Decision Making System using Machine Learning and Pearson for Heart Attack

Chandrasegar Thirumalai, IEEE Member,
School of Information Technology and Engineering,
VIT University, Vellore, India.
chandru01@gmail.com

Anudeep Duba
School of Information Technology and Engineering
VIT University,
Vellore, India.
anudeep58@gmail.com

Rajasekhar Reddy
School of Information Technology and Engineering,
VIT University,
Vellore, India.
rajashekar.singareddy@gmail.com

***Abstract*— This informational collection is utilized to anticipate the odds of an event of heart assault for a patient. In the season of cutting edge smartphones contributing 12 attributes is not feasible. We play out the product metric examination on the given informational collection. In view of the investigation of information we try to bring the total number of attributes into a small figure and in the end, we may be able to choose which property can be considered and which characteristic can be disregarded.**

Keywords- Regression, Pearson, Attributes, Boxplot

I. INTRODUCTION

The main goal is to bring down the number of attributes by establishing and quantifying the relationships between attributes. By dropping the number of attributes many rigorous calculations can be made easily with the help of very few attributes. Initially, box-plot and control chart techniques are used to remove unwanted samples from the information collection by pointing outliers which play fewer roles in any calculations with regard to the dataset. Then Pearson model [1], [5], [9], [11], [17], [19] is utilized to discover the relationship between the characteristics in view of the estimation of r. Pearson helps us to discover how intently a trait is related with different qualities. In the end, Linear regression model [21], [22] is for demonstrating the relationship between a scalar dependent variable y and at least one systematic factors meant by x. We have a dataset of almost 300 patients portraying their heart condition with the assistance of 12 characteristics. From the data analysis exploration [4], [7] we can choose which quality can be considered and which property can be disregarded.

II. Boxplot

A box plot is a graphical interpretation of measurable information in view of the base, first quartile, middle, third quartile, and greatest. Box plot lets us find the median of the data and also gives the median (lower tail) of lower quartile and the median (upper tail) of upper quartile. This helps us in pointing out the outliers which play a minimum role in calculations and can be neglected. Outliers eliminate the irrelevant samples in our sample data set. We shall choose an initial or a root attribute which has a considerable amount of outliers. In our scenario we choose age has root attribute and continue are a process. We can also apply control chart on the collection as a secondary measure in order to make sure that the outliers found in Boxplot match with the outliers found with control chart.

TABLE I. SAMPLES OF HEART-RELATED DATA – PART 1

S.No	*Age*	*Gender*	*Chest Pain*	*Blood Pressure*	*Cholesterol*
1	63	1	1	145	233
2	67	1	4	160	286
3	67	1	4	120	229
4	37	1	3	130	250
5	41	0	2	130	204
6	56	1	2	120	236
7	62	0	4	140	268
8	57	0	4	120	354
9	63	1	4	130	254
10	53	1	4	140	203
11	57	1	4	140	192
12	56	0	2	140	294
13	56	1	3	130	256
14	44	1	2	120	263
15	52	1	3	172	199
16	57	1	3	150	168

TABLE II. SAMPLES OF HEART-RELATED DATA – PART 2

FBS	*ECG*	*MHR*	*Peak BP*	*Slope BP*	*CA*	*THAL*
1	2	150	1	3	0	6
0	2	108	2	2	3	3
0	2	129	2	2	2	7
0	0	187	1	3	0	3
0	2	172	1	1	0	3
0	0	178	1	1	0	3
0	2	160	1	3	2	3
0	0	163	2	1	0	3
0	2	147	1	2	1	7
1	2	155	2	3	0	7
0	0	148	1	2	0	6
0	2	153	1	2	0	3
1	2	142	2	2	1	6
0	0	173	1	1	0	7
1	0	162	1	1	0	7
0	0	174	1	1	0	3
0	0	168	1	3	0	7
0	0	160	1	1	0	3
0	0	139	1	1	0	3
0	0	171	1	1	0	3
0	2	144	2	2	0	3
1	2	162	1	1	0	3

FBS – Fasting Blood Sugar, ECG – Electro Cardio Gram, MHR – Maximum Heart Rate, CA – Number of major vessels, Thal - Thalassemia.

