ISSN:0975-9646

Slope Switch Variable Value Prediction using Two Dimensional Curve Fitting for a Limited Data Mining Set

Venkatesh Suvarna

Computer Engineering Department, NMIMS University
Mumbai, India

***Abstract*— Data mining is a very vast field, being used for a variety of uses, mainly in businesses. It is the science of extracting meaningful information from an existing set of information. One such use could be predicting the sales for the next year, from the previous ten years sales figures. Such objectives could be achieved using Data Prediction Algorithms. Many people have designed such algorithms and they should be as accurate as possible, and they usually give better results, as there are more data points available. But, designing such algorithms are difficult, when you have limited data points available. Here, we aim to propose a new method to predict the value of a variable using few data points available, the method being highly accurate and simple to implement.**

***Keywords*— Data Mining, Curve Fitting, Interpolation, Variable Prediction, Data Set, Programming.**

I. INTRODUCTION

This sub-field of data mining, especially variable prediction, could be done using the technique of Curve Fitting. Curve Fitting is the process of formulating a curve that passes through the available data points as closely as possible. We shall first go through the basic techniques for curve fitting and then proceed with the new technique. The techniques available for curve fitting are:

1) Interpolation
2) Single Curve Fitting

1) Interpolation

Interpolation is done by connecting a single line between every two points. The example of interpolation of a set of data points is as shown below.

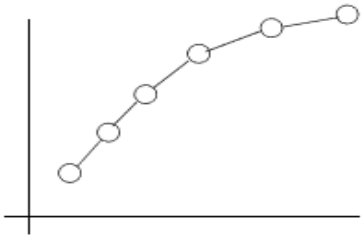

Fig 1. Interpolation on a set of plotted data points

This has limited use as a general function f(x). But since it is really a group of small f(x)'s and connecting one point to the next, it fails to work for data that has built in random error(scatter). And if we use this technique, we defer from the aim of defining a general curve that adheres to the data points with maximum accuracy possible. So we use the second technique, the single curve fitting.

2) Single Curve Fitting

The above technique formulated to a number of functions, for N data points, there would be N-1 f(x)'s which are very difficult to implement as a computer program. The technique. The below figure shows the comparative diagram for fitting a linear curve for a set of data points, using interpolation and single curve fitting.

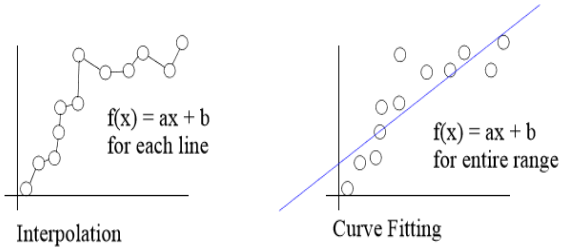

Fig 2. Comparative diagram of interpolation and curve fitting on a set of data points.

This approach is much better than the previous approach and would return with a single function f(x) that tries to pass through the data points as closely as possible.

The linear curve which is fitted through the data points, rarely passes through the points and would be more inaccurate if the data points had shown an increasing and then a decreasing nature. In that case we could then fit a 2-degree curve given by $f(x) = ax^2 + bx + c$. Some data points that exhibit more complex behavior, will require higher order functions to fit in the points more suitably, so selecting the appropriate equations that will fit in the data points with higher accuracy, requires a calculation parameter that will allow us to select the appropriate equation type. Once that is done, we could substitute the data points to get our required equation. Then, to calculate the value of the independent variable at a certain value of the independent variable, x in this case, into the obtained f(x) equation. The value of f(x) would be the expected value of the dependent variable.

Now, that we have understood the process of curve fitting, we shall go through the proposed method, which allows for maximum accuracy for variable value prediction.

II. PROPOSED METHOD

This method involves the following steps:

1) Extracting the data points
2) Calculating the slopes between points
3) Calculating number of sign changes in slopes

4) Calculating order of equation
5) Fitting the data points in equation
6) Put value of independent variable in equation to get predicted value of dependent variable

Now that the steps are clear, we shall now see in detail, how each steps are calculated.

