Journal of Computer Science

Original Research Paper

An identification Blemish in Co-Variance Study of Disease using Statistical Data Mining

Dr. Aneeshkumar, A.S.

Assistant Professor, Department of Computer Science, Alpha Arts and Science College, Porur, Chennai, India

Article history
Received: 16-04-2017
Revised: 23-04-2017
Accepted: 10-06-2017

Email: aneesh_kumar777@yahoo.com

Abstract: Data mining is an established technology for identifying relevant factors and predicting the occurrence of associated aspects for the estimation and future planning in various fields. It is about finding insights which are statistically reliable, significant and previously unknown data. Health care management is one among the major user of the Data mining techniques for diagnosing the attributes for a variety of medical issues and treatment planning. Disparity in liver enzymes is a liable physical problem which affects other proportional activities of the body. The influence of Liver disorder symptoms in diabetes patients are used here for the supportive study of the relational recognition of Liver disorder and Diabetes Mellitus.

Keywords: CHAID, Accuracy, ROC, Correlation, Significance

Introduction

Sugar is an essential part for energy expenses and metabolic needs in the body. Insulin hormone controls absorption of the sugar from eatable substances. In case of excessive consumption of calories, body is unable to create enough or cannot successfully use insulin and it will lead to Diabetes Mellitus (DM) which is a chronic metabolic condition distinct according to the level of hyperglycaemia (WHO/IDF, 2006; Harris and Zimmet, 1997). Increased level of calories in the body than the need, causes temporary sugar storage as fat and it become energy source in emergencies like fasting. But this will be harm to the health if it becomes long term storage (Schneiderman, 2004). This condition often cause other complications like heart disease, stroke, high blood pressure, liver disease, kidney disease, neuropathy and the loss of some organs in the body (Sriphaew *et al*., 2012).

World Health Organization report in 2006 says that there are more than170 million peoples suffer with diabetes in worldwide and this is going to increase up to 266 million by the year of 2030 and in India 31 million (10%) people with diabetes and It will be around 31 million in 2030 (Yoon *et al*., 2006). Type I diabetes is normally seen in younger population and Asians have accurate genetic vulnerability for type II diabetes (Yoon *et al*., 2006). All the biological changes are linked with psychological stress and lifestyle of the human being. Balanced food and proper exercise will help to reduce most of the chronic diseases. Socioeconomic variables like education, income, occupation, social cohesion, ethnicity, race and behaviour activities like regular and long term usage of alcohol and smoking and psychodynamic factors like stress and anxiety have influence in such sever diseases.

Liver is a part of the body which is also engage in glucose metabolism and energy expenditure. The collected carbohydrates from the gastrointestinal are go through hepatic processing and converted as metabolism into amino acids or fatty acids (Baig *et al*., 2001).

Data mining applications are becoming more important in the field of data warehouse and big data and it involves classification approaches to predict the categorical marker of the selected domain data. Classification problem is concerned with the discovery of classification rules for predefined classes and this is useful for correct classification of the unknown attributes in future (Au *et al*., 2003). Classification process is divided into two steps as learning and testing. In learning some percentage of the whole data is used to develop the model and the remaining data is to conduct testing with the developed replica.

Data Description: There are two set of parameters used for this study namely physicals observations and biological observations done in hospitals. Table 1 includes physiological parameters where some factors having two values and some others having more than two. Biological parameters of 9 to 22 items in Table 2 carrying three instances. These all are the attributes of Alcoholic fatty liver disorder and Non-alcoholic fatty liver disorder patients.

Table 1. Physiological parameters of the patients

Factor name	Instances of the factor
Gender (GN)	Male, Female
Age (AG)	30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69
Itching (IC)	Yes, No
Alcoholic Consumption (AC)	Yes, No
Smoking habit (SK)	Yes, No
Head Ache (HA)	Yes, No
Acting Differently (AD)	Yes, No
Obesity (OS)	Below Average, Average, Above Average

Table 2. Biological parameters of the patients

Factor name	Instances of the factor
BILIRUBIN D (BD)	Below Average, Average, Above Average
BILIRUBIN T (BT)	Below Average, Average, Above Average
S.G.O.T. (SO)	Below Average, Average, Above Average
S.G.P.T. (SP)	Below Average, Average, Above Average
GAMMA GT (GG)	Below Average, Average, Above Average
ALKALINE PHOSPHATE (AP)	Below Average, Average, Above Average
TOTAL PROTEINS (TP)	Below Average, Average, Above Average
ALBUMIN (AM)	Below Average, Average, Above Average
GLOBULINS (GB)	Below Average, Average, Above Average
Plasma Glucose-F (PF)	Below Average, Average, Above Average
Plasma Glucose-R (PR	Below Average, Average, Above Average
Blood Pressure-Diastolic (PD)	Below Average, Average, Above Average
Blood Pressure-Systolic (PS)	Below Average, Average, Above Average
Triglycerides (TG)	Below Average, Average, Above Average

