World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:2, No:5, 2008

A Neural Computing-Based Approach for the Early Detection of Hepatocellular Carcinoma

Marina Gorunescu, Florin Gorunescu, and Kenneth Revett

***Abstract*—**Hepatocellular carcinoma, also called hepatoma, most commonly appears in a patient with chronic viral hepatitis. In patients with a higher suspicion of HCC, such as small or subtle rising of serum enzymes levels, the best method of diagnosis involves a CT scan of the abdomen, but only at high cost. The aim of this study was to increase the ability of the physician to early detect HCC, using a probabilistic neural network-based approach, in order to save time and hospital resources.

***Keywords*—**Early HCC diagnosis, probabilistic neural network.

I. Introduction

HEPATOCELLULAR carcinoma is a form of cancer that attacks the liver. Liver cancer is the fifth most common cancer in the world today [1]. The World Health Organisation (WHO) has estimated that there are approximately 400,000 new cases of liver cancer worldwide. About three-quarters of the cases of liver cancer are found in Southeast Asia (China, Hong Kong, Taiwan, Korea, and Japan). Liver cancer is also very common in sub-Saharan Africa (Mozambique and South Africa). The frequency of liver cancer in Southeast Asia and sub-Saharan Africa is greater than 20 cases per 100,000 populations [2]. In contrast, the frequency of liver cancer in North America and Western Europe is much lower, less than 5 per 100,000 populations. However, the frequency of liver cancer among native Alaskans is comparable to that seen in Southeast Asia. Moreover, recent data show that the frequency of liver cancer in the U.S. overall is rising [3].

In addition to the relatively large prevalence rate, the prognosis for hepatocellular cancer is also quite alarmingly poor. The prognosis is very poor, most patients are diagnosed with the disease during the later stages of its progression [4], [5]. The prognosis is extremely low, with patient survivorship less than 12 months once a firm diagnosis has been made. Generally, the diagnosis is based on evidence from ultrasound, computerized tomography (CT), and laboratory based diagnostic procedures. These procedures are cumbersome and costly and are not routinely performed. Recent use of cDNA micro-arrays have provided a rational genetic basis for identifying gene deregulation – but again this is generally at the later stages of the disease [6]. Although *alpha-fetoprotein* (AFP) is the most important tumor marker for the diagnosis of HCC, a considerable proportion of HCC's do not produce AFP, making early diagnosis difficult with this marker alone. Among the detection factors, the serum enzymes analysis is a fast and simple method, representing an early step in HCC diagnosis.

The field of medical informatics had a rapidly extension during the last years. Many works, using rule-based systems, statistical learning systems, evolutionary algorithms, neural networks and rough sets have been used to perform good decisions in this sensitive medical domain.

Machine learning approaches based on decision trees have been applied to the identification of HCC with some degree of success [7]. Despite these research efforts, little progress has been made towards the early identification of this disease. Therefore, any technique that can simplify the diagnosis both in terms of required procedures and cost would be of great significance. In this paper, we describe a technique that is very efficient and costs effective – based on machine learning algorithms.

An effective and easy to use method for addressing such task is represented by the Probabilistic Neural Networks (PNN). The PNN was developed by Specht, 1988 [8], as a supervised neural network, consisting in a 3-layer, feed-forward, one-pass training algorithm.

PNN are widely used in the areas of pattern recognition, nonlinear mapping, estimation of the probability of class membership and likelihood ratio. They are closely related to Bayesian decision rule and use Parzen or Parzen like probability density function estimators. They combine some of the best attributes of statistical pattern recognition and feed-forward neural networks.

The performance of PNN is strongly influenced by the smoothing parameter, various searching techniques such as incremental search, Monte Carlo search and genetic algorithms search being used.

II. Probabilistic Neural Network

A. Bayesian Classifier

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. Minimizing the probability of error, or the expected risk, represents the

Manuscript received October 15, 2006.

M. Gorunescu is with the University of Craiova, 200486 Craiova, Romania (e-mail: mgorun@ inf.ucv.ro).

F. Gorunescu is with the University of Medicine and Pharmacy of Craiova, 200486 Craiova, Romania (phone: 40-251-435620, e-mail: fgorun@rdslink.ro).

