An effort to optimize the error using statistical and soft computing methodologies

Gopal Purkait
Department of computer Sc. & Engineering
Pailan College of Management & Technolgy
Joka, Kolkata-700104, West Bengal, INDIA
purkait.gopal@gmail.com

Dhrampal Singh
Department of computer Sc. & Engineering
JIS College of Engineering, Block "A" Phase III
Kalyani, Nadia-741235, West Bengal, INDIA
dharmpal1982@gmail.com

Abstract: In today's era, decision making problems are getting more importance than other's problem. In this paper, an effort has been made to make a comparison on the performance of different statistical methods like least square techniques based on linear equation, exponential equation and logarithmic equation on iris data set. Initially, data preprocessing techniques applied on the data set to clean it in proper format and later on concept of factor analysis on preprocessed data set to finds total effect value. The methods least square techniques based on linear equation, exponential equation and logarithmic equation have been used on total effect value and logarithmic equation outperformed the other used techniques with average error of 6.32 %.

Keywords: Data pre-processing; factor analysis; linear equation; exponential equation and logarithmic equation.

Introduction:

Data mining is the process of analysing large data sets to identify patterns and establish relationships to solve problems through data analysis. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use and to predict future trends. Data mining techniques are used in research areas, mathematics, cybernetics, genetics and marketing. While data mining techniques are a means to achieve efficiencies and predict behaviours, in general, the benefits of data mining come from the ability to uncover hidden patterns and relationships in data that can be used to make predictions on a large data set.

Data mining information on different types of data can be used to build prediction models for future purposes. Before we can use different type's data mining algorithms, a target data set must be assembled. The target data set must be large enough so that, the data must contain uncover patterns. Before we apply certain algorithms, the data should be preprocessed. Pre-processing is essential task to analyse the multivariate data sets before applying data mining techniques. In this phase we handles out-of-range values, impossible data combinations and missing values. Thus, the representation and quality of data is achieved before running an analysis. If there exits irrelevant and redundant information or any type of noisy and unreliable data then data cleaning process is applied on the data set.

An efficient prediction system can be designed by applying different types of statistical methods on different data sets. An approach has been made in this paper by using different types of statistical methods and factor analysis. We choose iris data set from the uci [11] repository and then we design our prediction system.

1.1 Motivation of work

The main motivation of the given work is to design a prediction or decision making system with minimum errors which can give prediction for an unknown data set with minimum error. Initially, the total effect has been formed by factor analysis the concept of different types of statistical methods has been applied on the total effect. And we select the model with minimum error.

1.2 Literature survey

Different types of research work exist on designing of prediction system and few of them are surveyed in this section.

An improved Y-means algorithm of clustering techniques is proposed by[1].The authors applied the four different types clustering techniques like K-Means, Fuzzy c mean, Mountain clustering and Subtractive clustering in Iris flower data set and analyse the result in terms of accuracy, run time, time complexity. The authors opined that improved Y-means algorithm produces better result when compared to other clustering techniques with less computation time.

A process of developing Artificial Neural network based classifier which classifies the Iris data set is presented by [2]. The author's presented a way to classify iris plants on the basis of the sepal length, sepal width, petal length, and petal width by using Neural Network (NN).

The authors of [3] comprise the performance analysis of Fuzzy c mean (FCM) clustering algorithm with Hard C Mean (HCM) algorithm on Iris flower data set in the basis of time complexity and space complexity.

Classifying the three different types of IRIS data set by using the Multi-Layer Feed Forward Neural network is presented by [4].The authors used three different types of IRIS data set of 150 instances and opined that Multi-Layer Feed Forward Neural Network (MLFF) is faster in terms of learning speed and gave a good accuracy than the existing methods.

A Neural Network Association Classification system is presented by [5]. According to the authors this classifier is accurate and efficient. The authors also opined that the classifier build for iris plant dataset is more accurate than previous Classification Based Association.

The authors of [6] made a comparison among the algorithms Decision tree, multilayer perceptron, and Naive Bayes and Multiclass classifier on iris dataset with the following TP-rate, Fp-rate, Precision, Recall and ROC Curve parameters. According to the authors multilayer perceptron is more accurate and efficient than the other methods.

The author's [7] presented a consensus matrix using multiple runs of the clustering algorithm k-means to check whether an existing cluster can be split. The author's used the consensus matrix for clustering by using two spectral clustering methods- Fiedler Method and the min-max cut Method.

