*IOSR Journal of Computer Engineering (IOSRJCE)*
*ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 1 (Sep-Oct. 2012), PP 12-21*
*www.iosrjournals.org*

# Efficiency of Prediction Algorithms for Mining Biological Databases

A. Lekha[1], Dr. C V Srikrishna [2], Dr. Viji Vinod [3]
*[(1)]Research Scholar, Dr M G R Educational Research Institute, Chennai, India – 600095*
*[(2)]Professor, Department of MCA, PESIT, Bangalore – 560085*
*[(3)]Head, Department of MCA, Dr. MGR Educational and Research Institute, Chennai, India-600095*

**Abstract:** *The paper deals with the analysis of the efficiency of prediction algorithms for mining biological databases and also suggests possible ways of improving the efficiency for a given dataset. The study reveals that the efficiency of a mining algorithm is a function of many variables of the dataset. The study proposes a predictive model through a case study.*

**Keywords:** *Biological databases, Breast Cancer, Efficiency Predictive algorithms, Statistical Analysis.*

## I. Introduction

With the inception of Information Technology in all the spheres of human life, many new fields of study emerged. Bioinformatics is a field which has attracted many researchers because of its importance and influence in the modern world [1]. The introduction of computer science in the medical science has brought remarkable changes in the old practices used in bioinformatics. It has introduced many new aspects in the medical sciences like Genomics, Proteomics, and Cell Biology. These concepts's existences were difficult to believe in the past [2].Mining biological databases are one such option which transformed the medical science from its traditional reactive methods to modern proactive approach. The biological databases contain information about the life sciences [3] and are enriched with the research based knowledge of proteomics, metabolomics, genomics, phylogenetics [4] and microarray sciences. It also contains information about genes and protein sequence [5]. The information present in the database is the real and true information though it can be current or old. In other words, the databases contain both historical data and the current facts [6]. The information is saved in a predefined format.

Mining methods have enabled the experts to extract information from the database in a way that it provides altogether a new set of information. This information is the one which cannot be easily observed by human beings in generalized reports. This is why, data mining is also known as Knowledge Discovery in Database [7]. The new set of information assists the decision makers and researchers [8] in focusing the unexplored pattern or information of the field. Mining takes place with the help of artificial intelligence i.e. which explores information to extract new patterns. The extracted information is real but presents a perspective which is not considered so far.

## II. Prediction through Mining

Mining biological databases is important for prediction of the likely biological patterns in the species. It helps the medical science identify the risk of potential diseases in the generation to come. It helps researchers to prepare the treatment and medication for the likely disease. It is a proactive approach which can save the generations from painful sufferings of diseases whose treatment is made possible only through the prediction practices. In case the predictions are not made, the medical practitioners will not be able to diagnose the problem in patient, the prescription may be the wrong one and the condition of patient is bound to worsen.

Mining analyses information in a unique way that helps in prediction about future trends and likely patterns to occur [9]. The major point of concern is about the validity if predicted information, which is largely determined by the efficiency of prediction algorithm.

## III. Prediction Mining Algorithms

In Bioinformatics, prediction is made through a well defined series of steps known as algorithm [10]. The algorithm contains the set of rules and procedures which are followed in the information analysis process. In order to predict reliable information, it is mandatory that the underlying algorithm takes into consideration all the possible variables and their interactions [11] which can cause changes in the predicted behaviour [12].

## IV. Statistical Analysis of Mining Algorithm

Prediction is usually based on historical data; certain statistical methods are used in analysing the data. The statistical methods are embedded in the algorithms [13] and they increase the efficiency of prediction algorithm [14]. For example, correlation throws light upon the type of association between variables and the strength of that relationship but it does not reveal the causality [15]. In mining biological databases, where major health related developments are to be made based on the predictions, it is important to measure the causality as well.

By using the linear regression model, the prediction is made easy but the accuracy of predicted value is highly questionable due to approximations and assumptions[16][17]. The prediction on the basis of one variable tends to a misleading prediction. In reality, there are always multiple independent variables leading to change in the values of dependent variables. It is crucial to analyse the combined effect of the variables involved so that the most influential one can be controlled which warrants for a better statistical model like multiple correlation and, or multiple regression analysis to predict the current and future trends. The method usually employed to predict the value of an unknown variable is the regression analysis.

