Development and Testing of an Efficient Artificial Neural Network Algorithm and Its Effectiveness for Prediction of Insurance Claims

First Ashwini Bapat and Second Dr. Prakash Bapat

Abstract - **This paper presents the development of an efficient Back -propagation Artificial Neural Network (ANN) Algorithm suitable for prediction of occurrence of motor insurance claims. So, first an ANN Algorithm is created and then tested for standard engineering curves for efficiency and effectiveness. Then the motor insurance claims data based on past claim settlement records is sorted and a representative data is built. Using this data a predictive model is created by applying ANN Algorithm. Then the model is validated with the help of a new data set again representing the motor insurance claims sample set. It is concluded that the ANN prediction model developed here can be effective tool for prediction of the future motor insurance claims based on the pre-existing data. There is a scope for improvement of the Algorithm for faster convergence and higher accuracy. At the same time development of systematic procedures for selection of representative data is required.**

Index Terms - **Back propagation, motor insurance claims, predictive model, Steepest Descent Optimization.**

I. INTRODUCTION

This paper presents developing an Artificial Neural Network Algorithm and testing it and its application for prediction of motor insurance claims. In this work, a computer algorithm has been developed and tested for its performance for implementation of training process of multiplayer perceptrons. The ability of computer program in identifying the non-linear character of certain curves, which are frequently encountered in mechanical engineering, has been tested.

For testing of the algorithm, standard curves utilized in Mechanical Engineering are used. The curves taken here for verification are P-H curve of NH3 refrigerant, Mollier chart and Temperature vs. viscosity characteristic of SAE10 bearing oil.

The algorithm was successfully tested and applied to developing a predictive model for Motor Insurance Claims. Artificial Neural Network is popularly utilized in insurance sector as a predictive modeling tool by many workers in this field. The sample data used for this purpose is of motor insurance claims data

For every business, statistical analysis is an important part. In insurance sector, statistical analysis plays the most important role. Insurance business is all about speculating the frequency and severity of the risk covered. The more the accurate forecast the better the performance of the business.

In this paper, the back propagation Neural Network training is done by the method of Steepest-Descent. This method requires information about the gradient vector (first derivative in the local neighborhood) of the operating point. First a training set is provided for the algorithm to learn and then it has been tested.

The results of the training process for the three cases selected for performing testing are presented in the chapter on computer simulation. The time required for training is studied as the accuracy is varied. The effect of varying the number of neurons in the hidden layer is also studied.

It is observed that the computer software developed in this work is quite effective in identifying the non-linear character of the physical curve. The computer program can be made more efficient by implementing faster and heuristic optimization algorithm.

There are some ANN softwares available for use. But, they are like black box as the source code is not available. Suitable changes in such softwares are not possible and the flexibility is less. Hence, a job specific algorithm needs to be developed. So, in this work it is proposed to develop an ANN algorithm which will be more effective for insurance prediction work.

II. ARTIFICIAL NEURAL NETWORK ALGORITHM DESCRIPTION

The whole algorithm may be divided into two parts:

A. The Back-propagation Model
B. The Steepest Descent Optimization Technique

A. The Back-propagation Model

For the back-propagation model, here multilayer feed forward method is used. Back-propagation is a supervised learning technique used for training artificial neural networks.

Manuscript received on May 27, 2010.

F. Ashwini Bapat is a student of National Insurance Academy School of Management, Pune-45, India. (Contact No: +91 9226930122, e-mail id: ash.bpt@gmail.com).

S. Dr. Prakash Bapat is Professor, Mechanical Engg. Department, Cummins College of Engineering for Women, Pune-52, India. (Contact No: +91 9273303942, e-mail id: pakyabapat2@yahoo.co.in)

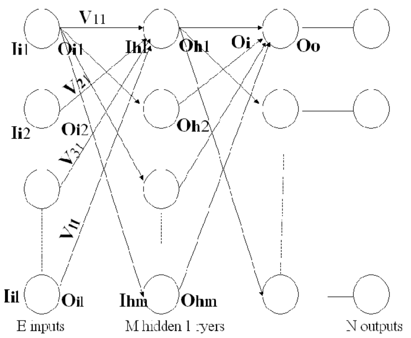


Figure [1]: Multilayer Feed-forward Model

Multilayer Feed-forward (MLFF) network with back-propagation (BP) learning network is sometimes called multilayer perceptron. The structure of the MLFF is as shown in figure [1].

