Mining the Time Series for Financial Gain

Rajesh Kumar
943A/28,Bharat colony, Rohtak,Haryana,India.

ABSTRACT

Control charts are widely used in finding the process out of control. In the context of financial time series ,change points occurrence is dependent on the sentiments of the traders, hence identification of change point in the financial time series is generally subjective. In this information age, emphasis is on the algorithmic trading where machine has to take trading decisions. In this paper a model is proposed which will take in to the consideration the sentiments of traders, hence volume weighted moving average of ten days is used in identification of sell or purchase signal. Results of the model has been taken in the consideration of worst case, only the closing prices of the month is recorded and trading decision is taken on the restricted data.

Keywords

Timeseries, cusum charts, control charts.

1. INTRODUCTION

Time series is a sequence of data points at a regular interval or it can be interpreted as sequence of values obtained from a process over a time[2]. This kind of data stream can be financial or real valued meteorological data or any other data produced at a regular interval . Data volume is huge, hence inference is taken from a summarized data of weekly, monthly , yearly or on daily basis. Like data mining, time series analysis deals in extracting meaningful information from the huge data. Data of time series is different from other data because in times series temporal ordering of data is mandatory, where as in other sampling , temporal order is not necessary. Owing to the different nature of discrete stream data and real time series data. There is little intersection between stream data and data series data[1]. There is always interest of researchers in the time series domain of indexing, clustering, classification ,summarization and anomaly detection. In indexing a time series, relevant features are kept under study and rest are dropped. Some times it leads to fuzziness, even the result of a query are not matched optimally but the match is satisfactory[2]. In fast subsequence matching in time series data, distance preserving transformation is used to extract features from the time series. This is adopted in R tree[2]. Second class of methods of searching is tolerant to noise so it is using the fuzzy sets in approximate searching[2]. In querying shapes of histories, time series is transformed into text, then shape definition language is defined which can be used to define the query on text. In approximate queries , time series is approximated by mathematical function[2]. Clustering is the finding the similarity on the basis of similarity index[3]. Classification is the assignment of time series on a predefined classes[4].Summarizing is pruning the large time series by keeping the essential features to avoid the bottle necks[5]. Anomaly detection is the finding abnormal behavior at a point[6].

Rest of the paper is organized as follows.

Section 2. covers control charts.

Section 3. covers proposed model.

Section 4. covers data analysis.

Section 5. covers conclusion.

Section 6. covers references.

2. CONTROL CHARTS

A Statistical process control is to detect and eliminate non random variation as they arise when process is out of control. Control charts are the primary tool for statistical process control[7]. Generally it is assumed that process out of control will be observed as soon it occurs without any tool , but it is not true. A control chart consists of the following.

- Points representing a statistics of measurement like a mean, range, proportion at the different point of time. The mean of this statistic using all the samples is calculated. A centre line is drawn at the value of the mean of the statistic
- The standard error, which can be calculated by standard deviation/sqrt(n).Upper and lower control limits indicating the threshold at which the process output is considered statistically in control. Generally upper and lower limit is identified as 3 standard errors from the centre line

The chart may have the following optional features.

- Two standards above or below the central line.
- Division into zones, with the addition of rules governing frequencies of observations in each zone.

There are some possible risks involved owing to the possibility of sampling errors. In the operating characteristics curve , if more than C number is nonconforming, then entire shipment is rejected, other wise shipment is accepted. Number C is called the acceptance number[7]. For the large data set we should select the parameter n and C wisely, so that the acceptance or rejection avoids the type 1 or type 2 errors. Following Performance requirements are specified by the users[7].

AQL(Acceptable quality level or good quality).

LTPD(lot tolerance percent defective poor quality).

A trade off is to be made with the value of n and c, so that inspection is neither too stringent or not so loose. Following control charts are used to avoid the process that are about to be out of control[7].

2.1 X bar chart [7]

This chart has three important parameters, average of the past observation, upper control limit and lower control limit. According to central limit theorem sampling distribution approaches normality as the size of the samples increases, regardless of the distribution of the measurements of individual sample units. In X bar chart

$$UCL = X'' + KSx'$$

where Sx' is the standard deviation , K can be 1,2,3.which is selected on the precision requirements.

$$LCL = X'' - KSx'$$

where Sx' is the standard deviation , k can be 1,2,3, which is selected on the precision requirements. Trade off is to be made with α risks and β risks. In α risk, called the producers risk or type 1 error, model predicts that process is out of control but actually it was in control. Second type of risk is β called the consumer risk or type 2 error when incorrectly concluding that process is in control.

