Accuracy comparison between gene selection methods using NAIVE Bayes classifier for the microarray data of JEV infected Mus Musculus brain cells

Akhil R Pillai
Department of Mathematics & Computer Applications
MANIT, Bhopal, Madhya Pradesh, India

Dr. Chandan Kumar Verma
Department of Mathematics & Computer Applications
MANIT, Bhopal, Madhya Pradesh, India

Abstract- Japanese Encephalitis is the most important cause of epidemic encephalitis worldwide. From the reports by various sources, about 68,000 cases of Japanese encephalitis (JE) are estimated to occur each year [1]. A vaccine is available for Japanese encephalitis, which utilizes effectively killed inoculated bacteria, but it is expensive and requires a primary vaccination followed by two successive boosters. No successful cure is discovered till date. This paper describes the analysis of microarray data of Japanese Encephalitis Virus (JEV) infected Mus Musculus cells performing computational accuracy comparison between four gene selection methods using a classifier. The accuracy check is utilized for the effective selection of differentially expressed genes which can be a possible drug target for JE. The proposed methodology includes the comparison of gene selection methods; Logfold selection, t- test, Forward feature selection using fuzzy entropy and Sequential forward selection with the help of the Naive Bayes classifier. The result derived shown that the Forward Feature Selection method using Fuzzy entropy proves to be the maximum accurate gene selection method using the Naive Bayes classifier and in the case of this particular data used.

Keywords – Japanese Encephalitis, gene selection methods, Naive Bayes classifier

I. INTRODUCTION

Microarray data analysis is one of the most exercised methods in the genomic data interpretation nowadays. The sentinel method which can be used to carry out parallel execution of different experiments on multiple genes. Even though all of the cells in the human body consist of identical genetic material, the same genes are not active in every cell. Studying which genes are active and which are inactive in different cell types helps scientists to understand how these cells function normally as well as how they are affected when various genes do not perform properly. Before invent of DNA microarray technology, scientists have only been able to conduct these genetic analyses on a few genes at once. With the development of this technology, however, scientists can now examine how active thousands of genes are at any given time. [2].Microarray technology helps researchers to learn more about many diseases, including a variety of heart diseases, mental illness and infectious diseases.

The most important cause of epidemic encephalitis worldwide is Japanese encephalitis virus (JEV) From the reports by various sources, about 68,000 cases of Japanese encephalitis (JE) are estimated to occur each year[3]. It follows an endemic pattern throughout most of Asia and parts of the western Pacific, but local transmission has not been observed in Africa, Europe, or the Americas. After a JE vaccine was licensed in the United States in 1992, only seven cases of JE have been reported from among United States travelers [4, 5].This virus is a member of a serogroup of JE in the genus Flavivirus, family Flaviviridae. It is transmitted between vertebrate hosts, like pigs, by mosquitoes, principally by Culextritaeniorhynchusand Culex vishnui groups, which breed particularly in flooded rice fields . West Nile virus (WNV), which recently spread to cause outbreaks of encephalitis in North America, is another important member of the same serogroup of flavivirus .Japanese encephalitis affects the membranes around the brain. In majority of cases, human JEV infections are asymptomatic or cause a non-specific mild illness and less than 1% of JEV infections results in symptomatic neuroinvasive disease which is characterized by rapid onset of high fever, headache, neck stiffness, disorientation, coma, seizures, spastic paralysis and death. Fatality rate can be as high as 60% among those with disease symptoms and 30% of those who survive suffer from lasting damage to the central nervous system .In areas where the JE virus is common, encephalitis occurs mainly in young children because older children and adults have already been infected and are immune[6]

A vaccine is available for Japanese encephalitis, which utilizes effective killed inoculated bacteria, but it is expensive and requires a primary vaccination followed by two successive boosters. Another inexpensive live-attenuated vaccine is used in China, but is not available elsewhere. Vector control through chemical methodology is not a solution because of the extensive behavior of breeding sites (irrigated rice fields).In the rice production systems facing water shortages, certain measures (alternate wetting and drying) may be applied that reduce vector populations Personal protection (using repellents and/or mosquito nets) will be effective under certain conditions.Elimination of pig population is often a measure taken in the wake of outbreaks. Special care should be taken to avoid the introduction of pig rearing as a secondary source of income for rice-growing farmers in receptive areas. Due to the strict and high infective characteristics of the disease; no effective or successful treatment is available for JE infections. Researchers as well as scientists have been working hard to get a lead on a drug which would prove useful for successfully evading this disastrous epidemic.

The implementation of microarray data analysis on JE infected cells is a novel path to find an effective defense against infections. Identifying the genes which are of significance is one way of initiating this. The meekest way to identify genes of possible interest through several related tests is to search for those that are unfailingly either up- or down regulated. Up to this extend, the application of simple statistical analysis would be enough. However, categorizing patterns of gene expression and adding genes into expression classes may provide much greater insight into the biological function as well as their significance. Implementing gene selection methods and other feature selection properties to detect a possible drug/therapeutic target for JE infection from the differentially expressed genes is the most applicable problem.

