Data Mining to Facilitate the Trading

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

ABSTRACT
Financial sector is always full of insecurity, owing to volatility in the financial sector, most of the investors fails to book the profit. It has been observed in this study that maximum percentage of return of a security or indices follows the Benford's law when price of the security or indices breaks the volume weighted moving average in upper trend. Results of this study can be used by investors in taking intelligent decisions. This model can be used in machine learning, which can facilitate the investors in the decision support system.

Keywords
Moving Averages, Data mining, Classification, Genetic algorithm, Benford's law

1. INTRODUCTION
Globalization and information technology revolution has created a huge amount of data. To tap the potential of this data, data mining techniques are required. As data sets have grown in size and complexity, the need of special tools like neural networks, genetic algorithms , decision trees and support vector machines emerged. These tools has helped the data analyzers in taking a wise and informed decisions. Data mining is the process of analyzing the data in a different perspective and summarizing it so that it can be a useful information, with the intention of uncovering hidden patterns from a large data sets[6]. It uses the statistical methods and artificial intelligence algorithms in indexing and storing of data bases so that information retrieved from it can be rationalized with an efficiency. Knowledge discovery in databases field is concerned with the development of methods and techniques for making sense of data[7]. Data mining has been applied to a number of financial applications including development of trading models, investment decision, loan assessment, portfolio optimization, fraud detection, bankruptcy prediction, real-estate assessment, and so on. The competitive advantages achieved by data mining include increased revenue[9]. The goals of data mining are briefly described in section 1.1,1.2,1.3 [8].

1.1 Classification
In the data analysis , it is essential to put the instances under the test in a desired class. It categorizes the instance in a particular category. The ability of a classifier refers to the ability to correctly classifying the unseen data in a class by taking a clue from the training data set Following methods are being used in classification.

1.1.1 Rule based methods
Data mining system learns from examples. It formulates classification rules in order for the prediction of future. For instance, in customer database in a bank, a query is made whether a new customer applying for a loan is a good investment or not? Typical rule are as follows which may be produced by rule based systems .

if STATUS = married and INCOME > 10000 and HOUSE_OWNER and has a VEHICLE =yes then INVESTMENT_TYPE = good.

1.1.2 Neural Network
Neural network can be used in the classification purpose. They simulate the human brain . Artificial Neuron can be supervised or unsupervised. They are composed of many units called neuron. Artificial neuron require long training time and are black box which lacks explanation, but it has high tolerance to noisy data so it can classify untrained data . Being the tolerant to noisy data, neural network are widely used in industrial applications.

1.1.3 Bayesian classification
Bayesian classification predicts class membership using Bayes theorem, which further uses probability. Its performance is comparable to selected neural network and decision tree. They can facilitate decision making even on computational intractable problems.

1.1.4 Support Vector Machine
Support vector machine can classify both linear and non linear data..Data from two classes are separated by hyper plane, Support vector machine finds the hyper plane by using training data. Its training is slow but accuracy is very high and SVM can model non linear problems also.

1.1.5 Genetic Algorithm
Genetic algorithm has taken a queue from the natural evolution. Initial population is created using randomly generated rules. Each rule is represented by a string of bits. In next generation, survival of the fittest selects the fittest rules. Crossover and mutation are used in production of offspring. In cross over substring of a rule are exchanged with substring of another rule. In mutation randomly selected bits are inverted. It being an iterative purpose, a rule will get position in next generation, if it crosses a threshold. Genetic algorithm can be used in classification besides optimization purpose.

1.1.6 Case Based Reasoning
Case based reasoning stores the old instances in a database to classify the unseen instances as equal to stored instance, if it does not exist than it search for another very similar instance.

1.2 Association
Rules that associate one attribute of a relation to another attribute approaches are the most efficient means of discovering such rules like in supermarket database. If a certain percentage of all the records that contain items A and B also contain item C .the specific percentage of occurrences is the confidence factor of the rule .Association rule mining is useful in mining single dimensional Boolean association rule from the transactional databases, it can be further extended for mining multilevel rule from the transactional databases.

