ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
CONSIGLIO NAZIONALE DELLE RICERCHE

P. Di Giacomo, G. Felici, R. Maceratini,
K. Truemper

DIAGNOSIS OF HEPATOCELLULAR CARCINOMA VIA LOGIC-BASED NEW SUPERVISED LEARNING METHOD

R. 559 Novembre 2001

Paola di Giacomo – Center of Biomedical Research, "La Sapienza" University of Rome, Italy.

Giovanni Felici – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, Viale Manzoni 30 - 00185 Roma, Italy. Email :felici@iasi.cnr.it.

Riccardo Maceratini – Center of Biomedical Research, "La Sapienza" University of Rome, Italy.

Klaus Truemper – University of Texas at Dallas, Box 830688, Richardson, TX 75083-0688, USA,Email: klaus@utdallas.edu.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica, CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

Abstract

The hepatocellular carcinoma is one of the most widely spread malignant tumors in the world. The ability to detect the tumor in its early stages in a minimally invasive way is crucial to the treatment of patients with this disease. In recent years several methods and techniques from the fields of artificial intelligence, decision theory and statistics have been introduced in the medical management of patients (diagnosis, treatment, follow-up); computational prognostic models are increasingly used in medicine to predict the natural course of disease, or the expected outcome after treatment. Prognosis forms an integral part of systems for treatment selection and treatment planning. Furthermore, diagnostic and prognostic models may play an important role in guiding diagnostic problem solving. In this paper we describe the application of a new learning tool for the diagnosis of hepatocellular carcinoma. The method adopted operates in the logic domain and presents several interesting features for the development of medical diagnostic systems. We consider a database of 128 patients, 64 of which affected by hepatocellular carcinoma, while the others affected by cirrhosis but not from hepatocellular carcinoma. Each patient is described by a number of attributes measured in non-invasive way.

Key words: Supervised Learning, Automatic Diagnosis, Logic Programming, Hepatocellular Carcinoma

1. Introduction

In this paper we describe the application of a new logic domain learning tool, the Lsquare System, in automated medical diagnosis. The interest for such systems has raised in recent years due to the excessive cost of mass screening and to the development of efficient diagnostic methods that use a possibly very large set of attributes describing the health state of the patients. The hepatocellular carcinoma (HCC) is one of the most widely spread virus-related malignant tumors in the world. Each year, only in Italy, it's estimated that 12,000 patients fall ill. Before the introduction of liver ultrasound, it has been impossible to detect the tumor in its early stages. Now, it is possible with the new imaging techniques to detect also very small lesions, but in the case of chronic hepatitis with a suspected HCC mass of diameter less that 2 cm, it is similar to active regenerative cirrhotic liver area. The high risk that many patients present for the surgical trip and the high number of prognostic factors included led to the rearch of alternative mini invasive treatments. The liver tumors are the more frequent malignant neoplasms, with over one million of new cases per year in the world. HCC represent about the 50% of these tunors, with a ratio M/F of 4:1. The incidence is low in the USA (1.9 deaths per 100,000 inhabitants/year), medium in Europe and South Africa (4,9-20), high in China, Korea and Mozambique (23,1-150), with mortality at 5 years of 80%. In the last decade this incidence in Europe and in Japan is strongly increased, as well as the hepatitis C, which at present has to be considered as the main etiopathogenetic factor (or co-factor), in case is high the synthesis of DNA ([16]), and the chronic active hepatitis B (relative risk=10-21,3%, risk which is 98 times higher respect to patients HbsAg negative). The recent progresses is the non invasive diagnostic by imagines have brought a change of scenario: the ultrasonography (spiral, porto-TC) and the new method for the NMR with epato-specific means of contrast and evaluation of the enhancement curves allow at present to carry out an early diagnosis and a careful staging. Essential complements to the minimal criteria of standard staging (TNM) with regard to the therapeutical planning are the vascolarization, the istologic type [17] and the human hepato-specific alpha-phetoprotein dosage (HAFP-mRNA) [18] in the pheriperal blood, correlated (p less than 0.001) to the presence of the intra-hepatic micro diffusion, portal thrombosis and metastases. The liver neoplasms which are not treated have as survival less then three years in the 87% of cases. Their association with the chronic hepatitis may discourage an invasive surgical approach and can suggest mini invasive acts which, associated to a constant monitoring of associated diseases, obtain results of life expectancy and quality which are equivalent or superior to those of the resection of the transplant. The hepatic percutaneous ethanol injection (max three modules, Child A-B, max volume 3 cm) presents the 98% of survival at one year and the 48-56% at five years [19]. The consequence of this fact is that the number of patients with primitive cancer of the liver that have a longer life expectancy and, then, need a continue monitoring and multidisciplinary cares is in a strong increase. Three diseases are often concomitant: carcinoma, cirrhosis with or without portal ipertension, active hepatitis. The research in this field aims to develop expert systems for automated cancer screening, diagnosis end prognosis using computed imaging, neural networks, linear-programming-based machine learning approaches. In particular we have many examples in literature for the breast cancer, prostate cancer, lung cancer, oral cancer tissues, automated skin cancer diagnosis and immunological cancer diagnosis (interesting material on this matter may be found in the web pages [20], [21], [22], [23]. Although the ability to detect the tumor in its early stages is relevant for life expectancy, the automated diagnosis process for the liver cancers has not been significantly investigated. Here we attempt to accomplish this task by means of a new learning method that develops a disgnostic system with several interesting features.