This model has three layers; an input layer, and output layer, and a layer in between not connected directly to the input or the output and hence, called the hidden layer. For the perceptrons in the input layer, linear transfer functions are used and for the perceptrons in the hidden layer and the output layer, sigmoidal or squashed-S functions are used. The input layer serves to distribute the values they receive to the next layer and so, does not perform a weighted sum or threshold.

The sigmoidal function used for the MLFF model is as below:

S (x) = 1/ [1 + {(e^-a) * x}]

The coefficient is a real number constant. Usually in NN applications is chosen between 0.5 and 2. As mentioned above, the output of the input layer is the normalized value of the input. Then weights are attached to these values and are entered into the hidden layer. There it goes through the sigmoidal function and the output of the hidden layer is again attached with weights. These values are entered in the output layer. The output layer is again sigmoidal function. The output of the output layer is the final output which is compared with the actual output. If the error is above the required limit then it again goes through the optimization process.

The number of neurons (m) can be varied. As the number of neurons increases the accuracy increases. The algorithm is designed in such a way that the accuracy required can be achieved according to the user in terms of accuracy level or number of iterations to be performed.

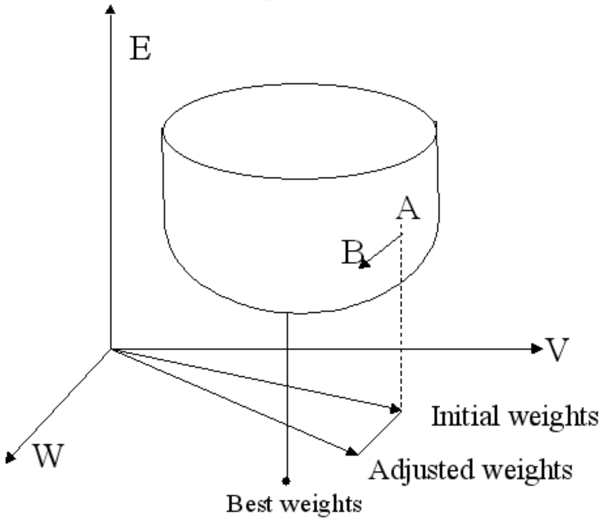


Figure [2]: Euclidian norm of Errors

B. *The Steepest Descent Optimization technique*

At the start of the training process, gradient descent search begins at a location with error value E determined by initial weight assignments W (0), V (0) and the training pattern pair (I^p, O^p) where

$$\boldsymbol{E} = \frac{1}{nset}\sum_{p=1}^{nset} E^p = \frac{1}{2\mathrm{X}nset}\sum_{p=1}^{k}(I_k^p - O_{Ol}^p)^2$$

During training, the gradient descent (shown in figure [2]) computations incrementally determine how the weights should be modified at each new location to move most rapidly

in the direction opposite to the direction of steepest ascent (a steepest descent). After the incremental adjustments to the weights have been made, the location is shifted to a different E location on the error weight surface. This process is repeated for each, training pattern (or each epoch [I^p, O^p], p = 1,2 …….. nset), progressively shifting the location to lower level until a threshold error value is reached or until a limit on the total number of training cycles is reached.

