Analysis of Data Mining Techniques on Real Estate

Geetali Banerji, Kanak Saxena

Abstract: Data mining techniques are broadly classified into two classes (i) Statistical Techniques and (ii) Knowledge Discovery. The continuing rapid growth of on-line data and the widespread use of databases necessitate the development of techniques for extracting useful knowledge and for facilitating database access. This paper analyzes the results of multilayer perceptron with pace regression and suggests a very efficient pattern which can be proved beneficial for knowledge discovery. The analysis is done using real estate data set which contains 5821 tuples and 43 attributes and determines that in India's scenario the demographic details of a person plays a very prominent role in identifying the investment behavior of a customer. In multilayer perceptron model, input layer is followed by two hidden layers. The first hidden layer contains 21 nodes as per various attribute weight age followed by second hidden layer which assigns re-processed weights to each of the 21 nodes. If we are discarding the demographic details then the model which is available consists of 13 Sigmoid nodes and there is a major change in error rate and correlation. We have used WEKA for analysis and found that in general multilayer perceptron(selected) is more efficient then pace regression(complete) in terms of statistical methods, but in Indian perception pace regression(complete) is more efficient than multilayer(selected).

Keywords: Multilayer Perceptron, Neural Network, Pace Regression

I. INTRODUCTION

This paper reflects the need of the investors or person in the interest of real estate era. Today, technology no doubt affects every aspect of person's perception in living, learning, playing, working etc. Even the person having their own house(s), one can be interested to renovate it, or to buy a new for their comfort or for the investment. The Real estate can be divided into three categories, commercial, residential and agricultural. Here we consider only predicting that the investor is having power to invest or not. The real estate professionals worked and found the below three factors to be very critical for their survival.

- Must continuously change the building design or land as per the current trend to attract more users.
- Innovative ideas to construct buildings and
- Relationship to be formed among every participant in the building life cycle.

Thus to select the participants is one of the key question in the real estate. Thus, across a wide variety of fields, data are being collected and accumulated at a dramatic pace.

There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. As large data sets encompass hidden trends, which convey valuable knowledge about the data set. The derived or acquired knowledge is very helpful in predicting the behavior of the user based on the data description. It can be expressed as rules or correlations highlight the associations that exist in the data [14]. For this we use the statistical analysis and knowledge discovery. With the statistical analysis we used many regression methods as linear, isotonic, pace and found that pace is more efficient among all [10]. For the machine learning process we use multiple perceptron for predicting the investment power with learning followed by the comparison among two.

The rest of the paper is organized as follows. Section II and III represents the introduction of multilayer perceptron and pace regression respectively. Section IV deals with the analysis of the real state data with was collected with 43 attributes and 5821 tuples. Section V concludes the paper with the future directions to continue the research in this direction.

II. MULTILAYER PERCEPTRON ARCHITECTURE

Neural networks are biological systems (a k a brains) that detect patterns, make predictions and learn. The artificial ones are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models from large historical databases. Despite the fact that scientists are still far from understanding the human brain let alone mimicking it, neural networks that run on computers can do some of the things that people can do. Because of the origins of the techniques and some of their early successes the techniques have enjoyed a great deal of interest. To understand how neural networks can detect patterns in a database an analogy is often made that they "learn" to detect these patterns and make better predictions in a similar way to the way that human beings do. Figure 1depicts a simple neural network. In our case the network takes in values for predictors for 42 various attributes of customers and predicts whether the person will be potential enough to invest in real estate.

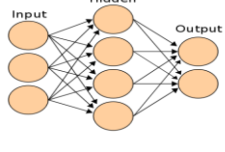


Fig.1 Neural Network

All neural networks have an input layer and an output layer, but the number of hidden layers may vary. Figure 2 is a

Manuscript received on July, 2012.
Geetali Banerji, Pursuing Phd. Banasthali University.
Kanak Saxena, Ph. D., computer science, Devi Ahilya University, Indore

diagram of a perceptron network with two hidden layers and four total layers:

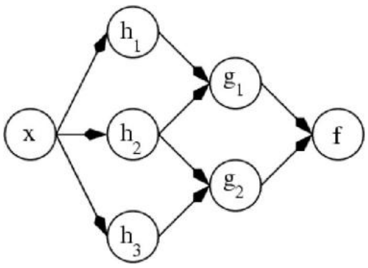


Fig. 2 Multilayer Perceptron

When there is more than one hidden layer, the output from one hidden layer is fed into the next hidden layer and separate weights are applied to the sum going into each layer.

A. Training Multilayer Perceptron Networks

The goal of the training process is to find the set of weight values that will cause the output from the neural network to match the actual target values as closely as possible. This can be done in following three steps.

- Train a neural network for classification problem using the data set.
- Prune the data to obtain optimized architecture in order to decrease the complexity.
- Generating the rules(Knowledge discovery and extraction)

B. Algorithm [11]

The algorithm for Perceptron Learning is based on the back-propagation rule discussed previously. This algorithm can be coded in any programming language. In this case we are assuming the use of the sigmoid function *f (net)* this is because it has a simple derivative.

i. Initialise weights and threshold.
Set all weights and thresholds to small random values.

ii. Present input and desired output
Present input $\mathbf{X}_p = x_0, x_1, x_2, \ldots, x_{n-1}$ and target output $\mathbf{T}_p = t_0, t_1, \ldots, t_{m-1}$ where n is the number of input nodes and m is the number of output nodes. Set w_0 to be $-ø$, the bias, and x_0 to be always 1. For pattern association, $\mathbf{X}_p$ and $\mathbf{T}_p$ represent the patterns to be associated. For classification, $\mathbf{T}_p$ is set to zero except for one element set to 1 that corresponds to the class that $\mathbf{X}_p$ is in.

iii. Calculate the actual output
Each layer calculates the following:
$y_{pj} = f\,[w_0x_0 + w_1x_1 + \ldots. + w_nx_n]$
This is then passes this to the next layer as an input. The final layer outputs values o_{pj}.

iv. Adapts weights
Starting from the output we now work backwards.