Model Building

Chi-squared Automatic Interaction Detector (CHAID) in classification is one of the oldest decision tree algorithm probably used to study the relationship between dependent instances and a set of predictor factors, based on adjusted significance (http://www.mu-sigma.com/analytics/thoughtleadership/cafe-cerebral-chaid.html). It was proposed by Kass (1980) in South Africa as part of his Ph.D. theses work (Aneeshkumar and Venkateswaran, 2012; http://www.obgyn.cam.ac.uk/cam-only/statsbook/stchaid.html). CHAID is an extension of US Automatic Interaction Detection (AID) and THeta Automatic Interaction Detection (THAID).

It builds non-binary trees for large datasets with complicated patterns, which include many categorical variables with multiple classes. CHAID is more suitable for process like market segmentation in various fields and it collects predictor's relations and which predicts the optimal value of the independent attributes. This algorithm follows three steps, which are:

- Preparing predictors- It creates categorical predictors of the dataset
- Merging categories- Determine least significant pair of predictors with respect to the dependent class and merge it together
- Selecting the split variable- Split predictor parameters according to smallest adjusted p value; until it reaches a terminal node

The total data set is spitted into two categories as 80 percentages for building the model and remaining to test the generated model.

Model Evaluation

The fitness of any classification model is identified with two markers which are sensitivity and specificity for n number of predictions (Gundogan *et al*., 2004). Determination of this depends on four factors which are True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) instances of the class. Sensitivity is the probability test to determine the proportion of patients with a positive result of the particular disease and which is identified with a formula of TP/(TP+FN). Specificity is a possibility proportion of the patients who is not a member of the specified disease group and got negative result also for that cluster and it can be find with TN/(TN+FP). This indicators lead to the fitness recognition.

Fitness of the classification = sensitivity* specificity.

The values from confusion matrix show that:

- Sensitivity of AFLD = 391/(391+0) = 1
- Specificity of AFLD = 359/(359+0) = 1
- Fitness of AFLD = 1*1 =1
- Sensitivity of NAFLD-M = 179/(179+0) = 1
- Specificity of NAFLD-M = 571/(571+1) = 0.998
- Fitness of NAFLD-M = 1*0.998 = 0.998
- Sensitivity of NAFLD-F = 179/(179+0) = 1
- Specificity of NAFLD-F = 571/(571+0) = 1
- Fitness of NAFLD-F = 1*1 =1

A. ROC Curve

Receiver operating characteristic curves are effective way to show positive and false items (http://en.wikipedia.org/wiki/CHAID). These factors are used to determine a cut off value for a class result or a particular decision threshold. ROC curve always lies in a unit square and it should pass through two points 0, 0 and 1, 1. The origin 0, 0 match when there is no classifier sensitivity and it reach to 1, 1 the classifier is maximum sensitive (Swets *et al*., 2000). ROC curve is formed with 1- specificity as X axis and sensitivity as Y axis and so the curve for AFLD and NAFLD-F lies in Y axis and for NAFLD-M it starts at 0.002 of X axis and grows towards Y axis to reach at 1. According to the accuracy value or ROC curve we can say that the specified model is fit for the classification of the given data.

