RESEARCH PAPER | Mathematics | Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

Multiple Regression Model and Similarity Analysis – A Comparison Study

KEYWORDS: Data Mining; Cosine Similarity; MLR

Mr. N.Senthil Vel Murugan
Department of Mathematics, Rohini College of Engineering and Technology, Kanyakumari

Dr. V.Vallinayagam
Department of Mathematics, St.Joseph's College of Engineering, Chennai

Dr. K. Senthamarai Kannan
Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli.

ABSTRACT *Data mining is an important role of biological research. Diabetes, obesity and High Cholesterol in the blood levels are closely related. These are considered as diseases which slowly threatens the human race with its presence particularly in a country like India. Based on the variables the risk of diabetes can be diagnosed. The aim of this paper is to analyze the Diabetics data and how the variables will affect quickly as possible using Singular Value Decomposition and Multiple Linear Regression model. The reasonable results verify the validity of our method.*

INTRODUCTION

Data mining is the discovery of useful knowledge from databases, Fayyad (1996). The steps of KDD process are collection, selection, transformation, visualization and evaluation of the extracted knowledge. Depending on the nature of the data as well as the desired knowledge there is a large number of algorithms for each task. All these algorithms try to fit a model to the data (Dunham, 2002).

Mathematical graphs and matrices have been successfully utilized in representing, characterizing, and analyzing biological sequences. A rigorous approach to gene expression analysis must involve an up-front characterization of the structure of the data. In addition to a broader utility in analysis methods, singular value decomposition and principal component analysis can be valuable tools in obtaining such a characterization, Wall Michael E (2001).

The Singular Value Decomposition (SVD) is one of the most important matrix decompositions used in computer vision. For any given matrix $A \in R^{m \times n}$ there exists decomposition $A = USV^T$ such that ***U*** is an ***m × n*** matrix with orthogonal columns, ***D*** is a ***n × n*** diagonal matrix with non-negative entries and V^T is an n ***× n*** orthogonal matrix. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regressions (MLR) is to model the relationship between the explanatory and response variables. The model for MLR, given n observations, is: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + e_i$ where i = 1,2, ..., n.

DESCRIPTION OF MODEL

Singular value decomposition and Principal Component Analysis are common techniques for analysis of multivariate data. In 1965 G. Golub and W. Kahan introduced Singular Value Decomposition (SVD) as a Decomposition technique for calculating the singular values and Pseudo-inverse of a matrix.

The Singular value decomposition closely associated to the companion theory of diagonaling a symmetric matrix. Hark back that if A is a symmetric real n×n matrix there is an orthogonal matrix V and a diagonal D such that

$A = VDV^T$ … (1)

Here the columns of V are latent vectors for A and diagonal entries of D are eigen values of A for Singular Value Decomposition begin with m×n real matrix. There are orthogonal matrices U and V and a diagonal matrix S, such that

$A = USV^T$ … (2)

Here U is m×m and V is n×n, so that S is rectangular with the same dimensions as A. The matrix S can be formatted to be non negative and in order of decreasing order. The columns of U and V are called left and right singular vectors for A, I.J. Good (1969).

RESULTS AND DISCUSSION

The data for the present study, i.e. the secondary data is collected from Hospitals and the diagnoses in the medical application area of diabetes are presented. This data consists of 383 samples and seven variables. Out of these variables we have consider six independent variables and one dependent variable. The dependent factor is age and the independent variables are Cholesterol, Stabilized Glucose, BMI, HDL, Glyhb and Bp levels. It is used in Latent Semantic Indexing to determine the rank of the Age and the independent factors. Before scoring the diabetics data with Latent Semantic Indexing we need to construct a matrix with the maternal variables available as "A".

TABLE - 1
DIABETIC FACTORS

Age	Chol. > 240	Stab.glu >126	BMI >30	HDL <50	Glyhb >6.5	BP >139
<25	1	0	9	15	0	3
25-35	9	3	24	38	3	16
35 - 45	17	11	17	66	7	32
45 - 55	17	13	9	34	14	40
>55	33	33	29	78	43	82

Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

The "Age, Y" is taken as the dependent variable and other variables are treated as independent variables X_1, X_2, X_3, X_4, X_5, X_6, where

X_1	:	Cholesterol level
X_2	:	Stabilized glucose level
X_3	:	Body Mass Index
X_4	:	High Density Lippo protein level
X_5	:	Glycholated Hemoglobin
X_6	:	Blood Pressure level

According to singular value decomposition theory, an arbitrarily real square matrix $n \times n$ can be decomposed into three matrices such that

$$A_{m\times n} = U_{m\times n} \; S_{m\times n} \; V^T{}_{m\times n} \quad \ldots(3)$$

Where U and V are orthogonal matrices and S is a singular matrix with eigen values as its diagonal entries, which are arranged in non-increasing order. The following analysis is done using MATLAB.