K. Revett is with the University of Westminster, London, UK (e-mail: revettk@westminster.ac.uk).

traditional goal for decision strategies. The Bayes decision rule can be summarized as follows: (*a*) Let D_k be the decision rule related to the state of nature B_k; (*b*) Given measurement x, the error related to B_k is defined by $P(error|x) = 1- P(B_k|x)$; Minimize the probability error; Bayes decision rule: "*Choose* D_k if $P(B_k|x) > P(B_j|x)$, $\forall j \neq k$" or, equivalently, "*Choose* D_k if $P(x|\ B_k)\ P(B_k) > P(x|\ B_j)\ P(B_j)$, $\forall j \neq k$". The Bayes decision rule is applied to PNN as follows. Consider the general case of the q-category classification problem, in which the states of nature are denoted by $\Omega_1, \Omega_2,\ldots, \Omega_q$. The goal is to determine the class membership of a multivariate sample data represented by a p-dimensional random vector $\mathbf{x}$ into one of the q possible groups $\Omega_1, \Omega_2,\ldots, \Omega_q$, that is to make the decision $D(\mathbf{x}) = \Omega_i$, $i = 1, 2,\ldots, q$, where $\mathbf{x}$ represents a sample. If the multivariate probability density functions $f_1(\mathbf{x}), f_2(\mathbf{x}),\ldots, f_q(\mathbf{x})$, the *a priori* probabilities $h_i = P(\Omega_i)$ of occurrence of patterns from categories Ω_i and the *loss* parameters l_i associated with all incorrect decisions given $\Omega = \Omega_i$, then, according to the Bayes decision rule, $\mathbf{x}$ is classified into the category Ω_i if the following inequality holds true:

$$l_i\, h_i\, f_i(\mathbf{x}) > l_j\, h_j\, f_j(\mathbf{x}),\ i \neq j \qquad (1)$$

The accuracy of the decision depends straight on the accuracy of estimating the corresponding p.d.f's.

The way to using the Bayes decision rule to PNNs is represented by the technique chosen to estimate the p.d.f's $f_i(\mathbf{x})$ corresponding to each decision class Ω_i, based upon the training patterns set. The classical approach uses a sum of small multivariate Gaussian distributions, centered at each training sample, that is:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\sigma^p} \cdot \frac{1}{m_i} \cdot \sum_{j=1}^{m_i} \exp\left(-\frac{\left\| x - x_j \right\|^2}{2\sigma^2} \right),\ i = 1, 2,\ldots, q \ (2)$$

where m_i is the total number of training patterns in Ω_i, $\mathbf{x}_j$ is the j-th training pattern from category Ω_i, p is the input space dimension and σ is the adjustable "*smoothing*" parameter using the training procedure. The smoothing or scaling parameter σ defines the width of the area of influence and should decrease as the sample size increases. The key factor in PNNs is therefore the way to determine the value of σ, since this parameter needs to be estimated to cause reasonable amount of overlap. Commonly, the smoothing factor is chosen heuristically.

B. Modified PNN Algorithm

This subsection introduces a simple yet efficient method of estimating the smoothing parameter σ.

Firstly, instead of using a sum of small multivariate Gaussian distributions (the classical Parzen-Cacoulos window classifiers), the exponential is replaced by the approximation Taylor polynomial. Consequently, the actual activation function consists in a sum of polynomials:

$$f_{T\,r}(x) = \frac{1}{(2\pi)^{p/2}\sigma^p} \cdot \frac{1}{m} \cdot \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{\left(-\frac{d\left(x, x_j\right)^2}{2\sigma^2} \right)^k}{k!},\ \text{for } r \geq 1. \quad (3)$$

In such a way, the running speed is doubled and the accuracy is maintained at the same level [9].

Secondly, another way to increase the running speed is to replace the heuristic search of σ in the whole real positive number set $\mathbf{R}_+$ by a Monte Carlo procedure estimating the best solution in the 99.7% confidence interval of the average distances between samples in each decision class.