2.2. R Chart [7]

R chart is widely used because of its computational simplicity and in its availability in tabular form. Since parameters are available in tabular form it can be programmed easily. Upper control limit denoted by UCL, and lower control limit denoted by LCL is given below.

$$UCL = DR'$$

$$LCL = CR'$$

C,D are obtained from the table where R' is the mean of all samples ranges.

Investigation is required in the following circumstances [7].

- One data point outs above the UCL or below the UCL.
- Two data points near the UCL or two data points near the LCL.
- Five data points successively above the central line or five data points successively points below the central line.
- Five successive data points on increasing or decreasing line.

2.3. Cumulative sum control charts[9]

CUSUM charts are the sequential analysis for the change detection in time series. CUSUM charts are not easy to operate but they can be used in detecting the small shifts from the mean, so they are better than Shewhart control charts when it is desired to detect shifts in the mean that are 2 sigma or less[9]. CUSUM charts uses tabular data hence can be easily used in spreadsheet program. In this tabular calculation two parameters H , K are used to calculate the S_{HI} ,$S_{lo.}$

$$SHI(i) = Max(0, SHI(i-1) + xi - \hat{\mu}_0 - k)$$

$$SLO(i) = Max(0, SLO(i-1) + \hat{\mu}_0 - k - xi)$$

Initially $S_{HI}(1)$ =0, $S_{lo}(1)$=0. When $S_{HI}(i)$, $S_{lo}(i)$ exceeds a threshold then process is assumed to be out of control.

3. PROPOSED MODEL

In this model, initially $S_{HI}(1)$ =0, $S_{lo}(1)$=0. Then $S_{hi}(i)$, $S_{lo}(i)$ are calculated as follows.

$$SHI(i) = Max(0, SHI(i-1) + xi - \text{VWMA}(10) - k)$$

$$SLO(i) = Max(0, SLO(i-1) + \text{VWMA}(10) - k - xi)$$

Where VWMA is volume weighted moving average, k value is taken as .3175 and H=4. This is a recursive process. Previous value of S_{hi}, S_{lo} are used in calculating the next values of S_{hi} and S_{lo}. Purchase signal is invoked when there is a transition of S_{HI} from 0 to threshold value H. similarly sell signal occurs when there is a transition of Slo from 0 to threshold value H.

4. DATA ANALYSIS

S_{hi} and S_{lo} of twenty five securities is calculated, in the following table only the summarized observation where change point is observed are shown. * mark shown indicates the purchase signal where as ** indicates the sell signal. Since Shi(1)=0 and $S_{lo}(1)$=0. Iteration will be effective from i=2. Only two instances moved in opposite direction when closing prices of the months were accounted. Average percentage change within the three months was 9.75. The OC curve for c=3 is given in Fig 1., as per the OC curve, Probability of accepting with c or fewer defects is .88.

Table 1. Experimental data showing the change points. Where * displays the purchase signal and ** signal displays Sell signal determined by the model.

Stock Name	Month of Year	Price	VWM Avg. 10day	S_{HI}	S_{LO}
SRF	April ,10	84	94	0	71.37
	May	*123	95	27.69	43.06
	June	111.75	92.5	46.63	23.5
	July	139	97	88.32	0
Kajaria	April,9	27	30.9	0	12.67
	May	*36	31.9	3.79	8.26
	June	30.7	31.48	2.7	8.73
	July	38	33.95	6.44	4.37
SBI	December, 09	2811	2676	4236.66	0
	January,10	**2642	2715	4163.35	72.69
	February	2630	2739	4054.04	181.38
	March	2765	2784	4034.73	200.07
Gold Bees	Dec,12	2903	2867	316.07	0
	January,13	** 2879	2885	309.76	5.69
	February	2797	2900	206.45	108.38
	March	2806	2906	106.14	208.07
ONGC	April,2013	311	309	50.07	0
	May	**290	309	30.76	18.69
	June	249	304	0	73.38
	July	267	303	0	109.07

