Journal of Statistical and Econometric Methods, vol.1, no.2, 2012, 39-51
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2012

Forecasting of Indian Stock Market by Effective Macro- Economic Factors and Stochastic Model

Jyoti Badge[1]

Abstract

The stock market patterns are non-linear in nature therefore it is difficult to forecast the future trends of the market. In this paper we have used different macro-economic factors of Indian stock market. Macro-economic factors include technical indicators. These technical indicators help to decide the patterns of the market at a particular time. There are hundreds of technical indicators are available, but all technical indicators are not useful. So we have obtained most effective technical indicators by applying Principal Component Analysis (PCA). Selected technical indicators are taken as input variable. Future prices are found through Hidden Markov Model (HMM). Hidden Markov Model is a very powerful stochastic model. In literature survey it was found that HMM gives better accuracy than other models. On the basis of experiment it was found that HMM with PCA performed well and gives Mean Absolute Percentage Error (MAPE) 1.77%.

Keywords: Macro-economic factors, Technical Indicators, Principal Component Analysis, Hidden Markov Model

[1] e-mail:jyoti.badge@gmail.com

Article Info: *Received* : May 12, 2012. *Revised* : June 23, 2012
Published online : July 30, 2012

1 Introduction

Present and past behavior of the stock market is considered while forecasting the trends of the stock market. The traditional prediction theories were based on linear time series models, but the patterns of stock market are not linear, it contains some non-linearity. Hidden Markov Model (HMM) plays a key role to find the non-linear patterns. It is a statistical Markov Model in which the system being modeled is assumed to be a Markov process with hidden states and it falls in the class of stochastic model. A stochastic model is a tool for estimating probability distributions of potential outcomes by allowing random variation in one or more inputs over time. The random variation is usually based on fluctuations observed in historical data for selected period measured using standard time-series techniques (Zastawniak, 1999).

Hidden Markov Model has proven to be useful in a wide range of applications for modeling highly structured sequences of data (Elston et.al. 2002). It has been extensively used in areas like speech recognition, handwriting recognition (Nag et.al. 1986, Kundu et.al. 1988, Matan et.al. 1992, Ha et.al. 1993, Schenkel et.al.1993, 1995, Bengio et.al. 1995) patterns recognizing in molecular biology (Krogh et.al., 1994, Baldi et.al. 1995, Karplus et.al. 1997, Baldi et.al. 1998) and fault-detection system (Smyth et.al. 1994) etc.

Bengio et.al. (2001) applied Input–Output Hidden Markov Model on financial time series data and performed number of comparative experiments aimed at measuring the expected generalization error of different types of model structures. In 2005 Hassan et.al structured Hidden Markov Model (HMM) for four states i.e. opening, high, and low and closing prices. In 2006 they presented a Hidden Markov Model based on fuzzy rule extraction technique for predicting a time series generated by a chaotic dynamical system. In 2007 they implemented a fusion model by combining Hidden Markov Model, Artificial Neural Networks and Genetic Algorithms for forecasting financial market behavior.

In literature survey it was observed that, stock prices were used as input parameter in Hidden Markov Model but not the technical indicators. So in this paper we have used technical indicators as an input variable which gives more information as compared to the stock prices.

2 Methodology

In this paper different computational methods were used. We have discussed in following subsections.

2.1 Markov Process

A Markov process is a stochastic process where the future event depends on

instantaneous preceding event. Markov process assumes that probability of the occurrence of an event solely depend on the occurrence of the current event. Any process that follows Markov property is called a Markov process.

According to Zhiyuan et al. 2010, a random sequence is called a Markov chain if

$$P(q_{t+1} = S_j / q_t = S_i, q_{t-1} = S_k, ..., q_1 = S_l) = P(q_{t+1} = S_j / q_t = S_i).$$

These probabilities are called transition probabilities and are denoted by $a_{ij}(t) = P(q_{t+1} = S_j / q_t = S_i)$.

2.2 Hidden Markov Model

Hidden Markov Model was first described by Leonard E. Baum in 1960s and has been used in analyzing and predicting time series phenomena. It is a generalization of a Markov chain, in which each state is not directly observable but variables influenced by the state are visible, also called "emission", according to given stationary probability law. In this case, time evolution of the internal states can be induced only through the sequence of observed output states.