$w_{ij}(t+1) = w_{ij}(t) + \tilde{n}\text{þ}_{pj}o_{pj}$, where $\tilde{n}$ is a gain term and þ_{pj} is an error term for pattern *p* on node *j*.

For output units

$\text{þ}_{pj} = ko_{pj}(1 - o_{pj})(t - o_{pj})$

For hidden units

$\text{þ}_{pj} = ko_{pj}(1 - o_{pj})[(\text{þ}_{p0}w_{j0} + \text{þ}_{p1}w_{j1} + \ldots. + \text{þ}_{pk}w_{jk})]$

Where the sum (in the [brackets]) is over the *k* nodes in the layer above node *j*.

III. PACE REGRESSION

Pace regression improves on classical least square regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regression. Its optimality in minimizing the expected prediction loss is theoretically established when the number of free parameters is infinitely large. [9] The Pace regression overcomes the dimensionality determination problem. It outperforms existing procedures for fitting linear models. Dimensionality determination, a special case of fitting linear models, turns out to be a natural by-product. Estimating a mixing distribution is an indispensable part of pace regression. It also has more general implications for empirical modeling. Pace regression is a best technique among other regression techniques [10].

IV. ANALYSIS

The neural network model is created by presenting it with many examples of the predictor values from records in the training set (in our case 5821) and the prediction value from those same records. By comparing the correct answer obtained from the training record and the predicted answer from the neural network it is possible to slowly change the behavior of the neural network by changing the values of the link weights.

A multiple perceptron model of real estate is shown in Figure 3. Here the training data set consists of 43 attributes and testing mode with cross validation with 10 fold. The 42 attributes are input and form 21 sigmoid Nodes, the summation of attributes multiplied with their reprocessed weights becomes input for linear node which further adds weight according to sigmoid node, which forms the final output.

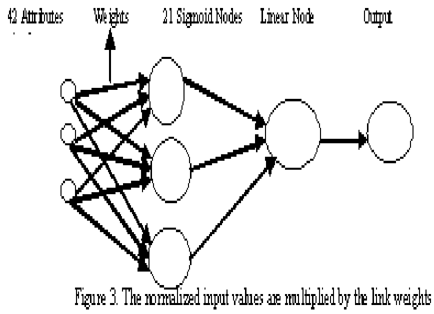


Figure 3. The normalized input values are multiplied by the link weights and added together at the output.

The factors on which comparison is done are Correlation Coefficient, Mean absolute error; Mean squared error, Relative Absolute Error, Root relative squared error and Time taken.

A. Comparative Study

Analysis is done on two different environments in both the techniques. In first case, we have taken all the attributes (complete) and compared the patters, results of both the methods. In second case, the same tests are performed, after discarding demographic details of customers (selected) and then compared the various results. It is found that if we are not taking care of demographic details we are not getting very accurate results as well as the error rate is increased.

The following paragraphs discusses the outcomes of various comparisons.

Table 1 shows the comparison between multilayer perceptron and pace regression on all as well as selected attributes. The results are generated using WEKA 3-6-2, open source software for regression analysis and data mining. It is very much clear that correlation coefficient is very high in multilayer perceptron as compared to Pace Regression. There is a major reduction in mean absolute error, root mean square error, relative absolute error and root relative square error. It proves that multilayer perceptron far better than regression methods. It also depicts that results on selected attributes are not very accurate.

In *pace regression*, the following pattern is generated, which considers all the attributes in general but a single pattern takes cares of various types of tuples.

IPC = -2.3609 + -0.0155 * CUSTYPE + -0.0143 * NOH + -0.0541 * ASH + 0.07*AS + -0.0085 * CMT + -0.0201 * NAT + -0.0138 * CS + -0.0175 * SCS + 0.102 * NR + 0.0646 * MRD + 0.1015 * LT + -0.1225 * OR + -0.0313 * SNG + 0.203 * HWTC + 0.0019 * HWC + 0.0222 * HLE + 0.0231 * MLE + -0.0162 * LLL + 0.0183 * HS + 0.0377 * BUS + -0.0247 * SRS + -0.0183 * MGMT + -0.0361 * TL + -0.0158 * UTL + 0.0155 * SCA + -0.0064 * SCB1 + -0.0293 * SCB2 + -0.0122 * SCC + 0.1113 * SCD + 0.1259 * RH + 0.0533 * HO + 0.0228 * C1 + 0.0325 * C2 + 0.2876 * NC + 0.2644 * PL + -0.0084 * OI + -0.0117 * INL + 0.0057 * INM + -0.0242 * INMH + 0.0447 * INH + 0.013 * INHH + 0.0112 * AI

Whereas in case of *multilayer perceptron*, 21 patterns are generated according to the weight age of various attributes in the data set proves to be very accurate and flexible. Some of the patterns generated for the sigmoid nodes, as per the weights are as follows