1.3 Sequence/Temporal
Sequential pattern functions identifies the collections of related records and detects frequently occurring pattern over a period of time under study .Difference between sequence rules and other rules is the temporal factor. For example - Retailers database can be used to discover the set of purchases that frequently precedes the purchase of a microwave oven or harvesting season.

Rest of the paper is organized as follows

- Section 2 covers Moving Averages
- Section 3 covers Benford's law.
- Section 4 covers Data Analysis .
- Section 5 covers Conclusion.
- Section 6 covers References.

2. MOVING AVERAGES

In statistics, a moving average ,rolling average or running average is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. A moving average may also use unequal weights for each data value in the subset to emphasize particular values in the subset[10]. Exponentially-weighted moving average tracks of all prior sample means. WMA weights samples in geometrically decreasing order so that the most recent samples are weighted most highly while the most distant samples contribute very little. In exponential weighted moving average smoothing scheme begins by setting S2 to y1, where Si stands for smoothed observation or EWMA, and y stands for the original observation. The subscripts refer to the time periods, 1, 2, ..., n. For the third period.

$$s3 = \propto y_2 + (1-\propto)s_2 \quad (1)$$

and so on. There is no S1; the smoothed series starts with the smoothed version of the second observation.

$$s_t = \propto y_{t-1} + (1-\propto)s_{t-1} \qquad 0 \propto \leq 1 \quad t \geq 3 \quad (2)$$

This is the basic equation of exponential smoothing and the constant or parameter $\propto$ is called the smoothing[12]. The speed at which the older responses are dampened (smoothed) is a function of the value of $\propto$. When $\propto$ is close to 1, dampening is quick and when $\propto$is close to 0, dampening is slow [12].

3. BENFORD LAW

Initially it seems that digits are equally likely to distribute in a number that forms an observations , but this conception was wrong and demystified by Benford law [15]. Benford law states that the probability of any digit D from 1 to 9 being the first digit is where distribution is not uniform is given by

$$log_{10}(1 + \frac{1}{D}) \quad (3)$$

Whereas probability at 2nd digit can be given by

$$\sum_{D1=1}^{9} \log_{10}\left(1 + \frac{1}{D1D2}\right) \quad (4)$$

Where D_2={0,1----9}

And probability of combination of 1st digit and 2nd digit can be given by the formula.

$$P(D1D2) = log_{10}(1 + \frac{1}{D1D2}) \quad (5)$$

Whereas D1D2={10,11------,99)

Table 1. Frequencies based on Benford's Law[13]

Digit	1st Place	2nd place	3rd Place	4th Place
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.19882	0.10097	0.1001
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.0994	0.09994
7	0.05799	0.0935	0.09902	0.0999
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.085	0.09827	0.09982

4. DATA ANALYSIS

Sixty five instances under the study were compared with ten days volume weighted moving average. Whenever price of an instance breaks the resistance of moving average, It has given the positive return and further it followed the Benford's law.

Table 2. Data of VWMA10 and security top price achieved.

Script	VWMA ten days	Top	Percentage Change
NIFTY	5750	6187	7.6
NIFTY	5921	6029	1.82401621
NIFTY	5673	6307	11.1757448
NIFTY	6088	6258	2.79237845
NIFTY	5453	5573	2.20062351
NIFTY	5627	5660	0.58645815
NIFTY	4975	5013	0.7638191
NIFTY	5068	5316	4.89344909
NIFTY	2807	3812	35.8033488
NIFTY	4373	4714	7.79785045
NIFTY	4921	5052	2.66206056
NIFTY	4998	5045	0.94037615
NIFTY	4995	5361	7.32732733
NIFTY	5145	5368	4.33430515
Reliance	800	835	4.375
Reliance	799	839	5.00625782

