Current Science International
ISSN: 2077-4435

Volume : 06 | Issue : 01 | Jan.- Mar. |2017
Pages**:** 243-258

Comparison between Traditional and Modern Models of Data Mining

Amel Ahmed Talaat

Statistics Dept., Faculty of Commerce, Girls Branch, Al-Azhar University, Cairo, Egypt.

Received: 21 February 2017 / Accepted: 27 March 2017 / Publication Date: 28 March 2017

ABSTRACT

This research introduced the various approach of data mining and there types.it also reviewed there various areas, in which these methods where successfully applied. The real data where used for application to compare between traditional methods, it was found that the ARIMA sessional model the most successful in application, then when comparing it with four of modern models for the duration of 60 months it was found that the NEURAL network is the most appropriate for the data used.

Key words: Neural networks model, traditional statistical methods, methodology

Introduction

Data mining is an important arena to talk about in this research and data analysis is one of its stages. Its software contains modern methods, but the use of these programs and these methods in the Arab researches is still minimal because of the scarcity of publications in Arabic, and difficulty to understand them. Therefore, most researchers have refrained from using such techniques in reconciliation the relationship between the dependent and independent variables using the multiple linear regression models because of easiness to understand and use. However, the modern applications demonstrated the weakness of the credibility of the multiple linear regression models to reconcile most of the contemporary nonlinear problems that contain interaction between variables due to large data sets. Methods of data mining can be classified into two basic types: parametric methods and nonparametric methods, in addition to the semi-parametric methods. Hence, the study will come in four parts: The first part will be devoted to the theoretical side as it contains the research design, problem, and previous studies. The second part will be devoted to the parametric methods and it will include the linear regression model, analysis of key components, factor analysis, discriminant analysis, discriminant correspondence analysis, cluster analysis, and models of probit and Logit models. The third part is devoted to nonparametric methods as it includes the analytical hierarchical process, the method of nearest neighbors, and generalized additive Models, sports programming, and divided harmonized regression functions, tree of decisions, neural networks, and nonparametric regression. Part four has been allocated to semi-parametric Methods and it contains semi-parametric regression. The research based on many studies, which include different methods of data mining to define these methods and ways of calculation, applications and conditions of simplified use, and a comparison between the preference mining.

Previous Studies

Some data mining studies dealt with the process of data mining (CRISP-DM 2003; Fayyad *et al.,* 1996 a,b,c,d).

It showed the steps that must be followed by the establishment in order to discover the important knowledge and patterns that were not known before. Some of them indicated to the methods of data mining and areas of use.Some studies focused on inventory and providing definitions to these methods, while others focused on the study of the theoretical aspects Christos (Stergiou and Dimitrios Siganos; CIPFA, 2008; Conjugate gradient method; 0,Conventional programming; Cornelius, 2002) (Generalized additive model; Giudici, 2003; Hastie and Tibshirani, 1993; Hastie *et al.,* 2009; Ichimura, 1993; Jackson, 1998; John, 2002; Koch et al., 1988; Mark, 2005; Mathematical optimization; Multivariate adaptive regression splines; Backpropagation, C4.5 algorithm;CHAID) and

Corresponding Author: Amel Ahmed Talaat, Statistics Dept., Faculty of Commerce, Girls Branch, Al-Azhar University, Cairo, Egypt. Email: d.amel___@hotmail.com

algorithms (Analytic Hierarchy Process; Analytic Hierarchy Process (AHP) Tutorial)(Neural Network Toolbox, Levenberg-Marquardt (trainlm) used in those methods for the estimation process. However, other studies interested in providing software (Neural Network Software *For researchers, data mining experts and predictive analysts)* and assisting tools. It was clear from data mining methods applications studies (Giudici, 2003; Hastie *et al.,* 1993), that the successful applications was as follows: Market Basket Analysis, assessment of creditworthiness, analysis of the stock markets, detecting fraud and malfunctions, and diagnosis of diseases.

Research Questions

The research provides answers to the following questions:
What are the methods of data mining?
When are they used? What are the successful applications of those methods?
That has been applied for the following: Model of linear regression, analysis of key components, factor analysis, discriminant analysis, discriminant correspondence analysis, cluster analysis, probit and Logit models, hierarchical analytical process, expert systems, the method of nearest neighbors, generalized additive models, sports programming, harmonized regression functions, tree of decisions, neural networks, nonparametric regression, and semi-parametric regression.

Parametric Methods

Multiple Linear Regression (MLR)

The multiple linear regression method is the oldest and most famous statistical methods used in data mining field. It is a predictive model used by the analysts if they wanted to assess the causal relationship between the quantitative variables and several other variables. The variable that we want to interpret its change or predict its value in the future has several names such as: The dependent variable, the response variable, or the explained variable and it takes the code of y_i . Other variables take the following names: The independent variables, the explanatory variables, predictors, features, covariates, or benchmarks and they take the symbol of x_{ij} , as x_{ij} refers to the Views, while $j = (1,2,...,p)$ refers to the variables.
The multiple linear regression model (matrix format) takes the following shape:

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{B} + \underset{n\times 1}{\ell} \qquad (1)$$

We can reach to its analytical form by estimating a vector of B features based on one of the estimation methods. The least square method (LS) is the most famous one where they choose a level that is totally lower than residuum squares ℓ . We can check the quality of model conciliation through diagnostic tools by drawing the residuum against the estimated values of the regression line $\hat{y}_i$ and then to consider the output format. If the regression was correct, the values of the dependent variable must be distributed randomly on the estimated line without forming any clear general trend. The quality of regression model reconciliation can be checked based on the summative index known as the selection factor R^2 as it takes a value ranges between zero and one, as indicated by the predictability of the values of y_i in a truer way depending on the relationship that ties them with values of $x_i's$. Finally, the overall moral of the model is tested using the F test, while the partial moral of the independent variables is tested using the independent test t.
If we have one set of data (one quantitative variable and several independent variables), the last will not depend on each other (ie, lack of linear multiple dualism multicolinearity), the multi-linear regression will be applied safely. In the case of the existence of the two sets of data (set of independent variables and set of dependent variables) or more, or if there was linear dualism in the case of one set of data, the linear regression will not fit, and one of the following methods can be applied.

