ISBN 978-93-5156-328-0
International Conference of Advance Research and Innovation (ICARI-2014)

Detection of Colon Cancer by Classification of Genes and Feature Selection using Microarray Data

Harshad Kokate, Neha Nair *, Kalyani Shete, Trupti Thakur, Sujit Ahirrao
Department of Computer Engineering ,Sandip Institute of Engineering and Management, Nashik, India

Article Info

Article history:
Received 2 January 2014
Received in revised form
10 January 2014
Accepted 20 January 2014
Available online 1 February 2014

Keywords
Microarray data
Classification
Feature selection
Gene identification

Abstract

The main aim of designing features selection learning algorithms is to obtain classifiers which use microarray data. It generally uses small number of attributes. These attributes gives differential future performance guarantees. The first goal is to the best of our knowledge, such algorithms that give theoretical bounds on the future performance. It doesn't have been proposed so far in the context of the classification gene expression data. PAC-Bayes learning settings for identifying a small subset of attributes perform reliable classification tasks. So, using PAC-Bayes approach we are dealing with colon cancer detection based on feature selection.

1. Introduction

The objective of designing feature selection learning algorithms is to obtain classifiers that depend on a small number of attributes and have verifiable future performance guarantee. We propose learning algorithms for building small conjunctions of decision stumps. We apply the PAC-Bayes approach for gene identification from DNA microarray data and compare our results to those of the well-known successful approaches proposed for the task. Our algorithm not only finds hypotheses with a much smaller number of genes but also having tight risk guarantees on future performance The traditional methods used for classifying high dimensional data are often characterized as either filters or wrappers depending on whether the attribute selection is performed independent of, or in conjunction with, the base learning algorithm. We apply the proposed approaches for gene identification from DNA microarray data. After finding the genes compare our results to the well-known successful approach proposed for the task. We show that our algorithm not only finds hypotheses with a much smaller number of genes while giving competitive classification accuracy but also having tight risk guarantees on future performance, unlike other approaches. The proposed approach is general and extensible in terms of both designing novel algorithms and application to other domains.

Corresponding Author,
E-mail address: nehanair13@gmail.com

2. Existing System HMM

Hidden Markov Model (HMM) provides a good probability method for modeling discrete sequences of data like DNA sequences (alphabet of four letters: A, C, G and T). To identify exons in DNA of genes P. falciparum is based on exon region in coding sequence (CDS). In principle, the model has start, exon, intron and stop genes and this structure of models are similar to the HMM structure in Daniel Nicorici's paper. The difference is that HMM system has the structure based on exons region in CDS and GT bases in starts of intron separated into G and T state and also AG in ends of intron region like in Fig. 2.1.

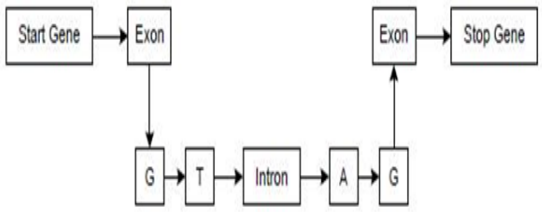

Fig: 2. Existing System: Hidden Markov Model

3. Proposed System

Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data generally used to detect the normal and defected genes from the given set of gene dataset that mostly used for disease detection. This technique basically makes use of PAC-Bayes Approach. With the help of this algorithmic approach the system develop to make comparison, identification and selection of normal and abnormal or defective genes. Normal genes refer to defect free genes and abnormal genes refers to defective genes. This is used to overcome the problem of previously existing system.

The system we are proposing based on four phase: Import Colon Dataset, Train Dataset, Test, Patient Information. An Import phase consists of importing of colon dataset which is further provided for training. Training consists of separating extra attributes from required ones. Then the main part testing includes the actual result generation which involves confusion matrix for detecting whether the person is positive or negative.

3.1 Gene

A gene is a molecular unit of heredity of a living organism. It is a name given to some stretches of DNA and RNA that code for a polypeptide or for an RNA chain that has a function in the organism. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring, although some organelles (e.g. mitochondria) are self-replicating and are not coded for by the organism's DNA.