This paper is organized as follows: Section II briefly reviews the proposed system and the different kind of gene selection methods that we have used to applying for a Naïve Bayes Classifier. Finally, Section III concludes results and its conclusion.

II. PROPOSED METHODOLOGY

A. Types of Different Gene Selection Methods–

The expression data of two types of cells over different time intervals. Data is available on NCBI GEO (Gene Expression Omnibus) database with GEO Accession No. GSE42942.It contains B16F10, a murine melanoma cell lines which is cytolytic and non-cytolytic to JEV infection. This variant cell line has two types of cells, one which is persistently infected JEV and another which is resistant to JEV infection. Data is already normalized by log base 2.
The Microarray data analysis can be carried out using different techniques. Computational interpretation is best carried out using software analysis. Gene selection techniques are utilized to identify the differentially expressed genes i.e. the genes having maximum importance. Here we analyze the efficiency of different feature selection methods using a specific classifier

1. t- Test

A t-Test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported[7]. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution.
The two sample t-test calculates a Student's t statistic and the associated probability to test whether or not the difference of the two sample means equals to μ_d (i.e. to test whether or not their means are equal, you can just test whether or not their difference is 0, $\mu_1 - \mu_2 = \mu_d = 0$

Hypothesis: $H_0 : \mu_1 - \mu_2 = \mu_d \qquad H_0 : \mu_1 - \mu_2 = \mu_d$

Consider two independent samples x_1 and x_2, of size n_1 and n_2 drawn from two normal populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2 respectively, we have:

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1j} \quad , \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2j}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (x_{1j} - \bar{x}_1)^2 \quad , \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

. Where $\bar{x}_1$ and $\bar{x}_2$ are sample means and s_1^2 and s_1^2 are sample variances. Then we can compute the t test statistic by: For equal variance is assumed, that is: $\sigma_1^2 = \sigma_2^2$ In this case the test statistic *t*:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{s_p\sqrt{(1/n_1 + 1/n_2)}}$$

As our data were already normalized, we directly move towards the Filtration in which non-significant gene was filtered out by applying t-test. We use the software Genowiz™ for performing the t test and retaining a list of 100 genes.

2. *Fold Change*

Fold change value can be a very effective parameter to filter the genes to find maximum fluctuating (significant) genes. The fold change represents the number of times a gene expression level is amplified considering the initial and the final expression level. The fold change value is identified by dividing the final value by initial value.

As the fold change values may have huge variations in values, it will be hard to analyze or perform studies on it. Therefore the data are normalized by applying a log2 on it. The value we derive by applying log to the values is logfold value. The logfold values represent the intensities of gene expression. Therefore the method of selecting the significant genes is by selecting the genes with maximum values, may it be either negative or positive.
The t- test retained a list of 100 genes and therefore to maintain equality 50 genes with highest negative logfold value as well as 50 genes with highest positive values were selected and compiled into one single list. This list was successfully saved as one of the inputs for the classifier.
There are two dentitions of fold-change in the literature. The standard dentition of the fold-change for gene i is

$$FC_i = \frac{\overline{x'_i}}{\overline{y'_i}} \qquad \text{------------------------ (1)}$$

Where x'_{ij} and y'_{ij} are the raw expression levels of gene i in replicate j in the control and treatment, respectively.

On the other hand, in Guo et al. (8) the fold-change for gene i is defined as

$$FC_i = \overline{x_i} - \overline{y_i} \qquad \text{--------------------(2)}$$

3. *Forward Feature selection Using Fuzzy entropy*

The FFS tool in MATLAB is developed utilizing the ‘forward feature selection using fuzzy entropy’. This tool effectively provides the option to the user for particularly selecting the number of genes to be retained. The data are first imported to the MATLAB user interface using the import button option. Here we specify the ‘MATRIX’ as the dataset which we want to input. After importing, the script file is selected and executed. The workspace computes the list of genes and displays it. As both the previous steps derived a list of 100 genes, here also a limit value is provided to display only top 100 genes from the list of whole genes. This list is further saved and is used as one of the input for the classifier.