Gabriel	October,13	20	20	0	12.07
	November	*24	20	3.69	7.76
	December	24	21	6.38	4.45
	January	20	21	5.07	5.14
Hero	June,2013	1663	1766	0	139.38
	July	*1819	1762	56.69	82.07
	August	2046	1792	310.38	0
	September	2009	1811	508.07	0
IDBI	January,13	107	98	21.38	0
	February	**87	97	11.07	9.69
	March	80	97	0	26.38
	April	88	97	0	35.07
IFCI	October,12	27	35	0	0
	November	**31	36	0	4.69
	December	33	34	0	5.38
	January	35	33	1.69	3.07
ACC	January,13	1323	1309	132.38	0
	February	**1159	1317	0	157.69
	March	1223	1319	0	253.38
	April	1218	1310	0	345.07
Aban	October, 13	253	272	0	93.38
	November	*377	316	60.69	32.07
	December	388	334	114.38	0
	January	448	366	196.07	0
Infy	June,13	2498	2525	0	26.69
	July	*2969	2574	394.69	0
	August	3105	2648	851.38	0
	September	3013	2694	1170.07	0
Tata Motors	August	136	142	0	16.38
	September	*161	145	15.69	0.07
	October	156	149	22.38	0
	November	163	151	34.07	0
Apollo Tyres	January,13	221	197	23.69	0
	February	**185	199	9.38	13.69
	March	167	197	0	43.38
	April	170	197	0	70.07
Unitech	March,12	28	27	0.69	0
	April	**26	27	0	0.69
	May	21	26	0	5.38
	June	22	26	0	9.07
NIFTY	March,09	3020	3245	0	1547.07
	April	*3473	3225	247.69	1298.76
	May	4478	3319	1406.38	139.45
	June	4241	3414	2233.07	0
Gammon	September ,13	**47	47.96	0	0.65
	October	40.75	47	0	6.59
	November	38	45	0	13.28
	December	38	41	0	15.97
NTPC	October,12	165	164.88	0	0
	November	**162	163	0	0.69
	December	156	161	0	5.38
	January	157	160	0	8.07
OIL	June,13	574	544	29.69	0
	July	**517	543	3.38	25.69
	August	434	537	0	128.38
	September ,	437	530	0	221.07
Praj Inds	March,13	80	79	2.38	0
	April	**70	78	0	7.69
	May	56	75	0	26.38
	June	58	74	0	42.07
KRBL	August,10,	24	24	0	3.38
	September	*24	23	0.69	2.07
	October	30	24	6.38	0
	November	31	25	12.07	0
GSPL	December, 10	117	105	11.69	0
	January	**102	107	6.38	4.69
	February	88	106	0	22.38
	March	99	107	0	30.07
Crisil	May,13	941	962	0	20.69
	June	*1109	1004	104.69	0
	July	1183	1069	218.38	0
	August	1112	1081	249.07	0
Nestle	August,12	*4635	4460	174.69	0
	September	4604	4477	301.38	0
	October	4700	4530	471.07	0
	November	4767	4573	664.76	0
ABB	March,11	796	803	0	159.38

	April	*854	801	52.69	106.07
	May	864	803	113.38	44.76
Cipla	May,13	370	386	0	0
	June	391	387	3.69	0
	July	400	394	14.38	0
	August	416	394	36.07	0

Table 2. Summarized result of proposed model.

Sample Size	Properly classified instances	Misclassified instances.	Percentage of instances with proper classification
25	23	2	92

Table 3.Probability of acceptance with acceptance number c=3 with a sample size n=25.[7].

Number of non conforming units in sample (n*p) ,where n is the sample size and p is fraction defective.	Probability of acceptance with c or fewer defects.
0	1
0.2	0.99996
0.3	0.9997
0.4	0.9994
0.5	0.998
0.6	0.997
0.8	0.995
1	0.98
2	0.88
3	.68

5. CONCLUSION.

The Volume Weighted Moving Average adds weight to a standard moving average based on the amount of volume in a given period of time. The idea behind volume weighted moving average is that the price should be given more weight in times of heavy trading activity, hence it takes in to consideration the sentiments of the traders. There are many ways of identification of turnaround of a stock based on the sentiments of the traders like candle stick charts. Main drawback of candlestick charts is its subjectivity. To avoid the subjectivity and considering the traders sentiments, proposed models is a suitable choice because it gives a probability of accepting with C =3 or fewer defects is .88, when closing prices of months is taken into consideration. Model will give even better results when opportunities for profit booking are searched on weekly or daily basis after taking a position in a stock.

6. REFERENCES.

[1] Keogh,E & Kasetty, July,200,. On the need for time series data mining benchmarks: *A survey and empirical demonstration,* In the proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, Alberta, Canada, pp 102-111.

[2] Henrik Andre Jonsson, 2002,Dissertation no 757, Indexing strategies for time series *data,* Department of computer and information science, Linkoping university ,Sweden,issn-0345-7524.