Principal Components Analysis (PCA)

When we come to statistically deal with any matter, we had to express the terms of so-called data mining in the data table. Data table is a matrix of $n \times p$ class, where matrix n indicates measurements taken in preview units of variables p under study.
The Principal Components Analysis aims to compress the data table in the shade of associated measurements expressed in a new set of associated (orthogonal) known as dimensions reduction. Then, it will be said that the new changes depend on the context or they are the principal components, factors, eigenvectors, singular vectors, or loadings. Each unit also represents (row) unit in a set of scores corresponding to its ingredients.
Analysis of principal components starts with calibration of all the variables in calculating the matrix of change. An iterative process aiming to reach k of the key components is applied where k<p. This process starts with getting the first principal component which describes all the variables in getting any vector (weights).

$$a_1 = (a_{11}, a_{21}, ...a_{p1})'$$

The outcome of solving the problem of maximizing the contrast in Y_1:

$$\max \text{var}(Y_1) = \max(a_1' S a_1)$$

Using Lagrange multipliers under constraint $a_1' a_1 = 1$, then we try to get a second component until we reach the k component i.e. getting the vector of weights.

$$a_k = (a_{1k}, a_{2k}, ...a_{pk})'$$

The outcome of solving the problem of maximizing the contrast in Y_k:

$$\max \text{var}(Y_k) = \max(a_k' S a_k)$$

Using Lagrange multipliers under constraints

$$a_k' a_k = 1,\ a_2' a_1 = a_3' a_2 = ... = a_k' a_{k-1} = 0$$

.

The principal components shall be drawn on the horizontal axis against the distinctive values on the vertical axis. It is known as scree plot. The component with a maximum height will be chosen.

Factor Analysis (FA)

Factor analysis is used in the reduction of dimensions. It sums up a variables to the lower number of variables during the process of data modeling. FA selects a subset of the variables from a larger group based on the highest correlations between the original variables with the key components factors. That shall be an entry for the treatment of multiple multicollinearity when reconciling multiple regression models because the resulting combination of factors are not related variables. Therefore, the factor analysis is used in the construction of the so-called models of latent variables. They are the non-viewed variables, which has no registered measurements. They are inferred
from other variables shown (through a mathematical model) with recorded measurements. They also used to discover the structure of relations between variables, which is known as the classify variables.

Factor analysis is divided into two types: Exploratory factor analysis and confirmatory factor analysis

- Exploratory factor analysis (EFA): It looks at the nature of the relations constructions affecting on the dependent variables (forms of models or its constructive structures).
- Confirmatory factor analysis (CFA): It tests any of these structures and affects on the dependent variables during prediction.

The factor analysis is related ton its predecessor principal components analysis, but they are not one thing. FA uses regression-modeling techniques to test the limits of error, while PCA is a statistical descriptive method.

Discriminant Analysis (DA)

Discriminant Analysis is used in statistics for pattern recognition and machine learning to find a linear combination of the quantitative independent variables that characterize the quantity or separate two or more categories of events (semi-dependant variable). In other words, DA is a way to classify the measurements in two or more groups. The main purpose of DA is to predict the group membership based on a linear combination of quantitative variables. The method begins with a set of views with known values and groups and it ends with a model that allows the prediction of group membership via independent quantitative variables only. The second purpose of discriminant analysis is to understand the data set via closer examination of the forecasting model to get an idea of the relationship between the independent variables used to predict the set membership.

For example, the admissions committee at any university may divide its graduates into two groups: Students who have completed the program in five years or less and other students. DA can be used to predict the successful completion of the study program for new students based on their scores in GRE score and their undergraduate grade point average. Examining the prediction model gives an idea of the extent of contribution of each variable (individually and in association with other variables) in the completion or non-completion of the program.

DA is similar to ANOVA and RA, which reflect the dependant variable in combination of the independent variables. However, the dependent variable in the last two methods required to be quantified unlike DA that shall be a classification. The logistic regression and probit regression are very close to DA. All of them are explained as variable rank. However, the first two methods are preferable in applications that do not assume that the independent variables follow the normal distribution, which is the basic assumption, upon which the DA is based.

DA is similar to both of PCA and FA in searching for the linear combinations of the variables that gives the origin of data interpretation. If DA is frankly trying to model the difference between the categories of data, PCA does not take into account any difference in the classes. FA also builds the combinations of variables on the differences rather than the similarities. DA also is different from the FA as it is not an interdependence technique: It differentiates between the dependent and dependent variables.

Finally, DA is applied when the independent variables takes constant quantity measurements. But when dealing with classified independent variables, the equivalent method will be the discriminant correspondence analysis.

Discriminant Correspondence Analysis (DCA)

As the name indicates, DCA is an extension of both of CA and DA. DCA, such as DA, aims to categorize the views in predefined sets and like CA as it can be used with a nominal variables. The basic idea behind DCA is to represent all group in total views and conducting simple CA on groups based on variables matrix. Original views shall be predicted and customize each expected view in the nearest group. Comparison between priori and posteriori classifications can be used to assess the quality of discriminant. Stability of analysis can be assessed using cross-validation techniques.

Cluster Analysis

The cluster analysis is one of the most famous descriptive (exploratory) methods for data mining. It is a method for grouping a certain set of views. If the data matrix consisted of n of views (cases or rows) and p of variables (fields or columns), the goal of cluster analysis will be represented in classifying or clustering the views in internal cohesion (coherent) and external non-cohesion groups (external separation). That can be explained as shortening of the dimensions in R^n space, but not in the same way of the main components. Cluster analysis in vertical reduction collects n views in g of sub-groups (as $g<n$) while the main components analysis converts the original variables p to k of new variables (as $k<p$).