Elements of Hidden Markov Model

Hidden Markov Model is a finite set of *states*, each of which is associated with multidimensional probability distribution. Transitions among the states are governed by a set of probabilities called *transition probabilities*. In a particular state an observation can be generated, according to the associated probability distribution.

HMM contains following elements

N is Number of hidden states

Q is Set of states $Q = \{1,2,...,N\}$

M is Number of symbols

V is Set of observation symbols $V = \{1,2,....,M\}$

A is State Transition Probability Matrix

$$a_{ij} = P(q_{t+1} = j / q_t = i) \qquad 1 \le i, j \le N \tag{1}$$

B is Observation probability matrix

$$B_j(k) = P(o_t = k / q_t = j) \qquad 1 \le k \le M \tag{2}$$

π is Initial state distribution

$$\pi_i = P(q_1 = i) \qquad 1 \le i \le N \tag{3}$$

λ is entire model

$$\lambda = (A, B, \pi) \tag{4}$$

2.3 Principal Component Analysis (PCA)

Principal Component Analysis is a statistical technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables that represents most of the information in the original set of variables. Its goal is to reduce the dimensionality of the original data set. A small set of uncorrelated variables is much easier to understand and can be useful in further analyses than a larger set of correlated variables (Dunteman et.al. 1989).

In Principal Component Analysis, the variance of a matrix (Z) is explained in terms of new latent variables which are called Principal Components (PC). The first Principal Component variable is the linear combination of matrix element that has the greatest variance. The second Principal Component Variable (PCV) is the linear combination with the next greatest variance among coefficient vectors of unit length that are orthogonal to the first coefficient vector. In this manner, one can obtain k possible Principal Component Variables. The calculated Principal Component is given by

$$t_1 = p_1' z \quad \text{Subject to } |p_1| = 1 \tag{5}$$

$$t_2 = p_2' z \quad \text{Subject to } |p_2| = 1 \tag{6}$$

$$\text{and } p_2' p_1 = 0 \tag{7}$$

The Principal Component loading vectors p are the eigenvectors of the covariance matrix $\sum$ of Z and the corresponding eigenvalues λ_i are the variances of the Principal Components. Using loading vectors the observation can be written as

$$Z = \sum_{i=1}^{k} t_i p_i' + E \tag{8}$$

where k, is the number of Principal Components obtained and E is the residual matrix (Abraham and Nair, 1998).

3 Experiment

HMM is constructed by estimating parameter set $\lambda(A, B, \pi)$. The states of HMM are: increasing state, decreasing state and no change state. Observation sequence is built using principal technical indicators. Initially parameter values are chosen randomly. The HMM was trained using training dataset, so that the values of A, B, π are re-estimated to suit the training dataset.

To predict the next day's closing price, we applied Baum-Welch algorithm (Baum et.al., 1996). Suppose the likelihood value for the day *x* is lcp_i. Now from the historical data set observation sequences are located that would produce the same or nearly same value of lcp_i (locate past days where the stock behavior is matched to the current day). HMM found many observations that produced the same likelihood value lcp_i. Then for each of the matched day, difference between

match day and it next day is calculated using (9)

$$Wd_k = \frac{\sum_m W_m diff_m}{\sum_m W_m} \tag{9}$$

where, W_m is weight assigned to day m
wd_k is weighted average of price difference for current day k
$diff_m$ is price difference between day m and $m+1$
then forecast the value for day $k+1$, by

$$fp_{(k+1)} = p_k + wd_k \tag{10}$$

where, $fp_{(k+1)}$ is the forecasted closing price and p_k is current day closing price.

4 Experimental Results

The experiment is conducted on *S&P CNX NIFTY* by taking six years daily data from 1-Mar-2002 to 31-Dec-2008. The total number of observations N is 1712. We divided the data into training and testing data

We considered several technical indicators as inputs for the model. They are Accumulation/Distribution Oscillator, Accumulation/Distribution Line, Chaikin Oscillator, Chaikin Volatility, Moving Average Convergence-Divergence (MACD), Stochastic Oscillator %K and %D, Williams %R, Williams Accumulation/Distribution Line, Negative Volume Index, Positive Volume Index, Relative Strength Index (RSI), Bollinger Band(Middle), Bollinger Band(Upper), Bollinger Band(Lower), Highest High, Lowest Low, Median Price, On-Balance Volume (OBV), Price Rate of Change, Price And Volume Trend (PVT), Typical Price, Volume Rate of Change, Weighted Close.