4. Sequential Forward Feature selection

The SFS Gene selection toolbox is successfully developed using sequential selection algorithm as the base[9]. Similar to FFS, the data are first imported using the import function and then the matrix data is assigned. The script file is executed and a list of significant genes is generated. According to the parameter, a list of 100 genes is generated. The list is finally extracted and saved as an input for the classifier. We can be calculated the Correlation and information-theoretic measures based on the rationale that good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other thus we can be calculated our measures as follows

Linear relation measures

Linear relationship between variables can be measured using the correlation coefficient

$$J(Y_m)=\frac{\sum_{i=1}^{m}\rho_{ic}}{\sum_{i=1}^{m}\sum_{j=j+1}^{m}\rho} \qquad \text{-----------------------(3)}$$

Where ρ_{ic} is the correlation coefficient between the feature i and the class label and ρ_{ij} is the correlation coefficient between features i and j

Non-linear relation measures

Correlation is only capable of measuring linear dependence A more powerful measure is the mutual information $I(Y_k:C)$

$$J(Y_m)=H(C)=H(C\,|\,Y_M)=\sum_{C=1}^{C}\int_{Y_M}\rho(Y_M,\omega_C)\log\frac{\rho(Y_M,\omega_C)}{\rho(Y_M)\rho(\omega_C)}dx \qquad \text{-----------(4)}$$

III. Experiment and Result

A . NAIVE BAYES classifier

A NAIVE BAYES classifier tool is developed using Bayes theorem for classification. The complete scripting is done in MATLAB, as it provides an effective platform for easy as well as efficient working of the classifier.

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[10,11,12].

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contain a continuous attribute, . Our first segment the data by the class, and then compute the mean and variance of in each class. Let μ_C be the mean of the values in x associated with class c, and let $\sigma_C{}^2$ be the variance of the values in x associated with class c. The probability density of some value given a class, $p(x=\upsilon\,|\,c)$ can be computed by plugging v into the equation for a normal distribution parameterized by μ_C and $\sigma_C{}^2$ That is,

$$p(x=v|c)=\frac{1}{\sqrt{2\pi\sigma_c^2}}\,e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

The method of application and classification is similar for all the four resultant gene lists which were derived from the various gene selection procedures. The steps are as follows:

1. Before importing the data file, the 100x 50 matrix data should be transposed to 50 x 100 matrixes. Now we have a data set having genes aligned horizontally and samples arranged vertically. A response variable (here 0 and +1) for different classes is added as another column along the samples.
2. Import the data using the import function
3. Assign the matrix value to the data
4. Import another excel class file containing the different class names and response variable and assign it as 'cell array'. In this case we selected two classes; INF for infected cytolytic cells and RES for resistant non-cytolytic cells.
5. After importing the data, he script is executed from the workspace and the toolbox starts calculating the accuracy value for all the possible combinations of genes.

At the end of the process the combinations with maximum accuracy rates are displayed. This result can be extracted and saved as a result file in text format. The result file of all the four data sets are compared and analyzed to find the maximum efficient gene selection method.

Here, from the tabular outputs, it is understood that each gene selection method has its own uniqueness. The outputs derived represent the accuracy value assigned by the Naïve Bayes classifier on different set of genes. Four different tables are derived four different but comparable results. The ranges of the accuracy rates vary between twenty units. The values below this measure are not considered apt by the classifier.

Table 1: list of genes derived from t- test and accuracy rate by the Naïve Bayes clasifier :

S.NO	SET OF GENES	ACCURACY RATE (in %)
1	(1 3)	62.00
2	(1 53)	62.00
3	(1 58)	74.00
4	(2 58)	76.00
5	(5 58)	78.00
6	(10 58)	78.00
7	(17 58)	78.00
8	(34 58)	80.00
9	(58 3)	80.00
10	(58 31)	80.00
11	(58 32)	80.00
12	(76 58)	80.00
13	(88 58)	80.00

Table 2: list of genes derived from Logfold value and accuracy rate by the Naïve Bayes :

S.NO	SET OF GENES	ACCURACY RATE (in %)

1	(1 1)	60.00
2	(1 21)	60.00
3	(1 26)	64.00
4	(1 36)	66.00
5	(1 69)	70.00
6	(2 52)	70.00
7	(21 69)	70.00
8	(26 52)	72.00
9	(37 1)	74.00
10	(52 88)	76.00

Table 3: list of genes derived from FFS and accuracy rate by the Naïve Bayes clasifier

S.NO	SETS OF GENES	ACCURACY RATE (in %)
1	(1 1)	50.00
2	(1 5)	60.00
3	(1 12)	60.00
4	(1 29)	66.00
5	(1 90)	72.00
6	(2 29)	74.00
7	(3 90)	76.00
8	(9 90)	78.00
9	(27 90)	78.00
10	(39 90)	80.00
11	(76 90)	82.00
12	(90 39)	82.00

Table 4: list of genes derived from SFS and accuracy rate by the Naïve Bayes clasifier

S.NO	SETS OF GENES	ACCURACY RATE (in %)
1	(1 3)	50.00
2	(1 16)	60.00
3	(1 19)	62.00
4	(3 18)	64.00
5	(3 19)	66.00
6	(21 34)	70.00
7	(27 12)	73.00

8	(39 70)	74.00
9	(44 63)	77.00
10	(53 87)	78.00

The results can be summarized below as:

1.Maximum accuracy value derived for the data set input from T test method is 80% and is distributed in the data sets (88 58) (76 58) (58 32) (58 31) (58 3) (40 58) (34 58) (21 58) and (18 58). All these sets have equal accuracy rates and therefore are of equal importance.

