CHAPTER II

LITERATURE REVIEW

This chapter describe about the fundamental theory used to conduct the research model.

2.1 Previous research

Hassan and Nath (2005) presented Hidden Markov Models (HMM) approach for forecasting stock price at the interrelated markets. Any fluctuation in market influences personal and corporate financial lives, and the economic health of a country. Hassan et al., (2005) consider 4 input features for a stock, which is the opening price, closing price, highest price, and the lowest price. The next day's closing price is taken as the target price associated with the four input features. The idea behind new approach in using HMM is that the using of training dataset for estimating the parameter set (A, B, π). Using the trained HMM, likelihood value for current day's dataset is calculated. For instance the likelihood value for the day is £ then from the past dataset using the HMM locate those instances that would produce the same £ or nearest to the £ likelihood value. Assuming that the next day's stock price should follow about the same past data pattern, from the located past day(s) simply calculate the difference of that day's closing price and next to that day's closing price. Thus the next day's stock closing price forecast is established by adding the above difference to the current day's closing price. The results show potential of using HMM for time series prediction.

Amri, (2008) discussing about Hidden Markov Model concerning about speech signal recognition using HMM-Neural Network (NN) method. The method is used to determine the sequence of speech signal data based on the initial and feature extraction from a batch of different word, NN is used to determine the success of speech signal recognition processes. In this research, a word is spelled by single utterance with different word and different number of words and being done with the set of 5-50 words in Bahasa. The condition applied is 50 number of words, NN structure used is amount to 5 layer hidden, 20 node, 10 node, 5 node, 10 node, 20 node for each.

Testing data using AT & T Database by the number of 40 individuals with three training images per individual and 7 images per individual testing. While the Yale Face Database with a number of 15 individuals with 4 images per individual and 7 training images per individual testing. Those models have successfully built applications embedded HMM-based face recognition with identification accuracy of 94.64% generalization (AT & T Database) and 77.14% (Yale Face Database). By the results it is known that AT & T database that does not require cutting face area is much better than the Yale Face Database. This is because the cutting area of the face in an image depends on the accuracy of face detection, so changes will take effect in the detection area of training and testing results.

Irfani et al., (2006) discussed about speech recognition. The development of speech recognition technology is one form of technological developments in the 20th century that utilize voice as input. The voice is an alternative method for humans to interact with computers. The computer will recognize the voice commands and perform the stretcher as a reaction to the command. Modern speech recognition systems are generally based on

Hidden Markov Models (HMMs). By HMMs sound signals can be characterized as a random process parameters, and parameters of the stochastic process which can be determined precisely. That statistical model then processed using the Viterbi algorithm. Viterbi algorithm is a dynamic programming algorithm to find possible hidden state sequence (commonly called the Viterbi path) which produced the series of observations of events, especially within the scope of the HMM. By viterbi algorithm processing statuses in the voice recognition system can be optimized.

On the other hand, the expansion of Hidden markov model was also discussed by Imam (2007). In this research, an application for face recognition is built based on embedded Hidden Markov Models (eHMM). eHMM are able to modeling image as 2 dimensional data better than ordinary HMM. eHMM will segmented face area on digital image into 5 super states (forehead, eyes, nose, lips, and chin) and into couple of embedded states inside those super states. Identifications are being done by comparing test image's observation likelihood with eHMM face model. Viterbi Algorithm is used to evaluate the best likelihood from the comparison of test image's observation with all people's eHMM. The accuracy of identification is known by introduction to each test image belongs to every individual.

The other study has been done by Fathoni, (2008) in Continuous hidden markov model review and its application on harvest unhusked. The harvest unhusked price is fluctuated according to some factors such as thai-rice import policy, fertilizer, harvest failure, natural disaster, political situation etc. Those factors assumed as the state of unobservable markov chain. The input data was the average at the producers in region I from January 2000 through March 2007. It is assumed that the price of unhusked raised by the random variables Y_k that spread in a certain distribution at interval changes $(\Omega, \mathcal{F}, P)$. To facilitate the search for the parameter estimators created a functional programming-based computing using Mathematica 6.0. The estimators obtained are used to calculate the expected value of the price of unhusked rice. From the results obtained, the continuous hidden Markov model is good enough explain the behavior of the price harvest. The more the cause value of the alleged incident, the better.