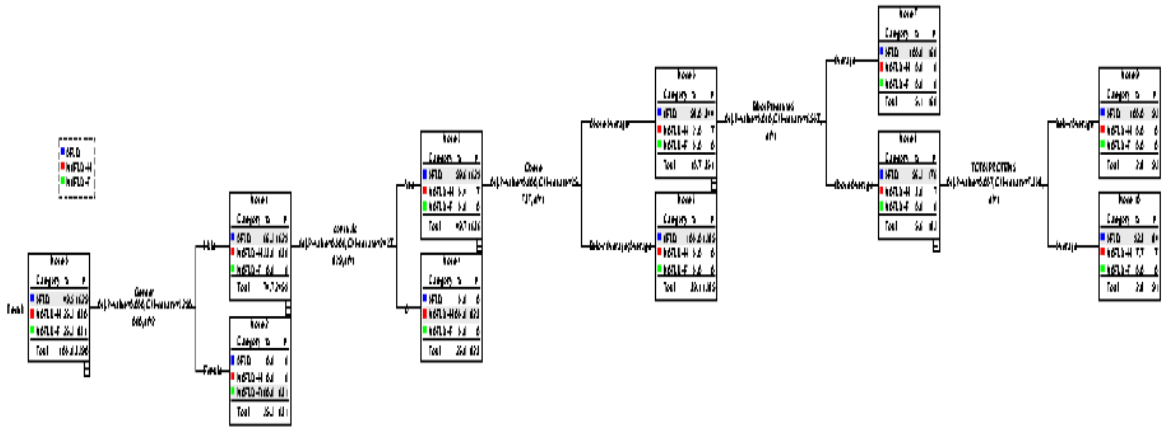

Fig 1. CHAID tree for the dataset

Table 3. Tree table representation of CHAID

Node	AFLD	NAFLD-M	NAFLD-F	Predicted category	Parent node	Attributes	Chi-Square	Split Values
0	1629	830	831	AFLD				
1	1629	830	0	AFLD	0	GN	3290.000	Male
2	0	0	831	NAFLD-F	0	GN	3290.000	Female
3	1629	7	0	AFLD	1	AC	2427.829	Yes
4	0	823	0	NAFLD-M	1	AC	2427.829	No
5	344	7	0	AFLD	3	OS	25.737	Above Average
6	1285	0	0	AFLD	3	OS	25.737	Below Average; Average
7	168	0	0	AFLD	5	P S	6.557	Average
8	176	7	0	AFLD	5	P S	6.557	Above Average
9	92	0	0	AFLD	8	T P	7.358	Below Average
10	84	7	0	AFLD	8	T P	7.358	Average

Table 4. Confusion Matrix of the classification

		Predicted		
		AFLD	NAFLD-M	NAFLD-F
Actual	AFLD	391	0	0
	NAFLD-M	1	179	0
	NAFLD-F	0	0	179

Table 5. Mean value and standard deviation of PF and PR in liver

	AFLD		NAFLD-M		NAFLD-F	
Factors	Mean	S.D.	Mean	S.D.	Mean	S.D.
PF	0.75	0.443	0.87	0.343	0.72	0.451
PR	0.80	0.402	0.78	0.412	0.87	0.335

Table 6. Correlation of PF and PR with other attributes among AFLD patients

Attribute	HPC	GPC	NPC	Z C	HNC	GNC	NNC
PF	BT, SP, AP, PD		IC, SK		OS, SO, PR	BD, TP	AG, PS
PR	OS, SO,SP, AP, TP,GB, PF,PD	PS	IC		AG, BT	SK, BD	

Table 7. Correlation of PF and PR with other attributes among NAFLD-M patients

Attribute	HPC	GPC	NPC	Z C	HNC	GNC	NNC
P F	AG, SK, HA, AD, SO, SP, TP, PR, PS	PD	IC, BT, GG		AP, AM		AC
P R	AG, SK, HA, AD, SO, TP, PF, PD	PS	IC, AC, BT		AM	GG	SP, AP

Table 8. Correlation of PF and PR with other attributes among NAFLD-F patients

Attribute	HPC	GPC	NPC	Z C	HNC	GNC	NNC
PF			AG, IC, SK, BT,AP, PD,PS			AM	AC, HA, BD, SO,SP, GG, TP, PR
PR			AC, SK, HA, BT, AP, TP, PS				AG, IC, BD, SO, SP, GG, AM, PF, PD

Note:

HPC - High Positive Correlation at one percentage of significance
GPC - Good Positive Correlation at five percentage of significance
NPC - Normal Positive Correlation with no significance
Z C - Zero Correlation means no correlation between the items
HNC - High Negative Correlation at one percentage of significance
GNC - Good Negative Correlation at five percentage of significance
NNC - Normal Negative Correlation with no significance

If A, B are correlated, it is necessarily entail that A cause B or B causes A (Aneeshkumara and Venkateswaran, 2015). For an example itching has correlation with plasma glucose F and plasma glucose R, so we can't conclude that itching is due to diabetics or diabetics is due to itching and so on. But both may linked with another parameter namely liver disorder or correlated multiple attribute have a strong relation with disease. But most of the attributes have either positive or negative correlation with Plasma Glucose F and Plasma Glucose R.

Result and Discussion

In this study, a total of 3290 dataset are classified in to three groups called Alcoholic fatty liver disorder, Non-alcoholic fatty liver disorder male and Non-alcoholic fatty liver disorder female. Figure 1 and Table 3 shows the diagrammatic and clear representation of CHAID tree with first set of independent attributes for three classes where gender, alcoholic consumption, obesity, blood pressure s and total proteins are considered as primary independent variables to build tree. The percentage of categorized data in each class and instances of attribute also present here. In 0th node, it covers hundred percentages of training data.