C. *Flowchart of the algorithm:*

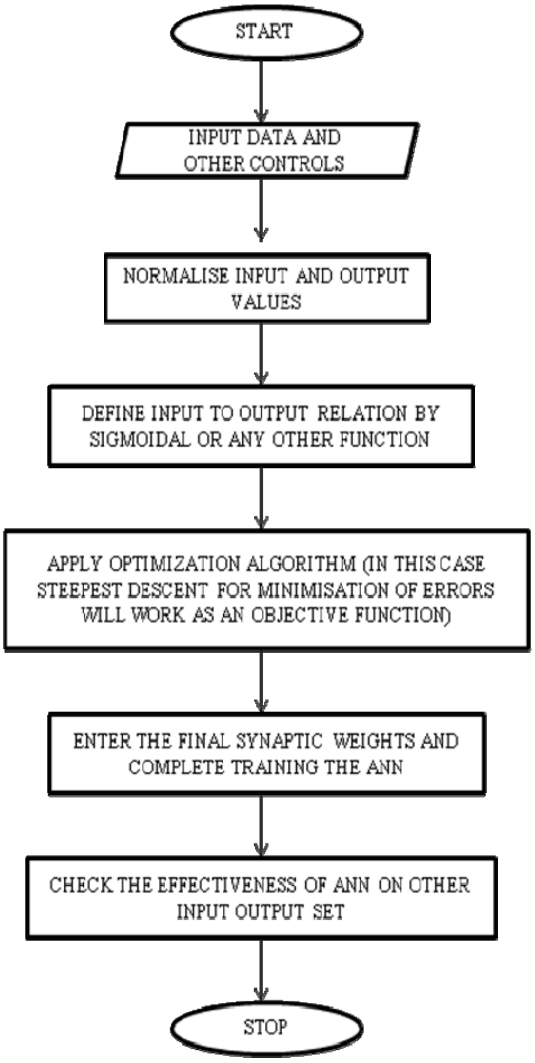

START
INPUT DATA AND OTHER CONTROLS
NORMALISE INPUT AND OUTPUT VALUES
DEFINE INPUT TO OUTPUT RELATION BY SIGMOIDAL OR ANY OTHER FUNCTION
APPLY OPTIMIZATION ALGORITHM (IN THIS CASE STEEPEST DESCENT FOR MINIMISATION OF ERRORS WILL WORK AS AN OBJECTIVE FUNCTION)
ENTER THE FINAL SYNAPTIC WEIGHTS AND COMPLETE TRAINING THE ANN
CHECK THE EFFECTIVENESS OF ANN ON OTHER INPUT OUTPUT SET
STOP

Figure [3]: Flowchart for Back-propagation program

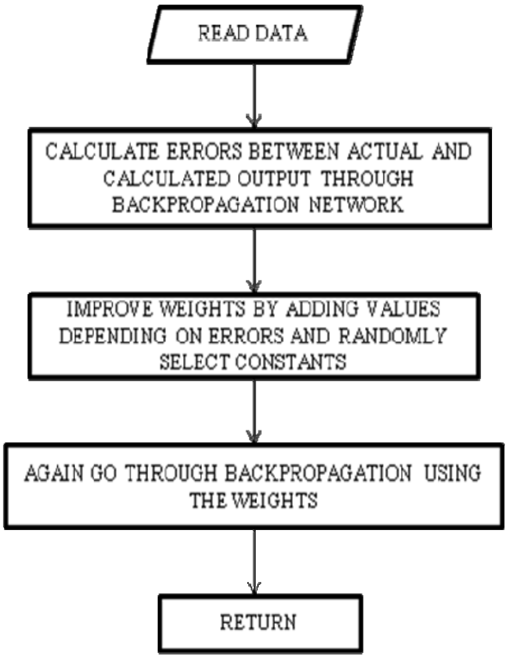

READ DATA
CALCULATE ERRORS BETWEEN ACTUAL AND CALCULATED OUTPUT THROUGH BACKPROPAGATION NETWORK
IMPROVE WEIGHTS BY ADDING VALUES DEPENDING ON ERRORS AND RANDOMLY SELECT CONSTANTS
AGAIN GO THROUGH BACKPROPAGATION USING THE WEIGHTS
RETURN

Figure [4]: Flowchart for steepest descent optimization technique

III. Testing of The Algorithm and Results

In this paper, the testing is done with the help of some curves popularly used in the mechanical field. Here three different characteristics are studied as Relation between temperature and viscosity of bearing oil SAE10, Relation between temperature and pressure, specific volume, enthalpy, (liquid and vapor) and entropy (liquid and vapor) for refrigerant NH3 and the Mollier Chart

The algorithm is designed in such a way that the accuracy required can be achieved according to the user in terms of accuracy level or number of iterations to be performed. Thus, for the purpose of testing the algorithm was run at various numbers of iterations and accuracy levels. The number of hidden layers also varied and the time required for the computation by the algorithm was noted. After testing the algorithm for these various conditions following results were observed. The results of the testing are consolidated in the table [1] below.

