CHAPTER III
METHODOLOGY

The methodology relating to the current study is discussed under the following heads

A. Selection of industries;

B. Period of study;

C. Selection of sick companies;

D. Selection of non sick companies;

E. Database of the study;

F. Hypothesis formulated in the study;

G. Concepts used in the study;

H. Tools of analysis applied in the study;

I. Models developed in the study;

J. Distinguishing the features of the models developed;

K. Methodology for testing the hypotheses and

L. Tabulation and analysis of data

A. Selection of industries

The study is confined only to large and medium industries. This is due to the fact that the intensity of the industrial sickness is more severe in these industries, since more than 83 percent of the outstanding bank credit is locked up in the sick units of these industries. Ten industrial

categories of steel, automobiles, chemicals, textiles, cement, fertilizer, sugar, tea, rubber tyres, and food and consumer products were chosen, since 72.8 percent of the total outstanding bank credit was locked up in the sick units of these categories. (Economic Survey, 2002-03)

B. Period of study

The study focused on the financial year 2000-01, since the year marked the beginning of the Second Generation reforms in India. Further in 2000-01, there was a severe set back in the industrial growth rate to 5.7 percent from 10.8 percent in 1999-2000. (Tenth Five Year Plan ,2002-07)

C. Selection of sick companies

Large and medium companies reported to have gone into sickness during the period chosen (2000-01) as per the Reserve Bank of India declaration and whose financial data for at least three years prior to sickness (1999-2000, 1998-99 and 1997-98), were available were alone selected for the sample of sick companies. Based on these criteria, 34 sick companies were identified in the ten industrial categories.

D. Selection of non-sick companies

The selection of non-sick companies was made independent of the industry category and size but based on the same financial year (2000-01) as that of their sick counter parts and whose three year financial data were available. Based on the above criteria, 38 non-sick companies were selected.

E. Database of the study:

The required data for the study were obtained from two sources:

1. Various issues of the Reports of Currency and Finance by the Reserve Bank of India, New Delhi.
2. Websites on Indian investments and industrial performance data

   (i) www.indiainfoline.com and

   (ii)www.indiastat.com

From the above sources, the balance sheets of the selected sick companies for one, two and three years prior to the year of sickness and of the non-sick companies for the same three years as that of their sick counterparts were obtained.

Based on the financial information available in the balance sheets– financial ratios, belonging to four major categories was estimated using a Microsoft excel spreadsheet. The four major groups were

(i) Turnover ratios which depict the operational efficiency of the company.

(ii) Liquidity ratios which show the company's ability to meet its short term financial obligation.

(iii) Solvency ratios which indicate the company's ability to meet its long term obligation and

(iv) Profitability ratios designed for the evaluation of the company's operational performance.

In the current study, 21 financial ratios comprising of eleven turn over ratios, three liquidity ratios, four solvency ratios and three profitability ratios were estimated and used. These ratios were selected for the construction of the prediction models based on the availability of the financial statement data.

F. Hypotheses formulated in the study:

The following three hypotheses were framed in this study

(1) There is no significant difference in the predictive accuracy of the PCA-MDA isolated and non-isolated models in the prediction of industrial sickness.

(2) There is no significant difference in the predictive accuracy of the PCA-ENN isolated and non-isolated models in predicting industrial sickness and

(3) There is no significant difference in the predictive accuracy of the PCA-MDA and PCA-ENN models in accurately predicting industrial sickness.

G. Concepts used in the study:

The following concepts were used in the study

(i) Non-SSI Sector: It refers to those industries whose total fixed capital investment exceeds Rs 3 crores. They are termed as large and medium scale units.

(ii) Prediction models: It refers to the techniques of forecasting sickness of a firm prior to it's occurrence.

(iii) Principal Components: These are derived orthogonal variables of reduced dimension which represents all the intrinsic content of the original variables.

(iv) Evolutionary Neural Networks: It is a soft computing model with the Multi-Layer Feed-Forward Neural Network as its host architecture and employs an evolutionary algorithm to determine its weights.

(v) Discriminant score: It is a statistical function which classifies companies into sick and non-sick classes on the basis of certain characteristics.

(vi) Chromosome: It is a set of character strings that defines a proposed solution to the problem that a genetic algorithm tries to solve.