It is difficult to find relevant technical indicators. It may happen that some indicators would provide excellent information for stock *A*, but they may not give any insight information for stock *B*. Thus, we needed a tool to choose the right indicators for each stock (Ince et.al. 2004). Also, our objective was to identify important indicators that can be used in HMM as input parameter. Therefore we applied principal component analysis for finding relevant technical indicator.

Principal Components are extracted using SPSS software. It is a component matrix helped to determine principal components as shown in Table 1.

The correlated components are Weighted Close, Williams %R, Volume Rate of Change and Chaikin Volatility in order of decreasing correlation. Therefore we selected Weighted Close, Williams %R, Chaikin Volatility and Volume Rate of Change for further analyses. These four indicators are used to make observation sequences in HMM.

Table 1: Component Matrix

Component Matrix				
	Component			
	1	2	3	4
Accumulation/Distribution oscillator	0.00203	0.44245	-0.2737	0.40415
Accumulation/Distribution line	0.95706	-0.055	-0.1046	0.06532
Chaikin oscillator	0.07894	0.82019	0.04924	0.16183
Chaikin volatility	0.04982	-0.254	**0.6084**	0.50738
MACD	0.07721	0.55753	0.57448	-0.2961
%K	0.08193	0.91037	-0.2152	0.15946
%D	0.09364	0.94111	-0.123	0.03153
Williams %R	0.09969	**0.95176**	-0.1344	0.07753
Williams Accumulation/Distribution line	-0.5397	0.09285	-0.1022	0.23323
Negative volume index	0.95092	-0.1393	-0.1902	0.09326
Positive volume index	0.4039	0.42693	0.56107	-0.2625
Relative Strength Index (RSI)	0.06885	0.88158	0.07665	-0.0531
Bollinger band(Middle)	0.9904	-0.0981	-0.0274	0.02818
Bollinger band(Upper)	0.98544	-0.1163	-0.0442	0.04841
Bollinger band(Lower)	0.99025	-0.0764	-0.0077	0.00443
Highest high	0.98924	-0.098	-0.0299	0.03923
Lowest low	0.99069	-0.0533	-0.0012	-0.0086
Median price	0.99382	-0.0192	-0.0164	0.00534
On-Balance Volume (OBV)	0.65798	0.02254	-0.0489	-0.0354
Price rate of change	0.01164	0.88665	-0.0263	-0.1446
Price and Volume Trend (PVT)	0.93757	0.1183	0.15054	-0.0898
Typical price	0.99383	-0.0172	-0.0176	0.0075
Volume rate of change	0.03227	0.05776	0.32695	**0.68775**
Weighted close	**0.99386**	-0.0162	-0.0181	0.00858
Extraction Method: Principal Component Analysis.				
4 components extracted.				

We have assumed three hidden states, Increasing, Decreasing and No change.

a) When $C_n - C_{n-1} > 0$, it is taken as increasing state.

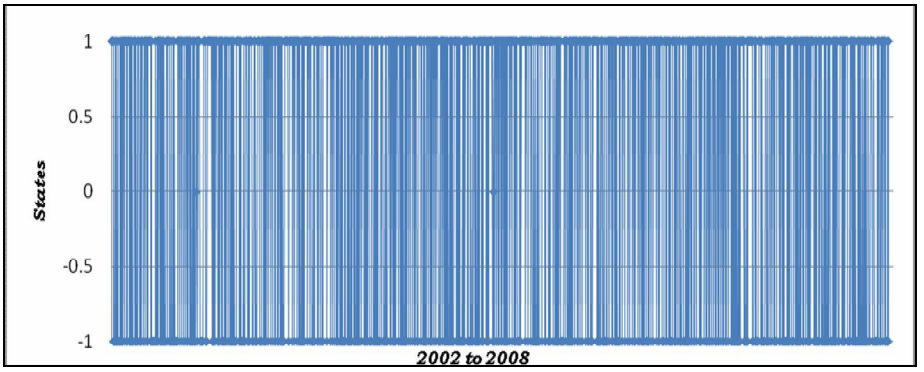
b) When $C_n - C_{n-1} < 0$ it is taken as decreasing state.

