ISSN : 0974 - 7435 Volume 8 Issue 7

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(7), 2013 [1019-1023]

Research on hepatitis virus identification based on improved BP neural network

Wang Zunfu[1], Wu Jianlin[1]*, Jiang Zhihua[2], Zhao Caiyun[3], Yu Bingxue[1]

[1]The First Affiliated Hospital, Guangxi Medical University, Nannning, Guangxi, 530021, (CHINA)

[2]Guang Xi Autonomus Region Center for Disease Contorl and Prevention, Nannning, Guangxi, 530021, (CHINA)

[3]The Clinical Laboratory, The Red Cross Hospital of Yulin City, Yulin, Guangxi, 537000, (CHINA)

E-mail: 407706483@163.com

ABSTRACT

Hepatitis virus identification plays a key role in clinical diagnosis and is one of the difficulties and hot research fields for the researchers related. The paper takes hepatitis B virus for example and presents a new model for hepatitis virus identification based on BP neural network and ant colony algorithm. First, the flow chart of hepatitis virus identification is designed based on the hepatitis virus image processing; Second, aiming at the shortages of the existing BP neural network algorithm of data-mining for hepatitis virus identification, ant colony and BP neural network algorithm are integrated and some improvements are advanced to speed up the convergence and simplify the structure and to improve identification accuracy of the original BP model. Finally, the model is realized by the data from three hospitals to carry out comprehensive hepatitis virus identification and the experimental results indicate that the model has favorable hepatitis virus identification results.

KEYWORDS

Clinical diagnosis;
Hepatitis virus identification;
BP neural network;
Ant colony algorithm.

INTRODUCTION

The life cycle of virus is believed to begin when virus infects the host cell membrane via its envelope proteins. The process of the attachment and the following the fusion of a viral envelope and cell membrane is thought to be very important for its successfully infecting host cells. Receptors on host cell membrane have the characters of specificity, high affinity, limited binding sites. Identification of the cellular macromolecules responsible for virus infection is useful to understand its life cycle and pathogenesis, to take measure of prevention and therapy of virus disease[1].

Hepatitis virus (HV) infection is a major worldwide public health problem. Despite considerable advances in the understanding of natural history of HV disease, the process of its replication and antigenic structure, most of the early steps in the virus life cycle remain unclear for lack of an in vitro infection system. HV attaching to permissive cells, fusing and penetrating through cell membranes and releasing subsequent genome are largely a mystery. Current knowledge on the early steps of HV infection shows that HV binding to the special reptors on human hepatocyte membrane via preS1 domain of large surface protein (LHVs) and then fusing with cell membrane trigger the viral infec-

tion. Thus the special recptors binding to HV is thought to be a potential target for the development of novel antivirals. Though many cellular molecules have been identified as putative receptors for HV attachment, the cellular protein has not been found yet. So identifying hepatitis virus successfully and effectively has become a hot spot research field for the researchers related[2].

LITERATURE REVIEW

Currently, technologies for HV identification at home and abroad include the following major methods. (1) Aerobic plate count (APC) is a method to detect the total number of virus generally used at home and abroad, which is to dilute the food to be detected into 2-3 kinds of appropriate dilution under sterile conditions, culture the same for 48 hours in the incubator of 36 °c, and carry out the counting of colony count artificially after taking out. The method is accurate in counting, simple in operating, easy in mastering, thus widely used till now. However, the method needs long detection time, high detection cost, and high detection environment requirement; (2) ATP luminous detection is to determine whether there is microbial infection in the sample food and the quantity of HV through the detection of the content of ATP substances in sample container. The method is easy and convenient in operating, but the detection results may have certain errors sometimes as it is unable to directly detect the food; (3) HV identification based on artificial intelligence technology is mainly the classifier technology based on feature recognition. The method has such advantages as high automatic degree, less manual intervention, high recognition precision, and etc. But different identification methods may have certain technical defects, for example, neural network is easy to fall into local optimization in the case of specific computation, thus leading to slow rate of convergence and long computing time[1,2].

The paper improve BP neural network with integration of ant colony algorithm to overcome the question of slow convergence speed of BP neural network and presents some improvements of genetic algorithm. In so doing a new algorithm for HV identification is advanced and try to speed up algorithm convergence and simplify algorithm structure.

FLOW CHART DESIGN

Flow chart of HV identification flow designed in the paper is as shown in figure 1, the steps of which are: (1) Collection of image information; (2) impurities processing on images, i.e. images preprocessing ; (3) adopting image segmentation technique to separate images and extract the targets to be recognized; what is used in the system is watershed algorithm; (4) making use of feature extraction technique to carry out feature extraction and optimization selection on feature vectors, thus feature vectors can be transferred to classifier to be recognized. Based on the self-learning and easy realization of programming of BP neural network, this thesis adopts BP neural network classifier. The designed the structure of water HV identification see figure 1.

