International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 3, June 2016, pp. 974 ~ 979
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i3.9155

Machine Learning Techniques on Multidimensional Curve Fitting Data Based on R- Square and Chi-Square Methods

Vidyullatha P, **D. Rajeswara Rao**
Departement of Computer Science & Engineering, KL University Guntur, India

Article Info

Article history:

Received Sep 7, 2015
Revised Nov 23, 2015
Accepted Dec 10, 2015

Keyword:

Chi-square
Curve fit
Interpolantlinear
Labfit
Surface fitting tool

ABSTRACT

Curve fitting is one of the procedures in data analysis and is helpful for prediction analysis showing graphically how the data points are related to one another whether it is in linear or non-linear model. Usually, the curve fit will find the concentrates along the curve or it will just use to smooth the data and upgrade the presence of the plot. Curve fitting checks the relationship between independent variables and dependent variables with the objective of characterizing a good fit model. Curve fitting finds mathematical equation that best fits given information. In this paper, 150 unorganized data points of environmental variables are used to develop Linear and non-linear data modelling which are evaluated by utilizing 3 dimensional 'Sftool' and 'Labfit' machine learning techniques. In Linear model, the best estimations of the coefficients are realized by the estimation of R- square turns in to one and in Non-Linear models with least Chi-square are the criteria.

Corresponding Author:

Vidyullatha P,
Departement of Computer Science and Engineering,
K L University,
Vaddeswaram 522502, Guntur District, Andhra Pradesh, India.
Email: lekana04cuty@gmail.com

1. INTRODUCTION

The analysis of data mainly focuses on the relationship of given variables. Statistically, the relationship is measured by using correlation. The standard method using in the correlation model is Pearson method in which coefficient are limited between minus one and plus one. There won't be any relation between data and yield variables if coefficient is zero. There is an absolute relation existing if coefficient is one. On the off chance that free variable(x) increases dependent variable(y) will also increments precisely in linear relation. Something else, if x diminishes y expands then the relation is called negative linear and the coefficient is minus one. Curve fitting is a more elevated amount of numerical structure than relationship. The usefulness of curve fitting mainly aims at formulation of a mathematical function by using type of input data. Different sorts of curves such as parametric curves, implicit curves and subdivision curves are utilized for fitting. Fitting a suitable curve or model to a progression of data points is a major necessity in numerous fields, for instance, computer graphics, image processing and data mining. There are different types of models to fit the curve. The different models are linear, polynomial of various degrees, power fit, logarithmic curve fit and non-linear curve fits. The curve fitting not just fits the unorganized data into various models additionally performs various tasks such as to reduce the noise, find the mathematical relationship among variables and assessment the qualities between data samples. For the curve fitting modelling processes, Wenni Zheng etal [1] proposed a B-spline curve fitting method in view of the L-BFGS optimization strategy on disorderly information purposes of foot point projection and demonstrates that it is the speediest technique in every cycle contrasted with traditional strategies. Yang etal [2] describes spline representation in curve fitting method. Weenie Zheng [3] proposed optimization method for fast fitting of B-spline curves to unorganized data points Francis etal [4] examined 3D parameter yield curve fitting method for predicting the

short and long terms structure of Government security yields. Paul Norman etal [5] evaluated the unit values from collected information like age and salary groups to make non-linear regression strategies of curve fitting models using the SPSS software. Fox j and Weisberg [6] portrayed fitting of non-parametric regression models utilizing R programming language which likewise incorporated smoothing systems on scatter plot strategies. Lian Fang etal [7] presented a method for producing a smoothen parametric curves on unordered data points focus on 2D data. Anandthirtha etal [8] worked on speech data, applied correlation coefficient and curve fitting method to focus the varying degrees of extreme speech disabilities to compare with normal child. In this paper, the surface fitting tool, which is a 3D representing data tool is utilized for fitting the unorganized environmental data points and plotting the graphs. Generally in MATLAB, basic fitting tool, curve fitting tool and surface fitting tool are available to fit the data. The first two fitting tools are 2D where the "Sftool", in Matlab R2010b and Labfit techniques are 3dimensional data fitting methods.

2. RESEARCH METHOD

2.1. Choosing a Curve Fit Model

Choosing a kind of curve fit normally depends upon the type of data. In this paper, for best curve fit, sftool in Matlab and Labfit are utilized. In linear model, for minimizing the sum of the squares, there are different methods namely Gauss-Newton method, Gradient descent method and Levenberg-Marquardt Method. In non-linear model, coefficients are calculated at minimum Chi-square value. Data fitting is the procedure of fitting models to information and investigating the exactness of the fit. Specialists and researchers use information fitting procedures, including scientific mathematical statements and nonparametric techniques, to model gained information. MATLAB® gives you a chance to import and picture your information, and perform essential fitting methods. The curve fit models are not only fit the data but also reduce the noise and smoothen the data [8].

