Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Evaluation of Prediction Models for Tumor Class Detection with Classification Algorithms Employing Chi-square and Information-gain Feature Selection

Sonal Saxena[*1]
[1]*M.tech Scholar*
Department of Computer Science & Engg.,
RCEW jaipur, India

Vineet Khanna[2]
[2.] *Asst Professor*
Department of Computer Science & Engg.,
RCEW jaipur, India

Abstract- Data mining has become a powerful tool for extraction of knowledge from voluminous databases. It helps in analyzing the data in such a way that the analysis proves to be very useful for a specific application or organization. From past few years Data Mining has also stepped into healthcare sector where it is been used as an application for medical diagnosis. One of such use is discussed in this paper where with the help of some data mining tools and techniques, some classes of tumors will be detected accurately in some unknown samples of human gene.

Keywords – 10-fold cross validation, chi-squared feature selection, prediction model, SVM, feature selection

I. INTRODUCTION

Data mining is actually a task of finding out correlations and existing patterns from large databases in various domains. This concept can be used successfully for detecting or diagnosis of some diseases in human beings. Such an implementation is done for detection of some specific classes of tumor present in a database having a collection of gene molecular information which are affected from specific tumor classes but it was unknown to us till the time of experiment[1][7]. Tumors are generally a lump or mass of tissues which are a result of enormous or abnormal division of cells in a living body. Usually, they are mistaken as cancer, but this is not so as Tumors are broadly of two types viz. Malignant Tumors or Cancer and Non-malignant tumors. The former ones are dangerous as they keep on growing and eventually lead to death if left untreated. But the latter ones are not dangerous as they don't have such threatening tendency due to which they do not essentially need treatment or cure [4].

II. DATA MINING TOOLS AND TECHNIQUES FOR DETECTION

For doing this implementation various data mining tools and techniques were used. WEKA is the most popular data mining tool for applying various techniques. It has got very many options like data preparations, data classification, cross validation, association rule discovery, data splitting and many more. In this particular experiment WEKA 3.6.4 version was used.

First of all the 10-fold cross validation was applied on the complete data. This cross validation is done to sample the database. The data is randomly partitioned into 10 mutually exclusive folds which are actually of almost equal size. After doing this the classifiers are applied on the records of each fold leaving the ones present in the current fold and then classifiers are applied on the current fold afterwards. The estimates of all the folds are then averaged to get the final cross validation accuracy (CVA) of the samples [2].

The classification algorithms used for this implementation are:

1. **Naïve Bayes**: It is based on Bayes's Theorem which is basically statistical classifier. Hypothetically, this algorithm is proved to be far better than other algorithms as it has been observed to give minimum error rate.
2. **J48 Decision Tree**: They follow a tree type structure where each leaf node represent a test applied on the attributes, every branch represent the result of the test applied and the external or terminal node represents the label of the class. These classifiers are good in the sense that they do not need prior knowledge of the domain.. Also they are capable of handling high dimensional data. The J48 classifier uses Information Entropy concept to split the data into smaller subsets. Then it used to examine the normalized information gain and the attribute with the highest information gain is used for decision making. This algorithm is recurred on smaller subsets. It also provides tree pruning concepts.
3. **Random Forest**: It consists of a set of decision trees and so it can work efficiently on bulkier data. It can handle the missing values and data and even results a good accuracy in case of this large missing values [3]. It uses to compute prototypes which are used to give information about the variables and the classification relation.
4. **Support Vector Machines (SVM)**: SVM's are considered to be very efficient supervised learning algorithm. It is based on the concept that defines decision boundaries. It is basically used for non linear classification of data as the classification problems are not always very simple, in fact they are more complex and complicated.[6]

III FEATURE SELECTION

Feature Selection is the way of selecting some appropriate features which are relevant to specific criterion. It removes noisy data, reduces dimensionality of space etc. It is sometimes considered to be NP-hard problems.[10] But this process can even improve the performances of diagnosis and detection accuracies. There are three types of feature selection methods:

1. Filters
2. Wrappers
3. Embedded Methods

This feature selection is done with a proper structure. It includes 4 steps:

1. **Subset Generation**- generate the features of subsets of original data for evaluation.
2. **Subset Evaluation**- the generated subsets are evaluated based on some specific relevant criterion.
3. **Stopping Criteria**- The feature selection process is required to be stopped with some criteria otherwise it could run exhaustively. Ex. pre-defined no. of iterations, pre-defined number feature selection.
4. **Validation-** The validation is done by testing both these selected and original features to see if there is any improvement.