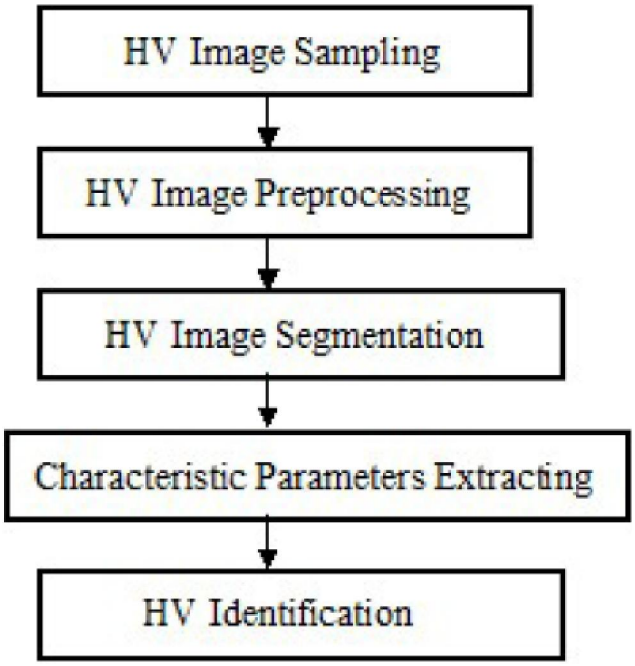

Figure 1 : The designed flow chart

The design of BP neural network shall be combined with projects. In this system, as what we need to solve is nonlinear problems in the closed interval, which can be solved with a hidden layer, three-layer BP neural network is hereby adopted. After extracting plenty of feature information of five kinds of characteristics, make use of conventional reservation method to train and detect classifier, where two thirds of the data are used to train the network, and one third of the data are used to detect the network, so as to detect the recognition rate.

ALGORITHM DESIGN

Working principle of BP algorithm

Up till now, hundreds of artificial neural network models are put forward from different views of research, among which multi-hierarchy feed forward error back propagation BP neural network is the most-widely used network model in actual research. Basic three-layer BP neural network structure is shown as Figure 2[3].

From the picture we can see that three- layer BP neural network is mainly comprised of input layer, hidden layer and output layer. Adjustable weight ω connects the layers. There can be several hidden layers, forming multi-layer BP neural network. The input of BP neural network is recorded as $x_i(k)$, the actual output of network is recorded as $y_j(k)$, the ideal output of network is recorded as $Y_i(k)$, the subscripts i, j indicate the nodes of input layer of network respectively, and k is the running iterations of BP neural network. Its approximation error is defined as Formula 1 in which L is the quantity of output layer nodes; in this way, the function characteristic of BP neural network can be described as Formula 2[4].

$$E = \frac{1}{2}\sum_{j=1}^{L}(Y_j(k) - \gamma_j(k))^2 \quad (1)$$

$$\gamma_j(k) = f(x_i(k), \omega) \quad (2)$$

In Formula 2, function f is obtained through the composition of weights of each network layer and node function, generally being very complicated non-linear function BP neural network training is to dynamically adjust the connecting weight ω to make Formula 3 workable. The learning of weight ω adopts the fastest grads descent principle, i.e. the variable quantity of weights is in proportion to the negative gradient direction of approximation error E. See reference 2 for specific calculation.

$$\lim_{k\to\infty} E = \lim_{k\to\infty} \frac{1}{2}\sum_{j=1}^{L}(Y_j(k) - \gamma_j(k))^2 = 0 \quad (3)$$

Ant colony algorithm design

Ant colony algorithm is a random search algorithm addressed by Scholar Dorigo and others from Italy in the 1990s, which solves TSP by manual simulation of ants search process, and achieves better results. It is an another intelligent heuristic search algorithm applicable for combinatorial optimization problems after the simulated annealing algorithm, genetic algorithm, tabu search algorithm, neural network algorithm, etc. Not only can Ant colony algorithm perform intelligent search and global optimization, but also have the features of robustness, positive feedback, distributed computation and easy combination with other algorithms. At the same time, such characteristics as discreteness and parallelism of ant colony algorithm are very applicable to deal with digital image, and its clustering features and image recognition process have greater similarities[6].