Reliance	867	881	1.61476355
Reliance	853	908	6.44783118
Reliance	878	895	1.93621868
Reliance	792	900	13.6363636
Reliance	731	863	18.0574555
Reliance	716	729	1.81564246
Reliance	729	816	11.9341564
eliance	808	912	12.8712871
Reliance	801	835	4.24469413
ONGC	315	340	7.93650794
ONGC	325	331	1.84615385
ONGC	282	289	2.4822695
ONGC	273	287	5.12820513
ONGC	278	299	7.55395683
ONGC	284	292	2.81690141
Kajaria	70	112	60
Kajaria	107	119	11.2149533
Kajaria	100	185	85
Kajaria	168	178	5.95238095
Kajaria	176	257	46.0227273
Kajaria	188	249	32.4468085
Kajaria	235	251	6.80851064
Kajaria	247	308	24.6963563
Gold ETF	1274	1334	4.70957614
Gold ETF	1258	1458	15.8982512
Gold ETF	1417	1471	3.81086803
Gold ETF	1068	1343	25.7490637
Gold ETF	1304	1391	6.67177914
Gold ETF	1338	1360	1.64424514
Gold ETF	1335	1399	4.79400749
Gold ETF	1400	1542	10.1428571
Gold ETF	1493	1519	1.74146015
Axis bank	1356	1536	13.2743363

Axis bank	1009	1218	20.7135778
Axis bank	1117	1279	14.5031334
Axis bank	928	1278	37.7155172
Axis bank	1015	1034	1.87192118
Axis bank	997	1110	11.334002
Axis bank	1026	1352	31.7738791
Axis bank	1256	1449	15.366242
Axis bank	1272	1337	5.11006289
Axis bank	1081	1103	2.03515264
Axis bank	367	778	111.989101
Axis bank	937	996	6.29669157
Asian Paints	2039	2847	39.6272683
Asian Paints	2652	2887	8.8612368
Asian Paints	2515	3167	25.9244533
Asian Paints	3015	3313	9.88391376
Asian Paints	3139	3263	3.95030264
Asian Paints	3700	3946	6.64864865

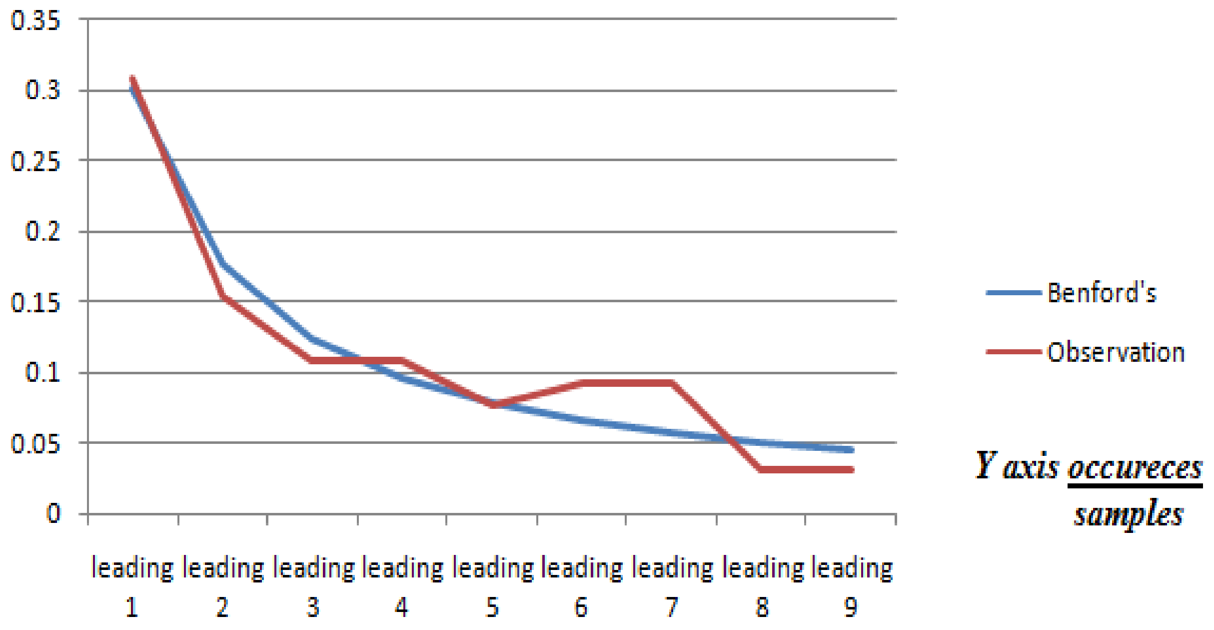

Fig. 1: Comparison of Benford's law with the returns leading digits occurrences/samples

