ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 4, April 2013

Data Mining: Estimation of Missing Values Using Lagrange Interpolation Technique

L.Sunitha **Dr M.BalRaju** **J.Sasikiran**

Department of computer Science and Engineering
Vidya Vikas Institute of Technology, Hyderabad, India

Abstract
In the real world Most of the datasets have missing data. The presence of missing values in a dataset can affect the performance of Mining Algorithms. In this paper we are using Lagrange interpolation method for prediction of missing values. This method is used to map a data item to a real valued prediction variable. In a dataset if one attributes is depending on other then by using known values we can find the unknown values.

KEYWORDS: Missing values, Interpolation, Prediction.

1. Introduction

In Data mining preprocessing [1] is important step .Pre processing means preparing data for mining. Missing data [2][3] is a common problem in all kids of research. The presence of missing values in a dataset can affect the performance of a classifier constructed using that dataset as a training sample. Several methods have been proposed to treat missing data and the one used more frequently is deleting instances containing at least one we can handle the missing values by ignoring data row, using global constant to fill missing value, using attribute mean to fill missing value missing value of a feature.

2. Common missing value mechanisms

In this paper we carry out experiments with datasets to evaluate the Effect on the misclassification.
Four methods for dealing with missing values:
1. **Removing the tuples**: If a missing value occurs on any of the p variables, eliminate the entire observation. This is the default method for most procedures. Problem with this method data set with even a modest amount of missing values scattered throughout can result in a substantial reduction in sample size.

2**. Filling the missing values manually**: This method is time consuming and not feasible for a large data set with many missing values.

3. **Use global constant**: Replace all missing attribute values by the same constant such as a label "unknown "or -∞.

4. **Use the attribute mean value**: Filling in the missing values of a variable. Substitution with a measure of central tendency, Mean, Median, Midrange, (Max + Min)/2 and Mode

2. Interpolation

Interpolation [4] is the process of finding unknown values from known values. Interpolation is one of the simplest methods which require knowledge of two point's constant rate of change. In computers, it is expensive to store

tabulated functions in its memory. For this, it is more convenient to use an algorithm to calculate the value of functions like sine for any arbitrary values of the argument. For instance, Take values (xi , yi), where i = 0, 1, ... , n of any function Y = f(x), The process of estimating the values of y, for any intermediate value of x is called interpolation. The method of computing the value of y,

3. Lagrange interpolation:

The Lagrange gave the following interpolation polynomial p(X) of degree n given n+1 points (xi, yi) .An interesting feature of this formula, and the feature we aim to preserve in generalizing it, is that
When we substitute xi for X, `i(xi) = 1 and `j(xi) = 0 (j 6= i), giving y = yi.

$$P(x) = \sum_{j=1}^{n} P_j(x),$$ where

$$P_j(x) = y_j \prod_{\substack{k=1 \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k}.$$

Complete formula is
(x-x1)(x-x2)…(x-xn)/(x0-x1)(x0-x2).(x0-xn)
*f(x0) + (x-x0)(x-x1)…(x-xn)/(x1-x0)(x1-x2).(x0
*f(x1)+(x-x0)(x-x1).(x-xn)/(x2-x0)(x2-x2).(x2-xn
Example:
Compute **f(0.3)** for the data

x0	**1**	**3**	**4**	**7**
f	**1**	**3**	**49**	**129** **813**

X=0.3
(0.3-1)(0.3-3)(0.3-4)(0.3-7)/(0.3-0)(0.3-3)(0.3-4)(0.3-7)
(0.3-0)(0.3-1)(0.3-4)(0.3-7) *3
+(0-1)(0-3)(0-4)(0-7)/(1-0)(1-3)(1-4)(1-7)
(3-0)(3-1)(3-4)(3-7) *49 +(0.3-0)(0.3-1)(0.3-3)(0.3-7)
(0.3-0)(0.3-1)(0.3-3)(0.3-4) *129 + *813
(4-0)(4-1)(4-3)(4-7) / (7-0)(7-1)(7-3)(7-4)
=1.831