(vii) Isolated database: It comprises of financial data of the companies prior to the year of sickness studied in an independent way.

(viii) Non-isolated Database: It comprises of the financial data of the companies prior to the year of sickness studied in an inter-related way.

(ix) Financial ratio: It is a quotient of two numbers, where both the numbers are financial statement values.

(i) Turnover ratios: These ratios the depict the operational efficiency of the company. The turnover ratios applied in the study are

a) Net sales to total assets;

b) Net sales to fixed assets;

c) Net sales to working capital;

d) Net sales to inventory;

e) Gross sales to depreciation assets;

f) Net assets to the total number of equity shares; (Net Asset Value)

g) Working capital to asset composition;

h) Total income to net sales;

i) Cost of material to net sales;

j) Employee cost to net sales and

k) Cost of sales to net sales.

(ii) Liquidity ratios: These ratios show the company's ability to meet it's short term financial obligations. The liquidity ratios used in the study are

a) Current assets to current liabilities;

b) Total long term debt to the shareholder's funds;

c) Income before interest and tax to the interest charges;

(iii) Solvency ratios: These ratios indicate the company's ability to meet its long term obligations. The ratios under this category included in the study are

a) Net profit before interest and tax to net worth;

b) Net profit before interest and tax to capital employed ;

c) Net profit after tax and preference dividend to number of equity shares and

d) Gross profit after tax and preference dividend to number of equity shares.

(iv) Profitability ratios: These ratios evaluate the company's operational performance. The study applied the following profitability ratios

a) Profit before interest, depreciation and tax to net sales;

b) Profit before tax to net sales and

c) Profit after tax to net sales.

(x) Expost sample data : It refers to testing the validity of the model using a within the sample period data termed as known instances.

(xi) Exante sample data: It refers to testing the reliability of the model using an out of sample period data referred to as unknown instances.

(xii) Classification accuracy rate : It is the percentage of correctly classified companies(sick or non-sick) out of its total companies(sick or non-sick)

a) Classification accuracy rate of sick companies=

$$\frac{\text{Number of correctly classified sick companies}}{\text{Total number of sick companies}} \times 100$$

b) Classification accuracy rate of non-sick companies=

$$\frac{\text{Number of correctly classified non-sick companies}}{\text{Total number of non-sick companies}} \times 100$$

(xiii) Type I error: It refers to sick companies misclassified as non-sick.

(xiv) Type II error: It indicates non-sick companies misclassified as sick.

H. Tools of analysis applied in the study

(i) Principal component Analysis (PCA)

PCA is a standard technique commonly used for data reduction, more specifically data reduction in statistical pattern recognition and signal processing. The reduced dataset retains most of the information content of the original data space. (Jollife, 1986).

PCA is a technique therefore to find the directions in which a cloud of data points is stretched most. These directions represent most of the information in the data and they allow to store the data in a compressed form and later reconstruct the data with minimal amount of distortion.

PCA underlines two facts

a) To estimate the co-efficient with which the original variable X is transformed into orthogonal variables i.e. the Principal components.

b) To establish some rule of decision about the number of principal components to be retained in the analysis.

The aim of PCA is the construction, out of a set of variables, $X_j (j = 1,2....n)$, new variables (P_i) called Principal Components which are linear combinations of the X's

$$P_1 = a_{11}x_1 + a_{12}x_2 \ldots\ldots a_{1n}x_n$$

$$P_2 = a_{21}x_1 + a_{22}x_2 \ldots\ldots a_{2n}x_n$$

$$\begin{array}{cc} . & . \\ . & . \\ . & . \\ . & . \end{array}$$

$$P_{K1} = a_{K1}x_1 + a_{K2}x_2 \ldots\ldots\ldots a_{Kn}x_n \qquad (1)$$

The PCA method can be applied by

(i) Using original values of X_J's

(ii) Taking deviations of the X_J's from the means

$$x_{j=}X_{J-}\overline{X_J} \qquad (2)$$

(iii) Taking standardized variables (deviations of X_j's from the mean and divided by the standard deviation)

$$Z_j = x_j / s_j \qquad (3)$$

The values of the principal components will be different depending on the way the variables are used.