The sensitive information of heart patient can be shared by the doctor and scan center using suitable methods [14], [16], [18], [20], [23], [24], [26]. Here we took 16 samples from the given information population and applied Box plot analysis and the same is plotted in the following Fig 1.

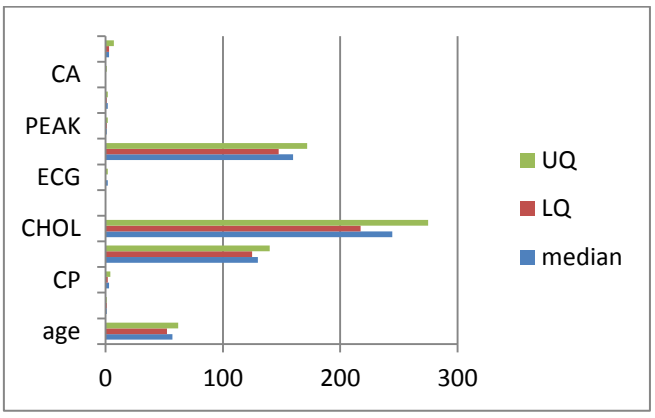

Figure 1. Box plot analysis of given data

TABLE III. BOX PLOT INFORMATION OF GIVEN DATA.

Attribute	*Median*	*LQ*	*UQ*	*Outliers*
Age	57	52.5	62	63,67,67,37,41,44,48,48,64, 49,52 – 11
Gender	1	1	1	0,0,0,0 --4
CP	3	2	4	1,1,1 --3
BPS	130	125	140	145,160,120,120,172,150,1 10,110,150,132,120,130 -- 12
CHOL	244.5	217.5	275	286,354,203,192,294,199,1 68,211,283,284,206 -- 11
FBS	0	0	0	1,1,1,1,1 -- 5
ECG	2	0	2	0
MHR	160	147.5	172	108,129,187,178,147,142,1 73,139,173,132 -- 10
PEAK	1	1	2	0
SLOPE	2	1	2	3,3,3,3,3 -- 5
CA	0	0	1	3,2,2,2,2 -- 5
THAL	3	3	7	0

We find out that age has 11 outliers and since we are choosing age as our root attribute we delete the corresponding 11 rows from all the attributes. Now our 16 samples will drop down to 6 samples, but we still couldn't reduce the number of attributes, so now we try to establish relationships between the attributes with a view to calculate one attribute value from other. These relationships are identified using Pearson

III PEARSON CORRELATION

The most well-known measure of correlation in statistics is the Pearson Relationship. The Pearson product-moment correlation coefficient (Pearson's correlation, for short) is a measure of the quality and heading of affiliation that exists between two factors measured on no less than an interval scale. The more grounded the relationship of the two factors, the nearer the Pearson correlation coefficient, r, will be to either +1 or - 1 relying upon whether the relationship is positive or negative, individually. Accomplishing an estimation of +1 or - 1 implies that every one of our data points is incorporated hanging in the balance of best fit.
Some of the former methods to forecast the decisions [25], [27], [28] based on their association of strength are Spearman [6], Analytical Hierarchical Process (AHP) [2], [8], [10].

A. Pearson based Attribute Clustering

1. Create a table between the different characteristics as Appeared in IV-A
2. Complete the table utilizing multiplication of attribute values.
3. Calculate the aggregate of every attribute independently.

4. Substitute every one of the traits in the equation offered underneath to get the Pearson coefficient.

$$r = \frac{N.\sum xy - \sum x.\sum y}{\sqrt{(N.\sum x^2 - (\sum x)^2).(N.\sum y^2 - (\sum y)^2)}}$$

In this case, we considered every relation with r under 0.8 as week relations and the related attributes are overlooked relations with r greater than 0.8 are taken as strong relations. This is our first step in decreasing the number of attributes.