3.2 Colon Dataset

The colon cancer data set was taken for the testing purposes from Kent Ridge Biomedical Data were analyzed with array which was complementary Repository Centre. The genes expression samples to more than 6500 human genes. The sample of 62people was taken from the dataset. Among them, 40 tumor biopsies are from tumor (labeled as"negative") and 22 normal (labeled as"positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels. The source of the dataset did not mention the number for the confidence of the expression levels. As shown in Fig.3.2.1 the data of all samples in a micro array are presented in a table constructing the gene expression matrix. The rows of the matrix correspond to the single genes and the columns to the single samples.

Dataset

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	Class
20.689...	86.81	51.46	289.42...	99.110...	67.561...	259.91...	138.89...	88.2325	39.667...	67.828...	75.6775	83.5225	28.701...	0.0
33.142...	321.33...	41.591...	263.36...	126.78...	92.238...	96.276...	150.59	82.2375	85.033...	152.195	186.56...	44.4725	16.773...	1.0
5.0160...	116.18...	50.873...	71.401...	151.87...	82.715	31.1025	193.92	76.9725	224.62...	31.225	42.656...	16.0925	15.156...	0.0
13.935...	239.27	29.558...	367.58...	152.59...	41.683...	5.925	183.00...	74.528...	67.710...	48.338...	42.52	49.9825	16.085	1.0
29.364...	52.591...	43.636...	56.0725	126.46...	76.603...	161.35	61.701...	54.563...	223.35...	73.098...	57.598...	7.48875	31.8125	0.0
3.553...	84.206...	47.478...	73.505	122.53...	47.535	147.23...	28.31	33.195	91.85	5.87875	36.2975	9.815	21.883...	1.0
36.623...	72.246...	53.3	154.84...	147.80...	51.0325	80.19	76.485	98.5375	54.621...	30.54	52.961...	37.236...	24.445	0.0
38.391...	60.5275	14.93	96.66	116.19...	46.673...	172.78...	51.823...	97.855	98.982...	24.196...	29.766...	44.376...	52.29	1.0
8.614...	110.21...	16.93	353.455	111.13...	42.146...	277.85...	176.10...	149.15...	209.91...	99.915	122.44...	40.391...	26.8475	0.0
34.685...	149.91	107.15	117.59...	401.77...	141.37...	215.88...	71.0375	93.521...	136.04...	43.23	67.15	65.61	44.0425	1.0
28.65...	66.685	217.95...	135.54...	306.47...	271.79...	554.19...	65.1975	277.20...	442.78...	377.83...	335.15...	116.66	58.663...	0.0
33.798...	174.555	54.2	231.73	337.13...	75.9225	351.73...	127.25...	104.88	178.56...	163.76	205.43...	58.243...	44.0575	1.0
20.307...	171.41...	31.2475	151.92...	185.71...	112.21...	189.05...	80.5	93.9075	161.81...	29.041...	45.5025	27.0725	34.6925	0.0
25.896...	88.106...	14.881...	313.30...	128.99...	59.29	74.723...	108.725	62.2	25.996...	32.7025	33.491...	38.7275	21.631...	1.0
39.798...	120.81...	65.438...	249.30...	254.91...	64.88	435.98...	144.28	71.161...	212.72...	51.461...	87.351...	47.898...	89.453...	0.0
7.192...	6.01	15.725	496.08...	149.62...	26.24	190.25...	68.4825	52.2175	12.063...	9.1875	18.015	8.4475	37.8225	1.0
7.4589...	23.396...	19.321...	168.44	5.95	26.715	244.27	30.775	23.968...	145.31...	45.931...	58.611...	17.256...	48.4925	0.0
7.141...	78.5	5.9675	96.816...	40.444...	32.4075	88.155	15.693...	52.9425	26.890...	34.633...	33.043...	36.8975	64.521...	1.0
3.2107...	75.251...	19.568...	197.93...	80.970...	29.935	155.76...	50.895	43.6575	118.46...	22.0475	62.686...	6.0	25.105	0.0
29.783...	46.575	45.098...	85.81	181.21...	69.3375	124.295	21.403...	67.866...	106.05...	26.408...	32.325	39.608...	36.0425	1.0

Fig: 2. Colon Dataset

3.3 DNA Microarray

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations 1called spots (or features).As shown in Fig.3.3.1 microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene.