Table 4 of confusion matrix show that, true positive value in AFLD, NAFLD-M and NAFLD-F are 391, 179 and 179 respectively. Only one patient of NAFLD-M is considered falsely as a member of AFLD. In Table 5, the mean value for Plasma Glucose F and Plasma Glucose R among alcoholic liver disorder patients is 0.75 and 0.80 respectively. In case of Non-alcoholic fatty liver disorder male group Plasma Glucose F value is 0.87 and Plasma Glucose R is 0.78 but when it reach to on non-alcoholic fatty liver disorder female, the mean value for Plasma Glucose F is decreased to 0.72 and Plasma Glucose R increased to 0.87. It consider as a reverse exploit with respect to its previous group. The standard deviation for all the categories lies between 0.350 and 0.450.

In Table 6 and 7 most of other attributes are highly correlated with plasma glucose f and plasma glucose r with one percentage level of significance, but in Table 8 of NAFLD-F group albumin is the only factor correlated with PF at five percentage significance level and the remaining attributes having normal correlation with two parameters of diabetes. But there are no factors with zero correlation. As an overall assessment, most of the attributes are correlated either positively or negatively to PF and PR.

The facts analysed here states that nearly 80 percentage patients among three liver disorder groups are suffering from diabetes. But there is a serious variation

between in some cases between Plasma Glucose F and Plasma Glucose R, that is some of them are in the boarder of Plasma Glucose F but Plasma Glucose R is identified as normal for them and vice versa. So these peoples are having chance to either increase the value of another Plasma Glucose or a serious other disorder and which should be analysed with the help of more biomedical data and health experts advice.

Conclusion and Future Work

The result indicates that liver disease may contribute to the causes of diabetes or diabetes may contribute to the causes of liver disease. This result is often useful to make awareness in both categories of patients about another one and in future it may lead to conduct more studies with numerous data to determine which one of this having influence on another.

Acknowledgement

We express our sincere thanks to the Director Dr.(Capt.) K.J. Jayakumar M.S., M.N.A.M.S., F.A.I.S. and Chief Manager Dr. R. Rajamahendran, B.Sc., M.B.B.S., D.M.C.H., D.H.H.M., P.G.D.H.S.C.(Diab.), F.C.D., Sir Ivan Stedeford Hospital, Chennai for providing permission to collect data. We are grateful to the chief Manager for his guidance and also would like to thank other hospital staffs for their valuable suggestions throughout the study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved

References

Aneeshkumar, A.S. and C.J. Venkateswaran, 2012. Estimating the surveillance of liver disorder using classification algorithms. Int. J. Comput. Applic., 57: 39-42.

Aneeshkumara, A.S. and C.J. Venkateswaran, 2015. Reverse sequential covering algorithm for medical data mining. Proc. Comput. Sci., 47: 109-117. DOI: 10.1016/j.procs.2015.03.189

Au, W., K.C.C. Chan and X. Yao, 2003. A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. Evolut. Comput., 7: 532-545. DOI: 10.1109/TEVC.2003.819264

Baig, N., S. Herrine and R. Rubin, 2001. Liver disease and diabetes mellitus. Clin. Lab. Med., 21: 193-207. PMID: 11321935

Gundogan, K.K., B. Alatas and A. Karci, 2004. Mining classification rules by using genetic algorithms with non-random initial population and uniform operator. Turk. J. Elec. Eng., 12: 43-52.

Harris, M. and P. Zimmet, 1997. Classification of Diabetes Mellitus and other Categories of Glucose Intolerance. In: International Textbook of Diabetes Mellitus, Alberti, K, P. Zimmet and R. Defronzo (Eds.), John Wiley and Sons Ltd, Chichester, pp: 9-23.

http://en.wikipedia.org/wiki/CHAID

http://www.mu-sigma.com/analytics/thoughtleadership/cafe-cerebral-chaid.html

http://www.obgyn.cam.ac.uk/cam-only/statsbook/stchaid.html

Kass, G.V., 1880. An exploratory technique for investigating large quantities of categorical data. Applied Stat., 29: 119-127. DOI: 10.2307/2986296

Schneiderman, N., 2004. Psychosocial, behavioral and biological aspects of chronic diseases. Curr. Direct. Psychol. Sci., 13: 247-251. DOI: 10.1111/j.0963-7214.2004.00318.x

Sriphaew, K., S. Pathomnop and M.L. Kulthon Kasemsan, 2012. Temporal data classification of diabetes mellitus on health examination data of factory employees. Int. J. Comput. Commun. Eng., 1: 31-34. DOI: 10.7763/IJCCE.2012.V1.10

Swets, J.A., R.M. Dawes and J. Monahan, 2000. Better decisions through science. Scientific Am., 283: 70-75. PMID: 11011389

WHO/IDF, 2006. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. Report of WHO/IDF Consultation.

Yoon, K.H., J.H. Lee, J.W. Kim, J.H. Cho and Y.H. Choi *et al*., 2006. Epidemic obesity and type 2 diabetes in Asia. Lancet, 368: 1681-88. DOI: 10.1016/S0140-6736(06)69703-1