In this current study the latter procedure (3) has been used as it is more general, where variables measured in different units can be handled.

The PCA is concerned with explaining the variance-covariance structure, through a few linear combinations. In general it satisfies two conditions.

(i) Principal components are uncorrelated (orthogonal)

(ii) The first principal components P_1 absorbs and accounts for the maximum possible proportion of the total variation in the set of all $X's$, the second principal component P_2 absorbs the maximum of the remaining variation and so on. Thus it transforms the original $X's$ into orthogonal artificial variables, the principal components.

Algebraically principal components are particular linear combinations of the random variables, $X_1, X_2 \ldots\ldots X_n$. Geometrically, these linear combinations represent, selection of a new co-ordinate system got by rotating the original system with $X_1, X_2 \ldots\ldots X_n$, as the co-ordinate axes. The new axes show the directions with maximum variability and provide a simple description of a co-variance structure (Koutsyannis, 1998). Thus principal components solely depend on the co-variance matrix.

Let X be an n-dimensional vector to be mapped to Y, which is in m-dimensional space $(m < n)$. If E is the mean square error equal to the sum of the variances of the components truncated from X for dimensionality reduction, then PCA finds an invertible transformation T such that

$$Y = T.\ X \qquad (4)$$

is optimum with respect to the mean square error.

Consider a population of *n*-dimensional pattern vectors $X = (X_1, X_2, X_3 \ldots\ldots X_n)^T$. The co-variance matrix of the pattern vector population is given by

$$C = E(XX^T) \tag{5}$$

Where *E(X)* is the expectation of *X*. *C* is a symmetric square *(n* x *n),* standardized matrix. Let W_i be the eigen vectors of *C* that corresponds to the *n* largest eigen values $\lambda_1, \lambda_2, \lambda_3 \ldots\ldots \lambda_n$.

Let $Q = (w_1, w_2, \ldots\ldots w_m)$ be the matrix of the *m* largest eigen vectors. Here m is the dimension for which the data space is to be reduced. Projecting the multivariate data vectors on to the space spanned by eigen vectors using

$$Y = X^T.Q \tag{6}$$

yields derived variables of dimension *m.*

The eigen vectors give the directions in which the data cloud is stretched most. The projections of the data on the eigen vectors are the principal components. The eigen values give an idea of the amount of information each principal component represents. The principal components of large eigen values represent greater information in the dataset. Thus eigen values provide a measure for the significance of abstract factors with respect to the original data. Thus principal components represent the absorbance data.

The eigen vectors associated with the largest eigen value has the same direction as the first principal component, the next largest eigen vector with second principal component and so on. Thus it shows

$$\lambda_1 > \lambda_2 > \lambda_3 .....\lambda_n > 0 \quad (7)$$

and $Var\,(Y_n) = a'_n \sum a_n \quad i = 1,2.....n$ (8)

$$Cov\,(\mathrm{Y_n}, \mathrm{Y_m}) = a'_i \sum a_k \quad i,k = 1,2.....n \quad (9)$$

The principal components are uncorrelated combinations $Y_1, Y_2 .....Y_n$ whose variances are as large as possible. Thus first principal component P_1 maximizes $Var\ (Y_1 = a'_1 \sum a_1)$. It is thus clear that $Var\ (Y_1 = a'_1 \sum a_1)$ can be increased by multiplying any a_1 by some constant.

Thus P_1 = Linear combination of $a'_1 x_1$ that maximizes $Var\ (a'_1 x)$ subject to $a'_1 a_1 = 1$

P_2 = Linear combination of $a'_2 x$ that maximizes $Var\ (a'_2 x)$ subject to $a'_2 a_2 = 1$

$Cov\,(a'_1 x, a'_2 x) = 0$

Therefore

$\mathrm{P_i}$ = Linear combination of $a'_i x$ that maximizes $Var\ (a'_i x)$ subject to $a'_n a_n = 1$ and covariance $(a'_n x, a'_m x) = 0$ where $m < n$. Thus the dataset X of dimension n is transformed into Y of dimension m.

Principal component analysis thus is more of a means to an end than an end in itself, because it frequently serves as an intermediate step in much larger investigations like multiple regression analysis.