$$A = \begin{bmatrix} 1 & 0 & 9 & 15 & 0 & 3 \\ 9 & 3 & 24 & 38 & 3 & 16 \\ 17 & 1 & 17 & 66 & 7 & 32 \\ 17 & 13 & 9 & 34 & 14 & 40 \\ 33 & 33 & 29 & 78 & 43 & 82 \end{bmatrix}$$

$$U = \begin{bmatrix} -0.08 & 0.29 & -0.24 & -0.11 & -0.92 \\ -0.26 & 0.54 & -0.64 & 0.35 & 0.32 \\ -0.44 & 0.61 & 0.62 & -0.22 & 0.09 \\ -0.34 & -0.22 & 0.30 & 0.84 & -0.21 \\ -0.78 & -0.46 & -0.24 & -0.35 & 0.03 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.25 & -0.09 & 0.12 & 0.33 & 0.87 \\ -0.21 & -0.27 & 0.06 & -0.34 & 0.26 \\ -0.24 & 0.30 & -0.89 & 0.18 & 0.06 \\ -0.67 & 0.61 & 0.31 & -0.27 & -0.08 \\ -0.25 & -0.47 & -0.30 & -0.64 & 0.04 \\ -0.57 & -0.49 & 0.07 & 0.51 & -0.41 \end{bmatrix}$$

$$S = \begin{bmatrix} 168.8 & 0 & 0 & 0 & 0 \\ 0 & 35.73 & 0 & 0 & 0 \\ 0 & 0 & 12.84 & 0 & 0 \\ 0 & 0 & 0 & 5.96 & 0 \\ 0 & 0 & 0 & 0 & 1.16 \end{bmatrix}$$

Dimensionality Reduction:

Computing U_k, S_k, V_k from U, S, V using MATLAB. Let us take the economic dimension K = 2, that is rank 2 approximation that means the first 2 columns of U and V and the first two rows and columns of S.

$$U_k = \begin{bmatrix} -0.08 & 0.29 \\ -0.26 & 0.524 \\ -0.44 & 0.61 \\ -0.34 & -0.22 \\ -0.78 & -0.46 \end{bmatrix} \quad S_k = \begin{bmatrix} 168.8 & 0 \\ 0 & 35.73 \end{bmatrix}$$

$$V_k = \begin{bmatrix} -0.25 & -0.09 \\ -0.21 & -0.27 \\ -0.24 & 0.30 \\ -0.67 & 0.61 \\ -0.25 & -0.47 \\ -0.57 & -0.49 \end{bmatrix} \quad S_k^{-1} = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.03 \end{bmatrix}$$

Dimensionality reduction has been done by choosing the rank to be 2 i.e. K = 2 is applied.

Using $X_i = X_i^T \; U_k \; S_k^{-1}$ ...(4)

We get the new coordinates of vectors in this reduced space the new set of coordinate vectors are given below:

X_1 = [-0.2453 -0.0952]

X_2 = [-0.2119 -0.2718]

X_3 = [-0.2376 0.2972]

X_4 = [-0.6664 0.6093]

X_5 = [-0.2497 -0.475]

X_6 = [-0.5688 -0.4895]

and Y = [-2.1718 -0.4249]

Vector determination using cosine similarity values, we rank results in decreasing order. The Cosine Similarity of two vectors (d1 and d2) is

defined as:

$$\cos(d_1, d_2) = Sim(d_1, d_2) = \frac{d_1 . d_2}{\|d_1\| \, \|d_2\|} \quad \ldots (5)$$

Hence,

Sim (Y, X_1) = 1.96 Sim (Y, X_2) = 1.97

Sim (Y, X_3) = 1.61 Sim (Y, X_4) = 1.83

Sim (Y, X_5) = 1.93 Sim (Y, X_6) = 2.07

These data are showed in the Fig. 1.

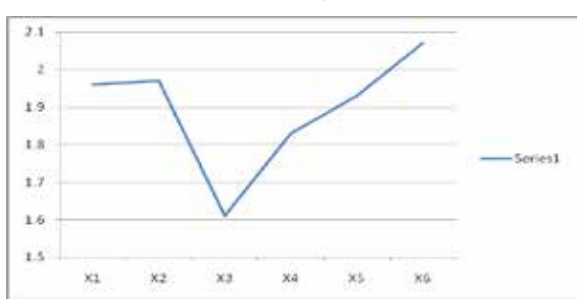

Fig.1 : Cosine similarity value

Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. So we consider only positive cases. From the above, it reveals that $X_6 > X_2 > X_1 > X_5 > X_4 > X_3$.

That means the value may be interpreted as the proportion of variability in Y that is explained by X_1, X_2, X_3, X_4 , X_5 , X_6. It reveals that Diabetics is very closed related to Blood Pressure and cholesterol level. The Proportion of Variability in Y can be arrived using the Cosine similarity value. It is seen that affected Diabetics are closely related.