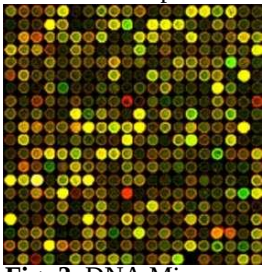

Fig: 3. DNA Microarray

4. Methodology

PAC-Bayes Approach:

The main disadvantage of HMM was it works on only positive samples and the generated result always varies. So, to overcome this drawback we are using PAC-Bayes approach. This approach is used to obtain sparse classifiers with minimum number of stumps. The sparsity is enforced by selecting the classifiers with minimal encoding of the message strings and the compression. The sparsity is enforced by selecting the classifiers. With minimal encoding of the message strings and the compression set in respective cases. We now examine if sacrificing this sparsity in terms of a larger separating margin around the decision boundary (yielding more confidence) can lead us to classifiers with smaller generalization error. The learning algorithm is based on the PAC-Bayes approach that aims at providing Probably Approximately Correct guarantees to Bayesian learning algorithms specified in terms of a prior distribution P (before the observation of the data) and a data- dependent, posterior distribution Q over a space of classifiers.

It suggest that the learner should try to find the Bayes classier B that uses a small number of attributes (i.e., a small k), each with a large separating margin (bk-ak), while keeping the empirical Gibbs risk Rs(G) at a low value. We utilize the greedy set covering heuristic for learning. In our case, however we need to keep the Gibbs risk on Slow instead of the risk of a deterministic classifier.

Since the Gibbs risk is a soft measure that uses the piece-wise linear functions instead of the hard indicator functions, we cannot make use of the hard utility function of. Instead, we need a softer version of this utility function to take into account covering (and erring on) partly an example. That is, a negative example that falls in the linear region of a(a,b)which in fact partly covered and vice versa for the positive example.

The main part of our algorithm is the confusion matrix which gives us the efficiency for our result. A confusion matrix contains information about actual and predicted classification done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier. The confusion matrix is more commonly named contingency table in which the matrix could be arbitrarily large. The number of correctly classified instances is the sum of diagonals in the matrix. Improved Genetic algorithm starts with an initial population which is created consisting of randomly generated rules. Each rule can be represented by a string of bits.

