Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Blemish identification in Co-variance of Disease using Data Mining Techniques

Dr. A. S.Aneeshkumar[#1], Dr.A.Ambeth Raja[*2], Dr.SripriyaArunachalam[#3]

[1]Assistant Professor & Head, Research Department of Computer Science & Applications, Research Coordinator & Research Advisor, AJK College of Arts and Science, Coimbatore, Tamil Nadu, India
aneeshkumar.alpha@gmail.com

[2]Head & Associate Professor,PG Department of Computer Science,ThiruthangalNadar College,Selavayal, Chennai -51,Tamil Nadu, India
arajacs1983@gmail.com

[3]Assistant Professor, Vels University, VISTAS, Chennai, Tamil Nadu, India
priyaarun02@gmail.com

Article Info

Page Number: 4354 - 4362

Publication Issue:

Vol 71 No. 4 (2022)

Article History

Article Received*: 25 March 2022*

Revised*: 30 April 2022*

Accepted*: 15 June 2022*

Publication*: 19 August 2022*

Abstract

Data mining is an established domain used to identify influenced or associated aspects and its occurrence for the forecasting and future predictions in several fields. These information are statistically consistent, substantial facts and formerly unknown. Health care is one of the major data mining application area. In health related organizations, the techniques of data mining are used to observe characteristics of a disease or a particular medical condition and formation of its treatment. Liver enzymes are useful for accountable physical and chemical activities in the body and its variations will affects other relevant functionalities of our body. Diabetes Mellitus is a state, where the body does not produce enough insulin to convert blood sugar to energy. The influence of diabetes in liver disordered patients are the base of this study. Chi-squared automatic interaction detector is applied in this paper to determine the contribution of Liver disease to diabetes mellitus and vice-versa.

Keywords: -Data Mining; CHAID; Correlation; Accuracy; Diabetes Mellitus; Liver disorder.

INTRODUCTION

In human body, blood sugar is essential for energy usage and metabolic control. Pancreas produces insulin hormone to control the absorption of blood sugar from edible elements. Body is inept to generate or use enough insulin successfully if the calorie consumption is excessive and continues. This situation will lead to Diabetes Mellitus (DM) stage. Based on the level of hyperglycaemia DM is known as a chronicmetabolic condition [1, 2]. Immediate and increased level of calories causes sugar storage as fat in the body. It became energy deposit in crises as if we are in fasting. However, this storage turn harm to the body if it is stagnant for a long period [3]. As a subsequent to this, it often lead to other complications such as high blood pressure, liver disorder, heart issues, stroke, neuropathy, kidney related diseases and the damage of other organs [4].

In 2006, World Health Organization (WHO) reported that, in worldwide more than one seventy million peoples are suffering from diabetes and it is going to grow up to two sixty six million in 2030. In India, thirty one million (i.e. 10% people) are living with diabetes mellitus. In 2030, it will

increase drastically about thirty one million [5]. Diabetes can be classified into two categories, where type I diabetes is usually affected in children and Asians are more genetic susceptibility for second category (i.e. type II diabetes) [6].

All biological changes in the body are associated with the lifestyle and psychological stress of human beings. Diet and regular workout will support to decrease chronic disorders and its associated symptoms. Socioeconomic factors such as education, occupation, income, social cohesion, race, ethnicity and behaviors like long term and regular usage of alcohol and smoking have impact in such diseases. Psychodynamic factors like anxiety and stress also influence the severity.

Liver is one of the large organ in our body, which is also involve in energy expenditure and metabolism. The carbohydrates gathered from gastrointestinal is undergone with a hepatic process to convert metabolism into fatty acids or amino acids [7].

Data mining techniques are more familiar in big data analysis, warehouse and data processing fields. Data mining embraces classification methods to forecast the categorical information of the designated data. It is the process of generating rules for predefined categories. The classification method is also applied to correctly differentiate unknown attributes in future [8]. Classification consists of two steps, which are learning and testing. In learning, a model is developed with a major part of the whole dataset. A small part of the dataset is used to test the model, which is developed in learning phase.