Sigmoid Node 13

IPC=-0.583 + CUSTYPE*-1.277119249 + NOH*1.047869233 + SH*1.556254728 + AS*-0.916643488+CMT*2.735987211+NAT*-2.03650943+CS*0.818124428+CS*-0.568618468+NR*2.716436701 + MRD*0.453650707 + LT*-1.758296458 + OR*0.429904235+SNG*1.794655244+HWTC*0.32104538+HWC*1.23526592+HLE*0.349264459+MLE*1.476429886+LLL*-0.784144661+HS*-0.992803801+ BUS * -0.679912914+SRS*1.740525176+MGMT*-0.47712932+TL*6.692549441+UTL*0.673512792+SCA*5.897074334+CB1*0.721763457+SCB2*0.923256826+SCC * 1.630738884+SCD*0.303496956+RH*0.097873626+HO*-1.455065218+1*3.440583033+C2*2.234604415 + NC*-0.398478131 + PL*-1.099401924 + OI*0.425862775 + INL*-0.954166497+NM*2.251016687+INMH*-0.932407326+INH*6.821764314+INHH*1.011448519 + AI*-2.02295018

Sigmoid Node 15

IPC=-0.591+ -1.112329668*CUSTYPE + -0.130789028*NOH + 0.486926326*ASH+ 1.037301865*AS + 9.424719766*CMT + 3.896196184*NAT + -0.113477392*CS+ 4.544551207*SCS + 5.398195486*NR + 0.211503376*MRD + 8.070215123*LT +-3.837488723*OR+4.404332891*SNG+488245265*HWTC + 1.697420837 * HWC + 3.264279362*HLE+9.769398474*MLE+-10.24017696*LLL+9.864817009*HS+0.37155292*BUS+15.74963025*SRS+13.39830713*MGMT+18.10346203*TL+1.581619927*UTL+-1.555745367*SCA+7.15414507*SCB1+-6.451647334*SCB2+8.966390522*SCC+0.066256378*SCD + -0.102647949*RH+1.091457004*HO+5.014749402 *C1 + 4.362545134*C2 + 0.19215151*NC + 0.344177704*P + 6.299699783*OI + -0.854619715*INL + 2.798172128 *INM + -1.007482916*INMH + -13.41339682*INH + 3.024666622*INHH + -1.492272154*AI

Following pattern is generated for Linear Node 0, which assigns various weights to 21 sigmoid nodes.

Node0=-0.02028+Node13*2.295616572+Node15*1.910500073+ode18*1.732994173* Node 16*1.232705576 + Node6*1.191171179 + Node20*1.157799721+ Node10 * 1.146473742 + Node 9*1.073451205 + Node12*0.716190231 + Node3 * 0.696801975 +Node1*0.621422146+Node7-0.513836101+Node19*-0.572744882+Node21*-0.735996424+Node17*-0.906671742+Node8*-1.044035863+Node5 * -1.09450818 + Node4 * -1.198409682+Node11*-1.366231779 + Node2*-1.497851544 + Node14*-2.101273786

In case of selected attributes, some of the generated patterns are as follows.

Sigmoid Node 5

IPC=4.673971076+-23.38120911*NOH+6.30970362*ASH+-0.074349188*AS+-0.084578269*HLE+ 1.587264036*MLE + 33.33796113*LLL + -16.54505083*HS+ -7.682547755 * BUS+17.63179733*SRS+36.86917468*MGMT+-3.366516483*TL + 0.775016298*UTL + 1.831659255* RH + 54.58787105*HO + -9.219738598*C1 + 25.04603678*C2 + 5.68270678* NC + -0.678004427*PL + 55.13946899*OI + 40.14737217*INL + 57.09035528 *INM + 1.1524926*INMH + 25.6428126*INH + 1.63798341*INHH + .089920973*AI

Sigmoid Node 4

IPC=3.753343295+ 23.13064539 * NOH + 113.8372897*ASH + 0.964216137*AS + 0.11608276 *HLE+ -0.213171326 * MLE+15.87432305*LLL+15.7577012*HS+ 0.80770913*BUS +-127.2374002*SRS+-6.399290041*MGMT+36.19748257 * TL+22.51465076*UTL+-0.602199533*RH+-55.25089704*HO + 4.665142876*C1 + -23.46065022*C2 + 9.450209595*NC + 6.512588126*PL + 62.53556473*OI + 68.47586267*INL + 12.62841251*INM + 0.867316552*INMH + 100.7488721 * INH +20.05246037*INHH + -0.117734593*AI

Following pattern is generated for Linear Node 0, which assigns various weights to 13 sigmoid nodes.

Node 0 = Sigmoid Node 1*-4.378494066 + Sigmoid Node 2*-5.508157156 + Sigmoid Node 3 *-6.261879372 + Sigmoid Node 4*3.753343295 + Sigmoid Node 5*4.673971076 + Sigmoid Node 6 *-2.931225697+Sigmoid Node 7*0.278155031 + Sigmoid Node 8*-3.121085375 + Sigmoid Node 9*-5.431623086 + Sigmoid Node 10*-0.847170019 + Sigmoid Node 11*-7.617227182 + Sigmoid Node 12*-1.700157286 + Sigmoid Node 13*-7.026772582

B. Analysis of various Graphs

Graph 1 shows a comparison between Multilayer patterns and pace regression on all attributes. It is found that in case of pace regression it treats each attribute more or less same, but in case of multilayer each pattern gives different weight age to different attribute according to the training set. We have taken only 11 sigmoid patterns to make the graph simple.

Graph 2, 3 and 4 shows the weight age of various attributes in case of pace regression (complete set), multilayered (on average of selected attributes) and pace regression (on selected attribute) respectively. Graph 5 and graph 6 depicts the weights assigned to various sigmoid nodes in case of selected attribute set and complete attribute set respectively. Graphs 7 to Graph 12 depict the comparison among multilayer (selected, complete) and pace regression (selected, complete).