The paper is organized as follows. Section 2 introduces the application of automated learning system in medical diagnosis. In some formal definitions of propositional logic are given. Section summarizes the concepts of separating sets and the solution algorithm. In section the Lsquare system is discussed. Section describes the architecture of the database for hepatocellular carcinoma, while section discusses the experiments that have been conducted. Finally, in section we present some conclusion and describe future work that will be done in the direction of producing a reliable, flexible and effective fully automated diagnostic system for hepatocellular carcinoma.

2. Learning Methods for Medical Diagnosis

A learning system can be formalized in the following way. Given a set of objects belonging to a set of classes, and a set of measurements on the objects, identify the class to which an object belongs to, based on the analysis of the measurements. In the medical context, we typically consider a set of patients that have a disease and a set of patients that do not have that disease (or that have a different one); from the analysis of the attributes of these patients the learning system "learns" some rules that separate the patients in the two sets. Once these rules have been learned and proved to be correct, they can be used to produce a diagnosis for new patients at a very low cost. Good learning systems have also the capability of identifying, in a large set of attributes, a reduced set from which it is possible to derive the correct diagnosis. Learning systems have been proposed by many researchers in the fields of Artificial Intelligence and Statistics; more recently, have appeared in the literature several approaches that adopt mathematical and combinatorial optimization models and techniques for learning systems, such as in [7], [8], [3], [5], [6], [9]. Methods based on neural networks ([10], [11]) have also been used to formalize and solve several learning problems. Lsquare is a particular type of learning system that operates in the logic domain. In this type of learning, the objects to be recognized are described by a number of logic attributes, that is, by the presence or absence of certain features. The method uses logic formulas to express separations amongst groups of data; it learns logic relations between the logic measures and the class to which an object belongs. Such logic relations can then be used to provide additional information on the context under study, or could be integrated with other knowledge bases expressed in the same logic domain. Such method is extremely interesting in the medical applications, as the knowledge condensed in the learned formulas can be understood and interpreted by the medical experts, thus providing also a compact explanation of the diagnosis. Lsquare has been developed in a recent collaboration between the Istituto di Analisi dei Sistemi ed Informatica del Consiglio Nazionale delle Ricerche (IASI-CNR) and the University of Texas at Dallas and is described in detail in [2] and [1].