c) When $C_n - C_{n-1} = 0$ it is in no change state.

where C_n is current closing price and C_{n-1} is the previous closing price. The states transition probability is expressed as

	Increase	Decrease	No Change
Increase	*Increase/ Increase*	*Increase/ Decrease*	*Increase/ NoChange*
Decrease	*Decrease/ Increase*	*Decrease/ Decrease*	*Decrease/ NoChange*
No Change	*NoChange/ Increase*	*NoChange/ Decrease*	*NoChange/ NoChange*

The transition states from 1st January 2002 to 31st December 2008 is shown in Figure 1.



* 1 indicates increasing state, -1 indicates decreasing state and 0 indicates no change state.

Figure1: State Transition from 1st January 2002 to 31th December 2008

It is difficult, to watch the movement of states in the figure because data size was too large. Therefore transition for year 2002 is shown in Figure 2. Each little point indicates a state and the connecting lines between points illustrate the transitions.

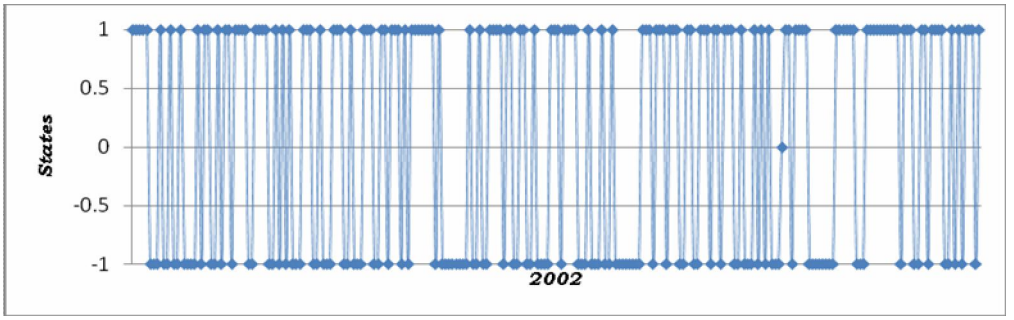


Figure 2: State Transition for 2002

For finding the trend of S&P CNX NIFTY, we need to find state transition probability.

The result is obtained in the following way

increase $\Rightarrow$ increase $\rightarrow$ 369 days
increase $\Rightarrow$ decrease $\rightarrow$ 396 days
increase $\Rightarrow$ no change $\rightarrow$ 1 day
decrease $\Rightarrow$ increase $\rightarrow$ 404 days
decrease $\Rightarrow$ decrease $\rightarrow$ 577 days
decrease $\Rightarrow$ no change $\rightarrow$ 0 day
no change $\Rightarrow$ increase $\rightarrow$ 1 day
no change $\Rightarrow$ decrease $\rightarrow$ 1 day
no change $\Rightarrow$ no change $\rightarrow$ 0 day

We get transition matrix as

$$\hat{A} = \begin{bmatrix} 0.485084 & 0.513619 & 0.001297 \\ 0.411824669 & 0.588175331 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

Increasing, decreasing and no change state percentage area of S&P CNX NIFTY for the year 2002 to 2008 is shown in Figure 3.