V CONCLUSIONS

Multilayer perceptrons are capable of generalisation, that is, they classify an unknown pattern with other known patterns that share the same distinguishing features. This means noisy or incomplete inputs will be classified because of their similarity with pure and complete inputs. Secondly they are highly fault tolerant. This characteristic is also known as "graceful degradation". Because of its distributed nature, a neural network keeps on working even when a significant fraction of its neurons and interconnections fail. Also, relearning after damage can be relatively quick. In this paper we have tested multilayer perceptron on real estate data containing 5821 tuples and 43 attributes and found that its outperformed than pace regression which is proved best among linear regression, isotonic and least median squares with various test modes [3].Further we have tested the same methods on selected attributes and found that demographic details of a customer matters a lot in the buying behaviour of a customer.

When compared the pace and multilayer with the complete data set and selected data set, we found that the pace gives more than +10 difference in the data set attribute (No car and policy), whereas multilayer gives (LLL, SRS and Mgt).

When compared with multilayer (Complete) and pace (selected) we found pace gives a big difference in MLE, BUS, HO,C1,NC,PL,NH where as the vice versa LLL,SRS,MGMT,TL and OI.

When compared multilayer and pace on selected dataset we found NC and Pl gives very high difference where as when compared with full dataset it gives difference in HWTC,SCD,RH,NC and PL.

Such analysis gives emphasis of general thinking of any Indian Scenarios what is the impact of the data which relates with caste and religion. Generally in India the question arises on three important things in a common man life food, clothes and house. Our analysis results on impact of Pace regression the way it works as it works on the weights preprocessing where as multilayer perceptron alters the weight at the time of processing. Thus, when compared Pace complete and multilayer perceptron selected, we found the optimum analysis according to the Indian condition is LLL, SRS, MGMT, TL and OI.

REFERENCES

1. Carbune P. L., "Expanding the meaning of an application for Data Mining", IEEE International Conference on Systems, man and Cybernetics, 1872-1873,2000
2. Han Jiawei and Kamber Micheline, "Data Mining: Concepts and Techniques, Second Edition", Morgan Kaufmann Publishers.
3. Hald A.,"Statistical Theory with Engineering Applications", A Wiley Publications in Applied Statistics
4. R. Bouckaert Remco, Eibe Frank et. al, "WEKA Manual for Version 3-6-2", January 11, 2010
5. Wang Yong and Witten H., "Pace Regression" Working Paper Series, ISSN 1170-487X, Working Paper 99/12, September 1999
6. Makhtar, M.; Neagu, D.C.; Ridley, M. "Predictive model representation and Comparison: Towards data and predictive mode governance", Computational Intelligence (UKCI), 2010 UK Workshop on, 2010
7. Somesha, Sowmya; Lee, David N.; Loddick, Sean J, "Gatecard reliability prediction analysis", Power Electronics and Applications (EPE 2011), Proceedings of the 2011-14th European Conference on, 2011
8. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth," From Data Mining to Knowledge Discovery in Databases", AI Magazine Volume 17 Number 3 (1996) (© AAAI)
9. Alex Berson, Stephen Smith, and Kurt Thearling, "An Overview of Data Mining Techniques", Excerpted from the book Building Data Mining Applications for CRM.
10. Geetali Banerji, Kanak Saxena, "Predictive Model- A Boon for real estate" , International Journal for Wisdom Based Computing Volume(1) 2, April 2012
11. sydney.edu.au/engineering/it/~irena/ai01/nn/8.html
12. Haykin, S. Neural Networks: A Comprehensive Foundation, Maxmillan, IEEE Press, 1994
13. [13] Negnevitsky, M. Artificial Intelligence: A Guide to Intelligent Systems, Addison Wiesley, England, 2002.
14. Joseph P. Bigus, Data Mining with Neural Networks: Solving Business Problems From application development to Decision support, McGraw-Hill, NY, 1996

AUTHORS PROFILE

Geetali Banerji M.C.A. She is an Associate Professor and Programme Director of BCA in Information Technology Department at Institute of Information Technology & Management affiliated to Guru Govind Singh Indraprastha University, New Delhi. At present she is pursuing Phd. from Banasthali University. Her research area is Data Mining.

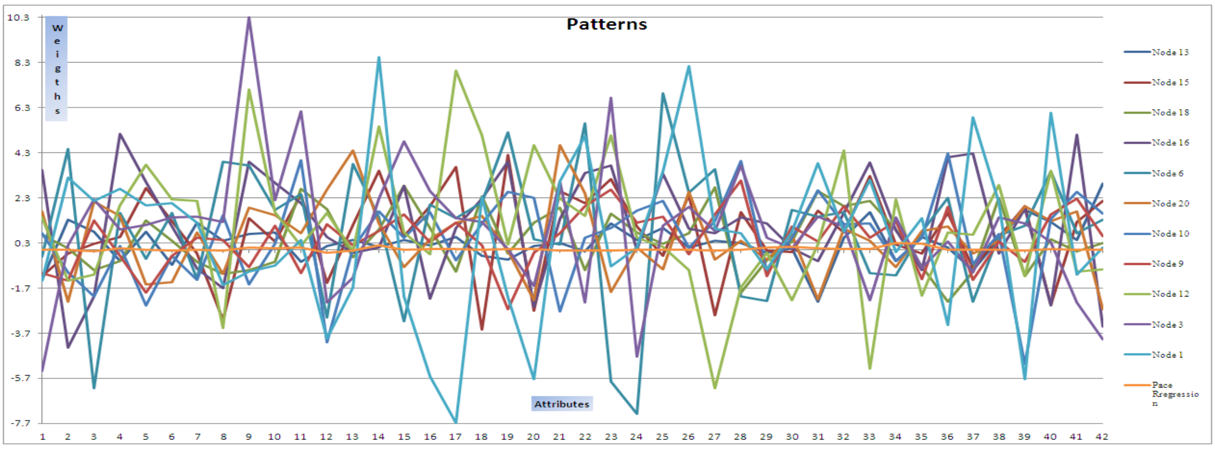
Kanak Saxena received her Ph. D. in computer science from the Devi Ahilya University, Indore, India. She is professor and head in the Computer Applications Department at the Samrat Ashok Technological Institute affiliated to Rajiv Gandhi Technical University, Bhopal. Her current research focuses on database systems, parallel computing, data uncertainty and design and other interests include network security and performance and software engineering. She is the member of the editorial board of international journals and member of various international committees of international conferences. She is a member of various professional societies and also a reviewer of various international conferences and international journals. She has also reviewed the books of various publications.

