2	Maxima	2080
		1722	-0.6	-76	-3	Maxima	
		1872	.33	-20	-2	Local Maxima	
		2080	.72	-32	1	Maxima	
10.	2008	2082	.6	-64	-10	Local Maxima.	2154.
		2154	.85	-82	-3	Maxima.	
11.	2076	2219	.83	-46	3	Maxima	2361
		2361	.001	-49	-8	Local Maxima	
12.	5088	5290	.5	-12	24	Maxima	5304
13.	3917	4248	.49	-104	8	Maxima	4504
		4171	.78	178	3	Local Maxima	
		4475	-.73	-179	43	Maxima	
14.	4866	5318	-0.16	-128	-36	Local Maxima	5564
		5333	-0.5	44	-7	Maxima	
15.	5879	6074	.24	-103	-4	Maxima	6074.
16.	5850	6144	-.8	-138	2	Maxima	6307

Table 3. Minima and Local Minima

Time series	**Purchase price at reversal**	**Price at signal**	**$\Delta p/\Delta v$**	**$\Delta^2 p$**	**$\Delta^2 v$**	**Prediction of model**	**Actual minima**
1.	1602	1406	.25	94	2	Local Minima	1262
		1262	.76	13	9	Local minima	
		1275	-0.5	33	-1	Minima	
2.	1266	1172	.22	105	-1	Minima	1172
		1178	-.3	10	2	Local Minima	
		1239	.25	-67	4	Minima	
		1236	.13	-1	-10	Local minima	
3	1320	1139	.75	4	-5	Minima	1101
		1140	-0.26	-19	-12	Local minima	
4.	1129	1097	.6	43	-5	Minima	1067
		1096	.25	2	2	Local Minima	
5	1796	1560	.66	200	1	Local Minima	1488
		1521	-0.5	65	-2	Minima	
		1508	.41	-26	-7	Local minima	
		1491	.5	-4	1	Minima	
		1488	.09	14	-2	Minima	
		1533	-0.7	-33	16	Minima	
6.	2015	1967	-0.75	82	-6	Minima	1967
		1988	-0.5	-64	12	Minima	
7.	3246	2890	-0.4	249	3	Minima	2890
8.	3938	3718	.2	204	-9	Minima	3708
9.	5705	4745	.15	426	-55	Minima	4573
10.	4228	2973	-0.6	-213	10	Minima	2693
		2755	-0.4	342	99	Local. Minima	
		2714	.3	-103	24	Minima	
11.	5036	4844	-0.6	-90	-4	Minima	4844
12.	5482	5059	-.02	-274	-30	Local Minima	4888
		5084	-0.4	6	17	Local Minima	
		4888	.93	-131	-10	Local Minima	

Table4. Percentage error between predicted maxima and actual maxima

Purchase Price at trend reversal.	Predicted Maxima at First Signal (e)	Percentage of return at first signal.	Actual Maxima	Percentage of Return on actual maxima.	Percentage error.
1418	1622	14.3	1756	23.8	9.5
1467	1509	2.8	1517	3.4	.54
1242	1312	5.63	1313	5.71	.08
1263	1378	9.10	1405	11.24	2.13
976	1067	9.3	1112	13.93	4.63
1123	1163	3.56	1187	5.69	2.13
990	1089	10	1089	10	0
973	1138	16.95	1138	16.95	0
2008	2154	7.27	2154	7.27	0
2076	2219	6.88	2361	13.72	6.84
5088	5290	3.97	5304	4.24	.27
3917	4248	8.45	4504	14.9	6.45
4866	5333	9.59	5564	14.34	4.75
5879	6074	3.31	6074	3.31	0
5850	6307	7.811	6307	7.811	0

Table5. The percentage error between predicted minima and actual minima

Purchase Price at trend reversal.	Predicted Minima at First Signal	Percentage of price gone down at first signal	Actual Minima	Actual percentage of price gone down	Percentage Error.
1602	1275	20.41	1262	21.2	.79
1266	1172	7.42	1172	7.42	0
1320	1139	13.71	1101	16.59	2.88
1129	1097	2.83	1067	5.49	2.66
1796	1521	15.31	1488	17.14	1.83
2015	1967	2.38	1967	2.38	0
3246	2890	10.96	2890	10.96	0
3938	3718	5.58	3708	5.84	.16
5705	4745	16.82	4745	16.82	0
4228	2973	29.68	2693	36.30	6.62
5036	4844	3.81	4844	3.81	0

