Data Mining on DNA Sequences of Hepatitis B Virus

1) Mr. Kiran P. Jagtap
Annasaheb Dange college of Engg. & Tech, Ashta.

2) Mr. A. B. Rajmane
Ashokrao Mane Group of Institution,Vathar,Vadgaon.

Abstract

Extraction of meaningful information from large experimental data sets is a key element in bioinformatics research. To identify genomic markers in Hepatitis B Virus (HBV) that are associated with Hepato Cellular Carcinoma (HCC) is one of the challenging task. An architectural framework of data mining which includes several molecular evolution analysis, clustering, feature selection, classifier learning, and classification algorithms. In the feature selection process, genetic markers are selected based on information gain theory for further classifier learning. Then, meaningful rules are learned by the algorithm is called the Rule Learning evolutionary algorithm. Also, a classification method by nonlinear integral has been developed. Good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral. The non-additively of the fuzzy measure reflects the importance of the feature attributes as well as their interactions.

1. Introduction

In Asia, infection of Hepatitis B virus (HBV) is a major health problem. At least 10 percent of the Chinese populations (120 million people) are HBV carriers, and up to 25 percent of HBV carriers died as a result of HBV related complications including liver cirrhosis and Hepato Cellular Carcinoma (HCC), i.e., liver cancer. Chronic infection by the HBV causes an increased risk of hepato cellular carcinoma (HCC).

In this paper the focus is to look into the clinical data prepared by the clinicians, and the HBV DNA genomes prepared by the biochemists of research group. Patients taking part in this study were selected by the clinicians carefully, according to their age, sex, and past clinical status. To reduce the noise of genotypic difference among the sequences collected, it proposes to analyze these DNA examples in each genotype separately. It identifies genetic marker(s) for liver cancer (HCC) from HBV DNA sequences.

Classification is one of the most studied data mining tasks. The objective is to predict the value (the class) of a user-specified goal attribute based on the values of other attributes, called the predictive (feature) attributes. The goal attribute might be the prediction of whether or not a patient has cancer, while the predictive feature attributes might be the mutation sites of the patient’s virus DNA. The aim of this study is to develop a data mining framework which contains an appropriate classifier for liver cancer based on HBV DNA and clinical data. The proposed work develops two new algorithms based on rule learning (RL). Further carry out a thorough comparative study on these two new models with existing classifiers. The classification model should have high sensitivity and acceptable accuracy and specificity for HCC diagnosis and prediction.

2. Related work

A case control study from Taiwan [2] suggested that genotype C HBV is more closely associated with cirrhosis and HCC in those who are older than 50 years, whereas genotype B is more common in patients with HCC aged less than 50 years [3]. Previous cohort study of 426 cases of chronic hepatitis B patients reviewed a higher risk of HCC and liver cirrhosis in genotype C infection [4]. The aim of this study is to find the genomic markers of the HBV and clinical information which are useful in predicting occurrence of liver cancer and response to therapy.

The extent and position of the single-stranded gap in DNA molecules from Dane [6] particles isolated from two donors of the serotype were determined by molecular hybridization and electron microscopic methods. The results showed that in each preparation more than 99% of the circular molecules are of uniform length and contain both single- and double-stranded regions. There are several common classification models such as Naive Bayesian Network [7], Decision Tree, Neural Networks, and Rule Learning using Evolutionary Algorithm [8]. The learning processes of Naive Bayesian Networks and Decision Tree are faster. However, they cannot cope well with feature interactions. Neural Networks are treated as black box learning and it is difficult for a human to understand or interpret the classification explicitly.

3. Proposed work

It includes following modules as below.

3.1 Data Collection for Genome Sequences

To collect HBV DNA sequences, either genotype B or C, from over 20 patients specifically for this dissertation. In the molecular evolution analysis and clustering, three subgroups have been identified in genotype C and a clustering method has been developed to separate the subgroups.

3.2 Clustering and splitting of datasets

To split of dataset for training and testing purpose based on 90-10 or 80-20 rule.

3.3 Feature Selection Algorithm

The main purpose of feature selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy. The main advantage of this method is that it produces a hierarchy of feature subsets with the best selection for each dimension.

3.4 Classification Algorithm

Make use of various algorithm based on Neural network, support vector machine & decision tree for Classification purposes.

3.5 Testing and Validation Algorithm

Create Testing dataset and apply the testing algorithm.

3.6 Executing the Hepatitis B Disease Detection using Neural Network

An Artificial Neural Network (ANN), or commonly just called neural network (NN), is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connection approach to computation. In most cases, an NN is an adaptive system that changes its structure or weights of the interconnections based on external and internal information (stimuli) that flows through the network. In more practical terms, NNs are nonlinear statistical data modeling for decision-making and classification tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. However, it is essentially a black box approach.

3.7 Executing the Hepatitis B Disease Detection using Support Vector Machine

SVMs are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper plane in that space, which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper planes are constructed, one on each side of the separating hyper plane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the neighbouring data points of both classes, since, in general, the larger the margin the smaller the generalization error of the classifier.