2.2. LAB Fit tool

The Labfit is programming package customized for testing and treatment of data. In LAB Fit, there are numerous favorable circumstances like treating of comparable information, non-practically identical data, discover probabilities for a few movements, focus engendered mistakes, plot 2D and 3D graphs, and execute a few estimations in an arrangement of linear comparison and Curve Fitting. The Labfit is mainly intended for curve fitting using nonlinear regression. The Labfit software can be utilized to fit curve up to six autonomous variables and one subordinate variable. There are about 280 functions with two free variables in Labfit library. The users of the software can also compose their own particular fit function in Labfit. It is helpful to fit curves in both 2D and 3D cases. The LabFit is used to treat distinctive sorts of data such similar data, non-practically identical data and error propagation. The Labfit has given 10 scientific mathematical equations which were shown in table 1 for better fitting of given data as indicated by Chi-square estimates. From the above table, it is clearly evident that the equation (1) best fits the given data since it has minimum Chi-square value. Subsequent to guaranteeing the best fit numerical mathematical statement, click on "Results: diagram" to get a 3D chart as indicated in Figure 5.

2.3. Surface Fitting Tool in Matlab

For fitting the curves and surfaces to the data points, surface Toolbox is one of the applications provided by Matlab [9]. The curve fit toolbox gives a chance to perform exploratory information analysis, for preprocessing and compare candidate models and remove outliers. It will provide the linear & non-linear models and also indicate custom mathematical equations. It also supports non-parametric modeling techniques such as splines, interpolation and smoothing of data points. Many statistical packages such as R and numerical software such as GNU Scientific Library, Maple, MATLAB, SciPy and Open Opt are useful for doing curve fitting in a various situations. There are additional programs particularly kept in touch with the curve fitting. The command Sftool opens curve fitting application which gives an adaptable interface and curve can be viewed the plots. Sftool [10] has numerous favorable circumstances, for example, to create, plot & compare multiple fits and also automatically produces the mathematical equations.

3. RESULTS AND ANALYSIS

One must choose right fitting tool prior to selecting a data set. By selecting a wrong tool, the user is going to get an inappropriate curve thus, it is ideal to pick right fitting model contingent upon the given information. In general, each model has its own presumptions for computing the fitting error to fit the curve. Contingent on the outcomes, three parameters are computed such as sum of squares of error (SSE), root mean squared error (RMSE) and R-square error (R^2). These three statistical parameters have the capacity to give

definite elucidation with respect to rightness of curve fit. If a model is in linear, the constraints such as constant, parameter and a predictor has a basic equation represented as y = a0 + a1x1 + a2x2. For linear equation, there is only one basic form whereas for nonlinear there are many different forms such as power, parabolic, exponential etc. In this research work, 150 data points of environmental variables are taken to represent a data model using surface fitting tool in Matlab [11] which results a linear data representation model. In this work, two dependent variables and one independent variable data points are taken into consideration for fitting the 3D graph which are shown in figures 2-4 having linear in function. Utilizing the surface fitting tool, it is easy to plot and examine the fits at the command line. At first sight it is obvious that a good fit should minimize the so called residuals. The R^2 measure is the most generally utilized and reported measure of error and goodness of fit for linear models. The estimation of R^2 measures goodness of fit which lies somewhere around 0.0 and 1.0. When R^2 reaches to 1, it is better to stop the curve fit process and the best fit results to make perfect predictions are shown in figures in 1 to 4. R^2 is determined by using two terms namely SSE (sum of squares of errors of regression) and SST (sum of squares of total). SSE is computed from the sum of the squares of the observed value and predicted value. This is also called the sum of squares of regression. SST is calculated from the sum of the squares of the observed value and mean value. If the curve fit is good, normally SSE is smaller than SST. Then, R^2 is calculated using this equation: R^2 = [1 - (SSE/SST)] =1.0-4165/62735= 0.9336. In general, Matlab gives functions in the form of arrays. It is difficult to find out functions at different points which are not covered in arrays. There are two methods to find out values between data points and beyond data points; interpolation and extrapolation respectively. A straight line can be defined between two data points (x1, y1) and (x2, y2) as given below:

$$y = y1 + \frac{(y2-y1)}{(x2-x1)}(x - x1).$$

This basic geometric formula is used to linearly interpolate between two data points. For better understanding, $y = (y_1 + y_2)/2$ if x is midway between x_1 and x_2. One should be cautious while using this formula. The linear approximation to the curved function represented by the dashed line "a" is pretty poor since the points x = 0 and x = 1 on which this line is drawn are far apart. By Adding a point between 0 and 1 at x = 0.5, then we get two-segment approximation "c" which is quite somewhat better. It can also be observed that that line "b" is a pretty good approximation since the function doesn't curve much. The same formula may be used for extrapolation also. It can be shown that the formula $f_{N+1} = 2_{fN} - f_{N-1.}$ If the end values in the array are f_{N-1} and f_N. Matlab is haviing its interpolation code as "interp1". For example, if we are having a set of data points {x, y} and we are having a different set of x-values $\{x_i\}$ for which we want to find out corresponding $\{y_i\}$ values, we can easily utilize the following three formulae for interp1 command:

y_i=interp1(x,y,x_i, 'linear')
y_i=interp1(x,y,x_i, 'cubic')
y_i=interp1(x,y,x_i, 'spline')

An example code is written below for data set representing the sine function.

```
clear;
% makes the data set with dx
dx = pi/5; x=0:dx:2*pi; y=sin(x);
% for a fine x-grid
xi=0:dx/20:2*pi;
% interpolate on coarse grid
% obtain yi values
% linear interpolation
yi=interp1(x,y,xi, 'linear');
% plot the data and the interpolation
Plot (x, y, 'b*', xi, yi, 'r-')
title ('Linear Interpolation')
% cubic interpolation
yi=interp1(x, y, xi, 'cubic');
% plot the data and the interpolation
figure
plot (x, y, 'b*',xi, yi, 'r-')
title ('Cubic Interpolation')
% spline interpolation
yi=interp1(x,y,xi, 'spline');
% plot the data and the interpolation
```

figure
plot(x, y, 'b*',x_i, y_i, 'r-')
title ('Spline Interpolation')
In case of 3-D interpolation on a data set of {x, y, z} to get the approximate values of z(x, y) at points {x_i y_i}, we use any of the following
z_i = interp2(x, y, z, x_i, y_i, 'linear')
z_i = interp2(x, y, z, x_i, y_i, 'cubic')
z_i = interp2(x, y, z, x_i, y_i, 'spline')

In this paper, the unorganized environmental data points are used to plot the curve. There are various interpolant fitting methods are available based on the type of curves and surfaces. For Non-linear model, almost 100 data points are taken to get a mathematical formula and 3D graph by using LAB Fit. This soft computing technique has evaluated best 10 mathematical equations and the equation Y=A*X2** (B*X1) has given best fit since it has got lowest Chi-square value for given data shown in the below Table 1.

Table 1. The List of Mathematical Functions in Labfit Method

S.No	Mathematical Computing Functions	Chi-square
1	Y=A*X2**(B*X1),where A=50.6,B=0.004	242.3
2	Y=A*(X1*X2)**B, where A=18.1, B=0.21	244.1
3	Y=A*X1**(B*X2),where A=55.2, B=0.002	271.6
4	Y=A*X2**(B/X1), A=78.2, B= -0.65	332.3
5	Y=A*X1**(B/X2), where A=75.7, B= -0.73	334.4
6	Y=A*(X1/X2)**B, where A=68.5,B= -0.005	352.2
7	Y=A*X1+B*X2**2,where A=2.8, B=0.009	575.2
8	Y=A*X2+B*X1**2,where A=1.7, B=0.03	583.8
9	Y=X1/(A+B*X2), where A=0.31, B= -0.001	616.4
10	Y=X2/(A+B*X1**2),where A=13.1,B= -0.03	4637.8