In this work, two feature selection methods were used for evaluation, which is discussed below:

Chi-square Feature Selection

Chi-square(X^2) feature selection is very common test. It evaluates the features by taking out the chi-squared statistic corresponding to the class. It assumes that the features are independent of each other. The formula for conducting this test is,

$$X^2 - \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Oij - Eij)^2}{Eij}$$

where,
Oij = observed frequency
Eij= expected frequency

Information Gain

This feature selection procedure is sometimes used interchangeably with the term ‘mutual information’. It is considered good method of feature or attributes selection. [9][10] It measures the capability of an attribute by information gain corresponding to a class. But one of its major disadvantage is that it does not consider the interactions among features which is sometimes required. It is given by:

$$\textbf{IG (Class, Attribute) = H (Class) – H(Class | Attribute)}$$

Where,
IG = Information Gain
H = Information Entropy
The information entropy is given by

$$\sum - p_i \log2\, p_i$$

where, p_i is the probability of i class.
H, the information entropy is one of the concepts of information theory which shows that how much information is contained by an event, in our case the information contained in an attribute. Usually, more random an attribute is, more information provides or simply, higher the value of Entropy, higher is the content of information. It quantifies the expected value of information and the unit of measurement is normally bits, nats or bans. It ranges from 0 to 1.

IV EXPERIMENTAL SETUP

The database of tumor affected gene values is available on the site of a very famous private cancer research institute,[8] which is a scientific community dealing with fundamental mechanism of cancer, analysis of the functionalities of cancer genes and the vulnerabilities of tumors.The community shares its ideas and data freely to let the interested people work on it. The data collected from this site, is a huge collection of gene expression information. [1]

The database taken from this site was already experiment ready, which means that some preparatory steps of data mining like data cleaning, data transformations etc. was not at all required to be applied on it. I had 69 samples of human genes of which 7071 attributes or features were present which makes it a very high-dimensional data. The whole data needs to be detected for presence of some specific category of tumors. Here we have 5 categories or classes of tumors viz.

1. Mediastinum (MED)
2. Macrophage Galactose Binding Lectin (MGL)

3. Extra Mammary Paget's Disease (EPD)
4. Juvenile Pilocytic Astrocytroma (JPA)
5. Recombinant Hemoglobin (RHB)

The file of all the 69 samples and approx. 7000 attributes is first needed to be converted into an arff format which is a compatible format for WEKA tool kit, a data mining tool.

First, the total database consisting of 69 samples and 7071 attributes were taken as full training data and then 10 fold cross validation is applied over it for the sampling purpose. With 10-fold cross validation, 3 different classification algorithms were applied viz. Naïve Bayes, J48 Decision Tree and Random Forest to find out the cross validation accuracy of the tumor prediction. The accuracies I have got in respective orders are 92.7536%, 76.8116% and 86.9565%. As clearly visible, when the complete set was taken the accuracy of Naïve Bayes is the highest as it detected 64 samples out of 69 very correctly. All these algorithms are good in different situations and scenarios.

After this, using libSVM tool(a library for support vector machine) to apply SVM algorithm on the full training data of 69 samples and 7071 attributes with 10-fold cross validation, the accuracy came to be 98.5507% which is the highest among all the four applied algorithms.

V. CROSS VALIDATION WITH CHI-SQUARE AND INFORMATION GAIN FEATURE SELECTION

We have already seen the cross validation, chi-square and information gain filtering or feature selection. Here are the results of these applications:

After applying both of them, 1% features were selected which counts to be total 71 number of attributes. So with 69 samples of data and 71 attributes the cross validation prediction accuracy for all the 4 algorithms comes out to be as shown in the given table:

TABLE 1
CVA WITH CHI-SQUARE FEATURE SELECTION

Algorithm	**Accuracy**
Naive Bayes	95.6522%
J48 decision tree	89.8551%
Random Forest	95.6522%
SVM	97.1014%.