An efficient data clustering mechanism with incremental clustering algorithm with genetic feature is proposed by [8]. The authors applied the said incremental clustering on IRIS dataset for optimal clusters. The authors considered the intra cluster variances for specifying the number of clusters instead of random selection of centroids from the data and measured the fitness through the mutation based distance for finding optimal cluster.

A back propagation neural network for iris flower dataset classification is designed by [9]. The authors trained the network for 1000 epochs with different number of neurons in the hidden layer. The authors measures performance of the developed network by plotting the error versus the number of iterations and opined that trained neural network classified the testing data correctly.

The author's [10] presents a multilayer feed- forward networks which is trained using back propagation learning algorithm. The authors used 500 to 5000 no. of epochs to train the neural network and opined that the accuracy ranges from 83.33% to 96.66%. The author's concluded that Multi-Layer Feed Forward Neural Network is faster in terms of learning speed and gave a good accuracy.

2. Methodology

In this paper we choose our data set as Iris dataset. Iris data base collected from uci repository. The data set contains 3 classes (Setosa, Versicolor and Verginica) of 50 instances each, where each class refers to a particular type of iris plant. Sepal length, sepal width, petal length and petal width are the four features used to classify each flower to its category. The three classes of the flower are Iris Setosa, Iris Versicolor and Iris Verginica. First we applied different data preprocessing techniques on the iris data set. Then we applied factor analysis on preprocessed data set and finds total effect value. We applied least square techniques based on linear equation, exponential equation and logarithmic equation on total effect value and analyse the performance of the above three techniques.

2.1 Data mining

Data mining is the process of analysing large data sets to identify patterns and establish relationships to solve problems through data analysis. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use and to predict future trends. Data mining techniques are used in research areas, mathematics, cybernetics, genetics and marketing. While data mining techniques are a means to achieve efficiencies and predict behaviours, in general, the benefits of data mining come from the ability to uncover hidden patterns and relationships in data that can be used to make predictions on a large data set.

2.1.1 Data preprocessing

Data pre-processing includes the following different techniques:

1. Data cleaning: Data cleaning is the process of finding and correcting corrupt or inaccurate data from the data source. In this process we fill the missing values in a data set and noisy data are smoothed.

2. Data integration: It is the process of integration of multiple databases, data cubes, or files in a single one.

3. Data transformation: This technique includes binning, regression, and clustering.

4. Data reduction: This technique is used to reduce of large volume of data to the meaningful one.

5. Data discretisation: This process is importance especially for numerical data.

2.2 Factor analysis

Factor analysis (Coster De, 1998; Wikipedia 2012) is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, that variations in fewer observed variables mainly reflect the variations in total effect. Factor analysis originated in psychometrics, and is used in behavioural sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data.

2.3 Statistical method

2.3.1 The least squares regression line

Linear regression finds the straight line, called the least squares regression line or LSRL that represents best observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population regression line is:

$Y = B_0 + B_1X$

Where, $B_0$is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

For a given a random sample of observations, the population regression line is estimated by:

$y = b_0 + b_1x$

Whereb_0 is a constant, $b_1$is the regression coefficient, x is the value of the independent variable, and y is the predicted value of the dependent variable.

2.3.1.1 Least square techniques based on linear equation

Let the equation be

$y = a + bx$

i.e

$\Sigma y = \Sigma a + \Sigma bx$

$\Sigma y = na + \Sigma bx$

(where n = number of terms)

Again

$y = a + bx$

$xy = ax + bx^2$(Multiplied both sides by x)

so, $\Sigma xy = \Sigma ax + \Sigma bx^2$

if x in chosen in such a way that $\Sigma x = 0$ then, $a = \Sigma y/n$ and $b = \Sigma xy/\Sigma x^2$

Putting the values of a and b in equation

$y = a + bx$ the equation of straight line becomes,

$y = \Sigma y/n + x\Sigma xy/\Sigma x^2$

For different values of x, the different values of y have been calculated.

The famous and unanimously least square based techniques linear equation; exponential equation and logarithmic equation have been applied on the data.

2.3.1.2 Least square technique based on exponential equation

Let the exponential function be of the form $y = A B^x$

Take the logarithm on both sides in above equation.

$\text{Logy} = \log(AB^x)$ or $\log y = \log(A) + x \log(B)$

Let us assume,

$a = \log(A)$, $b = \log(B)$ and $y = \log(y)$.