Table I Results of Multilayer Perceptron and Pace Regression on Selected / Complete Attribute Set

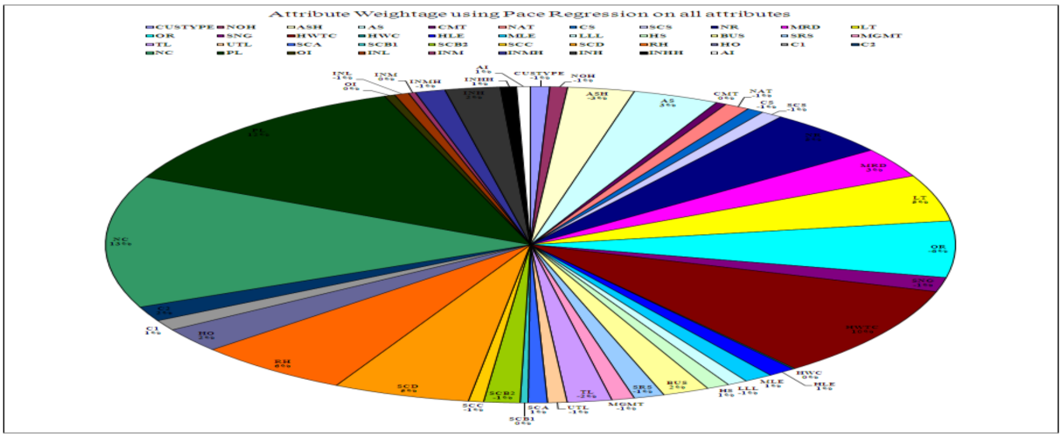
Factors	Multilayer Perceptron		Pace regression	
	All Attributes	Selected Attributes	All Attributes	Selected Attributes
Correlation coefficient	0.9004	0.7562	0.843	0.5929
Mean absolute error	0.2063	0.3726	0.3323	0.4935
Root mean squared error	.03605	0.5308	0.4248	0.636
Relative absolute error	31.2166 %	56.3781 %	50.2751 %	74.6714 %
Root relative squared error	45.6326 %	67.1908 %	53.7744 %	80.5055 %
Time (in seconds)	181.86	71.22	0.47	0.22

DATADICTIONARY

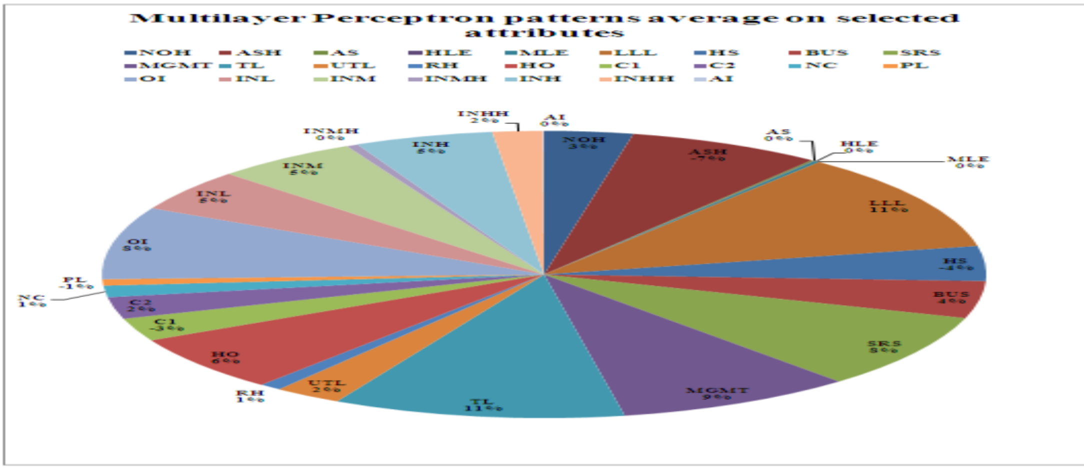
NO.	Name	Description	NO.	Name	Description
1	CUSTYPE	Customer Subtype	21	SRS	Service
2	NOH	Number of houses	22	MGMT	Management
3	ASH	Avg size household	23	TL	Trained labor
4	AS	Avg age	24	UTL	Untrained labor
5	CMT	Customer main type	25	SCA	Social class A
6	NAT	Nationality	26	SCB1	Social class B1
7	CS	Caste	27	SCB2	Social class B2
8	SCS	Subcaste	28	SCC	Social class C
9	NR	No religion	29	SCD	1Social class D
10	MRD	Married	30	RH	Rented house
11	LT	Living together	31	HO	Home owners
12	OR	Other relation	32	C1	1 car
13	SNG	Singles	33	C2	2 cars
14	HWTC	Household without children	34	NC	No car
15	HWTC	Household with children	35	PL	Policy Investment
16	HLE	High level education	36	OI	Other Investment
17	MLE	Medium level education	37	INL	Income < 20.000
18	LLL	Lower level education	38	INM	Income 20-55.000
19	HS	High status	39	INMH	Income 55-85.000
20	BUS	Business	40	INH	Income 85-125.000
41	INHH	Income >123.000	42	AI	Average income
43	IPC	Investment power class			