TABLE IV. PEARSON ATTRIBUTE ANALYSIS – PART 1

	Age	*Gender*	*CP*	*BPS*	*CHOL*
Age	1.00	0.66	0.88	0.99	0.95
Gender	0.66	1.00	0.68	0.66	0.48
CP	0.88	0.68	1.00	0.88	0.81
BPS	0.99	0.66	0.88	1.00	0.93
CHOL	0.95	0.48	0.81	0.93	1.00
FBS	0.30	0.31	0.45	0.62	0.41
ECG	0.63	0.31	0.45	0.62	0.62
MHR	0.99	0.67	0.86	0.98	0.29
PEAK	0.86	0.62	0.87	0.84	0.84
SLOPE	0.83	0.54	0.82	0.82	0.77
CA	0.41	0.27	0.46	0.35	0.31
THAL	0.83	0.78	0.85	0.83	0.71

The remaining details of Pearson relation coefficient are given in the next Table (Table V).

TABLE V. PEARSON ATTRIBUTE ANALYSIS - PART 2

	FBS	*ECG*	*MHR*	*PEAK*	*SLOPE*	*CA*	*THA*
Age	0.30	0.63	0.99	0.86	0.83	0.41	0.83
Gender	0.20	0.31	0.67	0.62	0.54	0.27	0.78
CP	0.18	0.45	0.86	0.87	0.82	0.46	0.85
BPS	0.35	0.62	0.98	0.84	0.82	0.35	0.83
CHOL	0.41	0.62	0.95	0.84	0.77	0.31	0.71
FBS	1.00	0.52	0.29	0.48	0.39	0.00	0.42
ECG	0.52	1.00	0.57	0.61	0.75	0.60	0.67
MHR	0.57	1.00	1.00	0.83	0.78	0.33	0.79
PEAK	0.48	0.61	0.83	1.00	0.79	0.40	0.87
SLOPE	0.39	0.75	0.78	0.79	1.00	0.45	0.79
CA	0.00	0.60	0.33	0.40	0.45	1.00	0.55
THAL	0.42	0.67	0.79	0.87	0.79	0.55	1.00

The relationships between all the corresponding attributes are quantified and now we can clearly differentiate week relations from strong ones. As we eliminate weak relations we must be able to produce one attribute from the other strongly related attribute. This can be easily done by linear regression method.

IV. LINEAR REGRESSION METHOD

Linear regression endeavors to demonstrate the connection between two factors by fitting a linear equation to watched information. One variable is thought to be an informative variable, and the other is thought to be a reliant variable. In statistics, linear regression is a strategy for evaluating the restrictive expected estimation of one variable y given the estimations of some other variable or variables x. A variable is, by definition, an amount that may fluctuate starting with one estimation then onto the next in circumstances where diverse examples are taken from a populace or perceptions are made at various focuses in time. In fitting factual models in which a few factors are utilized to anticipate others, what we want to discover is that the distinctive factors don't differ freely (in a measurable sense), however, that they have a tendency to shift together.

1. Make a relation among the attributes as in Table VI.
2. Fill in the estimations of x and y
3. Fill the table with $h_\theta(x^i)$ values
4. Fill the table with $h_\theta(x^i) - (y^i)$ values
5. Fill the table with $(h_\theta(x^i) - (y^i))^2$ values
6. Supplant the total values in the accompanying structure, to get the enhanced cost of θ.

$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$

The values are filled in only for strongly related attributes as week relations cannot be quantified.

TABLE VI. OPTIMIZED COST WITH ITS Θ VALUE

	Age	*Sex*	*CP*	*BPS*	*CHOL*
Age	1		5.23,0	274.76,0	9935.19,2
Sex		1			
CP	1301.07,2		1	8324,0	30127.5,0.5
BPS	68.69,0.5			1	2034.11,2
CHOL	1631.32,0			508.5,0.5	1
FBS					
ECG					
MHR	1074.62,2		11686.15,2	408.46,1	973.7,0.5
PEAK	0.97,0		0.2,0.5	0.96,0	0.96,0
SLOPE	1.69,0		0.3,0.5	1.69,0	
CA					
THAL	11.19,0		1.6,1.5	11.19,0	