Again it is visible from the above table that the prediction accuracy of SVM algorithm with chi-squared feature selection is the best (97.1014%).

The accuracy provided by Information gain feature selection is as shown:

TABLE 2
CVA WITH INFORMATION GAIN FEATURE SELECTION

Algorithm	**Accuracy**
Naive Bayes	100%
J48 decision tree	89.8551%
Random Forest	95.6522%
SVM	97.1014%.

VI. THE PREDICTION MODEL

After all the above prior experiments, the actual prediction models for each algorithm were built. Due to unavailability of sufficient data, the full training data of 69 samples with 7071 attributes was split into two parts where the first part having 46 samples and 7071 attributes were taken as the training data and 23 samples with 7071 attributes were taken as test data. The first sets of samples were trained to detect the tumor classes which were then applied on test data to see it working accurately. The Data Split feature was used in both WEKA and LibSVM tool.

The prediction accuracy of the 4 prediction models comes out to be as shown in the table:

TABLE 3
PREDICTION MODEL ACCURACIES WITHOUT FEATURE SELECTION

Algorithm	**Accuracy**
Naive Bayes	43.4783%
J48 decision tree	47.8261%
Random Forest	56.5217%
SVM	86.956%

The outcomes again show that in comparison to other algorithms, SVM has the highest prediction accuracy when no feature selection was applied to it (86.956%).

Then, chi squared feature selection was applied on the training and test samples and again the 4 algorithms were applied to see the prediction accuracy. The results we have got are shown below:

TABLE 4
PREDICTION MODEL ACCURACIES WITH CHI-SQUARE FEATURE SELECTION

Algorithm	Accuracy
Naive Bayes	56.5217%
J48 decision tree	56.5217%
Random Forest	73.913%
SVM	86.95%

So, again if we see the results, we will find that SVM with chi squared feature selection has given us the highest prediction accuracy till now undoubtedly (86.95%). However, the same accuracy was achieved when no feature selection was adopted, as shown in Table 3.

After this, with information gain feature selection, the accuracy of prediction models was:

TABLE 5
PREDICTION MODEL ACCURACIES WITH CHI-SQUARE FEATURE SELECTION

Algorithm	Accuracy
Naive Bayes	73.913%
J48 decision tree	65.2174%
Random Forest	91.3043%
SVM	99.181%

Again, SVM algorithm together with information gain feature selection gave the highest result among all other algorithms. As compared to chi-square, it has given almost 12-13% better result, as shown in Table 5.

VII. RESULTS

After all the implementations, this is the final result which we have got that the most efficient prediction model of all the four was of SVM and that too with information gain feature selection. If we compare the accuracies of all the algorithms with chi-square and information gain, then we will see that the latter has improved the accuracy in every case. So it can be said that this feature selection can be used for detecting those diseases where the domain expertise is missing. The results were found better than the work discussed in [9].The comparative graph is as shown below:

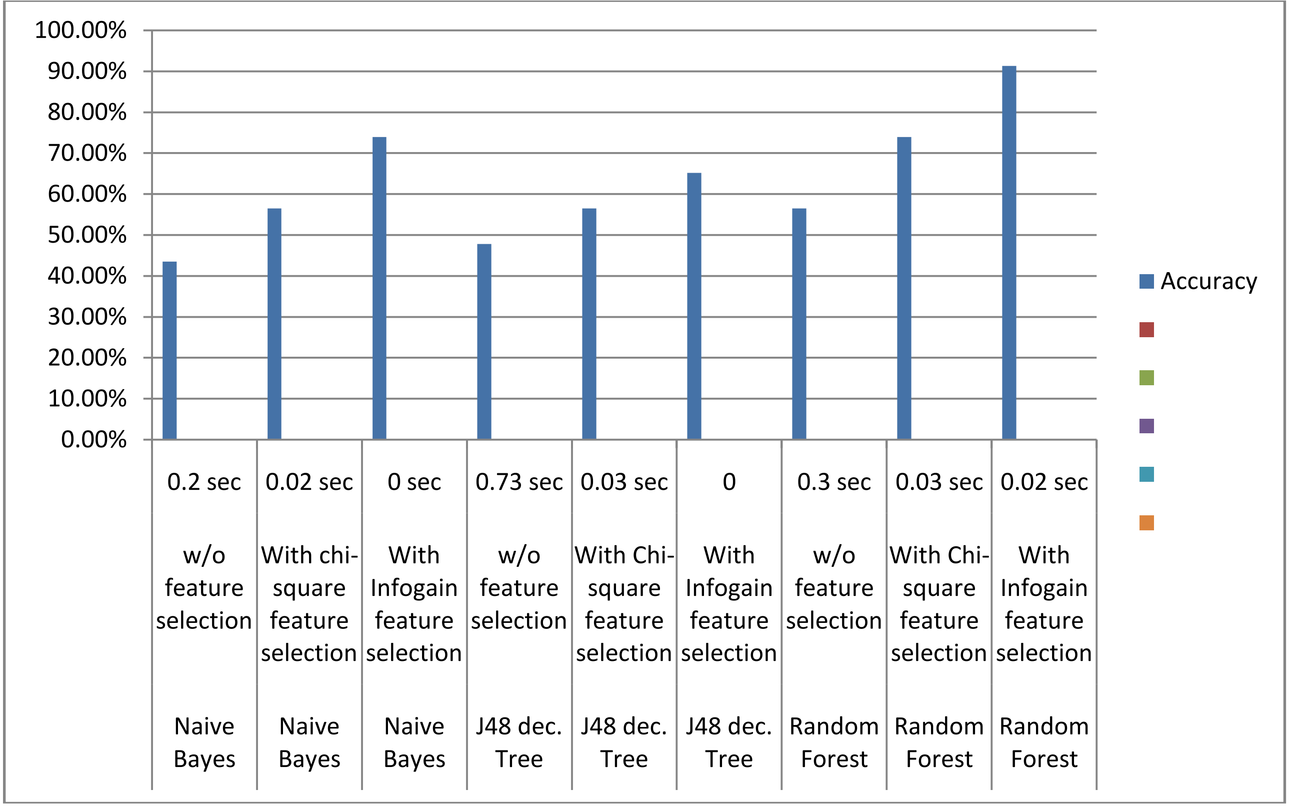

Fig.1 Comparative graph showing accuracies of various prediction models.