Algorithm: Lagrange interpolation

Step1: Read list of x value and corresponding y
Step2: Read Xp
Step3: Initialize sum to 0 and lf to 1
Step 4: for i=0 to n-1 do
Step 5: for j=0 to n-1 do
Step 6:if(i!=j)
Step 7:lf= (xp-xi)/(xi-xj)*lf
Step 8:sum=sum+lf*yi
Step 9:write ‘estimated value at xp’,sum
Stop

4. Implemetation and experimental Results

Example:

years(x)	3	8	9	6	11	1
salary(y)	30	57	64	43	59	20

How many record you will be enter: 6
Enter the value of x0 and y0 3 30
Enter the value of x1 and y1 8 57
Enter the value of x2 and y2 9 64
Enter the value of x3 and y3 6 43
Enter the value of x4 and y4 11 59
Enter the value of x5 and y5 1 20
Enter X for finding f(x): 10
f(10.0) = 66.729988
Therefore salary for 10 years is 66.729988

Inverse Interpolation: It is possible to find x value for given f(x) value this process is known as inverse interpolation [6]. In Lagrange formula just by substituting x=y,x[i]=y[i],x[j]=y[j]

Example: The following table gives the sales of a concern for the five years. Estimate the sales for year 1990.

Year(x)	1985	1987	1989	1991
Sales(in lakhs)	40	43	48	52

Results: we can enter (input) the above table to the program
How many record you will be enter: 4
enter the value of x0 and y0 1985 40
enter the value of x1 and y1 1987 43
enter the value of x2 and y2 1989 48
enter the value of x3 and y3 1991 52
Enter x for finding f(x): 1990
F (1990.0) = 50.312500

Result2: now we can find in which year 50 lakh sales will takes place
How many record you will be enter: 4
enter the value of x0 and y0 1985 40
enter the value of x1 and y1 1987 43
enter the value of x2 and y2 1989 48
enter the value of x3 and y3 1991 52
Enter f(x) for finding x: 50
f(50.0) = 1989.851807
Therefore In the year 1990(rounding) the sales are 50 lakhs.

how many record you will be enter: 3
enter the value of x0 and y0 2 1.414
enter the value of x1 and y1 3 1.732
enter the value of x2 and y2 4 2
Enter f(x) for finding x: 1.5793
f (1.6) = 2.494540.

5. Conclusion.

In this proposed work using Lagrange's Interpolation we can find the missing values in preprocessing. If two attributes are related to each other .x is independent and f(x) these are depending on x.from known values we find values .and also possible beyond the boundaries which are closer to boundaries also possible to estimate this is known as extrapolation .algorithm is implemented in c and test results are generated. Both for a give set of x and corresponding f(x) values are considered, we find f(x) for given x value .By using inverse interpolation we can find x for given f(x) value vise versa. The limitation is if the polynomial is existing then interpolation is used, if the f(x) not in uniform change we can go for regression model.

6. References:

1.http://www.iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

2.DataPreprocessing,n.wikipedia.org/wiki/Data_pre-processing

3. TREATMENT of Missing Data--Part David C. Howell, http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

4.http://mat.iitm.ac.in/home/sryedida/public_html/caimna/interpolation/lagrange.html

5.Chapter3Interpolationhttp://www.mathworks.com/moler/interp.pdf

6. New inverse interpolation methods, Alexandra opris¸ http://www.cs.ubbcluj.ro/~studia-m/2006-1/oprisan.pdf

Author profile.

Lingam sunitha received her MCA from Kakatiya University in 1999, and M.Tech (CSE) from JNTU, Hyderabad in 2009. She is now working as Associate Professor and also pursuing Ph.D in Computing Science and Engineering from JNTU Hyderabad, India. Her area of specialization is Data Mining.

Dr M.BalRaju received M.Tech (CSE) from Osmania University and Ph.D from JNTU Hyderabad in 2010.Now he is working as Professor and Principal in Vidya Vikas Institute of Technology, Hyderabad, India. His area of specialization includes Data Base, Data Mining, and Image Processing.

J.SasiKiran received M.Tech (CC) from Bharath University, ,Chennai.He registered Ph.D (CS) from Univ. of. Mysore in [2009-13].Now he is working as Associate Professor in Vidya Vikas Institute of Technology, Hyderabad, India. His area of specialization includes Network Security and Image Processing