TABLE VII. OPTIMIZED COST WITH ITS Θ VALUE - PART 2

	FBS	*ECG*	*MHR*	*PEAK*	*SLOPE*	*CA*	*THAL*
Age			1074.62,2	0.97,0	1.69,0		11.19,0
Sex							
CP			11686.15,2	0.2,0.5	0.3,0.5		1.6,1.5
BPS			408.46,1	0.96,0	1.69,0		11.19,0
CHOL			973.7,0.5	0.96,0			
FBS	1						
ECG		1					
MHR			1	0.96,0			
PEAK			0.96,0	1			2.73,2
SLOPE					1		
CA						1	
THAL				2.73,2			1

Now we can easily evaluate one attribute from other using TABLE VI and Table VII. In light of Linear Regression, we will get the enhanced θ value, and this θ value can proficiently replace the real attribute value (where r>0.8). This technique is utilized for cost improving which implies advancing the variable. This is one of the strategies utilized for accomplishing machine learning.

Now as we can clearly see how one trait can be evaluated from other for example BPS is usually the twice of CHOL value since the corresponding θ value is 2. So we can have only one of the either attributes. It goes similarly with other attributes too like age can be estimated from BPS, CHOL, MHR and much more so age also can be neglected.

In this way, we can drop the total number of attributes to a minute number. In this scenario, we can drop the attributes to 6-8 based on the past medical record of the patient.

So a new age application can predict the chances of heart attack by taking the minimalistic attributes possible from the user. And moreover based on these relationships many other characteristics of the user can be predicted and calculated from the initial inputs taken. This is a very small advancement in medical science but has a wide range of applicability

In any case, it is to be noticed that when a couple of traits is comparative in nature, then linear model creates the θ value as zero; though in Pearson r=1. From the perceptions of Table IV, it is certain that the Age versus Cholesterol credits is near each other to regress. So also, CP vs. Age, BPS, Cholesterol, and MHR are also fallen on the same cater. So we successfully regress them using the linear regression model and quantified their relationship.

IV CONCLUSION

Therefore the 12 characteristics can be dropped down. What's more, the machine has learned how attributes are identified with each other. This metric investigation helps us in the new time of Smart telephones where a phone acknowledges fewer sources of info and gives us an out of the bound outcome. In this situation, the odds of a heart assault can be resolved with only 8 or fewer qualities.

In the age of Internet of Things, this analysis is very useful in predicting the medical condition of the user within seconds just through a normal scan. And moreover in this busy world everyone is expecting spontaneous results and this is where this analysis comes handy. And particularly in a time where people want to achieve things just by using a mobile app, this analysis gives the solution to build an app which can literally deploy a detailed medical report just within seconds which usually takes a day time in a present day hospital.

REFERENCES

[1] Hauke J., Kossowski T., Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. Quaestiones Geographicae 30(2), Bogucki Wydawnictwo Naukowe, Poznań 2011, pp. 87–93, 3 figs, 1 table. DOI 10.2478/v10117-011-0021-1, ISBN 978-83-62662-62-3, ISSN 0137-477X.

[2] Piovani J.I., 2008. The historical construction of correlation as a conceptual and operative instrument for empirical research. Quality & Quantity 42: 757–777.

[3] P. Dhavachelvan, Chandra Segar T, K. Satheskumar, "Evaluation of SOA Complexity Metrics Using Weyuker's Axioms," IEEE International Advance Computing (IACC), India, pp. 2325 – 2329, March 2009

[4] Halstead Metric for Intelligence, Effort, Time predictions, DOI:10.13140/RG.2.2.17988.42881

[5] Fisher R.A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1: 3–32.

[6] Spearman C.E, 1904b. General intelligence objectively determined and measured. American Journal of Psychology 15: 201–293.