Thus the above equation is $y = a + bx$, which a linear equation. The values a and can be calculated as per least square techniques based on linear equation. Using the values of a and b, the values of A and B have been calculated as:

A = antilog a, B = antilog b .

If the values of A and B are put in the above equation, the different values of y can be obtained for different values of x.

2.3.1.3 Least square technique based on logarithmic equation

The logarithmic curve can be written either

$P = e^{a+bX} / (1 + e^{a+bx})$ or $P = 1 / (1 + e^{a+bx})$

where P is the probability of a, e is the base of the natural logarithm (about 2.718) and a and b are the parameters of the model. The value of a yields P when X is zero because the relation between X and P is nonlinear, b does not have a straightforward interpretation in this model as it does in ordinary linear regression. The graph and concept of logarithmic equation has been furnished as follows:

Let the given functionis of the form

$w = a/(1 + brx)$ or $w(1 + brx) = a$ or $wbrx = a - w$ or $a - w = wb\, ex(\log r)$

Taking log

$\log w + \log b + x \log r = \log(a{-}w)$

or, $\log(a - w) - \log w = \text{lob } b + x \log r = c2 + c1x$, where a is chosen in such a way $(a - w)$ is not negative.

Using least squares

$c2 = \Sigma(\log(a - w) - \log(w))/n$, $c1 = \Sigma(x\{\log(a - w) - \log w\})/\Sigma x2$.

Using the value of $c1$ and $c2$, the value of b and r can be calculated as follows:

$b = \text{antilog}(c2)$, $r = \text{antilog}(c1)$

If the values of *a*, *b* and *r*, are put in the above equation, for different values of *x*, different values of *w* will be produced.

3. Implementation

Step 1:

In this paper we choose our data set as Iris data set. First we applied different data preprocessing techniques on the iris data set. Then we applied factor analysis on preprocessed data set and finds total effect value which is depicted in the table 1.

Table 1
Factor analysis result

Sl. No.	A	B	C	D	Total Effect value
1	4500	2300	1300	300	6778.35
2	4300	3000	1100	100	7035.17
3	4400	2900	1400	200	7264.46
4	4400	3000	1300	200	7302.32
5	4400	3200	1300	200	7494.95
6	4800	3000	1400	100	7609.65
7	4600	3100	1500	200	7675.19
8	4600	3200	1400	200	7713.04
9	4700	3200	1300	200	7734.41
10	4800	3000	1400	300	7750.47
11	4900	3000	1400	200	7759.88
12	4900	3100	1500	100	7844.24
13	4900	3100	1500	100	7844.24
14	4900	3100	1500	100	7844.24
15	4600	3600	1000	200	7864.48
16	4800	3100	1600	200	7893.29
17	4700	3200	1600	200	7909.78
18	5000	3200	1200	200	7915.42
19	5000	3000	1600	200	7956.61
20	4600	3400	1400	300	7976.08
21	5000	3300	1400	200	8128.64
22	4800	3400	1600	200	8182.23

Step 2

In this step we applied least square techniques based on linear equation, exponential equation and logarithmic equation on the total effect value got from the table 1 and result is given in the following tables 2, 3 and 4 respectively.

Table 2
Least square techniques based on linear equation

Total effect value	*x*	*Y=a+x*b*	*Error %*	*Avg. error%*
6778	1	10623	57	14.42
7035	2	10600	51	
7264	3	10576	46	
7302	4	10553	45	
7495	5	10530	40	
7610	6	10506	38	
7675	7	10483	37	
7713	8	10459	36	
7734	9	10436	35	
7750	10	10412	34	
7760	11	10389	34	
7844	12	10366	32	
11498	-60	12053	5	
11549	-59	12030	4	
11549	-58	12007	4	
11630	-57	11983	3	
11633	-56	11960	3	
11666	-55	11936	2	

Table 3
Least square techniques based on exponential equation

Total effect value	*y =a+x*b*	*y=A *pow(B,x)*	*Error %*	*Avg error %*
6778	10646	4.2E+10	56.73	14.4
7035	10646	3.98E+10	50.67	
7264	10646	3.77E+10	45.59	
7302	10646	3.57E+10	44.52	
7495	10646	3.39E+10	40.49	
7610	10646	3.21E+10	38.06	
9342	10645	2.83E+09	1.1709	
9394	10645	2.68E+09	0.3552	
9414	10645	2.54E+09	0.1087	
9568	10645	2.4E+09	1.9609	
9697	10645	2.28E+09	3.5055	
9804	10645	2.16E+09	4.7927	