3.8 Executing the Hepatitis B Disease Detection using Decision Tree

A decision tree is a tree-structured classifier. The Decision Tree method learns decision tree using a recursive tree growing process. Each test corresponding to an attribute is evaluated on the training data using a test criteria function. The test criteria function assigns each test a score based on how well it partitions the data set. The test with the highest score is selected and placed at the root of the tree. The sub trees of each node are then grown recursively by applying the same algorithm to the examples in each leaf. The algorithm terminates when the current node contains either all positive or all negative examples.

3.9 Comparison Study for Different Platform of Disease Detection

Compare the obtained results from the different classification algorithm & shows which genotype is more responsible for disease detection.

4. Data mining framework

The typical data mining framework for genome sequencing and disease detection can be represented as shown below in the figure-1. The framework comprises of Training and Testing Algorithms which are preceded by splitting of datasets of training and testing purposes. In which it includes the different algorithms such as feature selection algorithm, molecular evolution analysis, clustering & several classification algorithms then evaluates it & validate the results & compare the results obtained by different algorithms & determine which genotype is more responsible for disease detection.

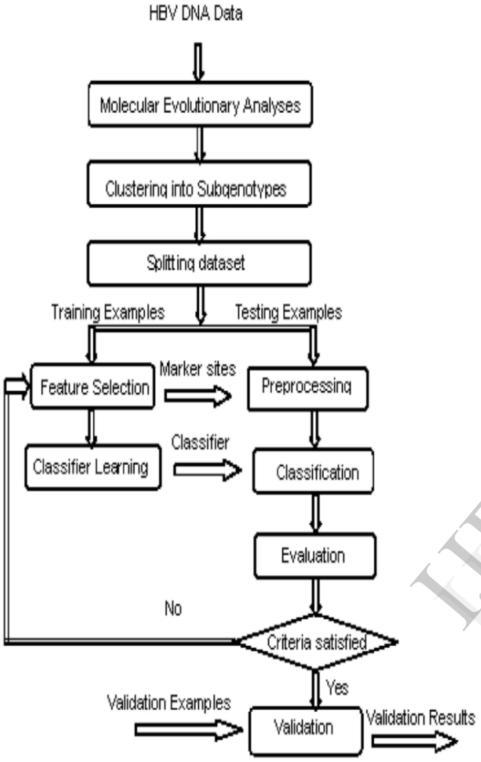

Fig.:- Data Mining Framework for genome sequences

5. Acknowledgments

Mr. Kiran. P. Jagtap, Author Department of Computer Science & Engineering, Annasaheb Dange college of Engg. & Tech., Ashta thanks Mr. A. B. Rajmane, project guide Department of Computer Science & Engineering, Ashokrao Mane group of institutions, Vathar Vadgaon, for his precious guidance, encouragement and continuous monitoring throughout the presented work.

6. References

[1] Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, Eddie Y.T. Ng, Henry L.Y. Chan, Stephen K.W. Tsui, Tony S.K. Mok, Chi-Hang Tse, and Joseph Jao-Yiu Sung, " Data Mining on DNA Sequences of Hepatitis B Virus" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 2, MARCH/APRIL 2011.

[2] R. P. Beasley, L. Y. Hwang, C. C. Lin, and C. S. Chien, "Hepatocellular Carcinoma and Hepatitis B Virus. A Prospective Study of 22 707 Men in Taiwan," Lancet, vol. 2, pp. 1129-1133, 1981.

[3] J. H. Kao, P.J. Chen, M. Y. Lai, and D.S. Chen, "Hepatitis B Genotypes Correlate with Clinical Outcome in Patients with Chronic Hepatitis B," Gastroenterology, vol. 118, pp. 554-559, 2000.

[4]H.L.Y.Chan et al., "Genotype C Hepatitis B Virus Infection Is Associated with an Increased Risk of Hepatocellular Carcinoma," Gut, vol. 53, pp. 1494-1498, 2004.

[5]Sequence of Hepatitis B Virus DNA Incorporated into the Genome of a Human Hepatoma Cell Line MARILYN ZIEMER, PABLO GARCIA, YOSEF SHAUL,t AND WILLIAM J. RUTTER* Department of Biochemistry and Biophysics, Hormone Research Laboratory, University of California, San Francisco, California 94143 Received 9 July 1984/Accepted 20 November 1984

[6]M.L. Wong and K.S. Leung, "Genetic Logic Programming and Applications," IEEE Expert, vol. 10, no. 5, pp. 68-76, Oct. 1995.

[7] D.M. Chickering, D. Heckerman, and D. Geiger, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," Machine Learning, vol. 20, pp. 197-243, 1995.

[8] A.A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery," Advances in Evolutionary Computation, A. Ghosh and S. Tsutsui, eds., Springer-Verlag, 2002.

[9] E. Orito et al., "Geographic Distribution of Hepatitis B Virus (HBV) Genotype in Patients with Chronic HBV Infection in Japan," Hepatology, vol. 34, pp. 590-594, 2001.

[10] K.S. Leung, Y.T. Ng, K.H. Lee, L.Y. Chan, K.W. Tsui, T. Mok, C.H. Tse, and J. Sung, "Data Mining on DNA Sequences of Hepatitis B Virus by Nonlinear Integrals," Proc. Taiwan-Japan Symp. Fuzzy Systems & Innovational Computing, Third Meeting (Keynote Speech), pp. 1-10, Aug. 2006.