Groupings and partitions or clusters can be configured via two ways:

- Hierarchical Methods: Number of clusters is estimated via succession method starting from *n* (which is the simplest situation, in which each view is treated as a separate group) until 1 (all views belong to one group.
- Non-hierarchical methods: The number of clusters is known in advance.

Probit and Logit Models

Probit model is a special type of regression, in which the dependent variable will be classification (binary) type. It will take two values only, success that is indicated by 1 and failure, which is indicated by 0 as follows for example: Married and unmarried, successful and unsuccessful, prefer not prefer, answer with yes or no and the presence or absence of a particular attribute.... etc.
Probit model takes the following form:

$$\Pr(Y = 1 \mid X) = \Phi(X'\beta),$$

Pr indicates to possibility and the symbol Φ to accumulated ordinary standard distribution function, while symbol β refers to the estimated parameters using the traditional great method of likelihood.
Logit model is a single/multivariate method that allows of estimating the probability of occurrence/non-occurrence of an incident to dependent binary variable, but it takes the following form:

$$y=\exp(b_0 + b_1*x_1 + \ldots + b_n*x_n)/\{1+\exp(b_0 + b_1*x_1 + \ldots + b_n*x_n)\}$$

The logistic regression model (LRM) is an example for this type of models. Logistics function always takes values ranging between zero and one, and it is known with the following model and figure:

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

Variable z (which can take any numerical values) indicates to the function inputs, while output is limited in values between zero and one. The variable z represents the exposure to a group of independent variables, while f (z) represents the possibility of the corresponding output in the light of the explanatory variable values. Variable z measures the total contribution of all the independent variables used in the model. It is called logit, and known via the following equation:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k,$$

Logistic regression is a useful way to describe the relationship between one independent variable or more such as age, etc. and it is a binary response variable that takes tow values only (success or failure).

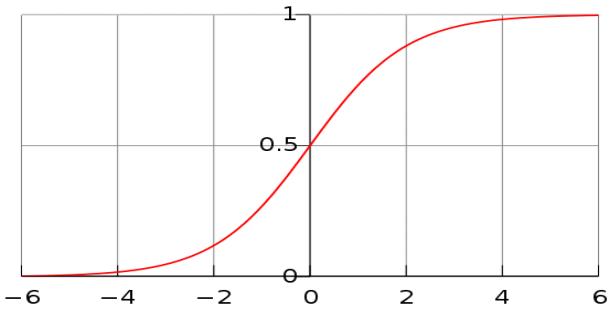

Fig. 1: Example of the Logistic Function

Nonparametric Methods

Analytical Hierarchy Process (AHP)

Analytical hierarchy process is known with structured technique to regulate and analysis of

complex decisions. It is a combination of mathematics and psychology presented by Thomas L.Saaty in the seventies to reach the best decision from among the alternatives available (Analytic Hierarchy Process)

- This process has been widely acclaimed by the decision-makers and in government, education, health, industry and businesses sectors. Instead of assuming that decision is "true", AHP helps the institutions to find the best decision in line with its objective and understanding of the problem. AHP provides a comprehensive framework for building the resolution problem, to represent and measure its elements, to link these elements to achieve the overall objectives, and to evaluate the alternative solutions.

The first thing done by users of AHP is to dismantle decision problem hierarchically into sub-problems that can be understood more easily so that they can be analyzed independently. The hierarchy elements may be related to any aspect of the problem, whether those elements were tangible or intangible, measured or estimated accurately or approximately and whether well or poorly understood.

Once the hierarchy is formed, the decision-makers assess its elements via comparing each element to other two elements every time in terms of its effect on the above item at the hierarchical division. When making comparisons, decision-maker can use real data from items may also use his judgment about the relative importance of the elements. Thus, the essence of AHP depends on the use of personal judgments as well as background information in conducting assessments. AHP turns these evaluations into numerical values (Analytic Hierarchy Process (AHP) Tutorial) that can be processed and compared to the full extent of the problem. The numerical weight or what is known as a priority is derived for each element of the hierarchy, allowing comparison to the diverse and non-measurable elements often with rational and consistent manner. That ability gives an advantage to AHP from other decision-making methods.

At the last step of the process, numerical priorities for each alternate decision will be calculated. These figures represent the relative ability of alternatives to achieve the target.

Expert Systems (ES)

Expert system in the field of artificial intelligence is a computer system that simulates the ability of human expert in decision-making (Jackso, 1998). Expert systems are designed to solve complex problems by reasoning acquired knowledge based on expert manner and not the developer method as in agreement programming (0,Conventional programming; Nwigbo *et al.,* 2011).The first expert system was emerged in the seventies, and then spread in the eighties of the last century (Cornelius *et al.,* 2002). The expert system consists of two parts: The first is a fixed and independent part from the system, which is the inference engine. The second variable represents the knowledge base .In the eighties, third part that allows the users to be connected was appeared, which is the dialog interface (Koch *et al.,* 1988). Expert systems were designed to facilitate tasks in the areas of accounting, law, medicine, process control, financial services, production, and human resources. Therefore, many applications supported them in the areas of fault diagnosis, medical diagnostics, and support decisions in complex systems, control of operations, educational programs, and knowledge management.

Method of the Nearest Neighbors (k-NN)

The nearest neighbors is one of the methods that depend on memory so it does not require, in terms of data mining, and training on (reconciling data model) unlike the other statistical methods. k-NN depends on an intuitive idea represents in nearby views must located in the same category. They are a method of classification that determines the class, in which we will create a new situation examines the number (k) in most similar cases or neighbors. The analyst resort to this method when making comparative analyzes using data reduction methods.

The application of this method requires (CIPFA, 2008). Calibration of all the variables and calculation of Euclidean distances between each pair of views. New views can be classified in one of four methods as follows: Papadakis method, which sometimes called Genesis (it depends on the calculation of residuum, and the use of covariance analysis) alongside with the method of correlation

(which is based on the use of the correlation between each pair of views through the least squares generalized if the information of the structure correlation was known), method of Wilkinson, and the method of least squares.