3. Preliminary Definitions

In this section we provide a quick overview of the main concepts of propositional logic used in the rest of the paper. A complete presentation of such material can be found in [12]. A *Boolean variable* is a variable that may take on only the value *True* or *False*. One or more Boolean variables can be assembled in a Boolean formula using the "not" operator ($\neg$), the "and" operator ($\wedge$), and the "or" operator ($\vee$). A Boolean formula where Boolean variables, possibly negated, are joined by the operator ($\wedge$) is called a conjunction; if the operator ($\vee$) is used, it is called a disjunction. A *conjunctive normal form system* (CNF) is a Boolean formula where some disjunctions are joined with the operator ($\wedge$). Each disjunction of a CNF system is called a CNF *clause*. A *disjunctive normal form system* (DNF), is a Boolean formula where

some conjunctions are joined with the operator ($\vee$). Each disjunction of a DNF system is called a DNF *clause*. A Boolean formula is satisfiable if there exists an assignment of *True/False* values to its Boolean variables so that the value of the formula is *True*. The problem of deciding whether a CNF formula is satisfiable is known as the *satisfiability problem* (SAT). In the affirmative case, one also must find an assignment of *True/False* values for the Boolean variables of the CNF that make the formula *True*. A variation of SAT is the *minimum cost satisfiability problem* (MINSAT), where rational costs are associated with the *True* value of each logic variable and the solution to be determined has minimum sum of costs for the variables with value *True*.

4. Logic Data and Logic Separation

We consider $\{0, \pm 1\}$ vectors of given length n each of which has an associated *outcome* with value *True* or *False*. We call these vectors *records* of *logic data* and view them as an encoding of logic information. A 1 in a record means that a certain Boolean variable has value *True*, and a -1 that the variable has value *False*. The value 0 is used when the *True/False* value of the variable is not known. The outcome is considered to be the value of a Boolean variable t (in typical medical applications, t indicates the disease that is to be diagnosed). We collect the records for which the property t is absent in a set A, and those for which t is present in set B. For ease of recognition, we usually denote a member of A by a, and of B by b. The *Lsquare* system deduces $\{0, \pm 1\}$ separating vectors that may be used to compute for each record the associated outcome, i.e., to separate the records in A from the records in B. A *separating set* is a collection of separating vectors. The separation of A and B makes sense only when both A and B are non empty, and when each record of A or B contains at least one $\{\pm 1\}$ entry. We also have to state a non trivial conditions for separation to be possible. Consider two records of logic data, say g and f. We say that f is nested in g if for any entry f_i of f equal to $+1$ or -1, the corresponding entry g_i of g satisfies $g_i = f_i$. It can then be proved the following (see [2] for proof) :

Theorem 4.1. *Let A and B be sets of $\{0, \pm 1\}$ records of the same length. Then a separating set S exists if and only if no record $b \in B$ is nested in any record $a \in A$.*

From the above theorem a clear characterization of separating vectors can be given: a $\{0, \pm 1\}$ vector s separates a record $b \in B$ from A if:

$$s \text{ is not nested in any } a \in A \tag{1}$$

and

$$s \text{ is nested in } b \tag{2}$$

Accordingly, we say that a separating set S separates A from B if, for each $b \in B$, there exists a separating vector $s \in S$ that separates b from A. The problem of finding a separating set for A and B is decomposed into a sequence of subproblems, each of which identifies a vector s that separates a non empty subset of B from A solving two logic minimization problems. To formulate these problems we introduce new logic variables, that are linked with the elements s_i of the vector s to be found. More precisely, we introduce Boolean variables p_i and q_i and state that $s_i = 1$ if $p_i = True$ and $q_i = False$, $s_i = -1$ if $p_i = False$ and $q_i = True$, and $s_i = 0$ if $p_i = q_i = False$. The case $p_i = q_i = True$ is ruled out by enforcing the following logic conditions:

$$\neg p_i \vee \neg q_i,\ i = 1, 2, \ldots, n \tag{3}$$

Consider now the separation conditions (1) and (2) in terms of the new Boolean variables p_i and q_i. For (1) we have that s must not be nested in any $a \in A$. Defining a^+ as the set of indices i for which the element a_i of a is equal to 1, that is, $a^+ = \{i | a_i = 1\}$ and, analogously, $a^- = \{i | a_i = -1\}$, $a_0 = \{i | a_i = 0\}$, we can summarize condition (1) writing that

$$\Big(\bigvee_{i \in (a^+ \cup a^0)} q_i\Big) \vee \Big(\bigvee_{i \in (a^- \cup a^0)} p_i\Big) \quad \text{for all } a \in A \tag{4}$$