Research in the field of prediction using Hidden markov model is a new breakthrough, because most of them are used for bio technology or recognition. Previous literature suggests that there has been no research on the gold market scientifically. Community observe the trends based on intuition and a simple calculation. Meanwhile they need to know the market trend for gold itself. This research was intended to study time series trend using Hidden Markov Model to give a scientific description by finding the most likely sequence.

2.2 Probability theory

The fields of statistics are related to the ways of data collection, processing, presentation, analytical and conclusion has been made based on data and analysis. Resulting conclusion is expected to be the description of population and their characteristics. Experiment is a method of data collecting. Set of all possible outcomes from a randomized trial called the sample space and its denoted by Ω. An event A is a subset of the sample space Ω.

2.2.1 Probabilty And Random Variabel

If an experiment obtain a continued sample, then the random variable that connected to those sample is called continued random variable. The probability spread of random variable is called probability density function (Lungan, 2006). A random variable x is a function of x: $\Omega \rightarrow \mathbb{R}$ which $\{\ \omega \in \Omega : X(\omega) \leq X\} \in$ field, for each $X \in \mathbb{R}$. It is not a variable but rather a function that maps events to numbers. A Random Variable is a function, which assigns unique numerical values to all possible outcomes of a random experiment under fixed conditions (Ali, 2000).

This example is extracted from Ali, (2000). Suppose that a coin is tossed three times and the sequence of heads and tails is noted. The sample space for this experiment evaluates to: S={HHH, IIHT, HTH, HTT, THH, THT, TTH, TTT}. Now let the random variable X be the number of heads in three coin tosses. X assigns each outcome in S a number from the set Sx={0, 1, 2, 3}. The table below lists the eight outcomes of S and the corresponding values of X.

Table 2.1 Probability sample space

Outcome	HHH	IIHT	HTH	THII	HTT	THT	TTH	TTT
X	3	2	2	2	1	1	1	0

X is then a random variable taking on values in the set Sx = {0, 1, 2, 3}. Mathematically, a random variable is defined as a measurable function from a probability space to some measurable space. This measurable space is the space of possible values of the variable, and it is usually taken to be the real numbers. The condition for a function to be a random variable is that the random variable cannot be multivalued. There are 2 types of random variables:

a. A Continuous Random Variable is one that takes an infinite number of possible values. Example: Duration of a call in a telephone exchange.

b. A Discrete Random Variable is one that takes a finite distinct value. Example: A number of students who fail a test.

a. Continuous random variable

If the random variable X is continuous with probability density function $f(x)$,

$$\mathrm{Var}\,(X) = \int (x-\mu)^2\, f(x)\, dx, \quad \text{..} \quad (2.1)$$

where μ is the expected value, i.e.

$$\mu = \int x\, f(x)\, dx, \quad \text{..} \quad (2.2)$$

b. Discrete random variable

If the random variable X is discrete with probability mass function $x_1 \rightarrow p_1, \ldots, x_n \rightarrow p_n$, then

$$\mathrm{Var}\,(X) = \textstyle\sum_{i=1}^{n} p_i \cdot\, (x_i - \mu)^{2} \quad \text{..} \quad (2.3)$$

where μ is the expected value, i.e.

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i \quad \text{..} \quad (2.4)$$

2.2.2 Conditional Expectation

In probability theory, a conditional expectation, also known as conditional expected value or conditional mean, is the expected value of a real random variable respect to a conditional probability distribution.