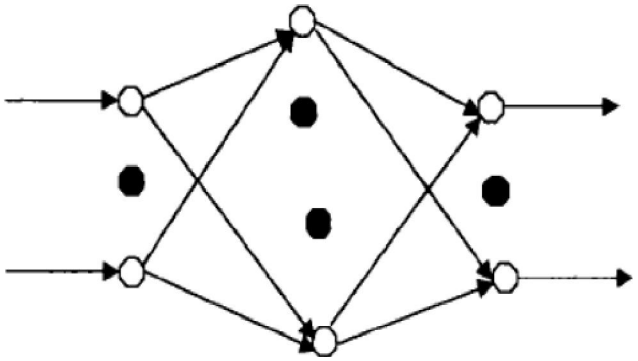

Figure 2 : Basic structure of BP algorithm

The study found that ants could leave a material called pheromone in the path to guide their own movement direction, and tend to move toward the high-strength material. Therefore, the collective behavior of a large number of ants can display the positive feedback of information: the more there are ants in certain path, the greater the probabilities that the subsequent ants select this path. Meanwhile the pheromone will disappear gradually over time, so the pheromone strength has the relations with the path length and the number of ants in the path. Ants search foods by such information exchange. Ant colony algorithm is a process to simulate the real ants to look for foods.

But ant colony algorithm uses a random selection strategy in the process of construction solution. This selection strategy reduces evolutionary rate and easily causes stagnation, that is, the solutions that all individuals found are exactly the same not to search the solution space further and find a better solution after the search reached a certain extent. As for the problem,

FULL PAPER

this paper adopts the selection strategy of the dynamic adjustment to improve the overall search speed and capabilities of ant colony algorithm.

The process of ants foraging is also a constant clustering process actually. Things are the cluster center. Ant colony algorithm is applied in the clustering problem, of which, main ideas are listed as follows.

If $X=\{X_i|X_i=(x_{i1},x_{i2},\cdots,x_{im}),i=1,2,\cdots,N\}$ is the data set to analyze the clustering, r is the cluster radius, $ph_{ij}(t)$ is the pheromone concentration in the path from data X_i to data X_j in the time point t, d_{ij} is the Euclid distance with weight of the samples and the cluster center, and p is the weighted factor that can be determined according to the impact degree in the re-clustering of each component. See Formula 4.

$$d_{ij}=\left\|P(X_i-X_j\right\|=\sqrt{\sum_{k=1}^{m}p_k(x_{ik}-x_{jk})^2} \quad (4)$$

If $ph_{ij}(0)=0$ is the initial information quantity, see Formula 5.

$$ph_{ij}(0)=\begin{cases}1 & d_{ij}\le r\\ 0 & d_{ij}\succ r\end{cases} \quad (5)$$

If $P_{ij}(t)\ge P(0)$, X_i is merged to X_j. If $C_j=\{X_k|d_{kj}\le r,k=1,2,\cdots J\}$, C_j is all data sets merged to X_j. See Formula 6 for the optimal cluster center.

$$\overline{C_j}=\frac{1}{J}\sum_{k=1}^{J}X_k \qquad X_k\in C_j \quad (6)$$

Through overall search improvement

Dynamic adjustment strategy: Stagnation is the fundamental cause resulting in the inadequacy of ant colony algorithm. Based on the deterministic and random selections, this paper adjusts the transition probability dynamically to build the selection strategy more conducive to the overall search[8].

The pheromone in the path occurs continuous change in the evolutionary process. The pheromone of better solution searched is strengthened to increase the selection possibility of next iteration, and some better solutions is forgotten gradually because fewer ants pass in the start-up phase so as to affect the overall search capabilities of the algorithm. If the ants are stimulated properly to try the path occasionally in the selection strategy, it is conducive for the overall search of the solution space. Thus, the inadequacy of basic ant colony algorithm is overcome effectively. See Formula 7 and Formula 8 for the improved selection strategy in this paper. When $q\le q_0$, $j\in allowed_k$ Formula 7 is used and $j\in allowed_k, others$ Formula 8 is used.

$$P_{ij}^{k}(t)=\arg\max\{\left|\tau_{ij}(t)\right|^{\alpha}\cdot\left|\eta_{ij}(t)\right|^{\beta}\} \quad (7)$$

$$P_{ij}^{k}(t)=\frac{\left[\tau_{ij}(t)\right]^{\alpha}\cdot\left[\eta_{ij}\right]^{\beta}\cdot X_{ij}(t)}{\sum_{k\in allowed_k}\left[\tau_{ik}(t)\right]^{\alpha}\cdot\left[\eta_{ik}\right]^{\beta}\cdot X(t)} \quad (8)$$

In Formula 7 and Formula 8, X_{ij} meets the requirements of Formula 9.

$$X_{ij}=\frac{m\cdot N_c}{m\cdot N_c+\delta\cdot Q_c(i,j)\cdot\eta(i,j)/\max\eta} \quad (9)$$

In Formula 9, m is the number of ants, N_c is the number of current iterations, $\max\eta$ is the maximum of heuristic function $\eta(i,j)$,and $Q_c(i,j)$ is total number of ants in the current path (i,j) from the first iteration. Q_c and η are considered in X. When previous iteration tends to suboptimal solution, the number $Q_c(i,j)$ of ants increases and its X value decreases constantly in spite of constant increase of the pheromone in the suboptimal solution. Therefore, another selection of the path can restrain the excessive increase of the pheromone to cause premature convergence, and is conducive to global convergence[8,9].