A comparison study is done between Singular value decomposition and multiple linear regression model to analyze the relationship between Age and Diabetics related variable using the linear regression model of the form "Age, Y " and other variables are treated as independent variables.

Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

TABLE - 2
REGRESSION COEFFICIENT

Predicator	Coefficient	t	Sig.
(Constant)	1.788	4.319	.000
cholesterol	.285	3.611	.000
stab.glu	.030	.414	.679
BMI	-.152	-1.091	.276
HDL	-.088	-1.032	.303
Glyhb	.301	4.326	.000
BP	.548	6.847	.000
a. Dependent Variable: Age			

From the above table, it is observed that the value of R^2 is 0.264. it includes the diabetic factors Age, Cholesterol, Stabilized Glucose, BMI, HDL, Glyhb and Bp levels. It seems to 26.4 % of the variation age is explained by the fitted model. The remaining 73.6 % of variation can be explained by the factors other than these variables like socio-economic factors.

Analyzing has been done for each Diabetic factor with the age as shown below:

TABLE 3
R^2 VALUES OF ALL INDEPENDENT VARIABLES

Dependent variable	Independent variable	R^2 value
Y	X_1	0.257
Y	X_2	0.192
Y	X_3	0.058
Y	X_4	0.086
Y	X_5	0.351
Y	X_6	0.397

From table 3, the Blood Pressure level is very close to Age as well as diabetic factors then Glycholoated Hemoglobin level and Cholesterol. These data are showed in the Fig. 2.

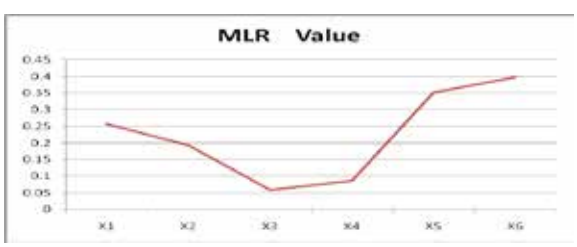

Fig. 2 : Multiple Linear Regression R^2 Value

The comparison value of Cosine Similarity and Multiple Linear Regression is as follows:

TABLE 4
COMPARISON STUDY

Variable related to Y	Cosine Similarity	MLR R^2 value
X_1	1.96	0.257
X_2	1.97	0.192
X_3	1.61	0.058
X_4	1.83	0.086
X_5	1.93	0.351
X_6	2.07	0.397

From the table 4, the blood pressure level is much related to age. The data are showed in the Fig. 3

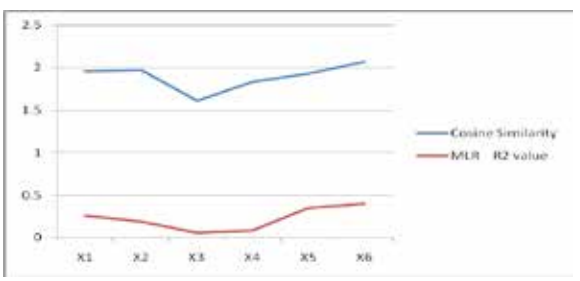

Fig. 3 : Comparison of Similarity and MLR

CONCLUSION

Data mining is a research area that aims to provide the analysts with novel and efficient computational tools to overcome the obstacles and constraints posed by the traditional statistical methods. Feature selection, normalization, and standardization of the data, visualization of the results and evaluation of the produced knowledge are equally important steps in the knowledge discovery process. The recent technological advances, have led to an exponential growth of biological data. New questions on these data have been generated. Scientists often have to use exploratory methods instead of confirming already suspected hypotheses.

From this study, the blood pressure level is much related to age then Glycholoated Hemoglobin level and Cholesterol level. Obesity once associated with beauty has now become an alarming factor to health as Cholesterol, Blood Pressure and Heart Attack are closely associated. The main impact of this work is to create awareness among patients and public by conducting this type of survey.

REFERENCE 1. Dunham, M.H. (2002). Data mining: Introductory and advanced topics. Prentice Hall, Upper Saddle River, New Jersey, USA | 2. Durbin, R., Eddy, S., Krogh, A., et al. (1998), "Biological Sequence Analysis", Cambridge University Press, Cambridge, UK. | 3. Fayad, U.M et al. (1996): Advances in knowledge discovery and data mining, AAAI/MIT press. | 4. Golub et. al (1965), "Calculating the singular values and Pseudo-Inverse of a Matrix, Numer. Anal. Ser. B, p.205-224, Vol. 2, No.2. | 5. Good, I.J (1969): "Some applications of the Singular value decomposition of matrix", Technometrics, vol.11, no.4, pp.823-831. | 6. Wall Michael E. et.al (2001), "Singular value decomposition analysis of Microarray data", Bioinformatics, vol. 17, pp.566 – 568. | 7. Zakaria Suliman Zubi, Marim Aboajela Emsaed (2010), "Using sequence DNA chips data to Mining and Diagnosing Cancer Patients", International Journal of Computers, Issue4, Vol.4, pg.201-214. |