2.2.3 Expected Value and Variance

The expected value, or mean, of a random variable is the weighted average of all possible values. The weights used in computing this average correspond to the probabilities in case of a discrete random variable, or densities in case of a continuous random variable. In probability theory and statistics, the variance is used as a measure of how far a set of numbers are spread out from each other. It is describing how far the numbers lie from the mean (expected value).

The variance is a parameter describing the actual probability distribution of an observed population. In the latter case a sample of data from a distribution can be used to construct the variance defined below.

If a random variable X has the expected value (mean) $\mu = E[X]$, then the variance of X is given by:

$$\operatorname{var}(X) = E\left[\left((X-\mu)\ \right)\right] \quad \text{...} \quad (2.5)$$

2.3 Time Series Analysis

A time series is a sequence of observations of a random variable. Hence, it is a stochastic process. Examples include the monthly demand for a product or the annual freshman enrollment in a department of a university. Forecasting time series data is important component of operations research because these data often provide the foundation for decision models. An inventory model requires estimates of future demands and a course scheduling and staffing model for a university requires estimates of future student inflow. Time series analysis provides tools for selecting a model that can be used to forecast of future events. Modeling the time series is a statistical problem. Forecasts are used in computational procedures to estimate the parameters of a model being used to allocated limited resources or to describe random processes. There are two main goals of time series analysis:

a. Identifying the nature of the phenomenon represented by the sequence of observations.

b. Predicting future values of the time series variable.

Both of these goals require that the pattern of observed data is identified and described. Once the pattern is established, we can interpret and integrate it with other data.

2.4 Stochastic Process

A probability space associated with a random experiment is a triple (Ω, $\mathcal{F}$, P) where:

(i) Ω is the set of all possible outcomes of the random experiment, and it is called the sample space.

(ii) $\mathcal{F}$ is a family of subsets of Ω which has the structure of a σ field:

a) $\emptyset \in \mathcal{F}$

b) If $A \in \mathcal{F}$, then its complement A^c also belongs to $\mathcal{F}$

c) $A_1, A_2, \ldots, \in \mathcal{F} \rightarrow \bigcup_{t=1}^{\infty} A_t \in \mathcal{F}$

(iii) P is a function which associates a number $P(A)$ to each set $A \in \mathcal{F}$ with the following properties:

a) $0 \leq P(A) \leq 1$,

b) $P(\Omega) = 1$

c) For any sequence $A_1, A_2, \ldots$ of disjoints sets in $\mathcal{F}$ (that is, $A_i \cap A_j = \emptyset$ if $i \neq$

d) $P\left(\bigcup_{t=1}^{\infty} A_t\right) = \sum_{t=1}^{\infty} P\left(A_t\right)$

The elements of the σ-field $\mathcal{F}$ are called events and the mapping P is called a probability measure. In this way we have the following interpretation of this model:

$$P(F) = \text{probability that the event } F \text{ occurs}$$

The set $\emptyset$ is called the empty event and it has probability zero. Indeed, the additivity property (iii,c) implies:

$$P(\emptyset) + P(\emptyset) + \ldots = P(\emptyset)$$

The set Ω is also called the certain set and by property (iii,b) it has probability one. Usually, there will be other events A T Ω such that $P(A) = 1$. If a statement holds for all ω in a set A with $P(A) = 1$, then the statement is true, or that the statement holds for almost all $\omega \in \Omega$. The axioms a), b) and c) lead to the following basic rules of the probability calculus:

$$P(A \cup B) = P(A) + P(B) \text{ if } A \cap B = \emptyset$$

$$P(A^c) = 1 - P(A)$$

$$A \text{ T} B \rightarrow P(A) \leq P(B)$$

Example : Consider the experiment of flipping a coin once.

$\Omega = \{\text{H,T}\}$ (the possible outcomes are "Heads" and "Tails")

$\mathcal{F} = P(\Omega)$ ($\mathcal{F}$ contains all subsets of Ω)

$P(\{\text{H}\}) = P(\{\text{T}\}) = \frac{1}{2}$

2.5 Markov Chain

Andrey Markov produced the first results (1906) for these processes theoretically. Markov chains are related to Brownian motion and the ergodic hypothesis, but Markov appears to have pursued this out of a mathematical motivation, namely the extension of the law of large numbers to dependent events. In 1913, he applied his findings for the first time to the first 20,000 letters of Pushkin's Eugene Onegin.