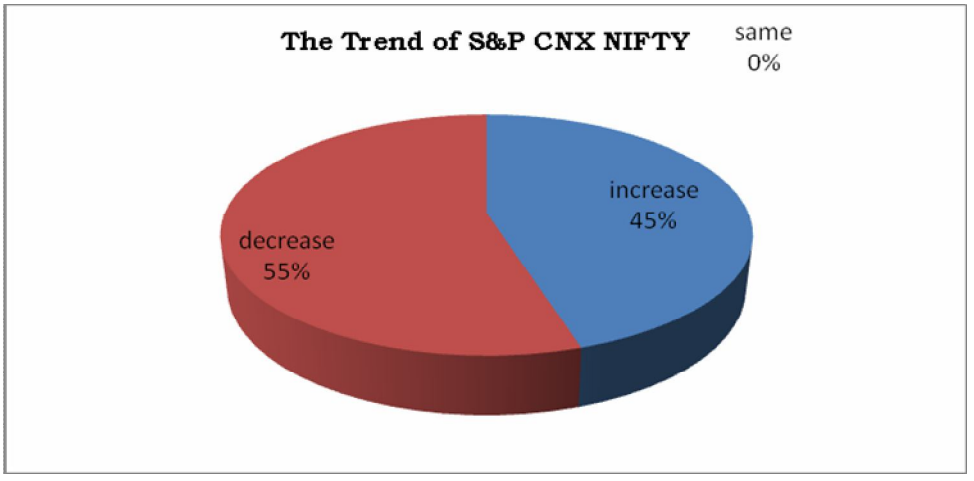


Figure 3: Percentage Area of States for S&P CNX NIFTY

The probability of increasing state during year 2002 to 2008 was 0.45,
The probability of decreasing state during year 2002 to 2008 was 0.55
The probability of same state during year 2002 to 2008 was nearly 0.

Observation sequence is generated through Principal Indicators. Selected indicators have some numerical values. We have transformed numerical values into symbolic values say (1, 2, 3), 1 means down, 2 means up and 3 means same.

The transformed sequence is formed by taking the difference of current day and previous day of each principal indicator.

a) When $I_n - I_{n-1} < 0$ assign 1,

b) When $I_n - I_{n-1} > 0$ assign 2 and

c) When $I_n - I_{n-1} = 0$ assign 3

where, I_n is the current day indicator value and I_{n-1} is the previous day indicator value.

Let us suppose that we have following sequence {1, 1, 2, 1}. Initially, $\pi = P(0,1,0)$. The observation emission probability matrix would be

$$\begin{array}{l} \\ \text{Increase} \\ \text{Decrease} \\ \text{Same} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0.25 & 0.75 & 0 \\ 0.75 & 0.25 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

HMM is trained with the help of Baum-Welch algorithm. A new model $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ is built which tries to maximize $P(O/\hat{\lambda})$. Using the trained HMM, likelihood value for current day's data set is calculated. Likelihood values for closing price are shown in Figure 4.

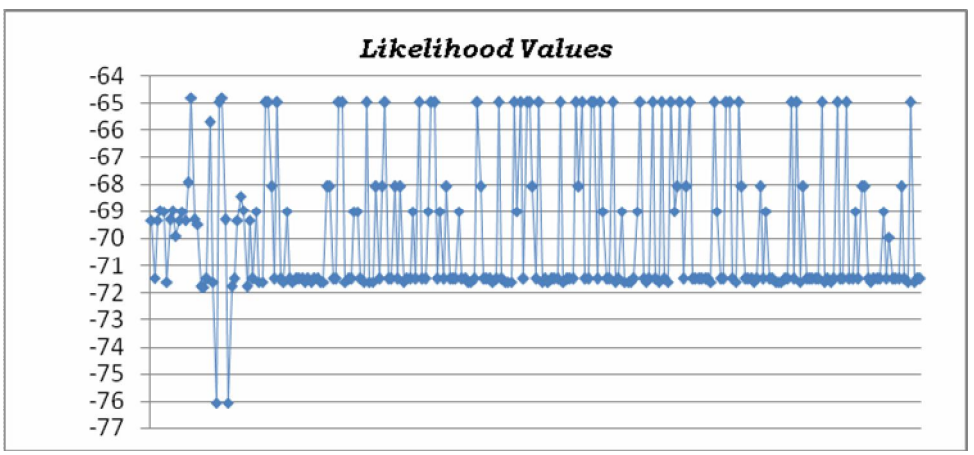


Figure 4: Log likelihood Values for Closing Price

Likelihood values helps to forecast the next day closing price. Suppose the likelihood value of 5-Jan-09 was -68.0787. From the past data set, we found all

those values which shall match the likelihood value of that day. We find a range of data vectors from the past training dataset which produce likelihood values closer to 5-Jan-2009. Matched likelihood values for 5- Jan-09 are shown in Table 2.