TABLE [1]: SUMMARY OF RESULTS

Object	No. of iterations	No of Neurons in hidden layers	Accuracy	Time of execution (min: sec: micro sec)
Bearing Oil SAE10	5,000	8	0.05	0:0:65
	10,000	8	0.01996	0:21:27
	10,000	10	0.01986	0:19:20
	50,000	8	0.01671	1:4:57
	50,000	10	0.0167	1:0:12
	19,335	10	0.0168	0:27:71
	50,000	20	0.028	11:19:00
Mollier chart	50,000	15	0.028	0.2:0.13
	10,000	20	0.0374	2:5:75
	10,000	15	0.02445	1:50:46
	968	15	0.03	0:13:02
	82	20	0.05	0:40:69
NH_3	50,000	5	0.023	180:30:00
	10,000	5	0.138	1:10:0
	10,000	8	0.1368	0.39:50

IV. Application of the Algorithm for Motor Insurance Claims

After the satisfactory testing of the algorithm, it was used for the commercial purpose where the curves are unknown and the relationship is needed to be established. So, the curve fitting is one of the important factors in Insurance sector for the estimation of frequency and severity of the claims to occur. Motor insurance claims data is taken into consideration for this work. The factors which are considered here are:

1. Age of the Vehicle
2. Sex of the driver
3. Age of the Driver
4. Driving Experience
5. Place of Repair
6. Claims

Here, the first 5 factors mentioned above were inputs and claim amount in INR was the output. The data size available is very large and the whole data cannot be utilized due to the limitation of the RAM size of computer. Thus, a part of data was needed to be extracted from the whole data which would represent the whole sample set.

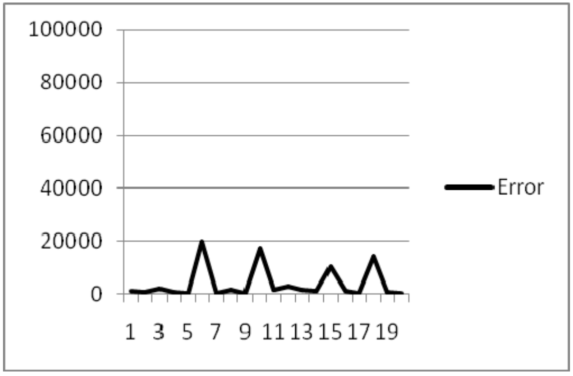
A. Development of Predictive Model for Motor Insurance Claims:

So, first the data was collected to cover all ranges of motor vehicle users. This included all vehicle models, geographical regions, vehicle ages etc for example collecting the cases of claims for all ranges of ages starting with ages 18-25, 45-60 and 60 and above. The same was considered for every field.

After collecting data rigorously, the next step was to clean the data to avoid any ambiguous data. So, various tests were applied to the excel sheet of data like the age of the driver should not be less than 18 years otherwise the claims will not be paid.

After cleaning the data, it needed preparation of a sample data for the learning by the algorithm to create weights to be attached to the inputs for establishing the relationship. So, a sample size of 20 values was created using random number generation. So, the data was arranged with a serial number. 20 random numbers were generated. With the help of these random numbers, 20 serial numbers were selected from the data and the values corresponding to these serial numbers were utilized to prepare the input file.

Using this data, first the supervised learning was done using 5000 iterations and 20 numbers of hidden layers. The error rate was 0.01332. The weights that have been found out were used to generate the relationship. Using these weights, the calculated outputs were generated. The errors thus calculated are plotted in rupees as shown in graph [1].



Graph [1]: Error in rupees as calculated during training by ANN

The data used for preparing the input file is shown in the table [2].

Here, the age of the driver column represents the age group. The ages are divided in different groups starting from 18.

Since the error rate was 1.3%, it can be seen that the error line is showing errors within a narrow gap and the maximum error is found to be 20,000 INR. So, the next step was to verify the relationship using another set of data.

B. *Validation of the Predictive Model:*

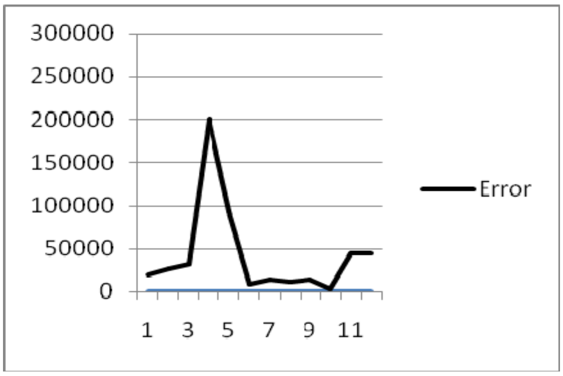
Again a small sample set of size 12 was created and was fed to the algorithm as inference data. Using the weights generated above, again outputs were calculated. This time the error rate was 0.14036. This means the accuracy level was around 86%.