The main applications of k-NN: Determine patterns and statistical classification, cluster analysis, retrieval of images based content from databases, and online shopping.

Generalized Additive Models (GAMs)

The generalized-added models (Generalized additive model) are one of the entrances to onparametric regression in the case of multiple independent variables. This method was emerged in the nineties by Trevor and Rob and Rob Tibshirani Hastie . This model mixes between the characterized generalized linear models and additive models. If the linear additive model takes the following form:

$$Y = b_0 + b_1*X_1 + ... + b_m*X_m$$

The GAM takes the different following shape:

$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m).$$

By comparing the two models, GAMs takes the from the multi-model its maintaining to the additive model, but it replaced the simple limits in the linear equation (b_i*X_i) at $f_i(X_i)$ functions. They are nonparametric functions for Xi unexplained variables. In other words, GAMs estimates nonparametric unspecified functions per each independent variable instead of the factors to reach a better prediction of the values of the dependent variable.

Fi (Xi) functions can be reconciled using (Mark, 2005). The following scatterplot smoother:

1) Cubic smoothing spline which is available in the SAS program.
2) LOESS method, which is also available on the previous program
3) Smoother Kernel, which is available in STATA program.
4) Thin-plate splines, which allow of interaction between the independent variables and is available in SAS programs and R.

Finally, GAMs is useful in the following cases (Generalized additive model :If the relationship between the variables is highly complicated and can not be reconciled via linear traditional model or any of the non-linear models (2 if prior reason is not available for the use of a particular model (3 if we want the suggest data in the appropriate function form. This means that models are suitable for the most modern applications that contain a large number of variables including the possible reactions in light of the large volumes of data, such as stock markets.

Mathematical Programming (MP)

Mathematical programming (Mathematical optimization) (in both mathematics and management science, computer science) refers to the optimization, which is the process of selecting the best solutions from among several alternatives under a set of constraints. The optimization in its simple form consists of maximizing or minimizing the real function via selecting the important values among a group of variables and calculates the value of the objective function. The optimization problem generalization creates variety of target functions an different types of ranges.

Add-in button (Mathematical Programming) in Excel allow of creating mathematical programming models using Solver Add-in. When Math Programming add-in is added, orders lines are added for each of the orders: Linear and integer programming, nonlinear programming, business network, and transport problems.

Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (Friedman, 1991; Multivariate adaptive regression splines) is one of the forms of regression analysis. They are a method of nonparametric regression that provide models to non-linear and interactions and builds models as follows:

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

They are a group of basis functions $B_i(x)$ weighted by the following fixed factors c_i. Each of basis functions takes one of the following three forms: 1) Constant 1 allows of the emergence the intercept in the model 2) hinge function on the form of max (0,x −const) or (0,*const* − *x*) max, where MARS selects the variables and the values of the contract automatically 3) the outcome of multiplied two functions or more.

MARS model can be reconciled in two stages based on the same approach used in the recursive partitioning when decision tree is reconciled:

1) The forward pass: It starts with model that contains the only intersection limit (average the dependant variable values) and then a pair of basis functions will be added to the form in every time until we reach to the maximum reduction in the error expressed in the total residuum squares. However, the front reconcile usually builds a overfit reconciliation model (model with a good reconciliation quality for the data used in its construction model, but its predictive performance for new data is weak).

2) The backward pass is completion of the previous phase to overcome the problem of overfit reconciling (to improve the predictive ability) via prune model by deleting its limits and one by one.

The limit with least influence is deleted to get the best sub-model. It compares the performance Sub-models by using the method of Generalized cross validation (GCV) to choose the best sub-model; where the least value of GCV indicates to better model . GCV is calculated based on the following formula:

GCV = RSS / (N * (1 – *Effective Number Of Parameters* /N)^2)

Neural Networks (NN)

Neural Networks is used to achieve many of the descriptive and predictive purposes during data mining (Genetic algorithm)- NN has emerged in the field of machine learning in an attempt to stimulate the neurological functions of the human brain through a combination of computational simple elements (neurons) in a very intersected system. NN has a special importance because it offers a highly efficient modeling of the complex problems (that contains hundreds or more of independent variables and many of the interactions and dependent variables) in a nonparametric manner from large databases. It can also be used in solving the problems of classification and regression, whether the data was incomplete or mutilated.

Components:

Figure (2) shows that the neural network consists of a range of primary computational units (known as nervous cells and they are connected to each other via weighted correlations.) Each cell is represented in a circuit, and take a natural number (from: 10 1 in this example). Links are represented using arrows and they take the symbol of w_{ij}, where i indicates to the number of the node, from which the arrow starts and j indicates to the number of end node. These units are organized in layers as each cell in the layer will be connected to all cells in previous and subsequent layer. The network begins with input layer (1:4 in this example) where each nod corresponds to an independent variable. Each node in a layer of inputs are connected to all the hidden layer (9: 5 in this example), and may hidden layers connect to another hidden layers (not shown in the diagram.) Layers end with the output layer (No. 10 at this example). It is one node or more that represent the dependent variable(s). They are the meeting point from another hidden layer.