5. CONCLUSION

Financial sector is always over shadowed by volatility. Investors has to take the returns by beating the volatility. To tap the potential of huge amount of data from the financial sector, this study can be very useful. Investor can use the heuristics provided by this study in taking the informed decisions that maximum probability of percentage returns provided by a security when it breaks the resistance of VWMA of ten days is a digit starting with one.

6. REFERENCES

[1] Rokach, Lior; Maimon, O. (2008 "Data mining with decision trees: theory and applications.", World Scientific Pub Co Inc. ISBN 978-9812771711.

[2]Varun Chandola et al,(2009)," Anamoly detection survey." , ACM Computing Surveys, Vol. 41(3), Article 15.

[3] Shyam Boriah, Varun Chandola and Vipin Kumar,(2008),"Similarity measure of categorical datas.", In Proceedings of SIAM Data Mining Conference, Atlanta, GA.

[4] Principles of Data Mining.(2007). doi:10.1007/978-1-84628-766-4. ISBN 978-1-84628-765-7.

[5]Varun Chandola, Shyam Boriah, and Vipin Kumar,(2009) In Proceedings of SIAM Data Mining Conference, Sparks.

[6]Kantardzic, Mehmed (2003)," Data Mining: Concepts, Models, Methods, and Algorithms.", John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

[7]Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,(1996) ,"To Knowledge Discovery in Databases." ,6, American Association for Artificial Intelligence. AI Magazine Volume 17 Number 3.

[8] Jiawei Han and Michelin Kamber ,(2006) "Data Mining Concepts and Techniques.",2nd edition, Morgan Kaufman.

[9] Dongsong Zhang and Lina Zhou,(2004) "Discovering Golden Nuggets: Data Mining,in Financial Application. ",IEEE Transactions on systems,man and cybernetics –Part C: Applications and review,Vol. 34, No. 4,

[10] pages downloaded from Engineering statistics handbook http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc431.htm

[11]Kalgonda , Koshti. , Ashokan.(2011)," Exponentially weighted moving average control chart.", Asian journal of management research,Vol2.

[12] Everette,Ronald ,(1995)", Production and operation management concepts models and",5th edition, Prentice Hall of India.

[13]Durtschi, Cindy and William Hillison and Carl Pachini. (2004) "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data", Journal of Forensic Accounting 1524-5586/Vol. V: 17-34.

[14]Nigrini, M. J. (1997),"Digital Analysis Tests and Statistics. ",Allen, Texas: The Nigrini Institute, Inc. Mark_Nigrini@classic.msn.com,1997

[15]Mark, j . Nigrini and Linda ,(1997),"The use of benford law as an aid in analytical procedure ",Auditing a journal of practice and theory ,vol16,no2.

[16]Agrawal, R., and Psaila, G. 1995. Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence.

[17]Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.;and Verkamo, I.(1996) Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds.

[18]Smyth, and R. Uthurusamy, (1993)" Detection of Abrupt Changes: Theory and Application.", Englewood Cliffs, N.J.: Prentice Hall, 514–560. Menlo Park, Calif.: AAAI Press. Basseville , M., and Nikiforov, I. V.

[19]Brachman, R., and Anand, T., (1996) "The Process of Knowledge Discovery in Databases: A Human-Centered Approach". In Advances in Knowledge Discovery and Data Mining, 37–58, Edition.

[20] Hall, J.; Mani, G.; and Barr, D. 1996 "Applying Computational Intelligence to the Investment Process" ,In Proceedings of CIFER-96: Computational Intelligence in Financial Engineering. Washington ,D.C.: IEEE Computer Society.