For condition (2) we have to enforce that, if s separates b from A, then s is nested in b. In order to do so we introduce a new Boolean variable d_b that determines whether s must separate b from A. That is, $d_b = True$ means that s need not separate b from A, while $d_b = False$ requires that separation. For the given $b \in B$, the separation condition is therefore:

$$\begin{array}{ll} \neg q_i \vee d_b & \text{for all } i \in (b^+ \cup b^0) \\ \neg p_i \vee d_b & \text{for all } i \in (b^- \cup b^0) \end{array} \tag{5}$$

We are now ready to formulate the problem of determining a vector s that separates as many $b \in B$ from A as possible (this amounts to a satisfying solution for (3)-(5) that assigns value *True* to as few variables d_b as possible). For each $b \in B$, define a rational cost c_b that is equal to 1 if d_b is *True*, and equal to 0 otherwise. Using these costs and (3)-(5), but omitting (3), the desired s may be found by solving the following MINSAT problem, with variables d_b, $b \in B$, and $p_i, q_i, i = 1, 2, \ldots, n$.

$$\begin{array}{ll} \min & \sum_{b \in B} c_b(d_b) \\ & \Big(\bigvee_{i \in (a^+ \cup a^0)} q_i\Big) \vee \Big(\bigvee_{i \in (a^- \cup a^0)} p_i\Big) \quad \text{for all } a \in A \\ & \neg q_i \vee d_b \quad \text{for all } b \in B, \text{for all } i \in (b^+ \cup b^0) \\ & \neg p_i \vee d_b \quad \text{for all } b \in B, \text{for all } i \in (b^- \cup b^0) \end{array} \tag{6}$$

The solution of problem (6) identifies a separating vector s and a subset $B' = \{b \in B | d_b = False\}$ that is separated from A by s. Using this information, we can formulate a second MINSAT problem where we select, amongst all separating vectors that separate B' from A, a vector with some desired characteristics. Let $c(p_i)$ and $c(q_i)$ be some costs associated with the variables pi and qi, for $i = 1, 2, \ldots, n$. Solving the following MINSAT problem we thus obtain a separating vector for which the cost of the logic variables used by that vector is minimized:

$$\begin{array}{ll} \min & \sum_{i=1}^{n} [c_{p_i}(p_i) + c_{q_i}(q_i)] \\ & \Big(\bigvee_{i \in (a^+ \cup a^0)} q_i\Big) \vee \Big(\bigvee_{i \in (a^- \cup a^0)} p_i\Big) \quad \text{for all } a \in A \\ & \neg q_i \quad \text{for all } b \in B', \text{ for all } i \in (b^+ \cup b^0) \\ & \neg p_i \quad \text{for all } b \in B', \text{ for all } i \in (b^- \cup b^0) \end{array} \tag{7}$$

A simple example of the role of cost values $c(p_i)$ and $c(q_i)$ is the following. Assume that $c(p_i)$ and $c(q_i)$ assign cost of 1 when p_i and q_i are *True* and cost 0 when they are *False*. The solution of (7) will then determine a separating vector that uses the minimum number of logic variables, that is, the minimum amount of information contained in the data to separate the sets. On the opposite, if we assign cost 0 for *True* and 1 for *False*, we obtain a solution that uses maximum amount of information to define the separating sets. This simple consideration has a

strong connection with the quality of the separations and with the error that will be made by the separating vectors on non classified records, as deeply discussed in [2] and [1]. The iterative process for determining a separating set is sketched below:

Algorithm FIND SEPARATING SET

Input: Sets A and B of $\{0, \pm 1\}$ records. For $i = 1, 2, \ldots, n$, cost functions $c(p_i)$ and $c(q_i)$.

Output: The largest subset B^* of B that can be separated from A, and a set S^* that accomplishes that separation.