Finally, the sum of training patterns that are classified in the right way is considered as cost function. Thus, it is easy to obtain the optimum value for σ by maximizing the cost function.

C. Classification Performance

In order to evaluate the classification performance, three metrics have been computed: accuracy along with the corresponding sensitivity and specificity of the classifier, where the parameters *a*, *b*, *c* and *d* are given in Table I.

TABLE I
CLASSIFICATION PARAMETERS

		Predicted class	
		Class A	Class B
Actual class	Class A	*a* (TP)	*b* (FN)
	Class B	*c* (FP)	*d* (TN)

As testing method, the *k*-fold cross-validation has been used.

Since prediction can be regarded as a Bernoulli trial with the possible outcomes: (*a*) correct prediction and (*b*) wrong prediction, given the accuracy denoted by Acc, evaluated on a certain test dataset, it is possible to predict the true accuracy of the model through its 95% confidence interval, using the Normal approximation for large enough test sets (e.g. $n > 30$) [10].

III. The Data

The model was fitted to real data concerning 269 individuals from the Department of Internal Medicine, Division of Gastroenterology, University Emergency Hospital of Craiova, Romania. This group of individuals consists of 239 patients without HCC and 30 patients with (HCC).

There are a lot of serum enzymes to be analyzed but, among them, some are the most important in HCC detection, such as: x_1 = TB (total bilirubin), x_2 = DB (direct bilirubin), x_3 = IB (indirect bilirubin), x_4 = AP (alkaline phosphatase), x_5 = GGT (gamma glutamyl transpeptidase), x_6 = LAP (leucine amino peptidase), x_7 = AST (aspartate amino transferase), x_8 = ALT

(alanine amino transferase), x_9 = LDH (lactic dehydrogenase), x_{10} = PI (prothrombin index), x_{11} = gamma, x_{12} = albumin, x_{13} = glycaemia, x_{14} = cholesterol. Along with another significant feature represented by the patient age, all these predictive factors may lead to an early HCC detection, helping the medical staff to save time and hospital resources.

The PNN-based classification algorithm has been applied to data in order to classify the initial group of individuals into two categories, depending on the diagnosis type: Ω_1 = HCC and Ω_2 = non-HCC. Each person in the data set is represented by a 15-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_{15})$, where the components represent some of the most important characteristics leading to the right medical diagnosis. Concretely, x_1 = TB, x_2 = DB, x_3 = IB, x_4 = AP, x_5 = GGT, x_6 = LAP, x_7 = AST, x_8 = ALT, x_9 = LDH, x_{10} = PI, x_{11} = gamma, x_{12} = albumin, x_{13} = glycaemia, x_{14} = cholesterol, x_{15} = age.

IV. Experimental Results

The key to obtain a good classification result using the PNN approach is to optimally estimate the misclassification costs and the prior probabilities. Unfortunately, there is no definitive science to obtain them and must be assigned as a specific part of the problem definition. In our practical experiment we have estimate them heuristically. Thus, as concerns the costs parameters, we have considered them depending on the average distances, inversely proportional, since for other choices the accuracy rate alters dramatically. As concerns the prior probabilities, they measure the membership probability in each group and thus we have considered them equal to each group size (that is $h_i = m_i$).

Since the searching procedure to estimate σ is stochastic, no test sample is available and the learning sample is too small to have the test sample taken from it, the 10-fold cross-validation has been used. Accordingly, the classification accuracy is computed 10 times, each time leaving out one of the sub-samples from the computations and using that sub-sample as a test sample for cross-validation, so that each sub-sample is used 9 times in the learning sample and just once as the test sample. The computation results are displayed in Table II.

TABLE II
Experimental Results

Division points	Accuracy (%)	Sensitivity (%)	Specificity (%)
100	81	21	88
200	83	40	88
300	87	40	92
400	86	41	91
500	87	43	92
600	89	45	95
700	87	70	88
800	92	70	95
900	92	90	92
1000	92	95	92

Moreover, the corresponding 95% confidence interval for Acc = 0.92 (value of the accuracy), produced in the cross-validation test, is given by (0.86; 0.97).

