ISSN 2277-3061

Identification Of Hepatocellular Carcinoma Using Supervised Learning Algorithms

Sagri Sharma, Sanjay Kadam, Hemant Darbari

Centre for Development of Advanced Computing, Pune, India

sagris@cdac.in

sskadam@cdac.in

darbari@cdac.in

ABSTRACT

Analysis of diseases integrating multi-factors increases the complexity of the problem and therefore, development of frameworks for the analysis of diseases is an issue that is currently a topic of intense research. Due to the inter-dependence of the various parameters, the use of traditional methodologies has not been very effective. Consequently, newer methodologies are being sought to deal with the problem.

Supervised Learning Algorithms are commonly used for performing the prediction on previously unseen data. These algorithms are commonly used for applications in fields ranging from image analysis to protein structure and function prediction and they get trained using a known dataset to come up with a predictor model that generates reasonable predictions for the response to new data.

Gene expression profiles generated by DNA analysis experiments can be quite complex since these experiments can involve hypotheses involving entire genomes. The application of well-known machine learning algorithm - Support Vector Machine - to analyze the expression levels of thousands of genes simultaneously in a timely, automated and cost effective way is thus used.

The objectives to undertake the presented work are development of a methodology to identify genes relevant to Hepatocellular Carcinoma (HCC) from gene expression dataset utilizing supervised learning algorithms & statistical evaluations along with development of a predictive framework that can perform classification tasks on new, unseen data.

Indexing terms/Keywords

Artificial Intelligence, Biomarker, Gene Expression Datasets, Hepatocellular Carcinoma, Machine Learning, Supervised Learning Algorithms, Support Vector Machine.

Academic Discipline And Sub-Disciplines

Applied Artificial Intelligence, Data Mining algorithms

SUBJECT CLASSIFICATION

Supervised learning algorithms, Support Vector Machine etc.

TYPE (METHOD/APPROACH)

Supervised methods represent a powerful approach as the prior knowledge is used to learn the method. There are reports describing multi-class classification of heterogeneous tumour types where support vector machine (SVM) method achieves the best performance. The ultimate goal of the efforts will be to decrease HCC aggressiveness and increase patient survival.

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol. 14, No. 3

www.ijctonline.com , editorijctonline@gmail.com

INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the most deadly problems worldwide. Scientists have been studying the molecular mechanism of HCC for years, but the understanding of it remains incomplete and scattered across the literature at different molecular levels. Chromosomal aberrations, epigenetic abnormality and changes of gene expression have been reported in HCC. Large amounts of data on genetic and epigenetic abnormalities, gene expression profiles, microRNA expression profiles and proteomics have been accumulating, and bioinformatics is playing a more and more important role.

The objectives to undertake the presented work are development of a methodology to identify genes relevant to HCC from gene expression dataset utilizing supervised learning algorithms & statistical evaluations; and to develop a predictive framework that can perform classification tasks on new, unseen data.

The current work shall prioritize potential disease-related genes using supervised learning algorithm support vector machine [1], to analyze the training data and produce an inferred function, which can be used for mapping new cases. An optimal scenario will allow the algorithm to correctly determine the class labels for unseen instances. The predictive framework will be able to classify microarray expression profiles into cancerous & non-cancerous and can be used as a diagnostic tool for the same.

In liver cancer, exploring gene expression patterns of samples from healthy patients and others infected with HCC has revealed a significant difference in the expression of some genes from normal to tumour samples. Genes having high variance between both classes of samples in their expression are informative features that should be used in any further analysis as suggested biomarkers. Early detection of HCC is a vital issue as it can help patients in receiving therapeutic benefits rather than curative surgery.

Briefly, an ideal biomarker is defined as a gene that has two discrete values, zero in normal samples and one in HCC samples. Genes with similar profile to the ideal biomarker are selected. Then they are ranked according to their similarity with the ideal gene using InfoGain Evaluator in WEKA™ that evaluates the worth of an attribute by measuring the information gain with respect to the class.

The paper elaborates upon the work done in the detection of Hepatocellular Carcinoma biomarkers through study of gene expression datasets using supervised learning algorithms.

The methodology chosen to reach the objectives is presented through the idea for attempting the problem, the dataset source, its various aspects, and the mechanism to derive the output etc.