Procedure:

1. Initialize $B^* = S^* = \emptyset$.
2. Solve (6) to get a largest possible subset B' of B that can be separated from A. If $B' = \emptyset$, output B^* and S^*, and stop.
3. Solve (7). Derive from the solution a separating vector s'', and add it to S^*. Add the records of B' to B^*. Redefine B as $B - B'$, and go to Step 2.

5. The Lsquare System

The *Lsquare* system ([13]) is a complete and easy-to-use tool to determine separating vectors by the algorithm FIND SEPARATING SET. The MINSAT problems (6) and (7) are potentially difficult as they belong to the class of *NP-complete* problems (see [14]). We solve the MINSAT instances encountered in Algorithm FIND SEPARATING SET with the aid of the Leibniz System, a software system for logic programming developed at the University of Texas at Dallas ([15]). Such solver is based on several decomposition and combinatorial optimization results and is rooted on theoretical results described in [12]. The system handles SAT or MINSAT problems with up to $6,000$ variables and $6,000$ clauses.

Lsquare is freely distributed and very simple to use. It refines the separating procedure described in the previous section by a special technique of resampling of the logic records available for training. From this resampling, a number of separating vectors is obtained, each one producing a vote for each record of A and B, indicating whether the record is classified in A or in B. The sum of all the votes so obtained is the *vote total* V, that, in the actual implementation of Lsquare, is an even integer that ranges from -40 to $+40$. If the vote V is positive (resp. negative), then the record belongs to A (resp. to B). If A and B are representative subsets of two families A and B, then the votes V for records of A (resp. for B) tend to be positive (resp. negative). We say "tend" since there cannot be a guarantee that this will be so unless A and B themselves are the sets A and B. To assess the reliability of the classification of records of A and B, Lsquare computes two estimated probability distributions. For any odd integer x, the first (resp. second) distribution supplies an α (resp. (β) of the probability that the vote V for a record of A (resp. B) is less (resp. greater) than x. For example, if one declares records with votes V greater than a given x to be in A, then α is an estimate of the probability that a record of A will be misclassified.

6. The Database for Hepatocellular Carcinoma

The first step has been the systematic collection of the clinical data. We have elaborated 37 forms, as shown in Table (1), each one proper to a class of risk factors and clinical fittings, to characterize a disease, which depends on environing, diet factors and life style.

8.

Abdominal pain	Asthenia
Fullness and anorexia	Jaundice
Vomiting	Bone pain
Ascitis	Splenomegaly
Fever	Spyder nevi
Hand erythema	Weight loss
Gynecomastia	Budd-Chiari syndrome
Hepatic bruit	Abdominal swelling
Paraneoplastic symptoms	Cirrhosis
Lab analysis	Hepatitis B test: HBV Ag
Hepatitis C test: HCV Ab	Hepatitis C test: HCV Riba test
Liver mass	Contraceptive pill
Carcinogen exposure	Alcohol assumption
Flushing	Other malignant diseases
Chest X-rays	Bone scintigraphy
Liver echo color doppler	Liver TC
Liver NMR	Liver ultrasound
Hepatitis C test: PCR RNA HCV	Hematemesis
IFN therapy	

Table 1: List of the forms

The second step has been the design and the implementation of a database to store the clinical data of patients. The coding of data, instead, follows the rules below (see also 2). For each patient, a string is given by the union of a number of substrings equal to the number of symptoms and/or diagnostic tests selected (i.e. 37, as mentioned above). The first element of each substring is -1, if the symptom and/or the test selected is absent, $+1$ if present, 0 if not available. Moreover, for each symptom and for each diagnostic test, as described above, we have a number of substrings equal to the number of parameters and each substring is given by the sequence of $\{\pm 1\}$ relative to the option selected.

The 37 forms were elaborated following some relevant clinical criteria. It has become apparent that hepatitis C virus (HCV) infection is a major risk factor for the development of hepatocellular carcinoma (HCC) worldwide. Evidence linking HCV with HCC includes the following:

1 A high proportion of patients with HCC have anti-HCV or HCV RNA detectable in serum. This is particularly apparent in southern Europe and Japan where 50-75 percent of patients with HCC have evidence of HCV infection (see [24], [25], [26]).

