An Approach of Data Mining for Predicting the Chances of Liver Disease in Ectopic Pregnant Groups

A.S.Aneeshkumar,
Research Scholar,
PG and Research Department of Computer Science, Presidency College, Chennai, India.

C.Jothi Venkateswaran, Ph.D.
Research supervisor & Dean, Department of Computer Science & Applications,
Presidency College, Chennai, India..

ABSTRACT
Diseases are the most serious social and expensive problem faced by the society. In the past decade, world has experienced a rapid increase in various Liver diseases and Ectopic Pregnancy. In this work we propose a novel approach to evaluate the increased tendency of ectopic pregnancy and liver disease among such groups, using data mining techniques. It's due to the modern adaptive life style and cultural changes of our society.

General Terms
Ectopic Pregnancy, Liver disorder.

Keywords
Data mining, regression analysis, hypothesis.

1. INTRODUCTION
The healthcare industry collects huge amount of health care data, which mined to discover hidden information for effective decision making. One of the major challenges facing health care organisation is the provision of quality services at affordable cost. Quality of service implies diagnosing patients correctly and administering treatments that are correctly with minimum clinical tests. They can achieve these results by employing computer based data mining and decision support system. Data mining has been around for more than two decades, its potential is being realised now and it combines statistical analysis, machine learning, algorithms, information retrieval and database technology to extract hidden patterns and relationship from large data base. Statistics suggest with current advances in early detection of diseases.

Before 19th century, the mortality rate from ectopic pregnancies exceeded 50%. But by the end of the century, the mortality rate dropped to five percent because of surgical intervention. Now a day's concurrently increasing the reported cases of ectopic pregnancy or eccysis, which means complication of pregnancy while the embryo implants outside the uterine cavity. Furthermore, they are dangerous for the mother, since internal haemorrhage is a life-threatening complication. Most ectopic pregnancies occur in the Fallopian tube (so-called tubal pregnancies), but implantation can also occur in the cervix, ovaries, and abdomen. An ectopic pregnancy is a potential medical emergency, and, if not treated properly, can lead to death [12].

In a normal pregnancy the fertilized egg enters the uterus and settles into the uterine lining where it has plenty of room to grow. About 1% of pregnancies is in an ectopic location with implantation not occurring inside of the womb, and of these 98% occurs in the Fallopian tubes. Now an ectopic pregnancy can be diagnosed very early using blood tests for HCG and vaginal ultrasound. Beta HCG is a very specific identification of pregnancy. A positive HCG level confirms that the patient is pregnant, but it will not provide the details about the position of the pregnancy. A vaginal ultrasound allows the doctor to locate the gestational sac of the early pregnancy. Normally the sac outside the uterus gives a positive diagnosis of ectopic in diagnosis. However, the sac cannot be seen clearly in ectopic pregnancies. Suppose the HCG level is more than 2000 mIU/ml. and the doctor cannot see a gestational sac, then the diagnosis is an ectopic case. Another blood test which can be helpful is a serum progesterone level, which is low (less than 15 ng/ml) in patients with ectopic pregnancies, as compared to normal pregnancies [13]. Lot of studies and researches going in this field, even though cannot judge the accurate reasons for increasing these problems.

Liver disorder can be identified from liver function test (LFT), in which the enzymes value may deviate from the normal pattern.

2. DATA SET
A long time collection of data in ectopic, liver dysfunction is also seen as a part of some cases. So this analysis is carried with 162 reported ectopic pregnancy cases of 13 months has been collected for this study, where 64 cases seen with abnormalities in liver function test. Normally Ectopic cases will be admitted in the bases of emergency and immediately conduct laparoscopy. Most of the obstetric experts refer for haematology and biochemistry with clotting time test, and therefore the other related study of data will not be collected by the medical institutions.