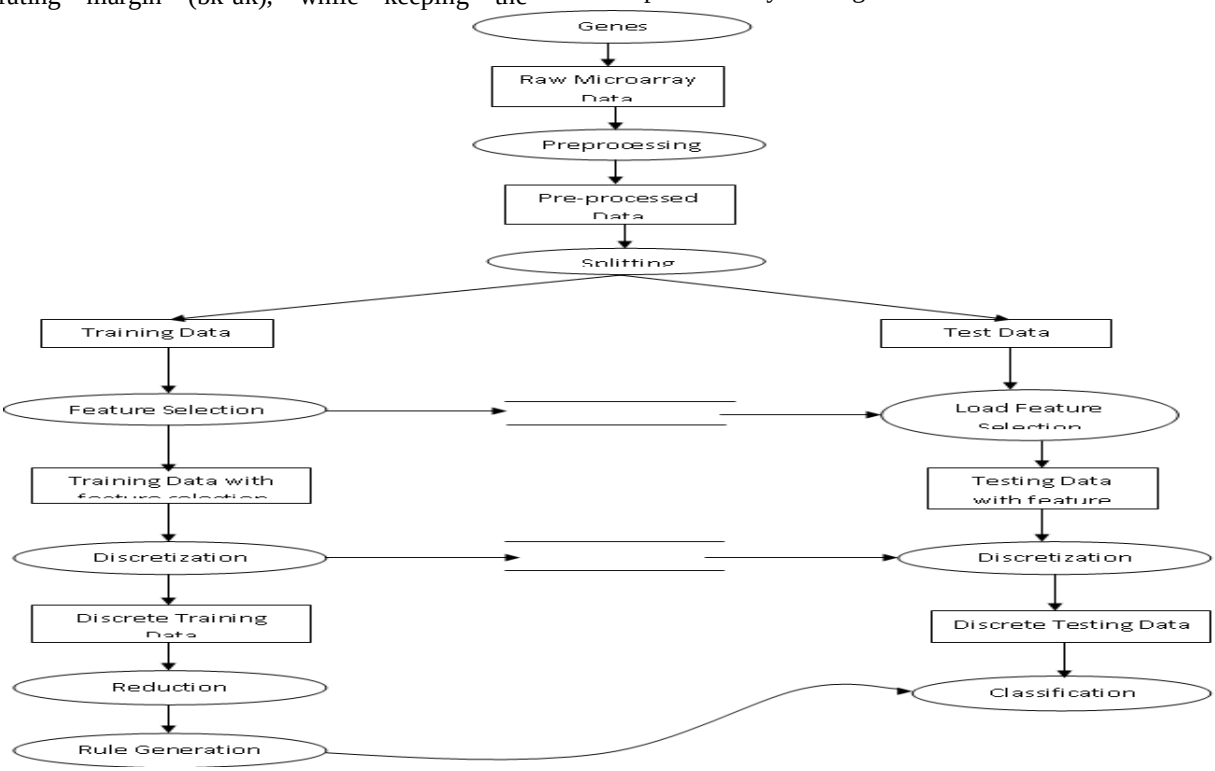

Fig: 4. System Architecture

5. System Architecture

Microarray (Dataset) contains both normal and abnormal genes with 2000 attributes for 62 patients. As shown in Fig.4.1, firstly it goes through pre-processing; as such large data contains extra attributes which we don't require for further processing. After going through the pre-processing, we get processed data which gets splitted into two parts: Training and Testing. Training data is our actual dataset through which we take attributes needs to detect that person is suffering from cancer or not. So, it goes through discrimination process which eliminates the need for extraction of data into standalone mining. Reduction process gives actual attributes for comparison and then rule generation contains PAC-Bayes approach. Test data contains microarray data with normal and infected genes.

6. Conclusion

Using PAC-Bayes learning algorithm for Colon cancer detection we are trying to eliminate the disadvantages of existing system like HMM. So, proposed system mainly focused on the feature selection and microarray decision stumps which boost up the performance in disease detection along with the efficiency provided with the help of confusion matrix. Hence the PAC-Bayes approach helps in biomedical field for the effective result generation with correct precision and accuracy.

References

[1] M. Eisen, P. Brown, "DNA Arrays for Analysis of Gene Expression", Methods in Enzymology,vol. 303, pp. 179-205, 1999

[2]Sridhar Ramaswamy, Sayan Mukherjee, "Multiclass cancer diagnosis using tumor gene expression signatures", Medical science, 98(26), 1514915154

[3] L. Song, J. Bedo, K. M. Borgwardt, A. Gretton, A. Smola, "Gene Selection via the BAHSIC Family of Algorithms", Bioinformatics, 23(13) 490-498, 2007

[4] T. S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, Bioinformatics", 16, pp. 906-914, 2000.

[5] Sanghamitra Bandyopadhyay, Ujjwal Maulik, "Gene Identification: Classical and Computational Intelligence Approaches", Bioinformatics, 35,451-235

[6] Mohak Shah, Mario Marchand,"Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data", 34(1), January 2012

[7] M. Marchand, M. Shah, "PAC-Bayes Learning of Conjunctions and Classification of Gene-Expression Data", Proc. Advances in Neural Information Processing Systems, pp.881-888, 2005