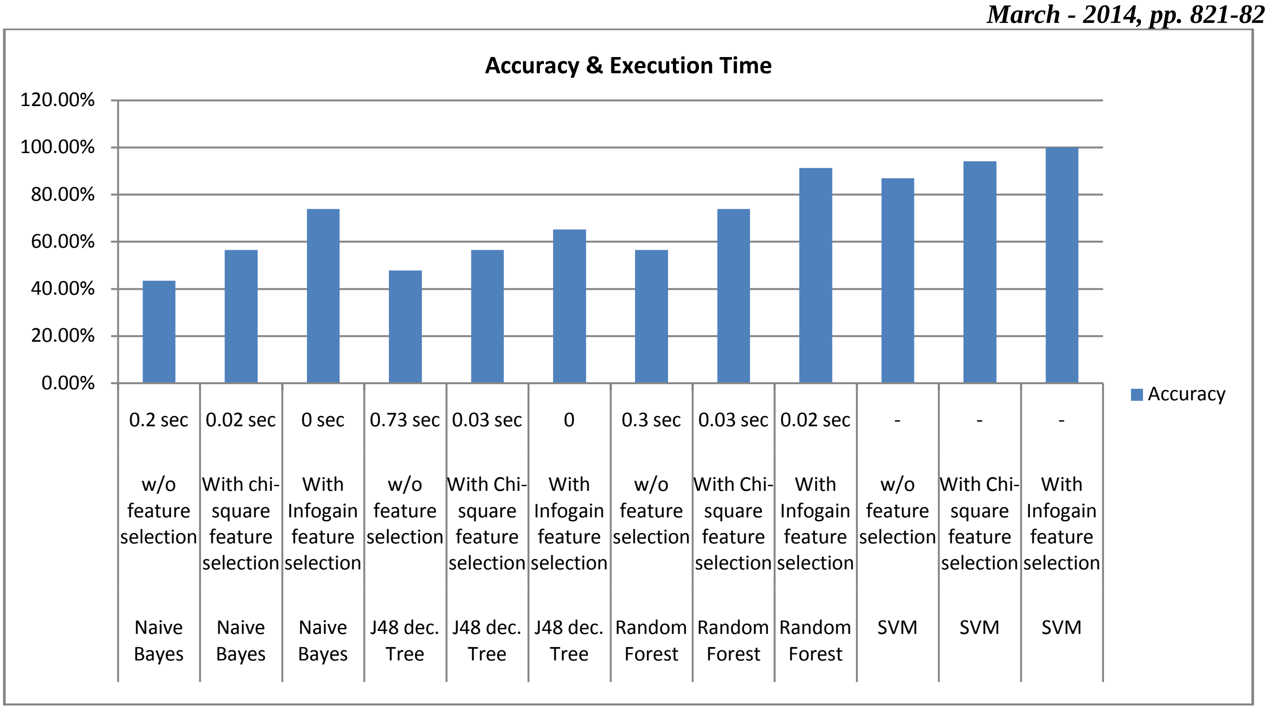

Fig.2 Prediction accuracies and execution time of all algorithms with and without feature selection

VIII. CONCLUSION

The experiments and implementations done in this work shows that SVM algorithm together with Information gain feature selection can be applied on higher dimensional data where even the domain knowledge is very low, for getting accurate and efficient results. Next, it can also be concluded that Random Forest algorithm can also be used for classification problem as it is also one of the efficient algorithm after SVM. The performance of Naïve Bayes has been satisfactory and it can provide efficient results in some cases with different criteria. But, J48 decision tree algorithm, as in this work has not proved to be very good enough for detection of tumor classes.

REFERENCES

[1] Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Second Edition, Morgan Kaufmann Publishers, San Francisco

[2] Eric P. Xing, Michael I. Jordan and Richard Karp.*Feature Selection for High-Dimensional Genomic Microarray Data.* Division of Computer Science, University of California, Berkeley, CA 94720 and Department of Statistics, University of California, Berkeley, CA 94720

[3] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1S): 5–32 doi: 10.1023/A: 1010933404324.

[4] Matthias E Futschik, Mike Sullivan, Anthony Reeve, Nikola Kasabov, *Prediction of clinical behaviour and treatment for cancers*, Departments of Information Science and Biochemistry, University of Otago, Dunedin, New Zealand

[5] Quinlan Ross J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

[6] Zhenqiu Liu, Dechang Chen, Guoliang Tian, Man-Lai Tang, Ming Tan,and Li Sheng, 2010. *Efficient Support Vector Machine Method for Survival Prediction with SEER Data.*. http://www.springer.com/978-1-4419-5912- 6. Advances in computational Biology,Arabnia

[7] R. L. Somorjai , B. Dolenko and R. Baumgartner,2002. *Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats,cautions* Institute for Biodiagnostics, National Research Council Canada, Winnipeg, MB,Canada R3B 1Y6.

[8] Data Courtsey: http://www.broadinstitute.org/scientific-community/data

[9] Elsheikh, Jarada, Nagi, Naji, peng, Karampelas, Ozyer, Kianmehr, Ridley, Rokne and Alhej, 2011 *Effectiveness of Feature Selection and Classification Techniques for Gene Expression Data Analysis*. ICIT 2011 The 5th International Conference on Information Technology

[10] Jasmina Novakovic, Perica Strbac, Dusan Bulatovic.(2011) *Toward Optimal Feature Selection Using Ranking Methods And Classification Algorithms.* Yugoslav Journal of Operations Research. 21 (2011), Number 1, 119-135. DOI: 10.2298/YJOR1101119N