3. REGRESSION ANALISIS
Regression analysis is used to predict the value of a dependent variable based on the value of at least one independent variable and explain the impact of changes in an independent variable on the dependent variable. The following equation gives a general approach of linear regression analysis.

$$B = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \ \ldots\ldots\ldots + \beta_k A_k + \varepsilon$$

Here the linear relationship between A and B, is distributed by some random error ε have a mean of zero. $\beta_0, \beta_1, \beta_2, \ldots . \beta_k$ are the constants, use to match the input samples in statistical estimation, where β_0 is the estimate of the regression intercept and other values are considered as estimate of the regression slope or regression coefficient.

Correlation and regression are intimately related and the correlation coefficient between A and B is,

$$r = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum(A_i - \bar{A})^2 \sum(B_i - \bar{B})^2}}$$

When B is regressed on A, the regression coefficient of A is,

$$\beta_1 = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sum(A_i - \bar{A})^2}$$

$$\beta_0 = \bar{B} - \beta_1 \bar{A}$$

as testing sets, which are used to identify the effectiveness of the model.

3.1.2. k-fold cross validation

K-fold cross validation measure has been used (Delen at al., 2005). In the basic structure, it consists of dividing the data into k- subgroups. Each subgroup is tested via the rule constructed from the remaining k-1 groups. Thus the k different test results are obtained for each train- test configuration [5]. Here for our study, using ten-fold cross validation. It will help us to increase the unbiased hypothesis and evaluate the overall performance of the prediction model. Table 2 shows the evaluation of predictive model for the training data set and repeated testing data set. Table 3 represents the achieved result after the application of ten-fold cross validation.

Where linear regression model is,

$$liver = 0.277 * ectopic + 1.4718$$

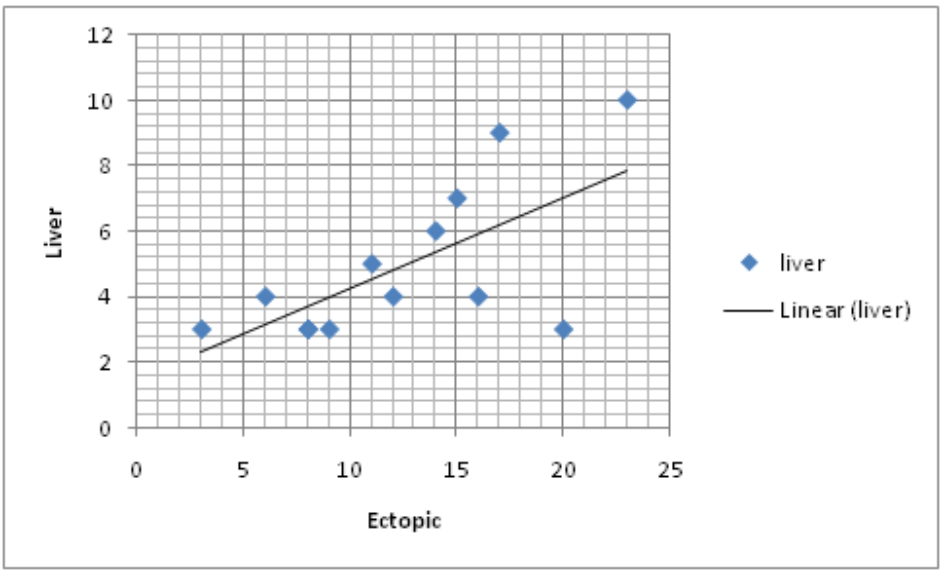

Figure 1: Positive linear relationship of data

TABLE I: Estimation of β value

β_1	β_0
0.2770	1.4713

3.1. WEKA Tool

Waikato Environment for Knowledge Analysis(WEKA) tool, is used to analyse the performance result obtained by using linear regression analysis.