Table 2: Matched Likelihood Past Values for 5- Jan-09

Date	Closing Price	Likelihood value
20-Nov-02	1001.6	-68.0787
11-Aug-03	1232.85	-68.0787
19-Dec-03	1778.55	-68.0787
17-Sep-04	1733.65	-68.0787
22-Jun-05	2187.35	-68.0787
27-Jun-05	2199.8	-68.0787
30-Mar-06	3418.95	-68.0787
16-Aug-06	3356.05	-68.0787
16-Nov-06	3876.85	-68.0787
22-Nov-06	3954.75	-68.0787
23-Apr-07	4085.1	-68.0787
20-Jul-07	4566.05	-68.0787

Using (9) we calculated weighted average of these price differences. The calculation of weighted average of the price differences of similar patterns is shown in Table 3.

Table 3: Calculation of Weighted Average of Price Differences of Similar Patterns

Date	I_m	W_m	*diffm*	W_m**diffm*
20-Nov-02	37.88276	9.68E-17	7.15	6.92E-16
11-Aug-03	46.62915	1.54E-20	1.9	2.92E-20
19-Dec-03	67.26875	1.67E-29	10.6	1.77E-28
17-Sep-04	65.57053	9.14E-29	-4.85	4.43E-28
22-Jun-05	82.73048	3.23E-36	-3.5	1.13E-35
27-Jun-05	83.20137	2.01E-36	-29.95	6.03E-35
30-Mar-06	129.3124	1.90E-56	-16.4	3.11E-55
16-Aug-06	126.9333	2.05E-55	-2.15	4.41E-55
16-Nov-06	146.6312	5.71E-64	-24.05	1.37E-62
22-Nov-06	149.5775	3.00E-65	-9.3	2.79E-64
23-Apr-07	154.5076	2.17E-67	56.7	1.23E-65
20-Jul-07	172.6982	2.73E-75	53.3	1.45E-73

Where, I_m is index number of matched day. Index number is obtained by (11)

$$\text{Index Number of the } j^{th} \text{ day} = \frac{CP \text{ of the } jth \text{ day}}{Average \text{ of } CP} * 100 \quad (11)$$

where,
CP is closing Price
Wd_k *is 6.91871E*-16/9.67765E-17
Wd_k is 7.149165205

Forecasted closing price of 6-Jan-09 is obtained by adding Wd_k in the closing price of 5-Jan-09. The forecasted closing price for 6-jan-09 is 2920.4. The actual close price of 6-jan-09 is 3112.8. Similarly we obtained forecasted values for other days. The actual and forecasted value of closing price is shown in Table 4.

Table 4: Actual and Forecasted Closing Price using HMM

Date	Actual Closing Price	Forecasted Closing Price	Percentage of Error
1-Jan-09	3033.45	2997.81	1.17%
2-Jan-09	3046.75	3057.08	0.34%
5-Jan-09	3121.45	3053.029	2.19%
6-Jan-09	3112.8	2920.4	6.18%
7-Jan-09	2920.4	2919.9	0.02%
9-Jan-09	2873	2916.87	1.53%
12-Jan-09	2773.1	2869.47	3.48%
13-Jan-09	2744.95	2775.5	1.11%
14-Jan-09	2835.3	2790.27	1.59%
15-Jan-09	2736.7	2739.39	0.10%

The error of HMM is measured in terms of Mean Square Error, Mean Absolute Percentage Error, Root Mean Square Error and Mean Absolute Percentage Error is shown in Table 5.

Table 5: Performance Measure of HMM

S.No.	Error Measuring Parameter	Value
1	Mean Square Error	57256.36
2	Mean Absolute Deviation	52.58
3	Root Mean Square Error	75.67
4	Mean Absolute Percentage Error	1.77%

The forecasting error in Hassan et.al. (2005) model was 2.01% and 1.92% in Hassan et.al. 2007 modified model. The forecasting error in our model is 0.15% less than Hassan et.al. (2007) model i.e. is 1.77%.

5 Conclusion

It was observed at the time of experiment that it is not necessary to consider all the technical indicators for analysis. Principal Component Analysis (PCA) can effectively be used to identify most useful technical indicators. These technical indicators can be referred as principal technical indicators. The result clearly shows that Hidden Markov Model (HMM) with principal technical indicators as input parameter performed well.

References

[1] Abraham, Bovas and Nair N. Unnikrishnan, *Quality Improvement through Statistical Methods*, Birkhauser, Canada, 1998.

[2] P. Baldi, Y. Chauvin, T. Hunkapiller and M. McClure, Hidden Markov Models of Biological Primary Sequence Information, *Proc Nat. Academy Sci. USA*, **91**(3), (1995), 1059-1063.