The results and observations acquired through the methodology application are also presented.

HCC and Machine Learning

Increased understanding of cancer biology and technological advances have enabled identification of a multitude of pathological, genetic, and molecular events that drive hepato-carcinogenesis leading to discovery of numerous potential biomarkers in this disease. They are currently being aggressively evaluated to establish their value in early diagnosis, optimization of therapy, reducing the emergence of new tumors, and preventing the recurrence after surgical resection or liver transplantation [2].

HCC Biomarkers are molecular indicators [3] of biological status, detectable in blood, urine, or tissue, useful for the clinical management of various disease states. Time and money can be saved by avoiding broad treatment approaches to diseases of particular organs or systems, and ideally, biomarkers could serve as a measurement tool to detect disease presence and progression and to guide more targeted therapy [4, 5]. HCC can benefit from tumor biomarkers' diagnostic, therapeutic, and prognostic capabilities.

With advances in understanding of tumor biology [6], along with the development of cellular and molecular techniques, the role of biomarkers related to early detection, invasiveness, metastasis, and recurrence has attracted great deal of research interest resulting in discovery and utilization of several novel markers in this disease [7].

Machine learning is an interdisciplinary field of research with influences and concepts from, e.g., mathematics, computer science, artificial intelligence, statistics, biology, psychology, economics, control theory, and philosophy. In general terms, machine learning concerns computer programs that improve their performance at some task through experience [8], i.e., programs that are able to learn. The machine learning field has contributed with numerous theoretical results, learning paradigms, algorithms, and applications.

Supervised machine learning algorithms are commonly used for many classification and regression tasks. Typically, a supervised algorithm, based on some input data provided to it, develops a model based on some mathematical and statistical methodologies. This model can be then used for performing the prediction on previously unseen data. A learning algorithm trains by observing known data, i.e., instances for which there is a correct supervisor response. According to some internal bias, it then generates a classifier. This classifier can then be used to classify new data of the same kind.

In supervised learning, the objective is to learn from examples, i.e., generalize from training instances by observing a number of inputs and the correct output [9].

Prior research suggests that among well-established and popular techniques for multi-category classification of micro-array gene expression data, support vector machines (SVMs) achieve the best classification performance, significantly

outperforming k-nearest neighbors, back propagation neural networks, probabilistic neural networks, weighted voting methods, and decision trees [10].

Related Work - Literature Survey

Microarray-based comparative genomic hybridisation has further increased the reliability and significance of the biological and clinical conclusions drawn from gene expression profiles. This will be the basis for developing new targeted therapies, an urgent need to reduce the mortality from hepatocellular carcinoma [11].

Expressing a gene means manufacturing its corresponding protein, and it has two major steps. In the first step, the information in DNA is transferred to a messenger RNA (mRNA) molecule by way of a process called transcription. During transcription, the DNA of a gene serves as a template for complementary base-pairing, which is then processed to form mRNA. The resulting mRNA is a single-stranded copy of the gene, which next must be translated into a protein molecule. During translation, the mRNA is "read" according to the genetic code, which relates the DNA sequence to the amino acid sequence in proteins. The mRNA sequence is thus used as a template to assemble the chain of amino acids that form a protein.

Gene expression dataset contains expression profiles for multiple samples. While the software supports multiple input file formats for these datasets, the tab-delimited GCT format is the most common. The first column of the GCT file contains feature identifiers. The second column contains a description of the feature (optional); Subsequent columns contain the expression values for each feature, with one sample's expression profile per column.

Supervised methods represent a powerful approach as the prior knowledge is used to learn the method. There are reports describing multi-class classification of heterogeneous tumour types where support vector machine (SVM) method achieves the best performance [12]. The ultimate goal of the efforts will be to decrease HCC aggressiveness and increase patient survival.

Supervised learning takes a known set of input data and known responses to the data, and seeks to build a predictor model that generates reasonable predictions for the response to new data Figure 1.