[7] Software metric Numerical Data analysis using Box plot and control chart methods, VIT University, DOI:10.13140/RG.2.2.27422.95041

[8] Vaishnavi B, Karthikeyan J, Kiran Yarrakula, Chandrasegar Thirumalai, "An Assessment Framework for Precipitation Decision Making Using AHP", International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[9] Griffith D.A., 2003. Spatial autocorrelation and spatial filtering. Springer, Berlin.

[10] Chandrasegar Thirumalai, Senthilkumar M, "An Assessment Framework of Intuitionistic Fuzzy Network for C2B Decision Making", International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2017

[11] Rodgers J.L. & Nicewander W.A., 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42 (1): 59–66.

[12] F. Fioravanti, P. Nesi, "A method and tool for assessing object-oriented projects and metrics management," Journal of Systems and Software, Volume 53, Issue 2, 31 August 2000, Pages 111-136

[13] Galton F., 1875. Statistics by intercomparison. Philosophical Magazine 49: 33–46

[14] Chandrasegar Thirumalai, Viswanathan P, "Diophantine based Asymmetric Cryptomata for Cloud Confidentiality and Blind Signature applications," JISA, Elsevier, 2017.

[15] Galton F., 1877. Typical laws of heredity. Proceedings of the Royal Institution 8: 282–301.

[16] Chandrasegar Thirumalai, Sathish Shanmugam, "Multi-key distribution scheme using Diophantine form for secure IoT communications," IEEE IPACT 2017.

[17] Galton F., 1888. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London 45: 135–145.

[18] Chandrasegar Thirumalai, Senthilkumar M, “Spanning Tree approach for Error Detection and Correction,” IJPT, Vol. 8, Issue No. 4, Dec-2016, pp. 5009-5020.

[19] Galton F., 1890. Kinship and correlation. North American Review 150: 419–431.

[20] Chandrasegar Thirumalai, Senthilkumar M, “Secured E-Mail System using Base 128 Encoding Scheme,” International journal of pharmacy and technology, Vol. 8 Issue 4, Dec. 2016, pp. 21797-21806.

[21] Yule G.U., 1897a. On the significance of Bravais’ formulae for regression, in the case of skew correlation. Proceedings of the Royal Society of London Ser. A 60: 477–489

[22] Chandramowliswaran N, Srinivasan.S and Chandra Segar.T, “A Note on Linear based Set Associative Cache address System” International J. on Computer Science and Engg. (IJCSE) & India, Engineering Journals & 0975-3397, Vol. 4 No. 08 / pp. 1383-1386 / Aug. 2012.

[23] T Chandra Segar, R Vijayaragavan, “Pell's RSA key generation and its security analysis,” in Computing, Communications and Networking Technologies (ICCCNT) 2013, pp. 1-5

[24] Chandrasegar Thirumalai, Senthilkumar M, Vaishnavi B, “Physicians Medicament using Linear Public Key Crypto System,” in International conference on Electrical, Electronics, and Optimization Techniques, ICEEOT, IEEE & 978-1-4673-9939-5, March 2016.

[25] Kalaiarassan G, Krishan, Somanadh M, Chandrasegar Thirumalai, Senthilkumar M, "One-Dimension Force Balance System for Hypersonic Vehicle an experimental and Fuzzy Prediction Approach," Elsevier, ICMMM - 2017.

[26] E Malathy, Chandra Segar Thirumalai, "Review on non-linear set associative cache design," IJPT, Dec-2016, Vol. 8, Issue No.4, pp. 5320-5330

[27] Sasikala, I., M. Ganesan, and A. John. "Uncertain data prediction on dynamic road network." Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, 2014.

[28] John, A., M. Sugumaran, and R. S. Rajesh. "Indexing And Query Processing Techniques In Spatio-Temporal Data." ICTACT Journal on Soft Computing 6.3 (2016).

Rajasekhar Reddy currently studying MS Software Engineering at VIT University, Vellore Campus, Vellore, India.

https://www.researchgate.net/profile/Rajasekhar_Singareddy

Anudeep Duba currently studying MS Software Engineering at VIT University, Vellore Campus, Vellore, India.

https://www.researchgate.net/profile/Anudeep_Duba