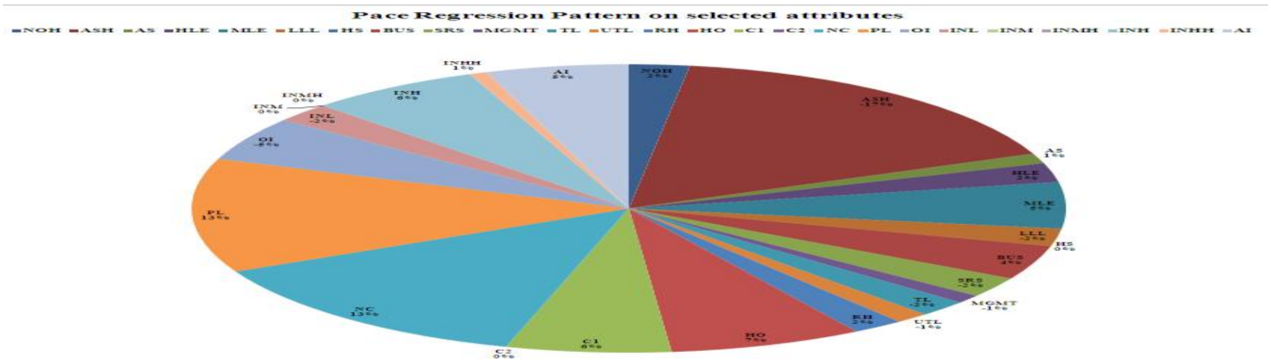
Graph 1. Comparison of Multilayer Patterns with Pace Regression Pattern



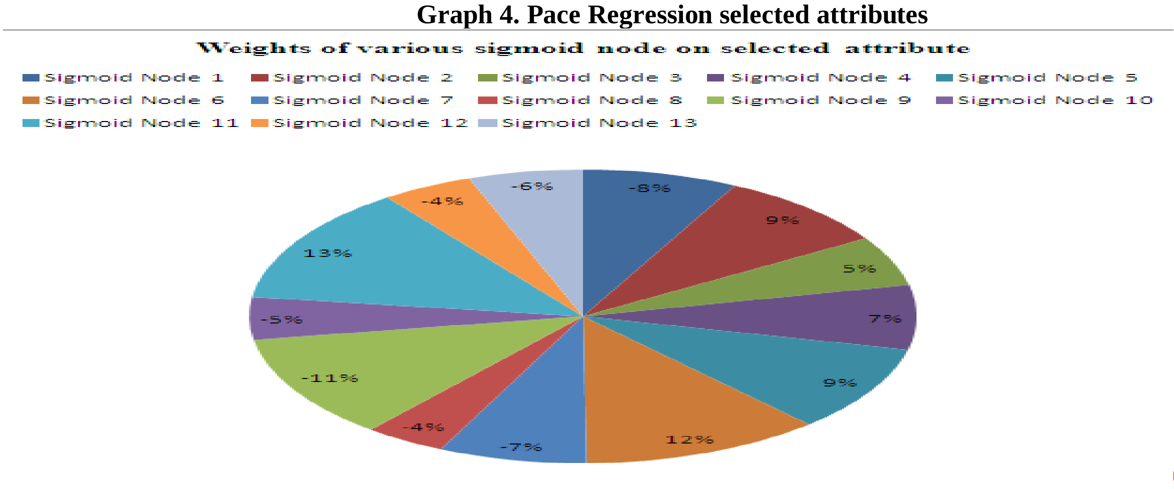
Graph 2. Pace Regression on complete attribute set



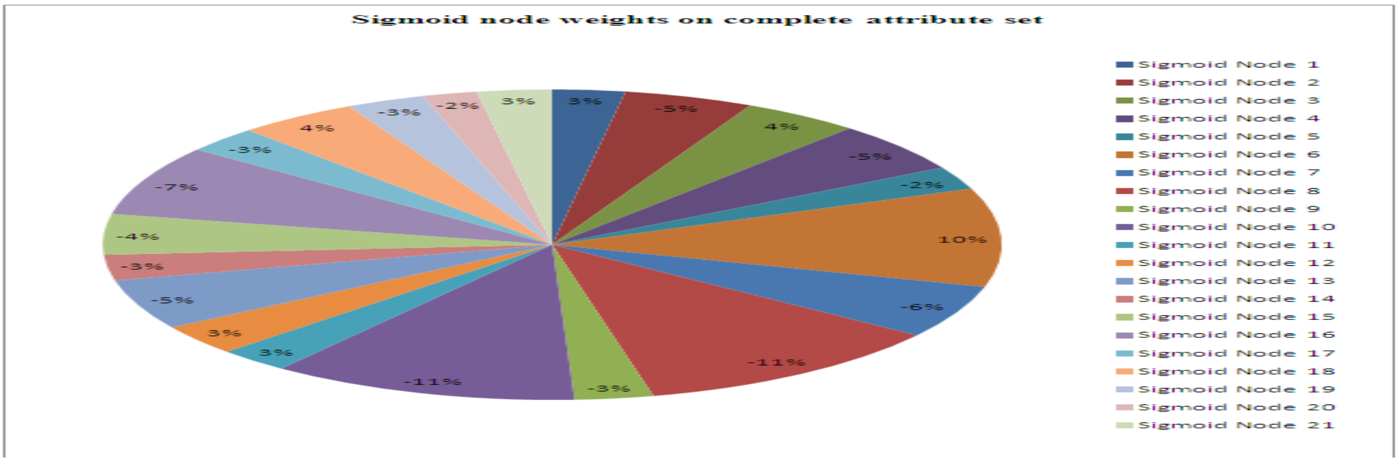
Graph 3. Multilayer Perceptron on selected attributes