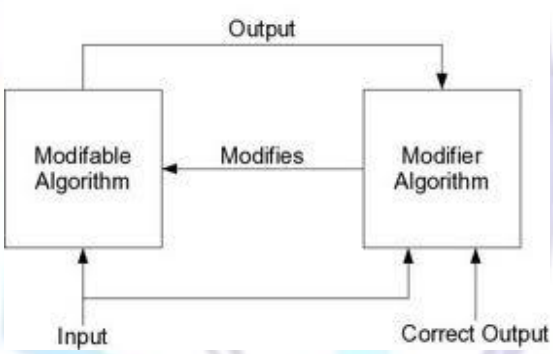

Fig. 1 Predictor Model for Response to new Data

The core of an SVM is a quadratic programming problem (QP), separating support vectors from the rest of the training data [13].

METHODOLOGY

Support Vector Machine – Algorithm

The goal is to use supervised learning to classify and predict HCC, based on the gene expressions collected from DNA analysis. Known sets of data will be used to train the machine learning protocol SVM to categorize patients according to their prognosis. The outcome of this study will provide information regarding the efficiency of SVM method the machine learning technique.

The basic tool used is SVM (Lib SVM) classifier which employs a set of mapping functions to map the input data into the reproducing kernel. The efficiency of classification depends on the type of kernel function that is used. So here analysis of the performance of various kernel functions used for classification purpose is done.

Specifically, SVM chooses the hyperplane that provides maximum margin between the plane surface and the positive and negative points. The separating hyperplane is optimal in the sense that it maximizes the distance from the closest data points, which are the support vectors [14].

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Kernel functions, which represent a dot product of input data points mapped into the higher dimensional feature space, that can be used are of the following types [15]:

1. Lineal Kernel

2. Multilayer perceptron kernel
3. Polynomial Kernel
4. Radial Basis function

The Classification Tool - R analysis

The R implementation is based on the S3 class mechanisms. It basically provides a training function with standard and formula interfaces, and a predict() method. In addition, a plot() method for visualizing data, support vectors, and decision boundaries is provided. Hyperparameter tuning is done using the tune() framework, which performs a grid search over specified parameter ranges.

There are five packages that implement SVM in R.

- e1071 [16]
- kernlab [17]
- klaR [18]
- svmpath [19]
- shogun [20]

For this work, e1071 package implementation is used because it is most intuitive.

The dataset

A HCC cancer dataset obtained from http://smd.princeton.edu/ cgi-bin//publication/viewPublication.pl?pub_no=107 & http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3500. There is a large collection of free microarray data sets for HCC and other cancers over the NCBI website.

The data for this experiment have already been normalized to the scale of (-4, 4).

The Method

As with any supervised learning model, first training of a support vector machine is done and then the classifier is cross validated. The trained machine is used to classify (predict) new data. Various SVM kernel functions can be used to tune the parameters, in order to obtain satisfactory predictive accuracy [21].

Therefore, the steps to be followed are:

- Training an SVM Classifier
- Classifying New Data with an SVM Classifier
- Tuning an SVM Classifier

The first row contains the labels and the remaining rows are the genes for each samples. Each column contains one sample. In the first row, for the labels, 1 stands for HCC and 0 stands for Liver sample (i.e. non-cancer). The dataset under consideration comprises of 153 samples with gene expressions levels as 24192. Due to the presence of several null values, the genes which have at least one “NA” in place of numeric number, is eliminated. So the dimension of table is reduced to 153 X 2915.

To create the trainset, 50 HCC samples and 50 non-HCC samples were extracted along with 2915 gene expression levels, after removing the NULL values. Thereafter, merging 50 hcc and 50 nohcc to form training set with 100 samples with 2915 gene expression levels. This section makes the Module 1.

So, Module 1, with a dataset of 100x2915 as train set and remaining 53x2915 as test set shall be classified using SVM and the results are produced. The SVM model is then tested using the test set and the exercise for module 1 is considered executed.

Module 2 is the dataset generated after extracting top set of genes on the evaluation of InfoGain Feature selection is considered to be an important step in the analysis of HCC.

After running the iterations for 20, 50, 100, 150, 200, 250 set of top genes, the set of top 100 genes gave the best output. Therefore, for module 2 top 100 genes set were considered.

Carrying out feature selection reduces the curse of dimensionality problem and improves the interpretability of the problem. Numerous feature selection methods have been proposed [22] and these methods rank the genes in order of their relative importance. Feature ranking methods namely information gain (InfoGain) for finding the minimum number of genes from gene expression dataset to achieve the high classification accuracy, is used for the present work.