[3] Kalpakis, K.Gada, Puttagunta, 2001, Distance measure for effective clustering of ARIMA time series . In the proceeding of the 2001, IEEE international conference on data mining ,San Jose ,CA, Nov-29,dec 2, pp273-280

[4] Geurt,P,2001, Pattern extraction for time series classification , In proceedings of the 5th European conference on principles of data mining and knowledge discovery,sept3 -7,Germany, pp 115-127.

[5] lin,J, Keogh,E, Lonrdi, S & Patel, 2002, Finding motifs in time series ,In the proceedings of the 2nd workshop on temporal data mining at the 8th ACM SIGKDD, International conference on knowledge discovery and data mining ,Alberta, Canada, PP 53-68

[6] Dasgupta, D & Forrest,1996, Novelty detection in time series data using ideas from immunology, in the proceeding of the international conference on intelligent systems.

[7] Everette E Adam,Jr. Ronald J.Ebert,1995,production and operations management ,5th edition, EEE,PHI,ISBN-81-203-0838-7.

[8] S. Chatterjee,Peihua Qiu,2009, Distribution free cumulative sum control chart using bootstrap control limits,The Annals of applied statistics, volume 3, no 1,349-369,DOI:10.1214/08-AOAS197.

[9] Pages downloaded from URL http://en.wikipedia.org/wiki/CUSUM

[10] Duda,R.O,Hart,1973,Pattern classification and series analysis, Wiley, New York.

[11] Ge,X & Smith,2001, Segmental semi markov models for end point detection in plasma etching, IEEE transaction on semi conductor engineering.

[12] Agrawal, R,Psaila,G ,Wimmers,E.L Zait,1995, Querying shapes of histories , In the proceedings of the 21st international conference on very large databases, Zurich, Switzerland, Sept 11-Sept 15, PP 502-514.

[13] Staden ,1989, Methods for discovering novel motifs in nucleic acid sequences, Computer application in bio sciences, Vol 5, PP293-298.

[14] Guha, Mishra, Motwani, Callaghan,2000,Clustering data streams, In the proceedings of the 41st symposium on foundations of computer science, Nov 12-14,Rodondo beach, C.A, PP 359-366.

[15] Huang ,Yu,P.S, 1999, Adaptive query processing for time series data. In the proceedings of the 5th international conference on knowledge discovery and data mining, San Deigo, C.A Aug 15-18,pp 282-286.

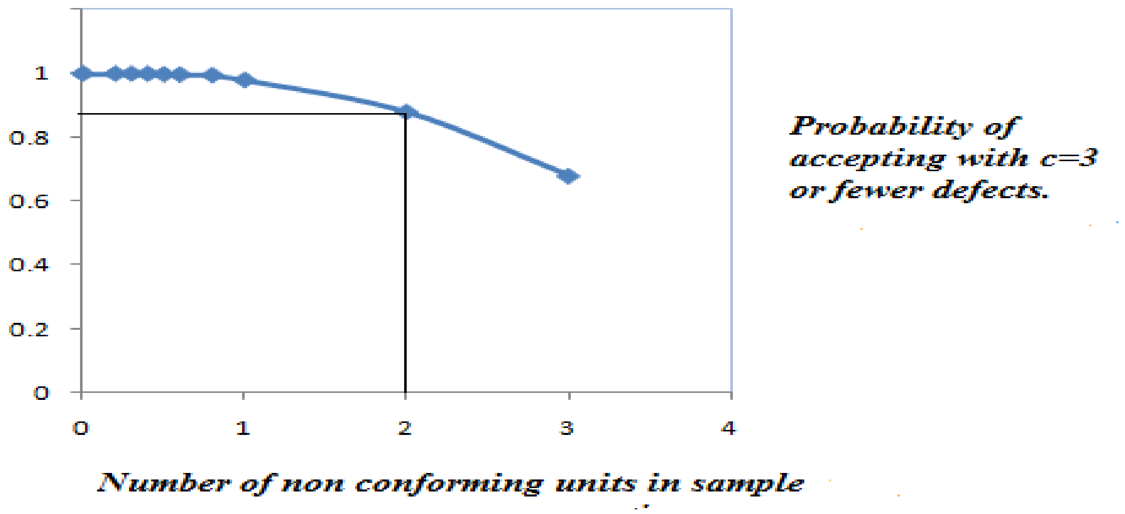

Fig 1. Operating Characteristics Curve for C=3.