Scientific Research and Essays Vol. 5(20), pp. 3059-3062, 18 October, 2010
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 ©2010 Academic Journals

Full Length Research Paper

Evaluation of risk of death in hepatitis by rule induction algorithms

Nilgün Ulutaşdemir[1*] and Özgür Dağlı[2],

[1]Yusuf Şerefoğlu School of Health, Kilis 7 December University, Kilis, Turkey.
[2]25 Aralık State Hospital, Gaziantep, Turkey.

Accepted 16 July, 2010

This study investigates the availability of rule induction algorithms (PART, J48, JRip) for the evaluation of death risk in hepaptitis based on a clinical database. Simple rules that are derived can be practically used to determine risk of death due to hepatitis. The results are quite satisfactory, where an accuracy of 84.5% is obtained for the prediction of death risk in hepatitis.

Key words. Hepatitis, rule induction algorithms, PART, J48, JRip, death risk.

INTRODUCTION

Hepatitis is inflammation of the liver from any cause. Hepatitis commonly results from a virus, particularly, one of five hepatitis viruses(A,B,C,D,E). Less commonly, hepatitis results from other viral infections, such as infectious mononucleosis and cytomegalovirus infection (Peksen et al., 2004). The major nonviral causes of hepatitis are alcohol and drugs. Hepatitis can be acute (lasting less than 6 months) or chronic; it occurs commonly throughout the world. (Chao and Chang, 2010) Acute viral hepatitis is the inflammation of the liver caused by infection with one of the five hepatitis viruses, for most people that are infected, the inflammation begins suddenly and lasts only a few weeks. Chronic hepatitis is the inflammation of the liver that lasts at least 6 months. It can persisist for years, even decades. Continued inflammation slowly can damage the liver, eventually producing cirrhosis and liver failure. (Keryn et al., 1998).

Many people have chronic hepatitis for years without developing progressive liver damage. For others, the disease gradually worsens. When this happens, the disease is the result of viral hepatitis B or C infection. There are antiviral agents that stop the inflammation. However, drugs are expensive, adverse effects are common, and hepatitis tends to recur once treatment is stopped (Souza and Glynn, 2004). Fibrosis in the liver may gradually worsen. Over a period of years, patients with hepatitis can develop cirrhosis, liver failure, or both. Hepatocelluar carcinoma also may be the final step before death. Regardless of the cause or type of chronic hepatitis, any complications such as ascites (fluid in the abdominal cavity) or encephalopathy (abnormal brain function) are steps of serious illness which will probably result in death (But and Loi, 2008). In severe hepatitis, especially in cirrhosis like in every disease history of the patient, signs, symptoms and laboratory values are important in progression and outcome of the disease. Risk of death increases in certain circumstances.

The aim of this study is to evaluate risk of death in hepatitis by decision based soft computing approaches using a clinical database mentioned in material and method.

MATERIALS AND METHODS

The clinical hepatitis datasets used in this investigation were obtained from the UCI Machine Learning Repository at the University of California (UCI) (Frank and Asuncion, 2010). The hepatitis database was donated by G. Gong, Carnegie-Mellon University, Ljubljana, Yugoslavia (1988). The data set includes 32 instances of one class (die) and 123 instances of another class (live). The dataset is described by 19 attributes of which 6 were numeric and 13 were nominal. Age, sex, use of steroids, use of antivirals, presence of fatique, malaise, anorexia, presence of spiders and ascites, hepatomegaly and splenomegaly in physical examination, bilirubin, laboratory values of alkalenphosphotase, SGOT, albumin and protrombin time, histology of the patient, were

*Corresponding author. E-mail: nulutasdemir@yahoo.com.