(ii) Multiple Discriminant Analysis (MDA):

Multiple Discriminant Analysis is a statistical technique which helps in classifying observations in one of the several pre-specified classes on the basis of certain characteristics (Cryer and Miller, 1991).

The discriminant function is given by

$$Z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + ...a_n x_n \qquad (10)$$

Where Z = discriminant index, X_i, $i = 1,2....n$ are the independent variables and a_j, $j = 0,1...n$ are the coefficients of variables. Thus MDA performs a multiple linear regression fit and returns the coefficients (a_j) to aid prediction.

(iii) Evolutionary Neural Networks (ENN):

The Evolutionary Neural Network employed in the soft computing model has a Multi-layer Feed-forward Neural Network (MLFNN) as its host architecture and employs an evolutionary algorithm to determine the weights. The network considered for the specific problem is a three-layered one with the configuration *l-m-n.* Here *l* is the number of input neurons, *m* is the number of hidden neurons and *n* is the number of output neurons. For the industrial sickness prediction problem, *l* represents the

principal components of the financial ratios and $n = 1$ is a single output neuron, which ultimately releases an output of sick or non-sick. The performance of the network was studied for varying values of m. The input and the hidden layers have one bias neuron each.

The number of weights to be computed is

$$w = (\ell + 1)m + (m + 1)n \tag{11}$$

with a gene length of g_i chosen for encoding weights, the length of the chromosome is given as

$$\ell_c = w.g_i \tag{12}$$

The algorithm makes use of decimal coded genes (non binary encoding with digits 0-9). An initial population of ℓ_c randomly generated chromosomes is first generated. To determine the fitness values for each of the chromosomes, the weights in the range of (-10, +10) are extracted from the chromosomes using the following formula where w_k is the extracted weight from the k^{th} gene ($k \geq 0$)

$x_{kg\ell+1}, x_{kg\ell+2}, \ldots\ldots x_{(k+1)g\ell}$ given $x_1, x_2, \ldots x_{g\ell}, \ldots x_{L_c}$ to represent a chromosome:

$$w_k = \begin{cases} + \dfrac{x_{kg_l+2}10^{g_l-2} + x_{kg_l+3}10^{g_l-3} + \ldots\ldots\ldots x_{(k+1)g_l}}{10^{g_l-2}}, \\ \quad if \quad 5 \le x_{kg_l+1} \le 9 \\ - \dfrac{x_{kg_l+2}10^{g_l-2} + x_{kg_l+3}10^{g_l-3} + \ldots\ldots\ldots x_{(k+1)g_l}}{10^{g_l-2}}, \\ \quad if \quad 0 \le x_{kg_l+1} < 5 \end{cases} \qquad (13)$$

The inverse of the root mean square of the error obtained while learning the training dataset is the fitness value for the particular chromosome.

For reproduction, the two-point cross over technique is employed with the help of which successive generations of population are generated. The training is stopped when a convergence criteria of 100 percent is achieved (i.e.), when 100 percent of the population have converged to the same fitness values.

(iv) Tests of significance

The following tests of significance have been applied in the study for making a decision on either to accept or reject the different hypotheses framed in the study.

a) Test of significance of classification accuracy rate :

This test (Z test) was applied to test the significance of the estimated classification accuracy rate of the different models. The formula used was

$$z = \frac{P_s - 0.5}{\sqrt{0.5x\dfrac{0.5}{n}}} \qquad (14)$$

where P_s = Proportion of correctly classified sick/non-sick companies and

n = Total number of sick/non-sick companies in the sample

b) Test of significance of single sample test :

This test of significance was applied to test the hypothesis whether there is significant difference in the population distributions for the sick and non-sick companies, pertaining to both the expost sample (large sample) and the exante sample (small sample). This was also called as the equality of means test.