[3] P. Baldi and S. Brunak, *Bioinformatics, The Machine Learning Approach*, Cambridge, MA, MIT Press, 1998.

[4] Leonard E. Baum and Ted Petrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, *The Annals of Mathematical Statistics*, **37**(6), (1966), 1554-1563.

[5] Y. Bengio, Y. LeCun, C. Nohl and Burges, C. Lerec: A NN/HMM Hybrid for On-Line Handwriting Recognition, *Neural Computation*, **7**(5), (1995), 1289-1303.

[6] Bengio Yoshua, Lauzon Philippe Vincent and Ducharme, Réjean, January, Experiments on the Application of IOHMMS to Model Financial Returns Series, *IEEE Transactions on Neural Networks*, **12**(1), (2001), 113-123.

[7] Dunteman H.George, Principal Component Analysis, **69**, Sage, (1989).

[8] Elston C. Robert, Olson M. Jane, Palmer Lyle, *Biostatistical Genetics and Genetic Epidemiology*, John Wiley and Sons, 2002.

[9] J.Y. Ha, S.C. Oh, J.H. Kim and Y.B. Kwon, Unconstrained Handwritten Word Recognition with Interconnected Hidden Markov Models, *Proc. 3rd Int. Workshop Frontiers Handwriting Recognition*, Buffalo, NY, (May, 1993), 455-460.

[10] Hassan Rafiul, *Hybrid HMM and Soft Computing Modeling with Applications to Time Series Analysis*, Master's thesis, 2007.

[11] Hassan Rafiul, Nath Baikunth and Michael Kirley, A Data Clustering Algorithm Based on Single Hidden Markov Model, *Proceedings of the International Multiconference on Computer Science and Information Technology*, (2006), 57-66.

[12] Hassan Rafiul, Nath Baikunth and Michael Kirley, Stock Market Forecasting Using Hidden Markov Model A New Approach, *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, (2005).

[13] Hassan Rafiul, Nath Baikunth and Michael Kirley, HMM based Fuzzy Model for Time Series Prediction, *IEEE International Conference on Fuzzy Systems*, (2006), 2120-2126.

[14] Hassan Rafiul, Nath Baikunth and Michael Kirley, A Fusion Model of HMM, ANN and GA for Stock Market Forecasting, *Expert Systems with Applications: An International Journal*, **33**(1), (July, 2007), 171-180.

[15] K. Karplus, K. Sjolander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm and C. Sander, Predicting Protein Structure Using Hidden Markov Models*, Proteins: Structure, Function, Genetics*, **1**(1), (1997), 134-139.

[16] A. Krogh, M. Brown, I.S. Mian, K. Sjölander and D. Haussler, Hidden Markov Models in Computational Biology: Applications to Protein Modeling, *Journal of Molecular Biology*, **235**, (1994), 1501-1531.

[17] Kundu and Bahl, L. R., Recognition of Handwritten Script: A Hidden Markov Model Based Approach, *Proc. Int. Conf. Acoust., Speech, Signal Processing*, New York, (1988), pp. 928–931.

[18] O. Matan, C.J.C.Burges, Y. LeCun and J.S. Denker, Multi-digit recognition using a space displacement neural network, *Advances in Neural Information Processing Systems*, **4**, (1992).

[19] R. Nag, K.H. Wong and F. Fallside, Script Recognition Using Hidden Markov Models, *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, (1986), 2071-2074.

[20] M. Schenkel, H. Weissman, I. Guyon, C. Nohl and D. Henderson, Recognition-Based Segmentation of On-Line Hand-Printed Words, *Advances in Neural Information Processing Systems*, **5**, (1993), 723-730.

[21] M. Schenkel, I. Guyon and D. Henderson, On-Line Cursive Script Recognition Using Time Delay Neural Networks and Hidden Markov Models, *Machine Vision and Applications*, **8**(4), (1995), 215-223.

[22] P. Smyth, Hidden Markov Models for Fault Detection in Dynamic Systems, *Pattern Recognition*, **27(**1), (1994), 149-164.

[23] Tomasz Zastawniak, *Basic Stochastic Processes: A Course through Exercise*, Springer-Verlag, London Limited, 1999.