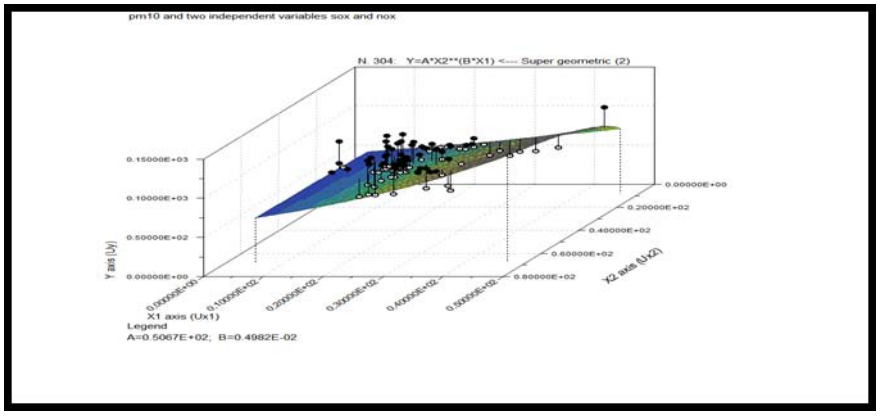

Figure 1. A 3D Graph By Labfit Method

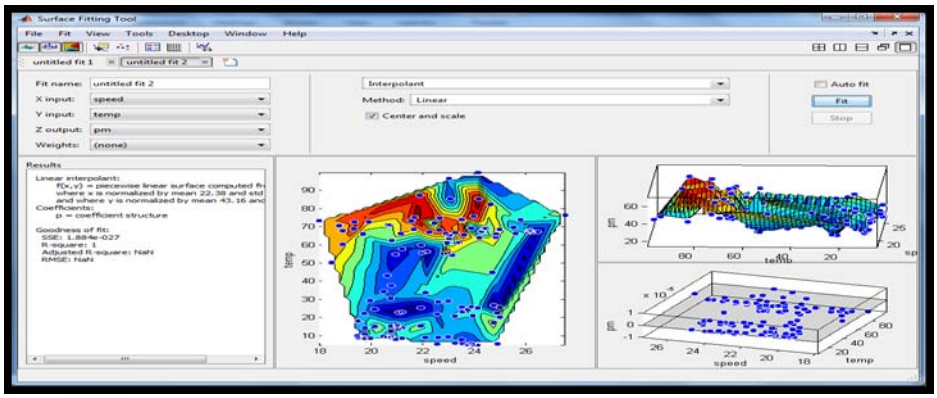

Figure 2. Interpolant Linear Relationship And Residual Plots Showing In Matlab

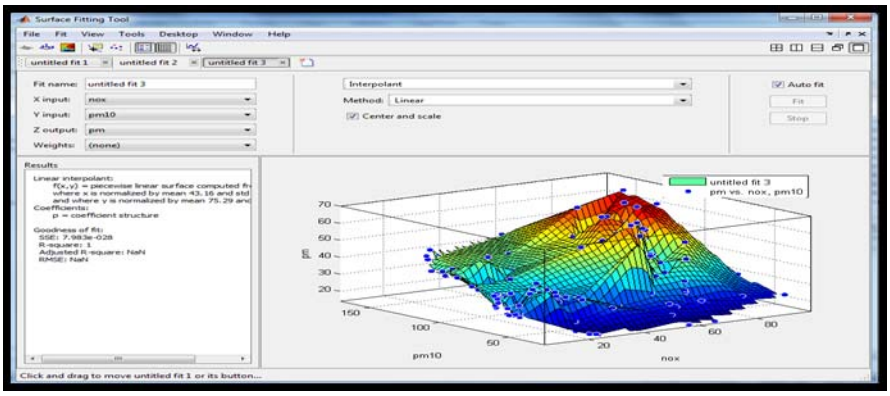

Figure 3. Interpolant Linear Relationship Showing R-Square Equals To One

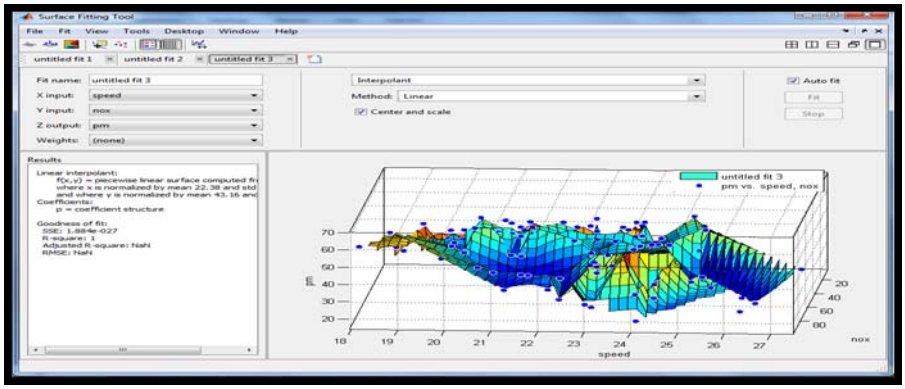

Figure 4. The Data Points Showing Linear Relationship in 3-D View

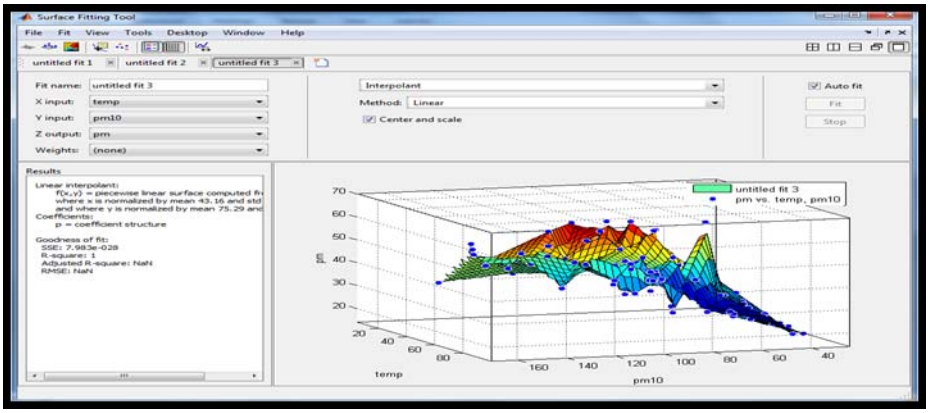

Figure 5. The Data Points Showing Linear Relationship In 3-D View Using Matlab

4. CONCLUSION

This research paper has described the data representation methods using MATlab and LABFit. Curve fitting gives detailed account of inter-relation of dependent variable with respect to independent variables. In this paper, one dependent and two independent variables are considered to evolve best fit model. In linear model, Curve fitting finds the values of the coefficients (parameters) which make a function match the data as closely as possible. The best values of the coefficients are known when the value of R-square becomes one. In Non-linear model, Curve fitting assesses and discovers the best scientific mathematical

statement whose Chi-square esteem is least. The fitting models and methods used here depend on the input data set. This paper portrays historical environmental data consists of linear variables which is fitted by information utilizing 3 dimensional. For data fitting, 150 non linear data points are collected from one of the coal consuming power generation plant and are used to plot the graph using "sftool" command in Matlab. For curve fitting, in the Matlab, the data is plotted in interpolant linear showing that R-square equals to one and it is represented in 3 dimensional methods for all variables. The given variables demonstrated that they are linear in structure so that, for