A Markov chain is a discrete random process with the property that the next state depends only on the current state. It is named for Andrey Markov, and is a mathematical tool for statistical modeling in modern applied mathematics, particularly information sciences. A useful heuristic is that of a frog jumping among several lily-pads, where the frog's memory is short enough that it doesn't remember what lily-pad it was last on, and so its next jump can only be influenced by where it is now.

Formally, a Markov chain is a discrete random process with the Markov property that goes on forever. A discrete random process means a system which is in a certain state at each step, with the state changing randomly between steps. The steps are often thought of as time (such as in the frog and lily-pad example), but they can equally well refer to physical distance or any other discrete measurement. The Markov property states that the conditional probability distribution for the system at the next step depends only on the current state of the system, and not additionally on the state of the system at previous steps:

$$P(X_{n+1}|X1,X2,\ldots..,Xn)=P(X_{n+1}|Xn) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.6)$$

Since the system changes randomly, it is generally impossible to predict the exact state in the future. However, the statistical properties of the system's future can be predicted. In many applications it is these statistical properties that are important. The changes of state are called transitions, and the probabilities associated with various state-changes are called transition probabilities. The set of all states and transition probabilities completely characterizes a Markov chain. By convention, assume all possible states and transitions have been included in the definition of the processes, so there is always a next-state and the process goes on.

A famous Markov chain is the drunkard's walk, a random walk on the number line where the position may change by +1 or −1 with equal probability. From any position there are two possible transitions, to the next or previous integer. The transition probabilities depend only on the current position, not on the way the position was reached. For example, the transition probabilities from 5 to 4 and 5 to 6 are both 0.5, and all other transition probabilities from 5 are 0. These probabilities are independent of whether the system was previously in 4 or 6.

However, the theory is usually applied only when the probability distribution of the next step depends non-trivially on the current state. A Markov chain is a sequence of random variables X_1, X_2, X_3, ... with the Markov property, given the present state, the future and past states are independent. Formally,

$$\Pr(X_{n+1}=\chi|X_1=\chi_1,X_2=\chi_2\ldots,X_n=\chi_n)=\Pr(X_{n+1}=\chi|X_n=\chi_n) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

The possible values of X_i form a countable set S called the state space of the chain. Markov chains are often described by a directed graph, where the edges are labeled by the probabilities of going from one state to the other states. A simple example is shown in the figure below, using a directed graph to picture the state transitions. The states represent whether the economy is in a bull market, a bear market, or a recession, during a given week. According to the figure, a bull week is followed by another bull week 90% of the time, a bear market 7.5% of the time, and a recession the other 2.5%. From this figure it is possible to calculate, for example, the long-term fraction of time during which the economy is in a recession, or on average how long it will take to go from a recession to a bull market.

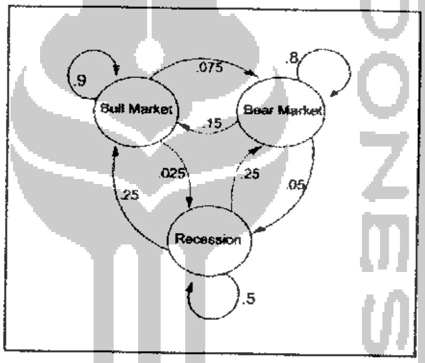

Figure 2.1 Example of state transition

The probability of going from state i to state j in n time steps is

$$P_{ij} = \Pr(X_n = j \mid X_0 = i) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.8)$$

and the single-step transition is

$$P_{ij} = \Pr(X_1 = j \mid X_0 = i) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.9)$$

For a time-homogeneous Markov chain:

$$P_{ij} = \Pr(X_{n+k} = j \mid X_k = i) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.10)$$