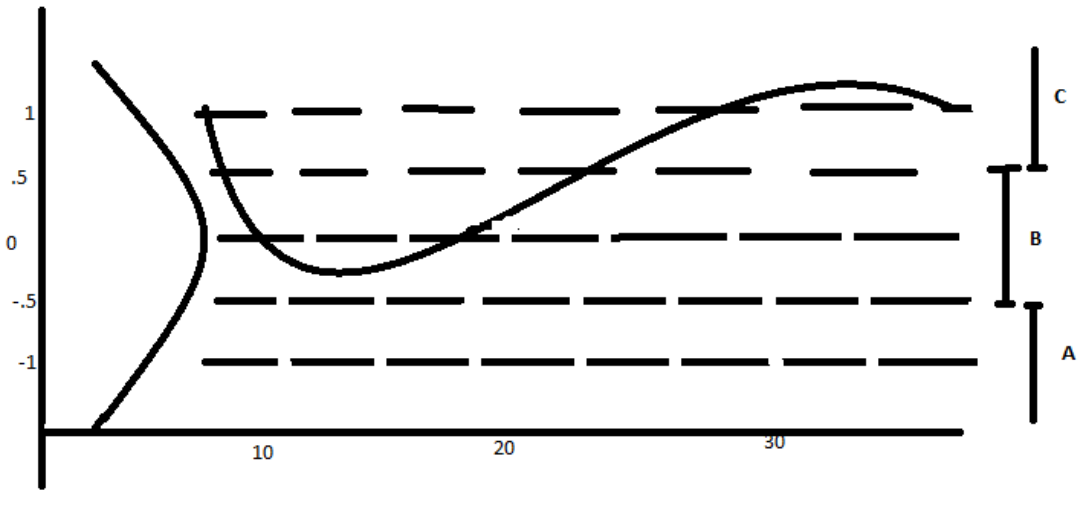

Fig1. SAX word conversion from real valued time series

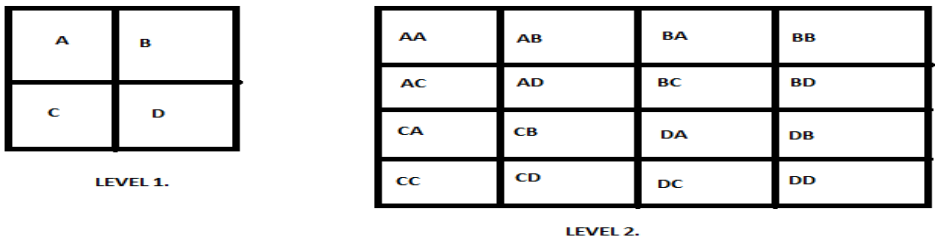

Fig2. Quad tree representation of a sequence over the alphabet (A,B,C,D)

7. REFERENCES

[1]R Agarwal, G .Psaila, E.l. Wimmers, M.Zait, 1995,Querying shapes of histories, in proc. Of the21st international conference on very large databases, pp 502,514.

[2] S. E Paraskevopoulou, D.Y Barsakcioglu, M. R Saberi, Amir Eftekhar, T.G. Constandinou, Dec 24, 2012,Feature Extraction using First and Second Derivative Extreme (FSDE) for Real-time and Hardware-Efficient Spike Sorting, Journal of neuroscience method pp 1-12

[3] Yang Z, Zhao Q, Liu W,2009, Improving spike separation using waveform derivatives. Journal of Neural Engineering 2009;6(4):2–12.

[4] H. Andre Jonsoon,2002,Indexing strategies for time series data, Linkoping studies in science and technology, diss. no 757,ISBN 91-7373-346-6.

[5] J.Lin,E. Keogh, S Lonardi, B.Chiu, june 13,2003,A symbolic representation of time series with implications of streaming algorithms. DMKD, San Diego C.A

[6]L. W. N.Kumar ,Venkata Loilla, E. Keogh, S. Lonardi ,C. Ann,Ratanamahatana, Assumption-Free Anomaly Detection in Time Series, partly funded by the National Science Foundation under grant IIS-0237918.

[7] J.lin,E. Keogh,S. Lonardi,J.P. Lankford,D.M Nystorm,2004, Visually Mining and Monitoring Massive Time Series, in proceeding of 10th ACM SIGKDD.

[8] Kumar, N.Lolla N., Keogh, E. Lonardi, S.Ratanamahatana, C. & Wei, L. 2005. Time-series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases. SIAM 2005 Data Mining Conference.

[9] R.Agarwal,C. Faloustos,A. Swami,Efficient similarity search in sequence databases,funded by national science foundation, grant no 895846.

[10] D. Abadiand ,Aurora, A data stream management system . In SIGMOD 2003.

[11] S. Guha, N. Kudos , Approximating a data stream for querying and estimation: algorithms and evaluation performance In ICDE 2002.

[12] F. Rasheed, M. Ashaalfa,R. Alhajj,2011, Efficient Periodicity Mining in Time Series Databases Using Suffix Trees, Ieee transactions on data engineering, vol. 23, no1, page 79-94