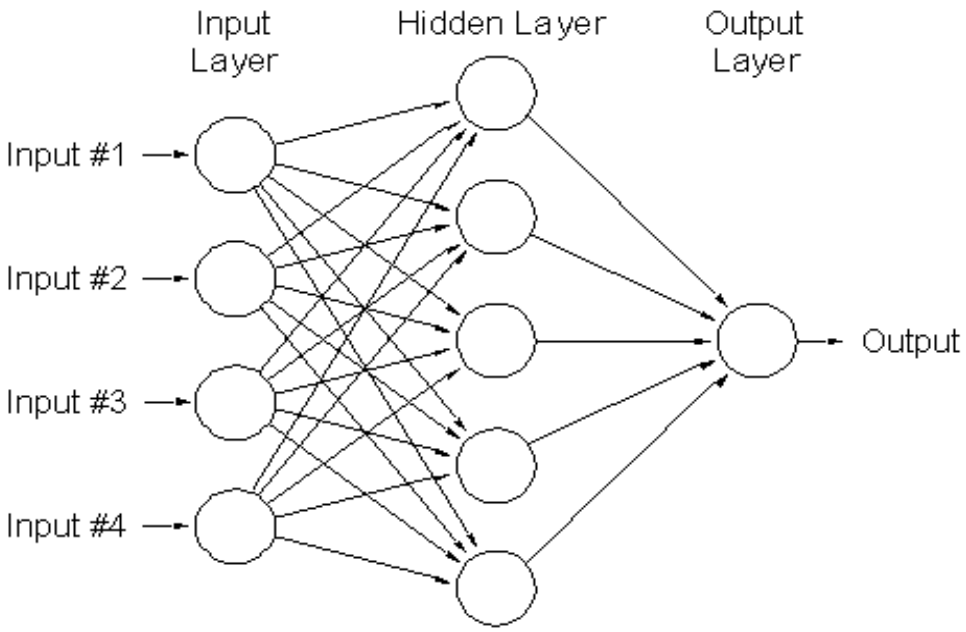

Fig. 2: Sample nervous network model

Weight w_{ij} is calculated via the total multiplication of the internal weights on the node by the values, from which such weights are launched. As an example, the value of weight correlation between layer 7 and 10 will be:

$$w_{7,10} = w_{17} * value\ of\ node\ 1 + w_{27} * value\ of\ node\ 2 + w_{37} * value\ of\ node\ 3 + w_{47} * value\ of\ node\ 4$$

Each node can be seen as an independent variable (nodes from 4: 1), or as a combination of the independent variables (the nodes 10:5.) Node No. 10 is a non-linear combination of values in the node No. 4: 1 because of the presence of activation function (the aggregated values in the hidden layer). It should be noted that if the activation function was linear, and there are no hidden layer, the neural network reduced to linear regression. The neural network is reduced to the logistic regression under non-linear activate functions with a particular form.

Potential:

Weights in the neural network as in the biological model reflect the adjustable factors in response to signals that travel through the network according to an appropriate learning algorithm and threshold value (also known as a bias) that resemble the intersection limit in the regression model. Cell j takes the final value θ_j and receives internal signals $x = [x_1, ..., x_n]$ from units (cells/nodes) connected to it from the previous layer. Each signal is associated to each particular weight $w_j = [w_{1j}, ..., w_{nj}]$.

The incoming signals, their weights, and final value for each cell will be studied through something called Combination Function. Combination Function of each cell produces one value called the potential or (Net Input). Activation function converts the potential into external signal.

Combination function is a linear one in normal. Thus, p_j potential consists of the total deviations of previous cells x_i balanced via w_{ij} weights emerged from the final value θ_j, which shall be expressed symbolically as follows:

$$p_j = \sum_{i=1}^{n} \left(x_i w_{ij} - \theta_j \right) = \sum_{i=0}^{n} x_i w_{ij}$$

As $x_0 = 1$ and $w_{0j} = -\theta_j$ Cell external reference j (i.e y_j) can be obtained by application of

activation function on the potential p_j to give:

$$y_j = f(\mathbf{x}, \mathbf{w_j}) = f(\mathbf{p_j}) = f(\sum_{i=0}^{n} x_i w_{ij})$$

Types of Activation Function:

There are many ways to activate cells in the neural network. The best-known one is linear, piecewise, Sigmoidal, and Softmax method:

1) Linear activation function: Linear activation function will take the following formula:

$$f(p_j) = \alpha + \beta\, p_j$$

Where p_j possibility belongs to a real numbers group, while α, β are constants. When the model requires that the cell exit be completely equal to its level of activation (potential), we shall put $\alpha = 0, \beta = 1$ and the linear function will be turned to unit function. See the strong similarities between the linear activation function and simple linear regression model, as the last can be seen as a simple type of neural networks.

2) Piecewise activation function: Linear activation function is defined in the following format:

$$f(p_j) = \begin{cases} \alpha & p_j \geq \theta_j \\ \beta & p_j < \theta_j \end{cases}$$

It is clear that it takes two values only according to exceeding the potential of final value. When $\alpha = 1, \beta = 0, \theta_j = 0$, we will be in front of a special case of piecewise activation known as sign activation function that takes a value of 1 in the case of positive potential when the value is equal to 0.

3) Exponential activation function:

It is the function that takes letter s, which is most commonly used in practical applications. This function produces positive values only in the interval [0,1]. It common use returned to being non-linear and its ability to be understand and differentiated easily. It is know in the following formula:

$$f(p_j) = \frac{1}{1 + e^{-\alpha p_j}}$$

Where α refers to positive parametric that organizes the inclination of the function.

4) The maximum smoothing function:

Used in normalizing the different nodes that has a relation among them. If the network has g nodes with v_j outputs number (where $j = 1,2,...,g$) the maximum smoothing function that prints v_j (with total equal to 1) will be:

$$soft\max(v_j) = \frac{e^{v_j}}{\sum_{j=1}^{g} e^{v_n}}$$

This function is used to solve the supervised classification problems when dependant variable takes g number of levels.

Training Methods:

Training in terms of neural networks is meant to learn the network how to accomplish a task. In the language of the statistics, it means the method used to estimate the network weights (unknown parametric.) Network can be trained or educated in several methods such as the best known and the

most widespread method of Backpropagation that seeks to update weights using several algorisms such as: The gradient drop algorithm (Delta rule (gradient descent) and comparative gradient algorithm (Conjugate gradient method;–Multilayer Perceptron Neural Networks), quasi-newton algorithm (Neural Network Toolbox, Levenberg-Marquardt (trainlm), Levenberg-Marquardt, and Genetic Algorithms.

Types of Neural Networks:

Neural networks can be classified according to the number of layers into two types: Single-Layer Perceptrons network, and Multi-Layer Perceptrons. Neural networks can be classified according to the direction of information flow into: Feedforward Networks where information moves from layer to the next without possibility of allowing to return backwards and Feedback Networks; where information moves from layer to the next while allowing them to return to the previous layers.