And

$$P_{ij} = \Pr\,(X_{k+1} = j \mid X_k = i) \quad\text{..}\quad (2.11)$$

so, the n-step transition satisfies the Chapman–Kolmogorov equation, that for any k such that $0 < k < \mathrm{n}$,

$$p_{ij}^{(n)} = \sum_{r\epsilon s} pir^{(k)}\, prj^{(n-k)} \quad\text{..}\quad (2.12)$$

The marginal distribution $\Pr(X_n = x)$ is the distribution over states at time n. The initial distribution is $\Pr(X_0 = x)$.

2.6 Hidden Markov Model

A *Hidden Markov Model (HMM) is a finite state machine which* has some fixed number of states. Hidden Markov models were introduced in the beginning of the 1970's as a tool in speech recognition. This model based on statistical methods has become increasingly popular in the last several years due to its strong mathematical structure and theoretical basis for use in a wide range of applications. If the parameters of the chain are known, quantitative predictions can be made. In other cases, they are used to model a more abstract process, and the theoretical underpinning of an algorithm. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are hidden to the outside, hence the name Hidden Markov Model

2.6.1 Characteristic of Hidden Markov Model

Hidden Markov Model is characterized by the following

a. number of states in the model (N / Q)

   $Q = \{q_1; q_2; : : : ;q_T\}$ - *set of states*

b. number of observation symbols (M / O)

   $O = \{o_1; o_2; : : : ;o_T\}$ - *set of symbols*

c. state transition probabilities (a_{ij})

   $a_{ij} = P(q_{t+1} = j|q_t = i)$

d. observation emission probability distribution that characterizes each state (b_j)

   $b_j(k) = P(o_t = k|q_t = j) \quad i \leq k \leq M$

e. initial state distribution (π)

As mentioned above the HMM is characterized by N, M, A, B and π. The a_{ij}, $b_j(O_t)$, and π_i have the properties:

$$\sum_j aij = 1, \qquad \sum_t bi(Ot) = 1, \qquad \sum_i \pi i = 1 \quad and$$

$$a_{ij},\ b_j(O_t)\ ,\ \pi_i \geq 0 \text{ for all } i,j,t. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.13)$$

2.6.2 Main issues using HMM

Most applications of HMMs are finally reduced to solving three main problems. These are:

a. Evaluation problem

Given the HMM $\lambda=(A, B,\pi)$ and the observation sequence $O=o_1\ o_2\ \dots\ o_K$, calculate the probability that model M has generated sequence O. Trying to find probability of observations $O=o_1\ o_2\ \dots\ o_K$ by means of considering all hidden state sequences.

b. Decoding problem

Given the HMM $\lambda=(A, B,\pi)$ and the observation sequence $O=_{o1\ o2\ \dots\ oK}$, calculate the most likely sequence of hidden states s_i that produced observation sequence O.

c. Learning problem

Given some training observation sequences $O=o_1\ o_2\ \dots\ o_K$ and general structure of HMM (numbers of hidden and visible states), determine HMM parameters $\lambda=(A; B, \pi)$ that best fit training data.

$$a_{ij}=\mathrm{P}(s_i\,|\,s_j) = \frac{\text{number of transition from state } Sj \text{ to state } Si}{\text{number of transitions out of state } Sj}$$

$$b_j(v_m)=\mathrm{P}(v_m\,|\,s_i) = \frac{\text{number of time observation } Vm \text{ occurs in state } Si}{\text{Number of times in state } Si}$$

$$\pi_i=\mathrm{P}(s_i) = \text{Expected frequency in state } s_i \text{ at time } k=1$$

In a Hidden Markov Model, three parameters need to be re-estimated, which is:

a. Transition probabilities (a_{ij})

b. Initial state distribution (π)

c. Emission probabilities [$b_j(o_t)$]

a. Re-estimating Transition Probabilities

What's the probability of being in state s_i at time t and going to state s_j, given the current model and parameters?