It is worth to remark that, since this method is based on the analysis of small or subtle rising of serum enzymes levels, sometimes very difficult to be noticed by the human observer, the PNN classifier has natural limitations concerning the accuracy and stability related to the random choice of the dividing points used from the Monte Carlo approach. On the other hand, the proportion between HCC samples and non-HCC samples in the dataset is 30/239 (12.5%), straight reflected in the balance sensitivity/specificity.

V. Conclusion

In this paper we have developed and provided evidence supporting the applicability and suitability of a PNN-based model for decision-making in the diagnosis of hepatic cancer.

These results were based on a relatively small sample set that was unbalanced with respect to the number of samples in each class (239 non-HCC and 30 HCC). Despite these issues, the classification accuracy was on the order of 92% with high sensitivity and specificity values.

PNN learn by examples so the details of how to recognize the disease are not needed. What is needed is a set of examples that are representative of all the variations of the disease. The dataset that we utilized in this study was obtained from a local hospital. The biochemical markers used in this study represents typical measurements that are employed in the diagnosis of early stage HCC. We used raw data (without any pre-processing) and we obtained reliable results suggesting the PNN ability and flexibility to learn from raw examples.

Although the early diagnosis of HCC is based on biochemical tests, modern approaches also use imaging tests (i.e. transabdominal ultrasound and/or spiral computed tomography). Therefore, another way to enlarge this heuristic approach in medical research is represented by the replacement of the Euclidian distance in the PNN algorithm with a general mixed-weighted measure of similarity. Such an approach will strengthen the decision process by using different types of attributes of the training patterns.

Clearly, much work still needs to be done to improve this methodology and thus providing physicians with an effective computational support in early HCC diagnosis.

Acknowledgment

The authors wish to thank Dr. Adrian Saftoiu, MD, Chief of Endoscopy Laboratory, Department of Internal Medicine, University of Medicine and Pharmacy Craiova, for supplying the data and for his valuable comments too.

References

[1] Z. Y. Tang, "Hepatocellular carcinoma-Cause, treatment and metastasis", *World J. Gastroenterol*, vol. 7, pp. 445-454, 2001.

[2] T. M. Block, A. S. Mehta, C. J. Fimmel, and R. Jordan, "Molecular viral oncology of hepatocellular carcinoma", *Oncogene*, vol. 22, pp 5093-5107, 2003.

[3] S. S. Thorgeirsson, J. W. Grisham, "Molecular pathogenesis of human hepatocellular carcinoma", *Nat. Genet.*, vol. 31, pp. 339-346, 2002.
[4] J. R. Bloomer, "Serum alpha-fetoprotein in nonneoplastic liver diseases", *Dig. Dis. Sci*., vol. 25, pp. 241–242, 1980.
[5] J. Bruix, A. J. Hessheimer, A. Forner, L. Boix, R. Vilana, and J. M. Llovet, "New aspects of diagnosis and therapy of hepatocellular carcinoma", *Oncogene*, Vol. 25, pp. 3848-3856, Jun. 2006.
[6] L.H. Zhang, and J.F. Ji, "Molecular profiling of hetaptocellular carcinomas by cDNA microarray", *World J. of Ger*., vol 11(4), pp 463-468, 2005.
[7] F. Ciocchetta, R. Dell'Anna, F. Demichelis, A. Sboner, A. P. Dhillon, A. Dhillon, A. Godfrey, and A. Quaglia, "Knowledge discovery to support hepatocellular carcinom early diagnosis", in *International Joint Conference on Neural Network – Special Session*: *Knowledge Discovery, and Image and Signal processing in Medicine*, 2003.
[8] D. F. Specht, "Probabilistic neural networks for classification mapping or associative memory", in *Proc. IEEE International Conference on Neural Networks*, vol. 1, 1988, pp. 525-532.
[9] F. Gorunescu, "Benchmarking Probabilistic Neural Network algorithms", in *Proc. 6-th Intl. Conf. on Artificial Intelligence and Digital Communications*, Thessaloniki, Greece, 2006, pp.1-7.
[10] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005, ch. 4.