2. Maximum accuracy value derived for the data set input from Logfold change method is 76% and is shown by the data set (52 88).Here a single and maximum value is derived but having a lesser value than the above result. Therefore the accuracy of this method is considered to be a bit less.

3. Maximum accuracy value derived for the data set input from Forward Feature Selection method using Fuzzy entropy is 82% and is exhibited by the data sets (76 90) and (90 39).This shows that both these data sets have maximum accuracy and are equally important. Here the accuracy is the maximum compared to both results mentioned above. Therefore, it is evident that this method is having the maximum efficiency and accuracy while applying gene selection.

4. Maximum accuracy value derived for the data set input from the Sequential Feature selection method is 78% and is displayed by the data set (53 87).The result is unique and significant. This result shows a degraded value while comparing to that from FFS method, but proves to be more efficient than T test and Logfold change methods.

Comparing all the above result data, it is evident that the most efficient and accurate method for gene selection is Forward Feature Selection method using Fuzzy entropy. It derives an accuracy of 82% and provides two data sets with the same value. This is the maximum value in all the results.

IV.CONCLUSION

The results derived can be compiled to the point that the following gene data sets, play an important role in the JEV infection:

Table 5:Gene numbers and their names

(88 58)	(Adam38 and Clec4a4)
(76 58)	(BC080695 and Clec4a4)
(58 32)	(Clec4a4 andA4galt)
(58 31)	(Clec4a4 and Zfp366)
(58 3)	(Clec4a4 and BC117090)
(52 88)	(Olfr132 and Adam38)
(76 90)	(BC080695 and Taar5)
(90 39)	(Taar5 and Nlrp4e)
(53 87)	(Olfr467 and Naaladl1)

Among these genes the maximum accuracy and therefore maximum significant genes are BC080695, Taar5 and Nlrp4e.

These sets of genes are differentially expressed and have the maximum intensities. The expression values are compared from two classes of genes, i.e. infected and resistant. After the infection, there is a significant change in expression values of these genes, may it be positive or negative.The positively expressed genes are termed as over expressed and the negative ones are termed as under expressed. The over or under expression of genes may be due to their significant role in infection or maybe because of their association with a pathway which is related to the infection. As these genes are differentially expressed, they may prove to be suitable candidates as therapeutical/drug targets and can be analyzed to understand important information about the infected system and cure.

REFERENCES

[1] TSAI, T. F. 2000. New initiatives for the control of Japanese encephalitis by vaccination: minutes of a W. H. O./CVI meeting, Bangkok, Thailand, 13–15. Vaccine 18(Suppl. 2):1–25 October 1998.

[2] G.G. LENNON AND H. LEHRACH. Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics*, 10:314–317, 1991.

[3] TSAI, T. F. 2000. New initiatives for the control of Japanese encephalitis by vaccination: minutes of a W. H. O./CVI meeting, Bangkok, Thailand, 13–. Vaccine 18(Suppl. 2):1–25, 15 October 1998

[4] Centers for Disease Control and Prevention (CDC). Japanese encephalitis in two children--United States, 2010. MMWR Morb Mortal Wkly Rep 2011; 60:276.

[5] Hills SL, Griggs AC, Fischer M. Japanese encephalitis in travelers from non-endemic ountries, 1973-2008.; 82:930 , Am J Trop Med Hyg2010

[6] http://www.who.int/water_sanitation_health/diseases/encephalitis/en/ [4/18/2014]

[7] Zimmerman, Donald W. "A Note on Interpretation of the Paired-Samples t Test". Journal of Educational and Behavioral Statistics 22 (3): 349–360 (1997)..

[8] Guo et al " A comparison of fold-change and the t-statistic for microarray data analysis" Third Conference on Uncertainty in Artificial Intelligence (2007)

[9] Parkash et al. O.M. Parkash, P.K. Sharma, R. Mahajan "New measures of weighted fuzzy entropy and their applications for the study of maximum weighted fuzzy entropy principle" Information Sciences, 178 (11) pp. 2389–2395 (2008)

[10] Mitchell, T. Machine Learning, McGraw Hill (1997).

[11] Vangelis M., Ion A., and Geogios P. Spam Filtering with Naive Bayes - Which Naive Bayes Third Conference on Email and Anti-Spam. (2006)

[12] George H. John and Pat Langley. Estimating continuous distributions in Bayesian classifiers the Eleventh Conference on Uncertainty in Artificial Intelligence(1995).