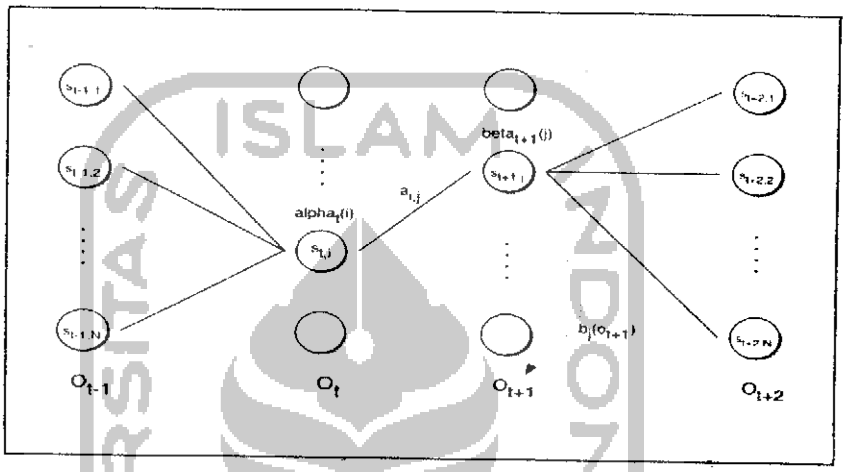

Figure 2.2 Reestimating transition probabilities

Given :

$$\xi_t(i,j) = P(q_t = s_i, q_{t+1} = s_j \mid O, \lambda) \quad ,$$

So that,

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{i,j}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{i,j}b_j(o_{t+1})\beta_{t+1}(j)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.14)$$

The intuition behind the re-estimation equation for transition probabilities is:

$$\hat{a}_{i,j} = \frac{\text{expected number of transitions from state } s_i \text{ to state } s_j}{\text{expected number of transitions from state } s_i}$$

$$=\hat{a}_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{j'=1}^{N} \xi_t(i,j')} \quad \text{.......................................} \quad (2.15)$$

Defining

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \quad \text{....................................} \quad (2.16)$$

As the probability of being in state s_i, given the complete observation O and can be written as:

$$\hat{a}_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad \text{.......................................} \quad (2.17)$$

b. Re-estimation of Initial State Probabilities

Initial state distribution π_i is the probability that s_i is a start state

$\hat{\pi}_i$ = expected number of state s_i at time 1

Thus can be written:

$$\hat{\pi}_i = \gamma_1(i) \quad \text{..} \quad (2.18)$$

c. Re-estimation of Emission Probabilities

Emission probabilities are re-estimated as:

$$\hat{b}_i(k) = \frac{\text{expected number of times in state } s_i \text{ and observe symbol } v_k}{\text{expected number of times in state } s_i}$$

$$\hat{b}_i(k) = \frac{\sum_{t=1}^{T} \delta(o_t, v_k)\gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \quad \text{.... .... (2.19)}$$

Where $\delta(o_t, v_k) = 1$, if $o_t = v_k$, and 0 otherwise

d. Updated Model

Coming from $\lambda = (A, B, \pi)$ we get to $\lambda' = (\hat{A}, \hat{B}, \hat{\pi})$ by the following update rules:

$$\hat{a}_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad \text{...................................... (2.20)}$$

$$\hat{b}_i(k) = \frac{\sum_{t=1}^{T} \delta(o_t, v_k)\gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}, \quad \text{...................................... (2.22)}$$

$$\hat{\pi}_i = \gamma_1(i) \quad , \quad \text{...................................... (2.23)}$$

2.6 Percentage Error

The percentage error for each number obtained by Percentage Mean Error (PME) stated below:

$$\left(\sum_{k=1}^{T} \frac{(Y_k - \hat{Y}k)}{Y_k} \text{ x } 100\% \right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ (2.23)$$

Where

Y_k = actual data

$\hat{Y}_{\mathrm{k}}$ = simulation data