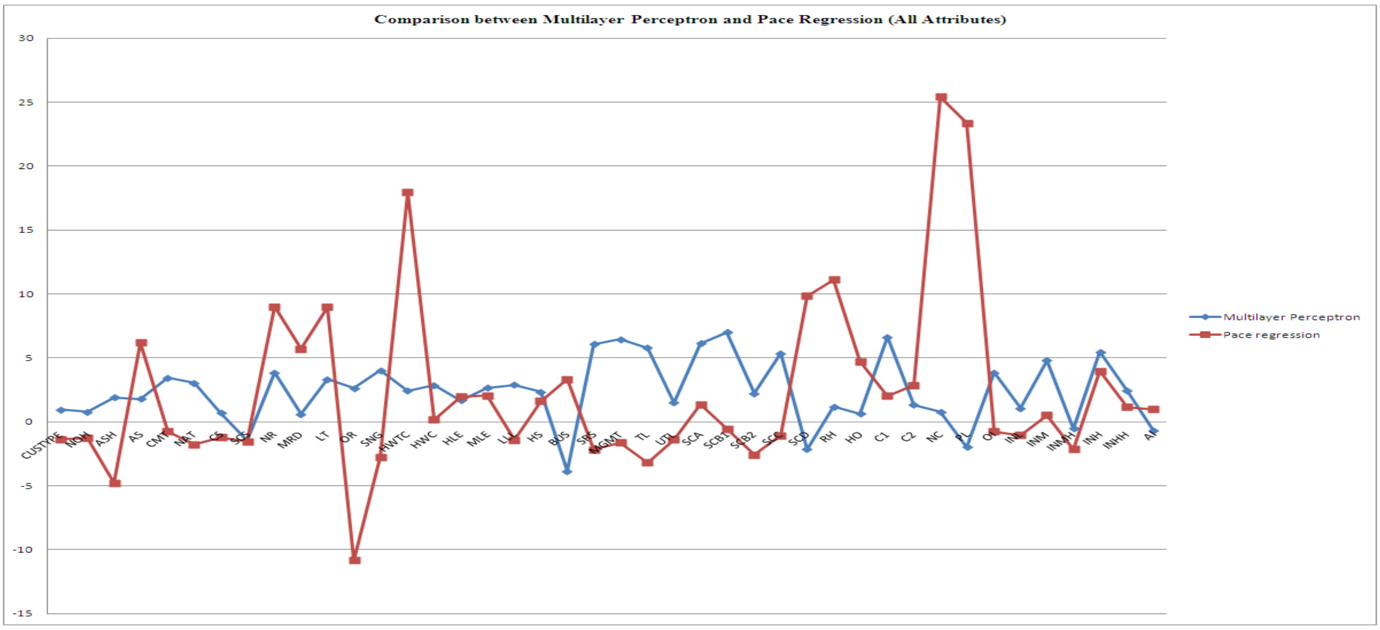
Graph 4. Pace Regression selected attributes



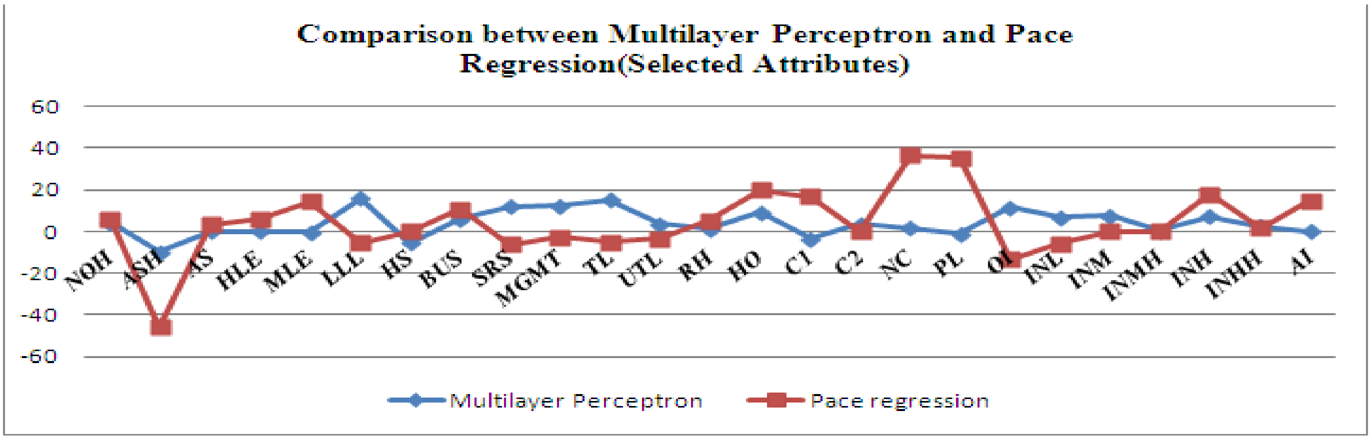
Graph 5. Weights of sigmoid node on selected attribute