forecast examination, the multi linear regression is the better decision which follows the mathematical equation Z= a+ b1x1 + b2x2+ b3x3. For Non-linear model, almost 100 data points are taken to get a mathematical formula and 3D graph by using LAB Fit. This soft computing technique has evaluated best 10 mathematical equations and the equation Y=A*X2**(B*X1) has given best fit. The ability to do curve fitting is an extremely helpful expertise for forecasting purposes. The future scope is reached out to gather more information focuses to speak to in 4D perspectives.

REFERENCES

[1] Z. Mei, *et al.*, "Curve fitting and optimal interpolation on CNC machines based on quadratic B-splines," *SCIENCE CHINA Information Sciences*, vol/issue: 54(7), pp. 1407–1418, 2011.
[2] X. Yang, "Curve fitting and fairing using conic splines," *Computer-Aided Design*, vol. 36, pp. 461–472, 2004.
[3] W. Zheng, *et al.*, "Fast B-spline curve fitting by L-BFGS," *Computer Aided Geometric Design*, vol. 29, pp. 448–462, 2012.
[4] F. X. Diebold and C. Li, "Forecasting the term structure of government bond yields," *Journal of Econometrics*, vol. 130, pp. 337–364, 2006.
[5] P. Norman, *et al.*, "Estimating detailed distributions from grouped sociodemographic data: 'get me started in' curve fitting using nonlinear models," *J Pop Research*, vol. 29, pp. 173–198, 2012.
[6] J. Fox, "Nonparametric Simple Regression: Smoothing Scatterplots", SAGE: Thousand Oaks, California, 2000.
[7] L. Fang and D. C. Gossard, "Multidimensional curve fitting to unorganized data points by nonlinear minimization," *ComPuter-Aided Design,* vol/issue: 27(1), pp. 48-58, 1995.
[8] http://lab-fit-curve-fitting-software.soft112.com/
[9] http://cda.psych.uiuc.edu/matlab_pdf/curvefit.pdf
[10] http://www.graphpad.com/guides/prism/6/curve-fitting/
[11] http://cn.mathworks.com/help/pdf_doc/curvefit/curvef