2 In patients with chronic HCV infection, progression can be noted from milder forms of hepatitis to cirrhosis and eventually, to HCC. This progression may take decades to occur. The precise mechanism by which HCV causes HCC is not known. Unlike the hepatitis B virus (HBV), HCV is not a DNA virus and does not become integrated within the genome of hepatocytes.

It is more likely that HCC occurs against a background of inflammation and regeneration, associated with liver injury due to chronic hepatitis. Most, but not all, cases of HCV-related HCC occur in the presence of cirrhosis, suggesting that it is the underlying liver disease per se that is the risk factor for HCC rather than HCV infection such as in [27].

Parameters	Options	Relative Substrings
Time interval	up to 1 week	-1 -1
	more than 1 week-1 month	-1 +1
	more than 1 month	+1 -1
Nature	increasing	-1 -1
	decreasing	-1 +1
	intermittent	+1 -1
Intensity	low or medium	-1
	high	+1
Type of pain	pyrosis	-1 -1
	relapsing	-1 +1
	impulsive	+1 -1
	under pressure	+1 +1

Table 2: Example of form for abdominal pain

The prevailing hypothesis has been that some cirrhotic nodules which grow larger than others (referred to as adenomatous hyperplasia) were the precursor for HCC. Recently, however, it has been suggested that foci of transformed hepatocytes may arise in between cirrhotic nodules and grow to become adenomatous hyperplasia and, eventually, HCC such as in [28]. Host factors which have been implicated in increasing the risk for development of HCC include age, male gender, and severity of underlying liver disease. Viral genotype may be important, although early suggestions that infection with genotype lb is more likely to result in the development of HCC have not been confirmed in larger studies such as in [29]. Although viral load has been related to the severity of liver disease, no clear link has been established between serum levels of HCV RNA and progression to HCC. Some external factors that might add to the risk for HCC in patients with HCV infection include alcohol consumption, coexistent HBV infection, and porphyria cutanea tarda, although this latter condition is found only in some geographic areas. The risk for a patient with HCV infection developing HCC cannot be calculated with any precision. It is known that approximately 20% of patients with chronic HCV infection develop histologic evidence of cirrhosis over a 10-year period. Furthermore, among patients with established cirrhosis due to HCV infection in screening programs, it has been found that 31 percent per year develop HCC, at least for the first 4-5 years of screening. By extrapolation, after 20 years of infection, 6-8% of patients with chronic hepatitis C can be expected to have developed HCC, although these calculations need to be validated by more prospective studies. Studies from Japan have found that the mean interval from HCV infection to the development of HCC is approximately 25 years, but these periods have a very wide range of variation. In the United States, HCC has been described as soon as 5 years from the onset of HCV infection. Typically, HCC carries a poor prognosis, with survival times from diagnosis measured in months. Screening studies have shown that small amounts of HCC can be detected at an early stage when it may be more amenable to curative therapy. At present surgical resection offers the best hope for prolonged disease-free survival. This may take the form of partial or total hepatectomy. Unfortunately, partial hepatectomy for HCC is associated with a very high recurrence rate (approximating 25% per year) while total hepatectomy implies liver transplantation. Thus, the true cost-effectiveness of screening for HCC remains uncertain. It has been suggested that progression to HCC can be halted or slowed down by treatment of the underlying hepatitis C. Recent studies from Japan, for example, have suggested that patients with cirrhosis due to

10.

Figure 1: Votes generated by Lsquare

Figure 2: Probability distributions generated by Lsquare

chronic HCV infection have a significantly lower risk of HCC if treated with alpha interferon than patients who were not treated. This improvement in risk was noted both in patients who had a good response to interferon as well those who did not. Data from these studies await confirmation in Europe or the United States. Then for each form (i.e. for each symptom and/or clinical test) we have established some parameters like time interval, nature of pain, intensity, type, level, site, presence of masses with relative characteristics etc. and for each parameter we have associated some options, each one corresponding to a sequence of $\{\pm 1\}$ following the rules described below (as shown in 2), in the case of abdominal pain).