Table 4
Least square techniques based on logarithmic equation

Total effect value	*log(a- y)- log(y)*	*1+b r^x*	*W=a/ (1+br ^x)*	*Error %*	*Avg. error %*
6778	0.13	2.41	6628	2.22	6.32
7035	0.11	2.39	6695	4.83	
7264	0.08	2.37	6764	6.9	
7302	0.08	2.34	6832	6.44	
7495	0.05	2.32	6900	7.94	
7610	0.04	2.3	6969	8.42	
7675	0.04	2.27	7037	8.31	
7713	0.03	2.25	7106	7.87	
7734	0.03	2.23	7175	7.23	
7750	0.03	2.21	7244	6.53	

4. Result analysis

It has been observed that linear equation; exponential equation and logarithmic equation estimated the average error of 14.42 %, 14.40 % and 6.32 % respectively. Therefore, logarithmic equation has been considered as the preferable optimizer as compared to the other used algorithms. Furthermore, the error of the logarithmic equation is still high and soft computing methods will be applied to optimize the error as the proposed work of this paper.

5. Conclusion

In this paper, an effort has been made to a compared the performance of least square techniques based on linear equation, exponential equation and logarithmic equation of statistical methods to estimate the average error for optimization of error. From the result analysis, it has been observed that logarithmic equation outperformed the linear equation and exponential equation with average error of 14.42 %, 14.40 % and 6.32 % respectively. Therefore, logarithmic equation has been considered as the preferable optimizer as compared to the other used algorithms.

References:

[1] V. Leela, K. Sakthipriya and R. Manikanda “Comparative Study of Clustering Techniques in Iris Data Sets” World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques): 24-29, 2014, ISSN 1818-4952.

.

[2]Shrikant Vyas and Dipti Upadhyay “Identification of Iris Plant Using FeedForward Neural Network On The Basis OfFloral Dimensions” International Journal of Innovative Research in Science,Engineering and Technology, ISSN: 2319-8753Vol. 3, Issue 12, December 2014, DOI: 10.15680/IJIRSET.2014.0312062

[3] Deepa Bhargava, VandanaMohindru and Manish Maan“Analysis of Clustering Algorithms on UCI Repository data set”International Journal of advanced Research in Computer Science” Volume-2,No-2, March-April 2011, ISSN No. 0976-5697.

[4]Madhusmita Swain, Sanjit Kumar Dash, Sweta Dash3 and AyeskantaMohapatra “An Approach For Iris Plant Classification Using Neural Network” International Journal on Soft Computing (IJSC) Vol.3, No.1, February 2012, DOI: 10.5121/ijsc.2012.310779.

[5] Ms.Prachitee Shekhawat, Prof. Sheetal S. Dhande “Building an Iris Plant Data Classifier Using Neural Network Associative Classification “International Journal of Advancements in Technology http://ijict.org/ ISSN 0976-4860, Vol. 2 No. 4 (October 2011)

[6] Kanu Patel, Jay Vala, Jaymit Pandya “Comparison of various classification algorithms on iris datasets using WEKA” International journal of Advance Engineering and Research Development (IJAERD) Volume 1 Issue 1, February 2014, ISSN: 2348 - 4470 @IJAERD-2014

[7] David Benson-Putnins, Margaret Bonfardin, Meagan E. Magnoni, AndDaniel Martin“Spectral Clustering And Visualization: A Novel Clustering of Fisher’s Iris Data Set”

[8] A Bhaskara Srinivas, B Vishnu Vardhan, L Ravi Kumar and Dr. J Rajendra Prasad
“An Efficient Data Clustering Algorithm over IRIS Dataset” International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 10, October 2013 ISSN: 2277 128X.

[9] R. A. Abdul Kadir, Khalipha A. Imam, M.B. Jibril “Simulation of Back Propagation Neural Network for Iris Flower Classification” American Journal of Engineering Research (AJER) 2017 American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-6, Issue-1, pp-200-205

[10] Madhusmita Swain, Sanjit Kumar Dash, Sweta Dash and Ayeskanta Mohapatra“An Approach For Iris Plant Classification Using Neural Network” International Journal on Soft Computing (IJSC) Vol.3, No.1, February 2012, DOI: 10.5121/ijsc.2012.3107 79.

[11]http://archive.ics.uci.edu/ml/datasets/Iris.