(1) For large samples

$$S_L = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (15)$$

where $\overline{X}$ = Sample mean;

μ = Population mean;

σ = Standard deviation of sample (large)

n = Sample size

(2) For small samples

$$S_s = \frac{\overline{X} - \mu}{s}\sqrt{n} \qquad (16)$$

where $\overline{X}$ = Sample mean;

μ = Population mean;

s = Standard deviation of sample (small)

n = Sample size

(c) Test of significance of two samples test :

This test was applied to test whether there was significant difference between the sample means of the discriminant score of two models in predicting sickness in companies. The test was used to compare the classification accuracy of the two models in classifying sick and non-sick companies, viz ; isolated and non isolated, pertaining to the expost sample (large sample) and the exante sample (small sample)

(1)For large sample

$$P_L = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}} \qquad (17)$$

where $\bar{X}_1$ and $\bar{X}_2$ = Means of the two samples(large)

s_1 and s_2 = Standard deviations of the two samples

n_1 and n_2 = Sizes of the two samples

(2) For small sample

$$P_S = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}} \cdot \sqrt{\frac{n_1 n_2}{n_1+n_2}} \qquad (18)$$

where $\bar{X}_1$ and $\bar{X}_2$ = Means of the two samples (small)

s_1 and s_2 = Standard deviations of the two samples

n_1 and n_2 = Sizes of the two samples

(I) Models developed in the study:

(i) PCA-MDA model

The input to the PCA is the financial ratio matrix $F_{N\times n}$ and the financial soundness Z of the companies (sick : Z = 1, non-sick : Z=0) where N is the number of companies (observations) and n is the number of financial ratios considered. The output of the PCA is the derived variable matrix D_{Nxm} where N is the number of companies and m ($m < n$) is the dimension chosen by the predictor. Now the derived value data set (D, Z) is used to estimate the discriminant function of the MDA. Here D represents the set of independent variable and Z the dependent variable. The output of the MDA is coefficient set $a_0, a_1, a_2 \ldots\ldots a_m$. Once the coefficients are got, equation (10) is used to predict the sickness of any company whose financial ratios are known. Thus if F' is the financial ratio vector of a company, it is standardized to a zero mean vector before obtaining its principal component D'. On evaluating equation (10) using D' and the known coefficients, $a_0, a_1, a_2 \ldots\ldots a_m$, the Z value is obtained. If $Z > 0.5$, the company is graded as sick and if $Z < 0.5$ it is graded as non-sick. The software IDL Ver 5.0 was used to estimate and choose the

appropriate principal components and thereafter determine the coefficients of variables in the MDA component of the model.

Learning phase of the model

In this phase the predictive capabilities of the PCA-MDA model are determined by first obtaining the coefficients of variables involved in the MDA component of the model.