Therefore, taking all top100 genes based on InfoGain, and creating a SVM Model with trainset of 100x101 (100 samples with 50 HCC and 50 non-HCC samples and 100 gene expressions).

The model, thus created, is tested upon by the 53x101 testset and result is obtained.

To obtain satisfactory predictive accuracy, tuning of the parameters of the kernel function is done by using svm.tune (function).

After training, the structured SVM model allows one to predict for new sample instances the corresponding output label; that is, given a natural language sentence, the classifier can produce the most likely parse tree. svm.predict() function is also used.

A confusion matrix [23] contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The table 1 shows the confusion matrix for a two class classifier [24].

A cross-tabulation of the true versus the predicted values yields the confusion matrix.

Table 1: Confusion Matrix

	Predicted	
Actual	True Positive(TP)	False Negative(FN)
	False Positive(FP)	True Negative (TN)

where:

TP: true positive, i.e. malign instances predicted rightly

FP: false positive, i.e. benign instances predicted as malign

TN: true negative, i.e. benign instances predicted rightly

|N|: total of benign instances

|P|: total of malign instances

Sensitivity = TP ÷ |P|

Specificity = TN ÷ |N|

Precision = TP ÷ (TP + FP)

Conclusions Results and Discussions

In this study the classifiers are tested with a high dimensional dataset for hepatocellular-carcinoma-data comprises of 153 samples with gene expressions levels as 24192. To create the train set, 50 HCC samples and 50 non-HCC samples were extracted along with 2915 gene expression levels, after removing the NULL values. The R command used to remove columns having NULL values and the rest of the samples were made as test data for creation of confusion matrix using model constructed through SVM. Out of the four kernel functions of SVM, The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

For this work, optimization was done using the four kernels and the results shown below in table 2 indicates that polynomial kernel is the best suited for our dataset.

Table 2: Optimization for kernel function using SVM

	Sensitivity	**Specificity**	**Precision**	**Support Vectors**
Linear	0.85	1.0	0.85	53
Polynomial	0.925	1.0	0.925	96
RBF	0.85	1.0	0.85	53
Sigmoidal	0.875	1.0	0.875	54

SVM parameters like Cost and Gamma are considered vital for accuracy point of view. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. As the polynomial kernel is the most fitting for the current work, the comparative counter for complete range of gamma and cost parameters is given in table 3.

Table 3: Gamma, Cost Parameter for Polynomial kernel

Gamma	C=10	C=100
1 e – 06	0.46	0.06
1 e – 05	0.06	0.04
1 e – 04	0.01	0.01
1 e – 03	0.09	0.08
1 e – 02	0.66	0.68
1 e – 01	0.67	0.69

By using the dataset with 153x2915 dimension and polynomial kernel function, the results obtained as mentioned here.

For Module 1:

Confusion table:

	True	
Predicted	HCC	non-HCC
HCC	34	0
Non-HCC	6	13

Sensitivity = 34 / (34+6) = 0.85

Specificity = 13 / (0+13) = 1.00

Precision = 34 / (34+6) = 0.85

For Module 2:

Confusion table

	True	
Predicted	HCC	non-HCC
HCC	35	0
non-HCC	5	13

Sensitivity = 35 / (35+5) = 0.875

Specificity = 13 / (0+13) = 1.00

Precision = 35 / (35+5) = 0. 875

Therefore, when we picked up the top 100 gene through the process of InfoGain Evaluator, the sensitivity as well as precision of the model increases by 0.025, which is a substantial increment.

Figure 2: Comparison between 2915 features and top 100 features

As a result of computational analysis by the way of tuning and preprocessing, near perfect classification of genes is achieved, but not with high confidence. Identification and analyses of a subset of genes from the HCC dataset whose expression is highly gain between the types of genes. The results are comparable to those previously obtained.

REFERENCES

[1] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 2010, 26(3):392-398.

[2] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 2010, 26(3):392-398.

[3] D. S. Chen, J. L. Sung, and J. C. Sheu, “Serum α-fetoprotein in the early stage of human hepatocellular carcinoma,” Gastroenterology, vol. 86, no. 6, pp. 1404–1409, 1984.