Applications of Neural Networks:

Neural networks are viable in the causal problems in the case of complex relationship between several independent variables (explanatory/ predictors/input) and one or more of the dependent variables (explanatory/predictors/outputs) where it is difficult to express those traditional links via traditional entries such as correlation and regression and differences between the groups. Examples of the problems, at which neural networks were successfully applied (Christos Stergiou and Dimitrios Siganos. Neural Network; Neural Network Software For researchers, data mining experts and predictive analysts) disclosure of medical phenomena, prediction of the stock market, and assessment of the creditworthiness of loan applicants.

Nonparametric Regression (NR)

If, for example, the researcher decided to use many cubic limits on the shape(Mark, 2005):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

During the reconciliation of the regression model which correlates between the variables y and x, that requires the correctness of the mathematical form supposed to express the data. It is said that the model is parametric because it depends on $\beta_1, \beta_2, \beta_3$ parameters. When there is not enough information to make such hypothesis, or if you just want to assume that:

$$y = f(x) + \varepsilon$$

In the light of normal hypotheses [that $f(x), f'(x), f''(x)$ is continuous] and calculating $f(x)$ based on the data, we use the nonparametric regression.

Nonparametric regression (NR) (Nonparametric Regression) is a form of regression analysis, in which the independent variable does not take a specific form, but it builds the information derived from the data. Therefore, NR requires a sample with larger size than the size required to calculate the parametric regression because the data suggests the structure of the model and parameters estimations. NR is estimated through (Hastie *et al.,* 2009; John, 2002).cubic smoothing spline functions that take the following formula:

$$RSS(f,\lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt$$

That means estimating the total residuum squares from all possible functions $f(x)$ under two continuous derivatives. λ refers to the fixed smoothing parameter while the first limit of the right side refers to the near data. The second limit measures the extent of function curvature but λ determines the differentiation between the two limits. If $\lambda = 0$, f can be any function, of which data shall be supplemented. If $\lambda = \infty$, f will be conciliated via least squares straight line due to the

lack of use from the second derivative.

Semiparametric Methods

Semiparametric Regression (SPR):

Semiparametric Regression is a combination of parametric and nonparametric regression. It is used if the full nonparametric model does not reflect the data perfectly and/or if the researcher wanted to use the parametric model, but he does not know the exact function formula for a subset of the explained variables or if the errors density was unknown. Where SPR contains parametric model, it relies on parametric assumptions. Therefore, they are subjected to two important problems: Misspecified (choose a wrong mathematical formula and/or deficiencies in introducing of the problem variables) and Inconsistent (distribution of capabilities away of the real value of the estimated parameter), as in the full parametric models.

There are many ways to estimate the SPR models, most notably:

1) Partially linear model defined as follows:

$$Y_i = X_i'\beta + g\left(Z_i\right) + u_i, \quad i = 1, \ldots, n,$$

Where Y_i refers to the dependent variable, and each of X_i, Z_i refer to the independent variables vector from $p \times 1$ grade. β refers to parameters vector from $p \times 1$ and $Z_i \in \mathrm{R}^q$ grade. Parametric vector β defines the parametric part of the model, while the unknown function $g\left(Z_i\right)$, part will be determined via an appreciated regression method.

2) Single index model, which is also known as Ichimura's method and takes the following form:

$$Y = g\left(X'\beta_0\right) + u,$$

The parameter β_0 is calculated by using the non-linear least squares method to minimize the function:

$$\sum_{i=1} \left(Y_i - g\left(X_i'\beta\right)\right)^2.$$

3) Smooth Coefficient\varying Coefficient Models that is known via the following formula:

$$Y_i = \alpha\left(Z_i\right) + X_i'\beta\left(Z_i\right) + u_i = \left(1 + X_i'\right)\begin{pmatrix} \alpha\left(Z_i\right) \\ \beta\left(Z_i\right) \end{pmatrix} + u_i = W_i'\gamma\left(Z_i\right) + u_i,$$

Where X_i refers to $k \times 1$ grade vector while $\beta(z)$ refers to a vector from smoothing non-determined functions in z.

Application for Comparison between the Models:

The monthly series of cases data considered by the Supreme Constitutional Court in Egypt during 2008-2013 (Supreme Constitutional Court, information and data management) are used to reconcile several models depending on the nature of the data, which explains the seasonal impact that reached its peak in June (611) and the lower level during August (809) as illustrated in the following diagram:

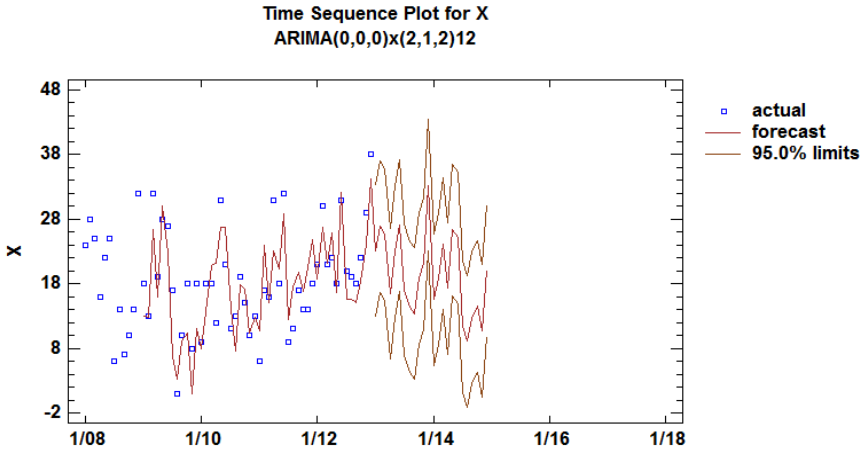

Fig. 3:The number of cases considered by the Supreme Constitutional Court during 2008-2013