Validation of the ANN model is done by creating another data file which is developed from the values shown in the table [3].

The errors are thus calculated in rupees and are plotted as shown in graph [2].

TABLE [2]: MOTOR INSURANCE CLAIMS DATA FOR DEVELOPMENT OF PREDICTIVE MODEL

SERIAL NUMBER	VECHMAKE	VEHICLE AGE	DRIVER SEX	DRIVER AGE	DRIVING EXP	PLACREP	CLAIMS
1	MARUTI ZEN	3.00	M	3.00	4.00	3	42600
2	HYUNDAI SANTRO	1.00	M	2.00	5.00	1	4600
3	MARUTI ZEN	4.00	M	4.00	2.00	3	51250
4	MARUTI 800	4.00	M	3.00	4.00	2	9030
5	HYUNDAI SANTRO	1.00	M	3.00	2.00	1	18700
6	TATA INDICA	1.00	M	2.00	5.00	3	15940
7	MARUTI 800	2.00	M	3.00	2.00	1	3948
8	MARUTI 800	1.00	M	2.00	4.00	1	14000
9	TATA INDICA	1.00	M	2.00	4.00	3	3800
10	MARUTI ZEN	1.00	M	2.00	5.00	3	9447
11	MARUTI ESTEEM	1.00	M	3.00	5.00	1	27400
12	MARUTI ZEN	2.00	M	1.00	3.00	2	16020
13	MAHINDRA ARMADA	2.00	M	1.00	2.00	1	31830
14	MARUTI 800	2.00	M	1.00	4.00	1	8242
15	DAEWOO MATIZ	1.00	M	2.00	4.00	1	12500
16	MARUTI ESTEEM	1.00	M	1.00	2.00	1	9050
17	HYUNDAI SANTRO	1.00	M	2.00	1.00	2	7061
18	MARUTI 800	1.00	M	2.00	4.00	1	16800
19	HONDA CITY	2.00	M	4.00	5.00	2	18100
20	MARUTI 800	5.00	F	2.00	1.00	3	11605



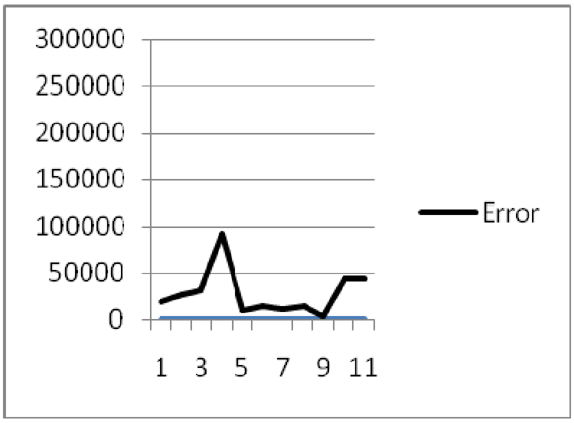
Graph [2]: Validation result – error in rupees

It can be seen from the graph [2] that the error corresponding to point 4 is very high. It seems to be the outlier which may be a moral hazard common to insurance field. Thus, after rejecting that value, again the values are plotted as shown in graph [3].

V. Conclusion and Discussion

This work was aimed at developing software for implementation of back-propagation algorithm and then studying its applicability to the Insurance field problems.

Artificial Neural Network utilizes not only linear relationship but non-linear relationship also for curve fitting. This helps to cover all aspects of mathematical relationships.



Graph [3]: Validation of results by removing outlier – Error in rupee

A. *Testing of Algorithm:*

Artificial neural network program developed for this work was tested for effectiveness on three cases as Relation between temperature and viscosity of SAE10 bearing oil, P-H, T-S and P-V curve for refrigerant NH3 and Mollier chart.