7. Analysis of Experiments

We have used a sample of 128 patients, of which 64 in class A (healthy patients but suspected HCC) and 64 in class B (ill patients). For each patient, we have considered all the 37 symptoms and/or diagnostic tests listed in 1. We have thus obtained for each record (i.e. for each patient) 280 of $\{0, \pm 1\}$. We have then reduced the number of symptoms and carried the tests. At this stage of the work, due to the limited number of patients data available, we focus the analysis on the training of the learning algorithm. First test. With the support of the physician, we have excluded: abdominal pain, asthenia, fullness and anorexia, abdominal swelling, jaundice, vomiting, bone pain, hematemesis, ascitis, splenomegaly, fever, Spyder Nevi, hand erythema, weight loss, gynecomastia, Budd-Chiari syndrome, hepatic bruit, liver mass, PCR-RNA HCV, HCV Riba test, contraceptive pill assumption, carcinogen exposure, alcohol assumption, flushing, other malignant diseases, mass chest X-rays, bone scintigraphy, liver NMR with a number of $\{0, \pm 1\}$ equal to 120, for each record. In this case, Lsquare has indicated the presence of records nested and it has stopped the computation. For the small number of patients, we have yet decided to reduce the number of symptoms and /or diagnostic tests and we have carried out a third test. Second test. The sample includes: HCV Ab, HBV Ag, liver ultrasound, lab analysis, cirrhosis and liver TC, equal to 58 of $\{0, \pm 1\}$ for each record. In this case Lsquare has been able to compute an exact diagnosis of hepatocellular carcinoma for the 64 patients in class B, in the training data file, really ill with cancer. This result is confirmed by the votes (Figure 1) given to each patient: if the vote is positive then the patient is not affected by hepatocellular carcinoma (class A) with probability ((see Figure 2) proportional to the value of the vote, while if the vote is negative means that the patient is ill (class B) with probability (that is proportional to the absolute value of the vote.

8. Conclusion

The method described presents several interesting characteristics for medical diagnostic decision support systems. The experiments conducted have produced a complete separation of the hepatocellular carcinoma database, proving that the system can be trained successfully with the available data. Future work will take into account the testing of the diagnostic system with new data to verify and validate the diagnoses automatically produced, the possible extension of the

set of attributes that describe the patients, and the use of error distribution to establish the confidence level of the diagnosis.

Acknowledgments

The authors thank doctor E. Pourabbas for her contribution in setting up of the experiments described in this paper. This research was supported in part by Office of Naval Research under grant N000145-93-1-0096.

References

[1] Felici, G., Sun, F.S., Truemper, K.: A Method for controlling errors in two-class classifications. In: COMP99: 23rd Annual International Computer Software and Applications Conference, Phoenix, Arizona, IEEE Computer Society, Los Alamitos, CA, (1999) 186–191.

[2] Felici, G., Truemper,K.: A MINSAT Approach for Learning in Logic Domains. INFORMS Journal on Computing, (2001) to appear.

[3] Mangasarian, O. L. : Mathematical programming in neural networks. ORSA Journal on Computing **5** (1993) 349–360

[4] Mangasarian, O.L., Setiono, R., Wolberg, W.H.: Pattern recognition via linear programming: theory and application to medical diagnosis. In: T. F. Coleman and Y. Li, eds., Large-scale numerical optimization, SIAM Publications, Philadelphia, PA (1990) 22–30

[5] Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast cancer diagnosis and prognosis via linear programming. Operations Research **43** (1995) 570–577

[6] Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programmino. SIAM News **23** (1990) 1–18

[7] Boros E, Hammer PL, Ibaraki T, Kogan A, Mayoraz E, and Muchnik I.: An Implementation of Logical Analysis of Data. RUTCOR Research Report 29-9. Rutgers University, NJ, July 1995.

[8] Kamath AP, Karmarkar NK, Ramakrishnan KJ, and Resende MGC.: A Continuous Approach to Inductive Inference. Mathematical Programming 1992: 57 pp. 215-238.