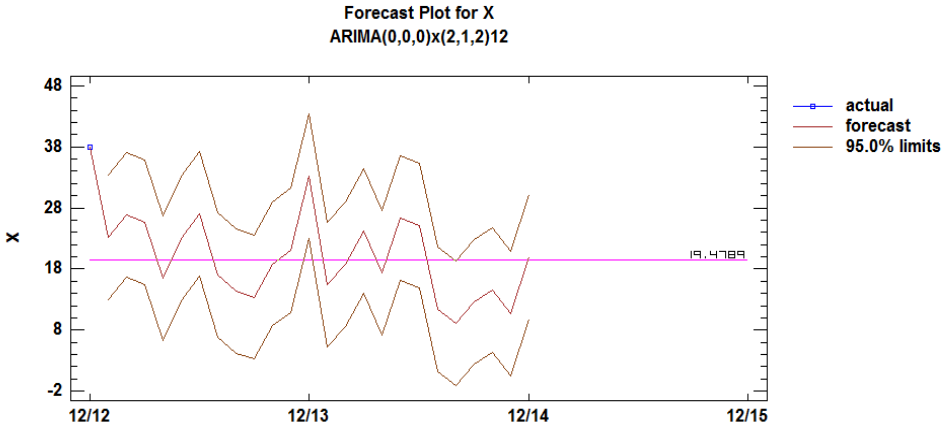

Fig. 4: Seasonal guide to the number of cases during 2008-2013

At the beginning, 18 ARIMA traditional models shown in the following table were reconciled. The best was ARIMA seasonal model (2,1,0)x(1,1,2) 12, as the first autocorrelation coefficient was 0.634.

After that, the best model from the previous ones shall be compared to four new models for 60 months. Table (1) shows the quality standards for the proposed model, which emphasizes the appropriateness of model quality to reconcile the data using statistical standards to measure the ability of the model to predict and use of the following criteria (Abbasi 2011, Wooldridge 2003):

1. Mean Square Error Root (RMSE)
2. Mean Absolute Percentage Error (MAPE)
3. Mean Absolute Error (MAE)
4. Theil Coefficient (T.C)
5. Coefficient of Determination (R^2)
6. Trade sign (TS)
7. The first autocorrelation coefficient (ρ_1).

The criteria used to judge the model preference proved that the neural network model is the best and most appropriate for the data used during the period 2008-2013.

Table 1: Comparison between quality indicators in traditional models and ARIMA model for number of cases during 2008-2013

Model	RMSE	MAE	MAPE
Random walk	9.06997	6.77431	62.3605
Random walk with drift = 0.0104027	9.16648	6.77481	62.3812
Constant mean = 18.8607	7.08446	5.35525	49.0912
Linear trend = -68.2119 + 0.119852 t	7.10294	5.2331	46.7143
Quadratic trend = 6693.6 + -18.5055 t + 0.0128185 t^2	6.58787	4.78467	41.0339
Exponential trend = exp(-2.00416 + 0.006681 t)	7.11337	5.11647	42.5103
S-curve trend = exp(7.54147 + -3406.72 /t)	7.12695	5.12222	42.6378
Simple moving average of 2 terms	7.60982	5.44652	51.2748
Simple exponential smoothing with alpha = 0.2303	6.7967	5.04166	47.5017
Brown's linear exp. smoothing with alpha = 0.1006	6.84085	5.03122	47.4461
Holt's linear exp. smoothing with alpha = 0.1247 and beta = 0.1451	6.90477	4.87344	42.7063
Brown's quadratic exp. smoothing with alpha = 0.0636	6.89329	5.07301	48.5641
Winter's exp. smoothing with alpha = 0.2086, beta = 0.0834,gamma = 0.1744	7.12028	5.28484	58.1989
ARIMA(0,0,0)x(2,1,2)12	4.76095	3.85657	28.3341
ARIMA(2,1,0)x(2,1,2)12	4.72003	3.64274	26.3177
ARIMA(0,0,1)x(2,1,2)12	4.8306	3.88196	27.5077
ARIMA(1,0,0)x(2,1,2)12	4.85863	3.90788	28.9779
ARIMA(1,0,1)x(2,1,2)12	4.80716	3.72139	28.5118

Table 2: Comparison of quality indicators for the five models used to reconcile the preparation of cases considered by the Supreme Constitutional Court during 2008-2013

Model	RMSE	MAPE	MAE	Coefficient of Theil	R2	TS	ρ1
Multiple Regression	32.490	76.933	-23.804	-0.836	14.20%	4.03	0.1831
Poisson Regression	-32.610	70.439	-23.921	-0.838	17.36%	4.01	0.2213
Time series ARIMA	-35.310	33.157	-26.139	-0.895	71.60%	2.39	0.2493
Neural Networks	-36.119	23.685	-27.229	-0.905	85.53%	3.49	-0.0714

Conclusion

The results of this work could be summarized in the following points: the appropriateness of neural networks model in prediction and drawing both long-term and short-term plans because of the speed and accuracy of the data, rather than the traditional statistical methods. By application of both traditional statistical models and artificial neural networks (ANN) is clear that the neural networks are better than the traditional statistical methods because of the methodology of not to rely on the linear data. The artificial neural networks are more accurate and efficient in prediction rather than the traditional statistical methods as the networks are characterized by high rate of precision. In addition, they are favorable in the prediction for long-time series, which does not have a clear influence of seasonal or self-correlation. Neural networks have to be used in studies that contain an outlook. They shall be analyzed using the modern statistical methods to achieve the maximum benefit as the network are speed and accurate. Neural networks exceed the traditional models significantly. In other words, the performance of neural networks is better than the conventional models due to the methodology of non-linear reliance. Neural networks can also be successfully applied to predict the monthly long time series characterized by seasonality or autocorrelation.

References

Al-Abbasi, Abdul Hamid Mohammed (2012), Kuwaiti government work force: Reality and influencing factors during 1993-2011, Statistical Egyptian Journal, Institute of Statistical Studies and Researches - Cairo - Egypt, Vol. 56 Issue (2), December 2012 P. 30-46.