1.1 Data Description

There are 22 major parameters used for this study. These parameters are collected through physical and clinical observations made in the hospitals or in the laboratories. Table 1 contains physiological constraints with two options. Apart from table, it includes gender, age and obesity. Gender (Ge) is classified as male and female. Age group (Ag) is collected as 8 categories started from 30 to 69, where the interval is 5. Obesity (Ob) is having three values, which are below average, average and above average. Biological parameters of table 2 is having three options (Clinical observation is in below normal level, normal and above normal level). These attributes are collected from alcoholic and non-alcoholic fatty liver disordered patients.

Table I: Physiological constraints

S. No.	Factor Name	Instance value	
1	Itching (It)	1	0
2	Alcoholism (A1)	1	0
3	Smoking (Sm)	1	0
4	Head ache (He)	1	0
5	Acting differently (Ac)	1	0

Table II: Biological constraints

S. No.	Parameter Name	Instance value		
		Below normal level	Normal	Above normal level
1	BILIRUBIN D (B1)	-1	0	1
2	BILIRUBIN T (B2)	-1	0	1
3	S.G.O.T. (S1)	-1	0	1
4	S.G.P.T. (S2)	-1	0	1
5	GAMMA GT (Ga)	-1	0	1
6	ALKALINE PHOSPHATE (A2)	-1	0	1
7	TOTAL PROTEINS (To)	-1	0	1

8	ALBUMIN (A3)	-1	0	1
9	GLOBULINS (Gl)	-1	0	1
10	Plasma Glucose –F (P1)	-1	0	1
11	Plasma Glucose –R (P2)	-1	0	1
12	Blood Pressure –Diastolic (B1)	-1	0	1
13	Blood Pressure –Systolic (B2)	-1	0	1
14	Triglycerides (Tr)	-1	0	1

II. MODELING

Chi-squared Automatic Interaction Detector (CHAID) is a powerful and traditional decision making algorithm in classification. It learn the association between dependent constraints and predictor instance [9]. Gordon V. Kass (1980) from South Africa has proposed CHAID as part of his Ph.D. dissertation work [10][11]. It is an extension of US Automatic Interaction Detection and THeta Automatic Interaction Detection.

CHAID figures non-binary trees for bigger and complex patterns of dataset. CHAID is generally used for the purpose of market segmentations. This method stores predictor's value and then determine the independent attributes' optimal value. The basic three steps of this algorithm are,

1. Preparing predictors- This step creates three categories of liver disorders from the given data.
2. Merging categories- Identify minimum significant symptoms of the particular liver disorder type and join it together.
3. Choosing the split value- Split identified symptom based on the lowest p value. This process continue until reach final node.

In this study, the medical dataset is divided into 2 groups. 80% of data is for the model building and remaining 20% is to test the classification model.