[9] Triantaphyllou E, Allen L, Soyster L, and Kumara SRT.: Generating Logical Expressions From Positive and Negative Examples via a Branch-and-Bound Approach. Computers and Operations Research 1994, 21 pp.185-197.

[10] Beale R, and Finlay J. : Neural Networks and Pattern Recognition in Human-Computer Interaction. Ellis Horwood Limited. Chichester, England, 1992.

[11] Nelson M.M., and Illingworth W.T.: A Practical Guide to Neural Nets. Addison-Wesley, Reading, MA, 1990.

[12] Truemper K. : Effective Logic Computation. Wiley-Interscience, New York, 1998.

[13] Truemper K.: Lsquare System for Learning Logic. University of Texas at Dallas, Computer Science Program. April, 1999.

[14] Garey M.R., and Johnson D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, San Francisco, 1979.

[15] Truemper K.: The Leibniz System for Logic Programming. Version 4.2, Leibniz Plano, Texas 75023, U.S.A., 1996.

[16] Tarao K., Ohkawa S., Shimizu A. et al.: Significance of hepatocellular carcinoma proliferation in the development of hepatocellular carcinoma from anti-hepatitis C virus-positive cirrhotic patients Cancer, 73, 1149-54, 1993.

[17] Yamashita Y., Matsukawa T., Arakawa A. et al. : US-guided liver biopsy: predicting the effect of interventional treatment of hepatocellular carcinoma Radiology, 196, 799-804, 1995

[18] Komeda T., Fukuda Y., Sando T. et al. : Sensitive detection of circulating hepatocellular carcinoma cells in periferal venous blood Cancer, 75, 2214-9, 1995.

[19] Ebara M., Kita K. Et al. : Therapeutic effect of percutaneous injection on small hepatocellular carcinoma: evaluation with CT Radiology, 195, 371-7, 1995.

[20] http://www-radiology.uchicago.edu/krl/toppage1.htm

[21] http://citeseer.nj.nec.com/326827.html

[22] http://researchportfolio.cancer.gov/cgibin/list.pl?Term=7&Category=0

[23] http://www.qub.ac.uk/cm/pat/researchgroups/quantitative/computer.htm

[24] Nishioka K, Watanabe J, Furuta S, Tanaka E, lino S, Suzuki H, Tsuji T, Yano M, Kuo G, Choo Q-L, Houghton M, Oda T. : A high prevalence of antibody to the hepatitis C virus in patients with hepatocellular carcinoma in Japan. Cancer 1991;67:429-33 .

[25] Bruix J, Barrera JM, Calvet X, Ercilla G, Costa J, Sanchez-Tapias JM, Ventura M, Vall M, Bruguera M, Bru C, Castillo R, Rodes J. : Prevalence of antibodies to hepatitis C virus in Spanish patients with hepatocellular carcinoma and hepatic cirrhosis. Lancet 1989;ii:1004 6.

[26] Di Bisceglie AM, Order SE, Klein JL, Waggoner JG, Sjogren MH, Kuo G, Houghton M, Choo Q-L, Hoofnagle JH. : The role of chronic viral hepatitis in hepatocellular carcinoma in the United States. Am J Gastroenterol 1991 ;86:335-8.

[27] Di Bisceglie AM, Simpson LH, Lotze MT, Hoofnagle JH.: Development of hepatocellular carcinoma among patients with chronic liver disease due to hepatitis C viral infection. J Clin Gastroenterol 1994;19:222 6.

[28] Tong MJ, El-Farra NS, Reikes AR, Co RL. : Clinical outcomes after transfusion-associated hepatitis C. N Engl J Med 1995;332:1463 6.

[29] Nishiguchi S, Kuroki T, Nakatani S, Morimoto H, Takeda T, Nakajima S, Shiomi S, Seki S, Kobayashi K, Otani S. : Randomised trial of effects of interferon-a on incidence of hepatocellular carcinoma in chronic active hepatitis C with cirrhosis. Lancet 1995;346:1051-5.