3.1.1. Training sets and Testing sets

Each model is built with a predetermined set of data and it is known as learning step or training phase, where the algorithm learning from the training sets [22]. Then another set known

TABLE II: Predictive model for training set

Time taken to build model	0 seconds
Correlation coefficient	0.6632
Mean absolute error	1.3921
Root mean squared error	1.7233
Relative absolute error	73.0621 %
Root relative squared error	74.843 %

TABLE III: Predictive model of 10-fold cross validation

Time taken to build model	0 seconds
Correlation coefficient	0.4651
Mean absolute error	1.5966
Root mean squared error	2.0957
Relative absolute error	76.4247 %
Root relative squared error	83.4599 %

3.2. Testing Differences between the Proportions

Z test is one of the major methods used in statistics for analysing the proportions between two data sets.

$$Z = \frac{p_1 - p_2 - p_3 - \ \ldots\ldots\ldots - p_k}{S.E.}$$

Where, $S.E. = PQ(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \ \ldots\ldots\ldots + \frac{1}{n_k})$

$$P\ = \frac{n_1p_1 + n_2p_2 + n_3p_3 + \ \ldots\ldots\ldots + n_kp_k}{n_1 + n_2 + n_3 + \ \ldots\ldots\ldots + n_k}$$

$$Q = 1 - P$$

TABLE IV: Proportional differences of the dataset

P	Q	S.E.	Z	|Z|
0.395	0.605	0.577	-8.7487	8.7487

3.3. Significance Test for Correlations

It is used to determine whether a specific predictor is significant or not. A hypothesis test, known as t-test is defined as,

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Where, degree of freedom (df) = 11

3.4. Goodness of Fitting the Model

In any statistical model analysis, the fitness of an observation set is explained by the goodness of its fit. By analysing the residuals, a majority of the tests for goodness of fit of a model are carried out[7]. The Chi-Square ($\varkappa^2$) test defines as,

$$x^2 = \sum\left(\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}\right)$$

Where, O_i denotes observed frequency of the joined event and E_i is expected frequency of the joined event.

4. RESULTS AND DISCUSSIONS

Implementation of these statistical methods, in the collected 164 ectopic and identified 64 liver disordered cases within that, gives a relation of positive linear regression from Figure 1. According to the table 1, we can see the estimate of the regression intercept, $\beta_0 = 1.4713$ and the regression slope value is 0.2770.

In linear regression analysis, with a total of training sets and another total of testing sets gives the result of correlation coefficient and other related error rates in Table 2. After apply the ten-fold cross validation, the result of the predictive model is shown in Table 3, where the positive value of the correlation decreased and simultaneously the error rates also increased. In both of these cases the time taken to build the model is 0 seconds.

Table 4 shows the difference between the proportions of both datasets for 13 months, where the standard error is 0.577 and the proportion is 8.7487. The significance test gives the value of 1.7419 and therefore the null hypothesis is rejected. Finally, in addition to that the chi-square used to assess the goodness of fit of the model. Here the degree of freedom is 12. For 1 df the chi-square value needed to reject the hypothesis at the 0.001 significance level is 10.828. The computed value is 668.9526 and hence the hypothesis is strongly rejected for this group of data.

5. CONCLUSION AND FURTHER STUDIES

The prediction of Liver disease from ectopic pregnancy is a very difficult job and the relation between both of these is not that much analysed or studied by the experts of medical fields, because the reason of ectopic may vary according to the cases. So it's peculiar to judge the correct reason behind it and liver disease is seen in some of them. But in the modern society the reported cases have a linear increment and a correlation. In future we will extend our work to identify the factors affecting the increased cases in recent time with the help of theory of evidence and fuzzy logic.

REFERENCES

[1] Huda Yasin, Tahseen A. Jilani and Madiha Danish, "Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications (975-8887), Volume 24- No.3, June 2011.

[2] Michael Morelli and Anthony J. DeSimone Jr. "Application of Dempster-Shafer Theory of Evidence to the Correlation Problem" ISFI@ 2002.

[3] P.Rajeswari, G.Sophia Reena, "Analysis of Liver Disorder Using Data mining Algorithm", Global Journal of Computer Science and Technology, vol.10 issue 14(ver. 1.0) November 2010.