Al-Abbasi, Abdul Hamid Mohammed (2010), modern analysis of time series using Eviwes, Institute of Statistical Studies and Researches, Cairo.

Al-Abbasi, Abdul Hamid Mohammed (2004) comparison between the use of neural networks and ARIMA to predict the numbers of monthly deaths from the traffic accidents in Kuwait, Arab Journal of Administrative Sciences, Kuwait, Vol. (3) Issue No. 11, P. 333-359.

The minister, Rizk Alsayed, Sameer, Hatem Abdel Wahid (2012), "Methods of data mining: Parametric and Nonparametric Methods" Institute of Statistical Studies and Researches, Egyptian Population and Family Planning Magazine, Vol. 45, Issue December 2012 P. 65-85

Analytic Hierarchy Process
http://en.wikipedia.org/wiki/Analytic_Hierarchy_Process

Analytic Hierarchy Process (AHP) Tutorial
http://www.cs.toronto.edu/~sme/CSC340F/slides/tutorial-prioritization.pdf

Backpropagation
http://en.wikipedia.org/wiki/Backpropagation

C4.5 algorithm
http://en.wikipedia.org/wiki/C4.5_algorithm

CHAID
http://en.wikipedia.org/wiki/CHAID

Christos Stergiou and Dimitrios Siganos. Neural Network
http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Neural Networks in Practice

[7] CIPFA (2008). Nearest Neighbours Model: Methodology Note and Instructions
http://www.cipfastats.net/default_view.asp?content_ref=2748

Conjugate gradient method
http://en.wikipedia.org/wiki/Conjugate_gradient_method

0,Conventional programming
http://www.pcmag.com/encyclopedia_term/0,2542,t=conventional+programming&i=40325,00.asp

Cornelius T. Leondes (2002. Expert systems: the technology of knowledge mana-gement and decision making for the 21st century, *Academic Press, pp. 1-22*.

CRISP-DM (2003), CRoss Industry Standard Process for Data Mining
http://www.crisp-dm.org.

Delta rule (gradient descent)
http://en.wikipedia.org/wiki/Delta_rule

Fayyad; U.M., Piatetsky-Shapiro; G., Smyth; P. and Uthurusamy; R (eds) (1996a), Advances in Knowledge Discovery and Data Mining, *AAAI Press*.

Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996b), From Data Mining to Knowledge Discovery: An Overview. In Fayyad; U.M., Piatetsky-Shapiro; G., Smyth; P. and Uthurusamy; R (eds), Advances in Knowledge Discovery and Data Mining , *AI*, *DDM*, *AAAI/MIT Press*, pp. 1-34.

Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996c), The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM, 39 (11), pp. 27-34.

Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996d), Knowledge Discovery and Data Mining: Towards a unifying framework, *AI*, *DDM*, *AAAI/MIT Press*, pp. 82-88.

Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996e), From data mining to knowledge discovery in databases, *AI Magazine*,17, (3), pp. 37-54.

Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines" *Annals of Statistics*, 19 (1): 1–67. doi:.%2%.%.%2%%.%2%/10.1214/aos/1176347963. MR1091842. Zbl .%.0765.62064.

Genetic algorithm
http://en.wikipedia.org/wiki/Genetic_algorithm

[20] Generalized additive model

http://en.wikipedia.org/wiki/Generalized_additive_model
Giudici; P. (2003), Applied Data Mining: Statistical Methods for Business and Industry, *John Wiley & Sons Ltd*.
Hastie; T., Tibshirani; R. (1993). "Varying-Coefficient Models" *Journal of the Royal Statistical Society , Series B, 55, pp. 757–796*.
Hastie; T., Tibshirani; R., Friedman; J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, *Springer Series in Statistics*.
Ichimura, H. (1993). "Semiparametric Least Squares (SLS) and Weighted SLS Esti-mation of Single Index Models" *Journal of Econometrics*, 58*, pp. 71–120*. doi:.%F-%%./-()10.1016/0304-4076(93)90114-K.
Jackson, Peter (1998). Introduction to Expert Systems, 3rd ed., *Addison Wesley*, *p. 2*, ISBN ----4----978-0-201-87686-4.
John Fox (2002). Nonparametric regression
http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf
Koch; C. G., Isle; B. A., Butler; A. W. (1988). "Intelligent user interface for expert systems applied to power plant maintenance and troubleshooting" *IEEE Transac-tions on Energy Conversion, PP. 3- 71*.
Mark E. Irwin (2005). Generalized Additive Models, *Harvard University*
http://www.markirwin.net/stat135/Lecture/Lecture34.pdf
Mark E. Irwin (2005). Non Parametric Regression, *Harvard University*
http://www.markirwin.net/stat135/Lecture/Lecture33.pdf
Mathematical optimization
http://en.wikipedia.org/wiki/Mathematical_optimization
Mathematical Programming
http://www.me.utexas.edu/~jensen/ORMM/frontpage/pdf/mathprog.pdf
Multivariate adaptive regression splines
http://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines
Multilayer Perceptron Neural Networks
http://www.dtreg.com/mlfn.htm
Neural Network Software *For researchers, data mining experts and predictive analysts*
http://www.alyuda.com/products/neurointelligence/neural-network-applications.htm
Neural Network Toolbox, Levenberg-Marquardt (trainlm)
http://www.caspur.it/risorse/softappl/doc/matlab_help/toolbox/nnet/backpr11.html
Newton's Telecom Dictionary (2010), Harry Newton, CMP Books,
http://www.cmpbooks.com.
Nonparametric regression
http://en.wikipedia.org/wiki/Nonparametric regression
Nwigbo Stella and agbo Okechuku Chuks(2011)."Exert system: a catalyst in educational development in Nigeria*," Proceeding of the 1stInternational Techn-ology, Education and Environment conference, (c) African society for* scientific *Research (ASSR)*
http://www.harmars.com/pics/261.pdf