Graph 6. Weights of sigmoid node on complete attribute set

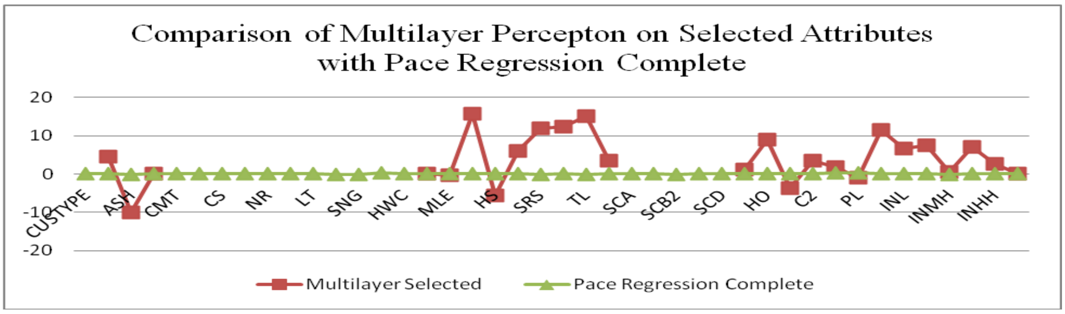


Graph 7. Comparison of Multilayer Perceptron and Pace Regression on all attributes

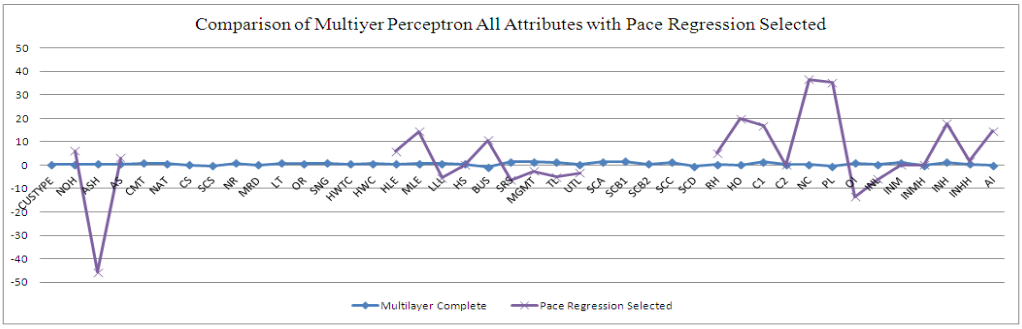


Graph 8. Comparison of Multilayer Perceptron and Pace Regression on selected attributes

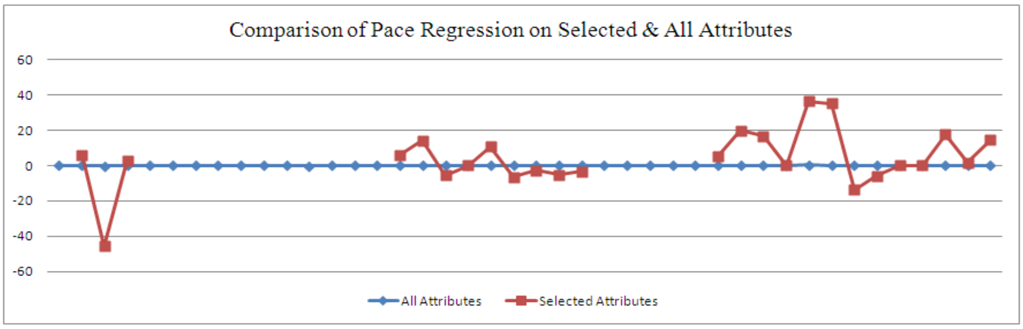
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



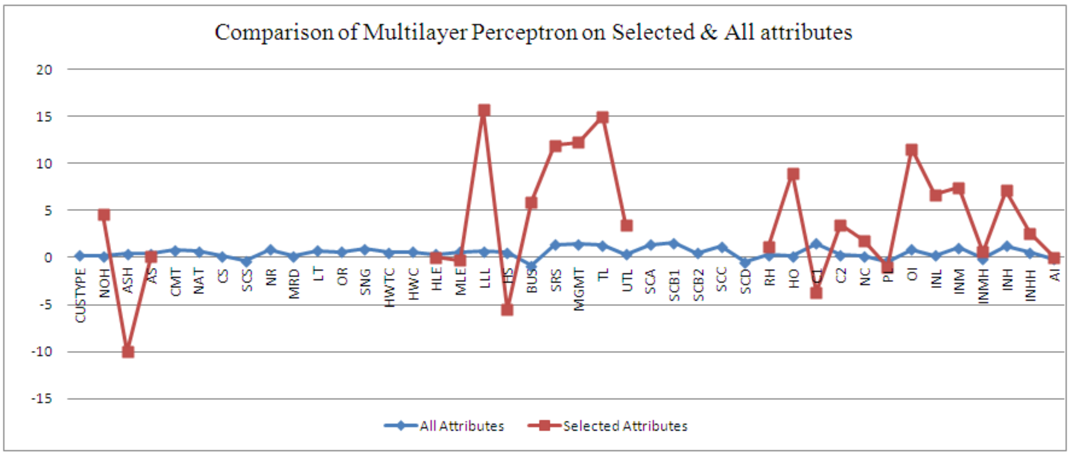
Graph 9. Comparison of Multilayer Perceptron selected and Pace Regression on all attributes



Graph 10. Comparison of Multilayer Perceptron on all and Pace Regression selected attributes



Graph 11. Comparison of Pace Regression on all and selected attributes



Graph 12. Comparison of Multilayer Perceptron on selected and on all attrib