[4] Polat.K and Gunes.S, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease", Digital Signal Processing, 17(4), 2007, pp. 702-710.

[5] Asha.T, Dr.S. Natarajan, Dr.K.N.B.Murthy, "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification", AIT, 2011

[6] Hosmer D. and Lemeshow S., "Applied Logistic Regression", John Wiley and Sons, 2nd edition, 2000.

[7] Ahmed Mohamed samir ali gamal eldin, Egypt, "(IJACSA) International Journal of Advanced Computer Science and Applications", Vol 2, No.12, Dec 2011.

[8] Sudheep Elayidom.M, Sumam Mary Idikkula and Joseph Alexander, "Applying Statistical Dependency Analysis Techniques In a Data Mining Domain", International Journal Of Data Engineering (IJDE), Volume (1): Issue (2).

[9] K. Rajeswari, Dr. V.Vaithiyanathan, Dr. P. Amirtharaj, Prediction of risk score for heart disease in India using machine Intelligence, 2011 International Conference on Information and Network Technology, IPCSIT vol.4(2011) IACSIT Press, Singapore.

[10] Sunita Soni, O.P.Vyas, Using Associative classifiers for Predictive Analysis in Health care Mining, International Journal of Computer Applications(0975-8887) Vol4-No.5, July 2010.

[11] Anne-Louise M Heath, Cynthia Reeves Tuttle, Megan SLSimons and Christine L Cleghorn, "Longitudinal study of diet and iron deficiency anaemia in infants during the first two years of life", *Asia Pacific J Clin Nutr (2002) 11(4): 251–257.*

[12] R Zuzarte, C C Khong, "Recurrent ectopic pregnancy following ipsilateral partial salpingectomy", Singapore Med J 2005; 46(9) : 476.

[13] http://www.drmalpani.com/book/chapter19.html

[14] Yun-Hsuen Lim, Soon P. Ng, Paul H. O. Ng, Ay E. Tan and Muhammad A. Jamil, "Laparoscopic salpingectomy in tubal pregnancy: Prospective randomized trial using endoloop versus electrocautery", *J. Obstet. Gynaecol. Res. Vol. 33, No. 6: 855–862, December 2007.*

[15] Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman, "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application", World Academy of Science, Engineering and Technology 50 2009.

[16] K.S.Kavitha , K.V.Ramakrishnan and Manoj Kumar Singh , "Modeling and design of evolutionary neural network for heart disease detection", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010, ISSN (Online): 1694-0814.

[17] Dasgupta, D. and F. Gonzalez, "An immunity-based technique to characterizeintrusions in computer networks", IEEE Trans. Evol. Comput. 6 (3), 1081-1088, June 2002.

[18] Leinbach SS, Bhat RA, Xia SM, Hum WT, Stauffer B, Davis AR, Hung PP, Mizutani S, "Substrate specificity of the NS3 series proteinase of hepatitis C virus as determined by mutagenesis at the S3/NS4A junction", Virology 1994, 204:163-169.

[19] Portnoy, L., E. Eskin, and S. J. Stolfo, "Intrusion detection with unlabeled data using clustering", In Proc. of ACM CSSWorkshop on Data Mining Applied to Security (DMSA-2001), Philadelphia. ACM, 5-8 November, 2001.

[20] G. Florez, SM. Bridges, Vaughn RB, "An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection", Annual Meeting of The North AmericanFuzzy Information Processing Society Proceedings, 2002.

[21] Lee, W., S. J. Stolfo, and K. W. Mok, " Mining in a data-flow environment: Experience in network intrusion detection," In S. Chaudhuri and D. Madigan (Eds.), Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, CA, pp. 114124. ACM,12-15 August 1999.

[22] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Published by Elsevier, second edition – 2006.

[23] http://en.wikipedia.org/wiki/Ectopic_pregnancy