EXPERIMENTAL VERIFICATION

Data collection and preprocessing

This paper takes the hepatitis B virus for example and adopts identification sample from three Chinese hospital to built sample database.

The design of BP neural network identification shall be combined with projects. In this system, as what we need to solve is nonlinear problems in the closed interval, which can be solved with a hidden layer, three-layer BP neural network is hereby adopted. After extracting plenty of feature information of HVV, make use of conventional reservation method to train and detect method, where two thirds of the data are used to train the network, and one third of the data are used to detect the network, so as to detect the recognition rate.

The system is to recognize the characteristics HVB according to its practical environment. This paper selects 10 characteristics in total, so 10 input neurons of network are required. Due to the feature of S -type function in BP neural network, the characteristic values shall be normalized, and the normalized values shall be limited within [0.1,0.9] As the system here is mainly to recognize five characteristics of HVB, the number of neurons of network output layer is designed as 5. Binary five-digit is used to indicate the output. (10000), (01000), (00100), (00010) and (00001) represent five characteristics respectively. Thus the mathematical corresponding relations of BP neural network are established. Input a ten-digit vector and output a five-digit binary number through function corresponding relations of network, so as to distinguish the classification of five kinds of micro-organism. Therefore, we have designed the rough input-output form of BP network classifier required by the recognition system of water microorganisms by and large[7].

Experimental results and analysis

As for the performance of the presented algorithm, this paper also realizes the application of the improved BP neural network, the ordinary BP neural network[9] and ordinary ant colony algorithm[8], evaluation performance of different algorithms is shown in TABLE 1. In table 1 evaluation results of training effects of different students are selected and compared with artificial evaluation to calculate the evaluation accuracy. And the calculation platform as follows: hardware is Dell Poweredge R710, in which processor is E5506, memory 2G, hard disk 160G; software platform is Windows XP operating system, C programming language environment.

TABLE 1 : The Application Performance of Different Algorithms

Algorthm	Improved Algorithm	Ordinary BP Algorithm	Ordinary Genetic algorithm
Accuracy Rate	95.77 %	88.66%	68.82%
Time Consuming	16s	571s	24s

CONCLUSION

The identification HV successfully and effectively plays a key role in clinical diagnosis. So, this paper, on the consideration of actual characteristics of identification HV, designs a flow chart for HV identification, and put forward a new HV identification model based on improving ant colony algorithm according to the identification requirement of multi-factor complicated system. Test results indicate the engineering practicability of the model on HV identification. In the next study, we shall pay attention to the combination of generality with individuality of HV identification.

REFENENCES

[1] D.N.Frick, A.M.Lam; The nonstructural protein helicase requires an intact protease domain. J.Biot. Chem., **279**, 1269-1280 **(2012)**.

[2] A.J.Syder, H.K.Lee; Small molecule scavenger receptor BI antagonists are potent HCV entry inhibitors, J.Appl.Viru., **54**, 48-55 **(2011)**.

[3] M.Patricia, H.F.Victor; Genetic optimization of neural networks for person recognition, TELKOMNIKA., **10**, 309-320 **(2012)**.

[4] X.Q.Wang; Study on genetic algorithm optimization for support vector machine. J. Info Sci., **4**, 282-288 **(2012)**.

[5] X.S.Yan, Q.H.Wu; An improved ant colony algorithm and its application. TELKOMNIKA., **10**, 1081-1086 **(2012)**.

[6] D.M.Zhang, S.Q.Wang; Minimizing network coding resources with BP algorithm. J.Comp., **7**, 464-473 **(2013)**.

[7] D.H.Li; Aquaculture monitoring system design based on BP neural network algorithm. J.Soft., **5**, 479-488 **(2013)**.

[8] S.F.Ding, W.K.Jia; An improved ant colony algorithm based on factor analysis. J.Conv. Tech., **5**, 103-108 **(2010)**.

[9] B.Z.Yang, G.Tian; Research on customer value classification based on BP algorithm. J.Mana Sci., **23**, 168-175 **(2011)**.