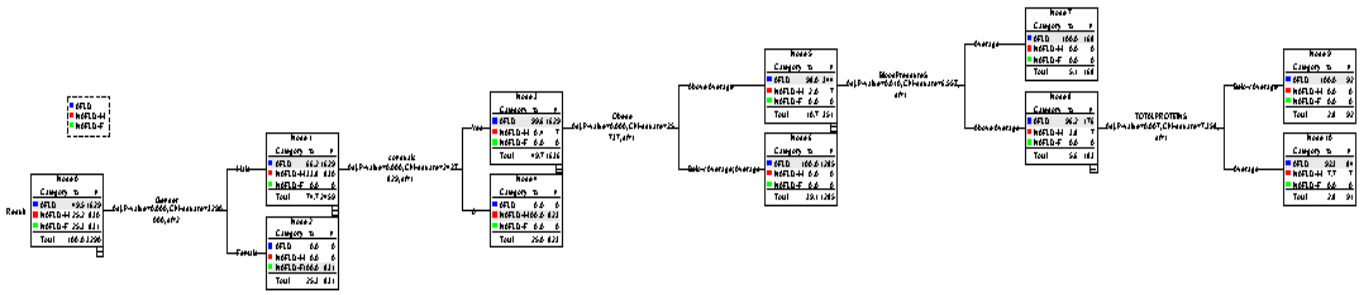

Fig 1: Generated CHAID tree for liver disorder

Table III: Tree table of CHAID

Node	AF		NAFm		NAFf		Total		Predicted disease	Parent Node	Primary Independent Variables				
	N	Percent	N	Percent	N	Percent	N	Percent			Symptom	Sig.[a]	Chi-Square	df	Split Values
0	1629	49.5%	830	25.2%	831	25.3%	3290	100.0%	AF						
1	1629	66.2%	830	33.8%	0	.0%	2459	74.7%	AF	0	Ge	.00	3290.0	2	Male

												0	00		
2	0	.0%	0	.0%	831	100.0 %	831	25.3%	NAFf	0	Ge	.000	3290.000	2	Female
3	1629	99.6%	7	.4%	0	.0%	1636	49.7%	AF	1	Al	.000	2427.829	1	Yes
4	0	.0%	823	100.0 %	0	.0%	823	25.0%	NAF m	1	Al	.000	2427.829	1	No
5	344	98.0%	7	2.0%	0	.0%	351	10.7%	AF	3	Ob	.000	25.737	1	1
6	1285	100.0%	0	.0%	0	.0%	1285	39.1%	AF	3	Ob	.000	25.737	1	-1;0
7	168	100.0%	0	.0%	0	.0%	168	5.1%	AF	5	B2	.010	6.557	1	0
8	176	96.2%	7	3.8%	0	.0%	183	5.6%	AF	5	B2	.010	6.557	1	1
9	92	100.0%	0	.0%	0	.0%	92	2.8%	AF	8	To	.007	7.358	1	-1
10	84	92.3%	7	7.7%	0	.0%	91	2.8%	AF	8	To	.007	7.358	1	0

III. MODEL EVALUATION

The fitness of any constructed classifier model is determined by sensitivity and specificity for any prediction [12]. These two factors can calculated with true positive (TP), false positive (FP), true negative (TN) and false negative (FN) instances of the classified dataset. Hence, the proportional probability of the liver disordered patients with a positive result of diabetes mellitus is known as the sensitivity. It is defined by a formula of TP/(TP+FN). Specificity specifies the possibility proportion of the liver disordered patients without diabetes mellitus and so it reflect in negative result group. Specificity can be calculated with TN/(TN+FP). These indicators of the classifier are recognized as the fitness recognition factors.

The classifier fitness = sensitivity* specificity

Table IV: Confusion Matrix of the CHAID classifier

		Predicted patients		
		AF	NAF m	NAFf
Actual patients	AF	391	0	0
	NAFm	1	179	0
	NAFf	0	0	179

The values from confusion matrix shows that,

Sensitivity of AF= 391/ (391+0) =1

Specificity of AF= 359/ (359+0) =1

Fitness of AF = 1*1 =1

Sensitivity of NAFm= 179/ (179+0) =1

Specificity of NAFm= 571/ (571+1) =0.998

Fitness of NAFm = 1*0.998 =0.998

Sensitivity of NAFf= 179/ (179+0) =1

Specificity of NAFf= 571/ (571+0) =1

Fitness of NAFf = 1*1 =1

A. ROC Curve

Receiver operating characteristic (ROC) curves are graphically effective way to differentiate positive and negative result groups with a clear cut off value for a class or adecision threshold [13]. An ROC constantly represent a unit square and this curve passes through two intervals, which are 0,0 and 1,1.