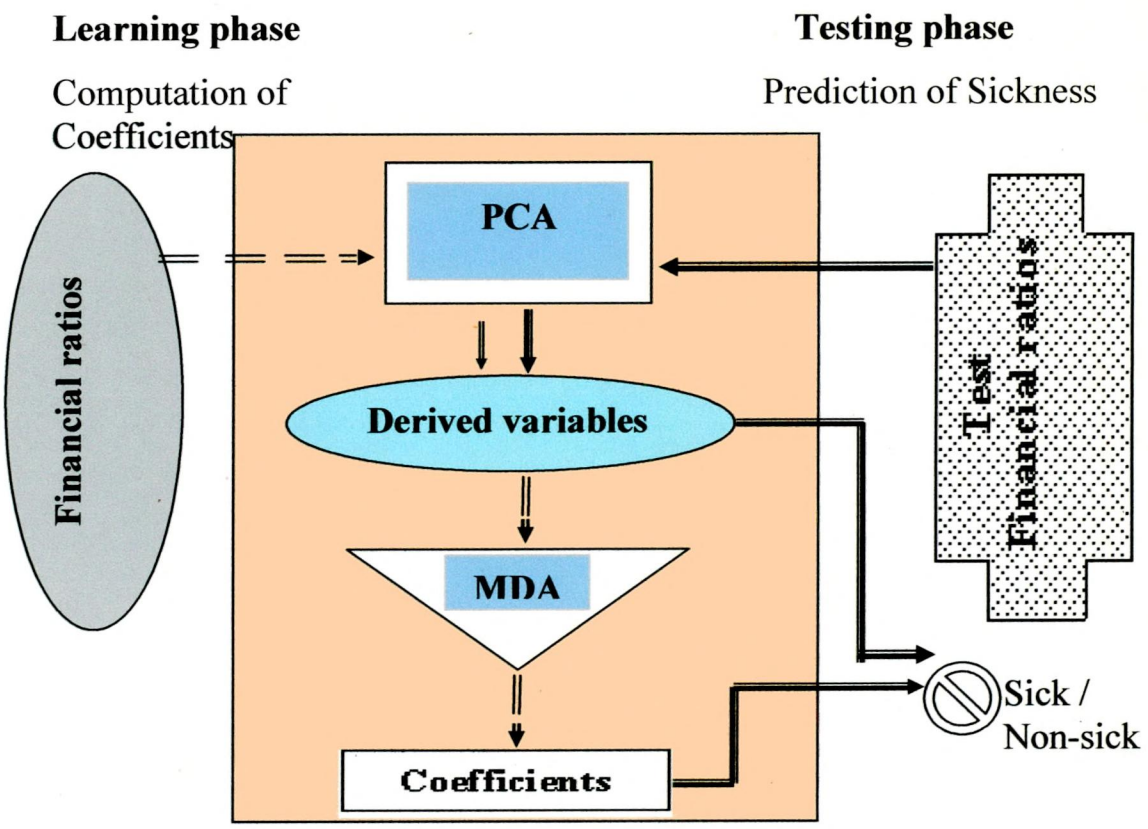
Testing phase of the model

The predictive capabilities are tested over the financial datasets of companies within the sample period (known data sets or *expost sample* data) and out of sample period (unknown data sets or *exante sample* data) used during the learning phase.

The flow of computation of the PCA-MDA model is depicted in Figure II

FIGURE II:

FLOW OF COMPUTATION IN THE PCA-MDA MODEL



The PCA-MDA model was *tested over the isolated and non-isolated financial data base*. The isolated data base comprised the financial data (21 financial ratios) of the 72 companies (34 sick and 38 non-sick) considered independently for the *learning phase* as one, two and three years prior to the year of sickness (2000-01) i.e. year1, year2 and year3 data sets only. The non-isolated data base for the learning phase constituted the financial data of the 72 companies over a period. With the observation period chosen as 3 years, the study employed data

sets pertaining to year1, years 1 and 2 considered together and years 1, 2 and 3 considered together for the determination of coefficients of variables. The *testing phase* involved two test sets out of which one included within the sample data sets pertaining to 72 companies (34 sick and 38 non-sick) for the financial year 1999-2000 only, termed as the *expost sample* (known instances) or validation test. The other test set included the out of sample datasets pertaining to 13 companies (7 sick and 6 non-sick) for the financial year 2000-01 only, termed as the *exante sample* (unknown instances) or forecast test.

(ii) PCA-ENN Model

The input data set to this model comprised of the chosen principal components (7) as per Kaiser's criterion, of the 21 financial ratios of the 72 Indian manufacturing companies (34 sick and 38 non-sick), as used in the PCA-MDA model.

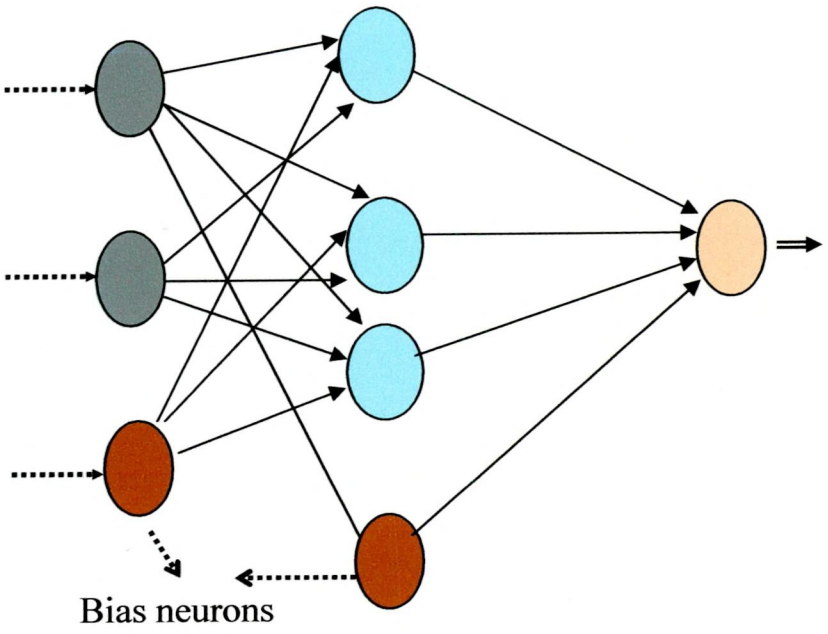
The ENN classifier used the same isolated and non-isolated data base used in the PCA-MDA model as it's training sets to train the network. A notable feature of the hybrid neural network is its fast convergence during the training session. In this particular case the network converged in less than 200 generations. The configuration of the network was *7-m-1* where m is the number of hidden neurons. For the study *m* was chosen to be 15.The principal components (7) chosen for the model represented the number of input neurons and 1, the number of