The first step involves the data points, for the case of 2 variables, in the form (xi,yi) would be given. The xi is the independent variable and yi is the dependent variable. Since, we have already stated that this method works best for variable value prediction only when the available data points are few in number. Once, the data points are available in the specified format, we can proceed with the next step.

The second step involves calculating the slopes between the points (divided into the section of two). The slope of the points say (x1,y1) and (x2,y2) are calculated as

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

The slopes between each pair of points is calculated, the same process that is carried out in interpolation. So, if there are N data points, then there will be N-1 slopes, each calculated using the above formula. The slopes calculated are stored together in an list for the next step.

The third step is to calculate the number of sign changes in the slopes. Let say there are N data points and there are N-1 slopes.We assign a sign change variable to zero. Let the slopes be $m_1, m_2, ..., m_{N-1}$. So we then start with slope m_1 with the next slope and check if there is any sign change. If there is a sign change, then we increment the sign change variable by 1. Else, if there is no sign change, then we do no change to the sign change variable. Once we are done scanning all the slopes, we obtain the final value of the sign change variable which is required for the next step.

The fourth step is the most important part of the process, deciding the order of the equation that would be most suitable for the behavior of the data points. Now, if the value of sign change variable is k, then the polynomial curve which is to be fitted to the data points will be of the order k, the same as the value of the sign change variable. So if k=2, then the equation that will best suit the data points would be:

$$f(x) = ax^2 + bx + c$$

And, if k=3, then the equation that will best suit the data points would be:

$$f(x) = ax^3 + bx^2 + cx + d$$

So, after we obtain the order of the polynomial equation, and its general form, as shown above, we would proceed to the next step.

The fifth step is to substitute the points in the equation obtained from the previous step, to get the value of the unknown variables. In the above equation, where k=3, x is the independent variable, y is the dependent variable, and variables a,b,c,d are unknown variables whose values are calculated by substituting the data points in the above equation. In general, the number of unknown variables would be equal to k+1. After substituting the data points we get k+1 equations which needs to be simultaneously solved to get the value of the unknown variables.

The sixth and the final step is to put the value of independent variable in the final equation obtained, to get the predicted value of the dependent variable. For example, if we assume k=3 and the polynomial equation when fitted to the data points gives us,

$f(x) = 7x^3 + 2x^2 + 4x + 10$, then to calculate the value of the predicted variable at x=7, then we put x=7 in the above equation to obtain the predicted value, $f(7)$, as 2537.

III. Performance Of The Method

This algorithm for variable value prediction using polynomial curve fitting, is highly accurate, as the method dynamically sets the order of the polynomial equation, to ensure that the equation passes through every data point. By this the error between the estimated data point and the actual data point is zero. This method works when the data points are limited in number say 10-15. The method could be implemented for higher number of data points but then the complexity of the algorithm would increase substantially, hence not recommended. This method could find a lot of applications in data mining in computer science, sales prediction and fuzzy logic.

IV. Method Implementation Using Programming

The proposed method could be implemented as a computer program, using any suitable programming language. The method uses the non-derivative approach for curve fitting, so could be implemented with ease. To store the data points, we could use arrays or lists to store them, we could store them in pairs, or a separate list for the independent and the dependent variable. The slopes could then be stored in another list, equation solving using the matrix method, and would ensure successful implementation of the method as a computer program.

IV. Conclusion And Future Scope

This method, involves a lot of mathematical computations, but finds immense applications in data miming for data prediction. This method was specifically targeted for variable value prediction when the available data points are few in number. As mentioned before, the complexity of the process increases with the increase in the data points, but it ensures higher accuracy. This research paper will serve as a baseline for researchers to allow for more such algorithms in this field, only to make it better.

Acknowledgment

I would like to acknowledge all my teachers, and my mother, for helping me during my research work. Without them, this paper would not be possible.

References

[1] Curve Fitting, web.iitd.ac.in/~pmvs/courses/mel705/curvefitting.pdf.
[2] *Curve Fitting, en.wikipedia.org/wiki/Curve_fitting*.