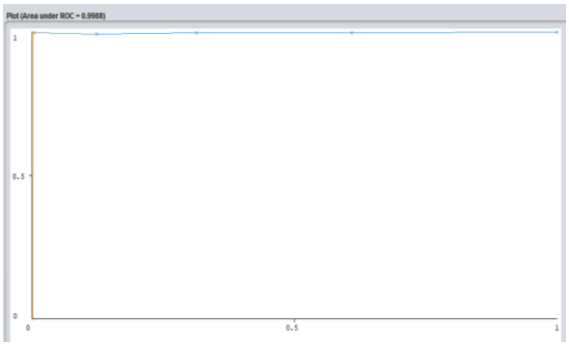

Fig.2: ROC curve of AF

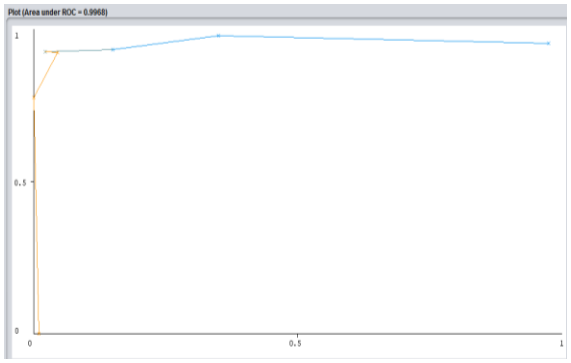

Fig.3: ROC curve of NAFm

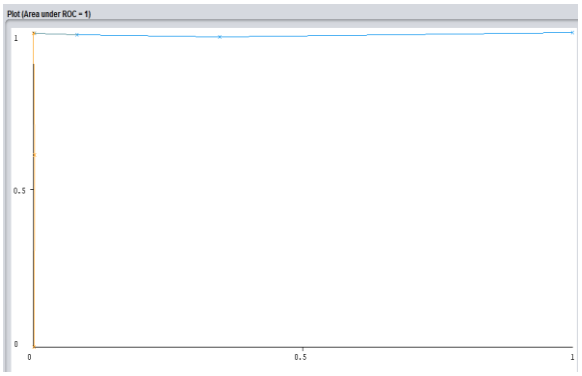

Fig.4: ROC curve of NAFf

The beginning point of it is 0,0 and it represents that there is no classifier sensitivity. The classifier shows its maximum sensitive when the curve reaches to 1,1 [14]. ROC is formed with X axis as specificity and Y as sensitivity. Therefore, the curve for AF and NAFf lies in Y axisupto 0.9988 and 1 respectively. ROC curve for NAFm starts at 0.002 of X axis and grows towards Y axis until reach 0.9968. According to the ROC curve, it can be determined that the indicated model is fit for the given data classification.

Table V: Mean and standard deviation of P1 and P2 in liver disordered patients data

Factors	AF		NAFm		NAFf	
	S.D.	Mean	S.D.	Mean	S.D.	Mean
P1	0.443	0.75	0.343	0.87	0.451	0.72
P2	0.402	0.80	0.412	0.78	0.335	0.87

Table VI: Correlation factors of P1 and P2 in AF patients

Attribute	HPC1	GPC5	NPC0	ZC 0	HNC1	GNC5	NNC0
P1	B2, S2, A2, B1		It, Sm		Ob, S1, P2	B1, To	Ag, B2
P2	OS, S1,S2, A2, To,G1, P1,B1	B2	It		Ag, B2	Sm, B1	

Table VII: Correlation factors of P1 and P2 in NAFm patients

Attribute	HPC1	GPC5	NPC0	ZC 0	HNC1	GNC5	NNC0
P1	Ag, Sm, He, Ac, S1, S2, To, P2, B2	B1	It, B2, Ga		A2, A3		A1
P2	Ag, Sm, He, Ac, S1, To, P1, B1	B2	It, A1, B2		A3	Ga	S2, A2

Table VIII: Correlation factors of P1 and P2 in NAFf patients

Attribute	HPC1	GPC5	NPC0	ZC 0	HNC 1	GNC 5	NNC0
P1			Ag, It, Sm, B2,A2, B1,B2			A3	A1, He, B1, S1,S2, Ga, To, P2
P2			A1, Sm, He, B2, A2,				Ag, It, B1, S1, S2, Ga, A3, P1,

			To, B2				B1

HPC1 - High Positive Correlation with 1% of significance.
GPC5 - Good Positive Correlation with 5% of significance.
NPC0 - Normal Positive Correlation without any significance.
ZC0 - Zero Correlation (without any correlation).
HNC1 - High Negative Correlation with 1% of significance.
GNC5 - Good Negative Correlation with 5% of significance.
NNC0 - Normal Negative Correlation without any significance.