output neuron which was either sick or non-sick. The training set was subjected to several runs for authenticating the results of the training sets. The software Salford C was applied to estimate the results under this model.

The architecture of the Evolutionary Neural Networks is depicted in Figure III

FIGURE III:

ARCHITECTURE OF THE EVOLUTIONARY NEURAL NETWORKS



The test sets considered here were the same as in the PCA-MDA model viz., expost sample and exante sample data sets.

J. Distinguishing features of the models developed in the study

(i) The technique of principal component analysis used in this study allows any number of financial ratios as inputs.

(ii) The selection of the sick and non-sick companies have been made irrespective of their size, industry category and capital employed.

(iii) The prediction models were tested not only on isolated data sets but also on non-isolated data sets.

(iv) The models were tested for external validity, using an out of sample period data (2000-01).

K. Methodology for testing the hypotheses

The first hypothesis was related to finding out whether there was significant difference in PCA-MDA model isolated and non-isolated model, in predicting the sick and the non-sick companies. Both the models were subjected to a validation test on a within the sample period, (expost sample dataset of 72 companies) as well as to a forecast test on an out of sample period (exante sample dataset of 13 companies). The PCA with the MDA as classifier was used to classify the companies into sick and non-sick companies. The predicted group of each company was compared with the actual group. The number of correct classifications was pooled into a classification matrix. Z test statistic was applied to test the classification accuracy. The hypothesis was then tested with a single sample test at 95 percent level of significance, individually for the isolated and non-isolated dataset models. Further while comparing the two models-isolated and non-isolated, the test of significance of

differences in the means of the discriminant scores of two samples were used. While the test of significance of large samples was applied to the expost sample dataset, the test of significance for small samples was applied to the exante sample dataset to test the hypothesis at 95 percent level of confidence.

The second hypothesis was related to finding out whether there was significant difference in PCA-ENN model, in predicting the sick and the non-sick companies pertaining to the isolated and non-isolated models. to the determination of the predictive ability of the soft computing PCA-ENN model pertaining to the isolated and the non-isolated dataset models. The two models were also subjected to the validation test and the forecast test on the same expost sample and the exante sample datasets, as applied in the PCA-MDA model. The PCA with the ENN as classifier was used to classify the sick and the non-sick companies. The predicted group was compared with the actual group of each company. The number of correct classifications was pooled in a matrix. Z test was applied to test the classification accuracy. The hypothesis later was tested applying the single sample test of significance at 95 percent level for the isolated and non-isolated models, individually. Subsequently while comparing the isolated and the non-isolated models, the test of significance of differences in the means of discriminant score of two samples was applied. The test of significance of large samples was applied to the

expost sample data and the test of significance of small samples to the exante sample data, to test the hypothesis at 95 percent level of confidence.

The third hypothesis is related to the comparison of the predictive accuracy of the two sickness prediction models, namely the PCA-MDA and the PCA-ENN models. The results of the classification accuracy of the PCA-MDA and the PCA-ENN models for the isolated and the non-isolated datasets for the same expost and the exante sample datasets were compared. The test of significance of differences in the means of the discriminant score of two samples was applied to test the difference between the two models in predicting industrial sickness. The hypothesis was thus tested using the test of significance of large samples for the expost sample and of small sample for the exante sample dataset, at 95 percent level of confidence.

L. Tabulation and analysis of data:

The data collected were tabulated and analyzed in the following chapter on "Results and Discussion".

****