It was observed that the algorithm is capable of effectively identifying non-linear nature of the curve and learn to predict all the results with high degree of accuracy. An accuracy level of up to 0.015 i.e. 1.5% in prediction can be obtained

TABLE [3]: MOTOR INSURANCE CLAIMS DATA FOR VALIDATION OF THE PREDICTIVE MODEL

SERIAL NUMBER	VECHMAKE	VEHICL E AGE	DRIVER SEX	DRIVER AGE	DRIVIN G EXP	PLACRE P	CLAI MS
1	MARUTI 800	2.00	M	4.00	2.00	1	3064
2	HONDA ACC	2.00	M	2.00	5.00	1	22000
3	HYUNDAI SANTRO	2.00	M	3.00	5.00	1	22380
4	TATA SUMO	3.00	M	2.00	5.00	3	8650
5	TATA SUMO	1.00	M	3.00	5.00	1	17700
6	MARUTI 800	2.00	M	4.00	2.00	1	6309
7	CONTESSA	2.00	M	4.00	5.00	3	42050
8	MARUTI ZEN	4.00	M	2.00	5.00	3	95000
9	HYUNDAI SANTRO	1.00	M	1.00	1.00	1	16349
10	MARUTI ZEN	2.00	M	3.00	5.00	1	11312
11	MARUTI OMNI	2.00	M	3.00	5.00	1	26500
12	DAEWOO MATIZ	1.00	M	2.00	4.00	1	13350

The other observation is that the algorithm learning time is quite fast when accuracy level of up to 3% is allowed. Higher accuracy level requires long convergence time. The number of neurons in the hidden layer also accelerates the learning rate.

The speed and effectiveness of the algorithm can be augmented by using other more effective optimization techniques.

B. *Algorithm Applied to Insurance field data:*

It can be seen from the data that even with just 5000 iterations the accuracy level was very high i.e. .0133. But, for the inference data the error rate increases to 14%. This is due to extreme non linear nature of the data and better accuracy in prediction can be achieved by increasing the accuracy level of the algorithm.

Artificial Neural Network helps to improve the accuracy to forecast the occurrence of future claims. This will help in pricing the insurance product better.

It helps to reduce the processing time for the data. Once the relation has been established using the training set, further data processing takes less time.

C. Limitations while doing project:

Required data was not available for other insurance sector.

Higher level of computer RAM was required which was not available. So, the program could not be run for more number of iterations. Thus the error rate was comparatively high.

Again because of the limitation of the RAM size, the sample size had to be smaller. This work is done on a dual core processor laptop with a 1 GB DDR2. Faster computing equipment may prove to be more effective.

D. Future Scope:

The data selected for model building and for testing needs to be cleaned further of outliers. It can be further extended to determine the influence of each parameter considered here. This can be done by deleting and adding the parameters and then repeating the same process. The accuracy level for this can be checked. It will help to find out the parameter which affects the output significantly. It will also help to decide how increasing or decreasing the factor affects the output. For example, in the above case, it indicates that as the age of driver increases claim amount decreases.

Thus, with the help of the results of this experiment, the algorithm can be further used for advanced purposes. In this paper for motor insurance claims data only 5 factors as mentioned above were considered. But, in India more than 10 factors are considered by various insurance companies like color of the vehicle, time of driving, geographical location, education of the driver. These factors can be helpful in developing an effective predictive model for better prediction of accidents.

Moral hazard is one of the difficulties in this statistical analysis as it corrupts the data. Such moral hazards are observed all over the world. So, further filtration of data with more scientific methods is required.

Motor insurance is only a part of insurance sector there are various other fields in insurance. One of the major parts is life insurance. With the help of better prediction, more precise premium rates can be applied.

In other fields of insurance where the consistency of data is very low and the sample size is very small, it will be difficult to use Artificial Neural Network Algorithms only. In such cases other modeling techniques such as Fuzzy and or Fuzzy Neural can be tried.

Although the initial work is done on laptop, final modeling may be developed on more powerful machines for higher accuracy of the model.

REFERENCES

[1] "Neural Networks, Fuzzy Logic and Genetic Algorithms – Synthesis and Application" S. Rajasekeran and G.A. Vijayalaksmi Pai, Prentice Hall India

[2] Neural Networks – A Comprehensive Foundation- Simon Haykin, Pearson Education Asia

[3] Understanding Neural Networks and Fuzzy Logic- Basic Concepts and Applications- Stamatios .V. Kartalopoulos, Prentice Hall India

[4] Elements of Artificial Neural Networks – Kishan Mehrotra, Chilukari .K. Mohan and Sanjay Raka, Penram International Publishing India.

[5] Optimization for Engineering Design- Algorithms and Examples- Kalyanmoy Deb, PHI.

[6] "Design Data Book" – B.D.Shivalkar, Central Techno Publications.

[7] "Refrigeration and Air-Conditioning" – R. S. Khurmi and J. K. Gupta

[8] Motor Insurance claims data from an Indian Insurance Company