If the symptom X and Y are correlated, it unavoidably states that X either causes for Y or Y causes for X [15]. Here for an example, age is correlated with plasmas glucose F and R, but we can't blindly conclude that age is the reason for diabetes or diabetes is the reason for aging. These two might have a relation with another factor like liver problem or correlated with multiple symptoms that have a solid relation with disease [16, 17]. The above tables shows that most of the collected instances have positive or negative correlation with P1 and P2.

IV. RESULT AND DISCUSSIONS

In this paper, 3290 patients' data are classified in to Alcoholic fatty liver disorder (AF), Non-alcoholic fatty liver disorder male (NAFm) and Non-alcoholic fatty liver disorder female (NAFf). Figure 1 shows CHAID tree and table 3 represents tree table of CHAID where, three types of liver disorders are classified. The percentage of categorical data in each type of liver disorder and its influencing attributes are also shown here. In 0^{th} node, it used 100% of its training dataset. In training phase, it is identified that there is no association between alcoholic fatty liver and gender (female) value. Similarly, based on significance and chi-square value, the null hypothesis is very strongly accepted for the relation between non-alcoholic liver disorder female and primary independent variable male, alcoholic fatty liver and non-alcohol consumption, non-alcoholic liver disorder male and alcohol consumption character, alcoholic fatty liver disorder and obesity (above normal, normal and below normal).

Confusion matrix in table 4 has shown that, true positive value of AF is 391, NAFm is 179 and NAFf is 179. In non-alcoholic fatty liver disorder female, only one patient is falsely classified as alcoholic fatty liver disorder. In table 5, the mean and standard deviation of plasma glucose F (P1) among AF patients are 0.75 and 0.443. The mean and standard deviation of P2 in AF patients are 0.80 and 0.402 respectively. In NAFm, the P1 mean is 0.87 and P2 mean is 0.78. The standard deviation of the same are 0.343 and 0.412. In NAFf, the mean value of P1 is decreased to 0.72 and standard deviation is increased to 0.451. P2 mean is equal to P1 mean of NAFm ie.,0.87. It shows reverse exploit to its previous group. The standard deviation of P2 in NAFf is 0.338.Tthe standard deviation of all groups are falls within the interval of 0.350 and 0.450. However, non-alcoholic fatty liver disordered female is having high deviation from the mean value.

In this study, it is found that most of the attributes are highly correlated with plasma glucose f (P1)and plasma glucose R (P2) with one percentage level of significance. In table 6 (AF patients),

Bilirubin T, SGPT, Alkaline Phosphate and Blood pressure-Diastolic are highly correlated with plasma glucose f. Plasma glucose r is highly correlated with obesity, SGOT, SGPT, Alkaline Phosphate, Total protein, Globulins, plasma glucose f and Blood pressure-Diastolic. Hence, SGPT, Alkaline Phosphate and Blood pressure-Diastolic are highly influencing the cause of P1 and P1. Since Itching is having positive correlation at 5% significance, it also considered as a common factor in both cases. Smoking habit may increase P1.

In NAFm patients (table 7), physiological parameters such as age, smoking habits, head ache, acting differently and biological symptoms SGOT, Total protein are seen highly in both P1 and P2 cases. However, SGPT, Plasma glucose-R and Blood pressure – Systolic are highly correlated with P and, Plasma glucose-R and Blood pressure – Systolic are the high symptoms of P2. Based on correlation analysis, P1 and P2 are positively related with each other.

In table 8, albumin (A3) is the only symptom that negatively correlated with P1 in NAFf at 5% significance level. All remaining factors shows positive or negative correlation with no significance for two types of diabetes. The traces analysed here identified that almost 80% of liver disordered patients are under the treatment of diabetics. In some cases, patients are having boarder value of plasma glucose F but R-value is calculated as normal and vice-versa. Therefore, these patients are having vulnerability to increase alternate plasma glucose or it may be the symptom of any other serious disorder. Such cases should be studied or analysed with the help of health experts and specified biomedical examinations.