[4] M. Capurro, I. R. Wanless, M. Sherman et al., “Glypican-3: a novel serum and histochemical marker for hepatocellular carcinoma,” Gastroenterology, vol. 125, no. 1, pp. 89–97, 2003.

[5] T. Nakatsura, Y. Yoshitake, S. Senju et al., “Glypican-3, overexpressed specifically in human hepatocellular carcinoma, is a novel tumor marker,” Biochemical and Biophysical Research Communications, vol. 306, no. 1, pp. 16–25, 2003. View at Publisher • View at Google Scholar.

[6] I. C. Weitz and H. A. Liebman, “Des-γ-carboxy (abnormal) prothrombin and hepatocellular carcinoma: a critical review,” Hepatology, vol. 18, no. 4, pp. 990–997, 1993.

[7] H. Shirakawa, H. Suzuki, M. Shimomura et al., “Glypican-3 expression is correlated with poor prognosis in hepatocellular carcinoma,” Cancer Science, vol. 100, no. 8, pp. 1403–1407, 2009.

[8] T. M. Mitchell. Machine Learning. Computer Science Series. McGraw-Hill, Singapore, international edition, 1997.

[9] S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Pren-tice Hall Series in Artificial Intelligence. Prentice Hall, 2nd edition, 2003.

[10] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005 Mar 1;21(5):631–43.

[11] An overview of hepatocellular carcinoma study by omics-based methods by Yunfei Pei, Ting Zhang, Victor Renault, and Xuegong ZhangActaBiochimBiophys Sin (2009) | Volume 41 | Issue 1 | Page 1-15

[12] Supervised classification of genes and biological samples by Adrian Tkacz, LeszekRychlewski, Paolo Uva, DariuszPlewczynski

[13] RANDOM FORESTS by Leo Breiman

[14] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[15] http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classificaton/SVM

[16] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel. e1071: Misc Functions of the Department of Statistics (e1071). TU Wien, Version 1.5-11, 2005. URL http://CRAN.R-project.org/

[17] A. Karatzoglou, A. Smola, K. Hornik (2009)."kernlab An S4 Package for Kernel Methods in R". URL http://www.jstatsoft.org/v11/i09/

[18] C. Roever, N. Raabe, K. Luebke, U. Ligges (2005). "klaR –

[19] Classification and Visualization." R package, Version 0.4-1. URL http://CRAN.R-project.org/

[20] T. Hastie. svmpath: The SVM Path algorithm. R package, Version 0.9,2004. URL http://CRAN.R-project.org/

[21] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN Machine

[22] Learning Toolbox. Journal of Machine Learning Research, 11:1799 1802, June 2010. URL http://www.shogun-toolbox.org/

[23] http://www.mathworks.in/help/stats/support-vector-machines-svm.html

[24] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., Gene selection for cancer classification using support vector machines, Machine Learning.,46 389–422, 2002 R. Kohavi, F. Provost: Glossary of terms, Machine Learning, Vol. 30,No. 2/3, 1998, pp. 271-274.

[25] http://en.wikipedia.org/wiki/Confusion_matrix

ISSN 2277-3061

Author' biography with Photo

Sagri Sharma born on 7th February is currently working at C-DAC and carries 10 years of experience working in the field of Health Informatics and ICT. She is a B.E. in Information Technology from DR. B.R. Ambedkar University, Agra and MS in Software Systems from BITS, Pilani.

Dr. Sanjay Kadam works as a Joint Director in the Evolutionary Computing and Image Processing Group at C-DAC, Pune. He has a M. Sc. in Mathematics from Pune University, an M.Tech in Computer Science from IIT, New Delhi, and a Ph.D in Computer Science from the University of London. His research interests include Image Processing, Parallel Processing, Neural Networks, and Soft computing.

An alumnus of the prestigious IIT – Roorkee, Dr Hemant Darbari is one of the founding members of Centre for Development of Advanced Computing (C-DAC), an R&D institute set up by the Department of Electronics and Information Technology, Govt. of India for carrying out advanced research in new and emerging technological domains.

Currently as the Executive Director, C-DAC Pune, he is primarily associated with Building capacity and capability for the National Supercomputing Mission and driving the R&D roadmap in the multi-specialty domains of Multilingual Computing, Software Technologies, Health Informatics, Disaster management, e-Governance and Education & Training.