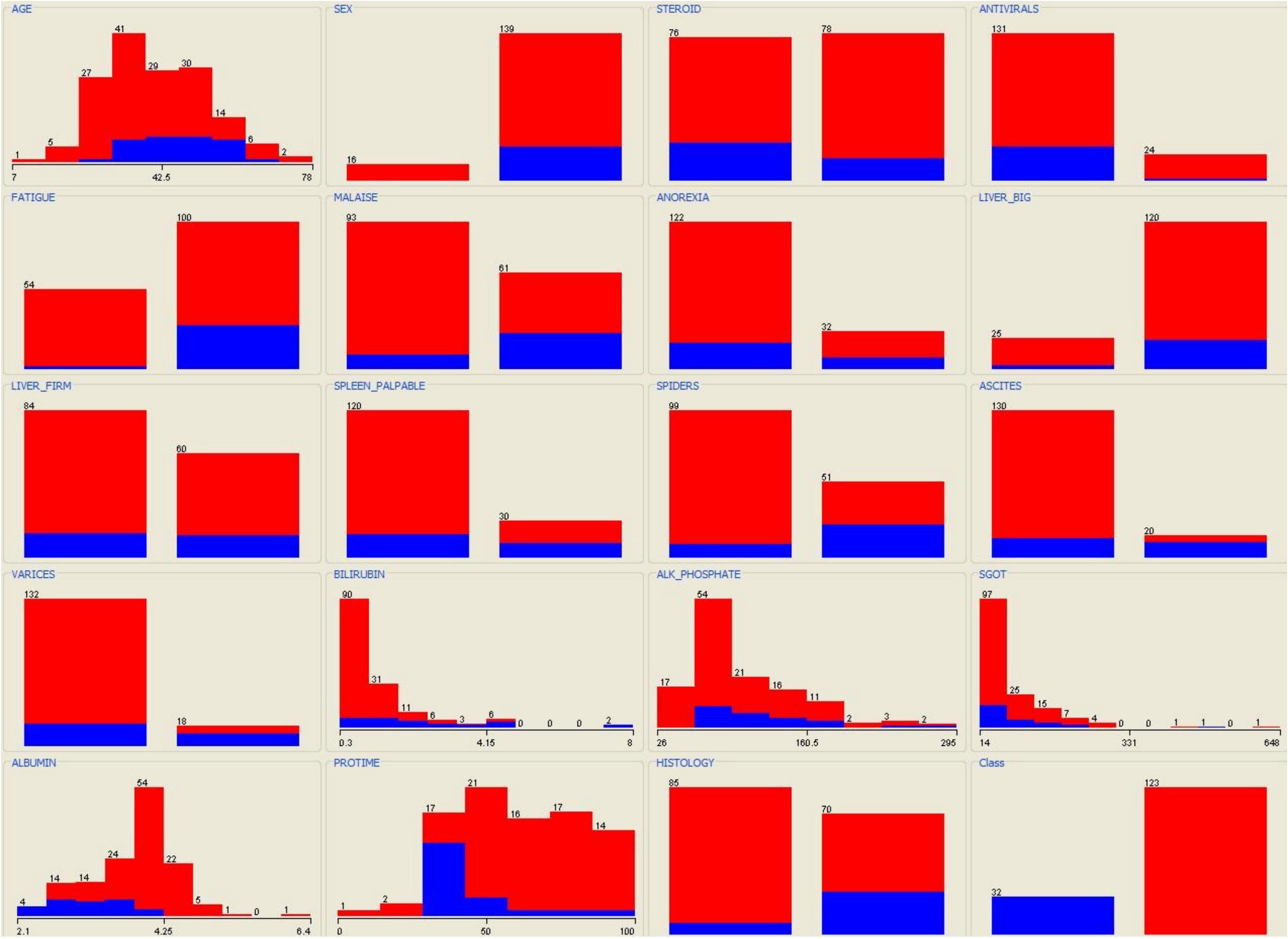

Figure 1. Visual description of hepapitis database.

known parameters of the 155 patients in this study. The dataset is given in visual form in Figure 1.

The analysis were performed using WEKA environment. Inside the Weka system, there exist many classification algorithms which can be classified into two types; rule induction and decision-tree algorithms (Witten et al., 1999). Rule induction algorithms generate a model as a set of rules. The rules are in the standard form of IF-THEN rules. Meanwhile, decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class outputs (Daud and Corne, 2007). The classification algorithms used in this study were PART, Jrip and J48. The data was divided into a testing and training sets by using crossvalidation (10 folds).

PART is a separate-and-conquer rule learner proposed by Witten et al., (1999). The algorithm generates sets of rules called 'decision lists' which are ordered set of rules. PART builds a partial C4.5 decision tree in each iteration and converts the "best" leaf into a rule. The algorithm is a mixture of C4.5 (Quinlan, 1993) and RIPPER (Cohen, 1995) rule learning. JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by Cohen1` (1995, 1996). Ripper constructs a ruleset by adding rules to an empty ruleset until all positive examples are covered. The J48 algorithm is the Weka implementation of the C4.5 top-down decision tree learner proposed by Quinlan (1993). The algorithm uses the greedy technique and is a variant of ID3, which It deals with numeric attributes by determining where thresholds for decision splits, should be placed (Daud and Corne, 2007).

RESULTS

The results of Jrip model are given as follows:
JRIP rules:

(ALBUMIN <= 3.8) and (ALBUMIN <= 2.8) => Class=DIE (13.0/2.0)
(PROTIME <= 42) => Class=DIE (15.0/7.0)
(SPIDERS = yes) and (BILIRUBIN >= 2) => Class=DIE (11.0/4.0)
ELSE Class=LIVE (116.0/6.0)

The results of PART model are given as follows:

PART rules:

ASCITES = no AND
SPIDERS = no AND

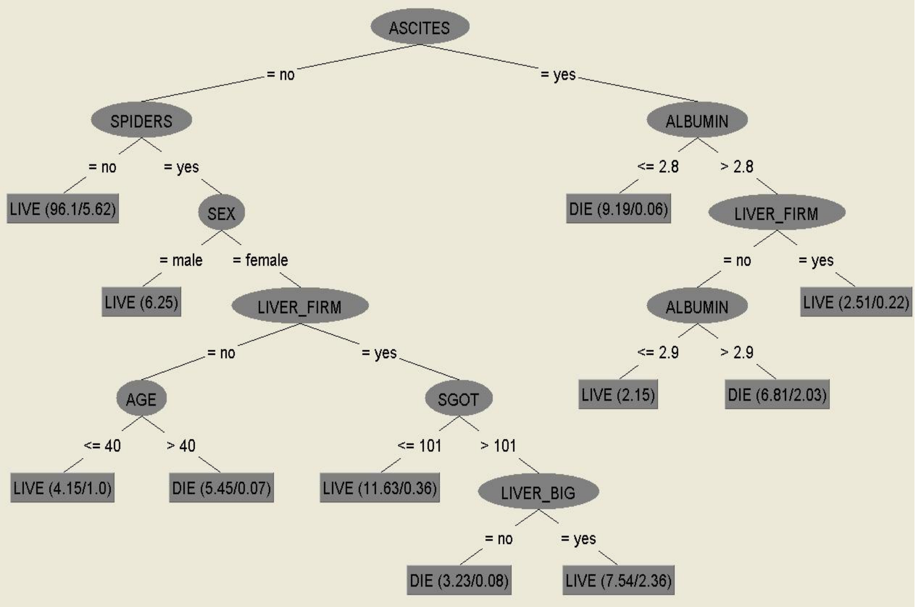

Figure 2. J48 pruned tree.

Table 1. Accuracy of Jrip model.

Correctly classified ınstances	121	(78.0645%)
Incorrectly classified ınstances	34	(21.9355%)

Table 2. Confusion matrix for Jrip model.

DIE	LIVE	
11	21	DIE
13	110	LIVE

Table 3. Accuracy of PART model.

Correctly classified ınstances	131	(84.5%)
Incorrectly classified ınstances	24	(15.5%)

Table 4. Confusion matrix for PART model.

DIE	LIVE	
21	11	DIE
13	110	LIVE

MALAISE = no: LIVE (72.09/1.0)

SEX = female AND
ALBUMIN > 3.6 AND
ANOREXIA = no AND
PROTIME > 46: LIVE (17.68/3.04)

SEX = female AND
ANOREXIA = no AND
PROTIME <= 51: DIE (18.79/2.21)

SEX = female AND
ALBUMIN > 2.8 AND
BILIRUBIN <= 2.5: LIVE (22.32/1.17)

SEX = male: LIVE (10.0)
ASCITES = yes: DIE (5.73/0.05)

ALK_PHOSPHATE > 168: LIVE (4.53/1.34)

ELSE DIE (3.85/0.67)

The results of J48 model are given in Figure 2.

DISCUSSION

Use of Evolutionary and soft computing for knowledge discovery in medical diagnosis is increasing day by day. This study is an application of rule induction algorithms which a branch of soft computing was used for the evaluation of risk of death in hepatitis. Algorithms used in this study were PART, JRip and J48, where analysis were performed by WEKA. The performances of the algorithms can be accepted as satisfactory in general and in terms of accuracy (PART (84.5%), JRip (78%), J48 (83.8%)) as shown in Tables 1, 2, 3, 4, 5 and 6. Although JRip has the lowest accuracy, the rule is quite simple, which can be practically used. On the other hand, PART has higher accuracy but with a comparativley complex rule algorithm.

Table 5. Accuracy of J48 model.

Correctly classified ınstances	130 (83.8%)
Incorrectly classified ınstances	25 (16.2%)

Table 6. Confusion matrix for J48 model.

DIE	LIVE	
14	18	DIE
7	116	LIVE

REFERENCES

But DY, Loi CL (2008), Natural history of hepatitis-related hepatocellular carcinoma. World J. Gastroeneterol, 14(11): 1652-1662.

Cohen WW (1995). Fast effective rule induction. In: Machine Learning. Lake Tahoe, California, pp. 115-123.

Cohen WW (1996). Learning rules that classify e-mail. In: Machine Learning in Information Access, pp. 18-25.

Chao J, Chang ET (2010). BMC Public Health, 25: 10-98.

Daud MNR, Corne DW (2007) Human Readable Rule Inductıon In Medıcal Data Mınıng: A Survey Of Exıstıng Algorıthms. In: WSEAS European Computing Conference, 2007, Athens, Greece, pp. 787-798.

D' Souza RF, Glynn MJ (2004), Improving general practitioners knowledge of chronic hepatitis C infection QJM. Aug, 97(8): 549-50.

Frank A, Asuncion A (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Hepatitis]. Irvine, CA: University of California, School of Information and Computer Science.

Keryn AG, Lane W, Kelly J (1998). Merck Manual of Medical Information 3rd Edition, pp. 18-571.

Peksen Y, Canbaz S, Leblebicioğlu H (2004). Primary care physicians approach to diagnosis and treatment of Hepatitis B and Hepatitis C patients. BMC Gastroenterol, 6: 4-3

Quinlan R (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo.

Witten IH, Frank E, Trigg L, Hall M, Holmes G, Cunningham SJ (1999). Weka: Practical machine learning tools and techniques with java implementations. In: Emerging Knowledge Engineering and Connectionist-Based Info. Systems, pp. 192-196.