V. CONCLUSION AND FUTURE WORK

This study indicates that both diabetes and liver disorder are having a good association. Either liver disorder may contribute to type II diabetes or type II diabetes may contribute to liver disorder. These results are often useful to spread awareness in public about the co-variance of diseases. In future, it may enlighten to conduct more researches with different data types to determine associated influence of diseases.

REFERENCES

[1]. “Definition and Diagnosis of Diabetes Mellitus & Intermediate Hyperglycemia”, Report of WHO/ IDF consultation 2006.

[2]. Harris M. and Zimmet P., “Classification of diabetes mellitus and other categories of glucose intolerance. In Alberti K, Zimmet P, Defronzo R, editors. International Textbook of Diabetes Mellitus. Second Edition. Chichester: John Wiley and Sons Ltd, p9-23, 1997.

[3]. Neil Schneiderman, “Psychosocial, Behavioral and Biological aspects of Chronic Diseases”, Current directions in psychological science 2004, Vol:13 No:6, pp. 247-251.

[4]. KritsadaSriphaew, SomphopPathomnop, and M. L. KulthonKasemsan, “Temporal Data Classification of Diabetes Mellitus on Health Examination Data of Factory Employees”,

International Journal of Computer and Communication Engineering, Vol. 1, No. 1, May 2012, pp.31-34

[5]. Yoon KH, Lee JH, Kim JW, Cho JH, Choi YH, Ko SH, Zimmet T and Son Y, "Epidemic obesity and type II diabetes in Asia", Lancet 2006;368, pp.1681-88.

[6]. Kun-Ho Yoon, Jin-Hee Lee, Ji-Won Kim, Jae Hyoung Cho, Yoon-Hee Choi, Seung-Hyun Ko, Paul Zimmet and Ho-Young Son, "Epidemic obesity and type 2 diabetes in Asia", www.thelancet.com Vol 368 November 11, 2006, pp:1681-1688.

[7]. Baig N, Herrine S and Rubin R, "Liver disease and diabetes mellitus", Clinical Lab Medicine 2001, 21, pp.193-207

[8]. W. Au, K. C. C. Chan and X. Yao, "A Novel Evolutionary Data Mining Algorithm With Applications to Churn Prediction", IEEE Transactions on Evolutionary Computation, Vol. 7, No. 6, 2003.

[9]. http://www.mu-sigma.com/analytics/thought leadership/cafe-cerebral-chaid.html

[10]. http://en.wikipedia.org/wiki/CHAID.

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

[12]. KorkutKorayGundogan, Bilal Alatas and Ali Karci, Mining Classi_cation Rules by Using Genetic Algorithms with Non-random Initial Population and Uniform Operator, Turk J Elec Engin, Vol.12, NO.1 2004, pp:43-52.

[13]. John A. Swets, Robyn M. Dawes and John Monahan, "Better decisions through science, Scientific American, 283(4), pp.70–75, 2000.

[14]. W. B. Langdon and S. J. Barrett, "Genetic Programming in Data Mining for Drug Discovery" Evolutionary Computing in Data Mining, Springer, pp. 211-235, 2004.

[15]. Jiawei Han and MichelineKamberData Mining: Concepts and Techniques, Second Edition, Elsevier, pp: 68, 2006.

[16]. A.S.Aneeshkumar, Dr. C. JothiVenkateswaran, "Relevance Study of Data Mining for the Identification of Negatively Influenced Factors in Sick Groups", Procedia Computer Science, Volume 47, 2015, Pages 101-108, DOI: https://doi.org/10.1016/j.procs.2015.03.188.

[17]. Dr. A.S.Aneeshkumar, J. AngelinJeba Malar, M. Anoop,"Effectiveness of Association Rule Mining Classification in HIV and HCV", 3rd International Conference for EmergingTechnology (INCET) 2022, Belgaum, India, IEEE Xplore, 15 July 2022, Page. 1-5, DOI: 10.1109/INCET54531.2022.9824414.