The prediction is made on historical data only and it may be possible that the value predicted may be misleading. If a prediction model is applied to generate the future trends [18] and the validity of the predicted values increase [7] then the algorithm is considered as efficient. The prediction model includes the factors like environment, possible changes in the past data and the context in which the predicted values will be used. For better prediction, the context in which past data is gathered also plays a vital role [19]. The ambient intelligence analyses the data accordingly and presents the trend which is workable for future decisions.

## V. Efficiency Of The Algorithm

The prediction model contains the built in mechanism to create a sequence of value observed over the period of time. The values are analysed with respect to the various identified variables. The algorithm with ambient intelligence uses multiple dimensions of value analysis. The algorithm creates multiple sequences based on complex computational methodology and presents the report to the decision maker. The complexity of the algorithm also determines the result's reliability [20]. The analysis of sequential dataset is the most common feature of prediction algorithms applied in the present era [21]. To support the claims made earlier we consider a case study in breast cancer.

Thus the paper analyzes the efficiency of predictive mining algorithm on the data set related to breast cancer. This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. M. Zwitter and M. Soklic have provided the data [22]. The Arff conversion of the data set was provided by Håkan Kjellerstrand[23]. This data set has 286 instances described by 9 attributes + one class attribute. The set includes 201 instances of one class and 85 instances of another class.

We consider two cases to make predictions – (i) classification by considering data as categorical (ii) classification by considering data as numeric and quantification is done by considering an identifier.

The main aim of the first experiment is to predict the class by finding the efficiency of different algorithms considering data as categorical. The experiment predicts whether the data being tested can be correctly predicted and classified as recurrence class or non-recurrence class. The experiment was done with five different algorithms namely Decision Tree, OneR, PART, JRip and ZeroR. Four different predictive methods – Cross Validation, Percentage Split, Testing data and Training data are used in this analysis. For the cross validation the method used is 10 fold. While using percentage split the number of instances gets reduced to 97 since it uses the 2/3rd method. In the first part the mean absolute error for each of the algorithm was measured since it is one of the widely used statistics for regression.

The following table gives the efficiency of the predictive algorithms using all the four methods in terms of mean absolute error.

| | **ZeroR** | **OneR** | **JRip** | **PART** | **Decision Table** |
|---|---|---|---|---|---|
| **Cross validation** | 0.418 | 0.342 | 0.379 | 0.365 | 0.374 |
| **Percentage Split** | 0.428 | 0.319 | 0.418 | 0.359 | 0.402 |
| **Training set** | 0.418 | 0.273 | 0.354 | 0.299 | 0.329 |
| **Testing Set** | 0.418 | 0.276 | 0.355 | 0.307 | 0.334 |
| **Difference** | 0.010 | 0.069 | 0.064 | 0.066 | 0.073 |

Table 1 – Mean Absolute Error

From Table 1 we can see that for all the algorithms the error is quantifiable. It is obvious that for a best accuracy we wish to obtain the smallest possible value for the error. ZeroR has a consistent high value of error for all the methods. It is observed that variations between the methods are not significant for any given method. We propose the following method to have a clear idea on the efficiency of these algorithms.

A confusion matrix is considered for each of the above algorithms which shows (i) True Positive (TP), (ii) True Negative (TN), (iii) False Negative (FN) and (iv) False Positive (FP).

| | Actual value | |
|---|---|---|
| Predicted Outcome | TP | FP |
| | FN | TN |
| Total | P | N |

Table 2– Confusion Matrix

Normally in medicine a false positive causes unnecessary worry or treatment, while a false negative gives the patient the dangerous illusion of good health and the patient might not get an available treatment. In this case study True Positive denotes the number of no-recurrence events of cancer correctly classified as no-recurrence events. False positive denotes the number of recurrence events of cancer incorrectly identified as no-recurrence events. True negative denotes the number of recurrence events of cancer correctly identified as recurrence events of cancer.

Using these values we calculate the different measures. The measures on which the prediction algorithm is analyzed are (i) number of correctly classified and incorrectly classified instances, (ii) accuracy, (iii) sensitivity, (iv) specificity, (v) positive predictive value (PPV), (vi) Negative predictive value (NPV). The meaning of each parameter is given below for immediate reference.

Sensitivity measures the proportion of the actual positives that are correctly identified as true. Specificity measures the proportion of the negatives that are correctly identified. We note here that a perfect predictor should have 100% sensitivity and 100% specificity. A high PPV means that the predictor always classifies correctly but has to be taken with the value of NPV since predictive values are inherently dependent upon the prevalence. A high NPV means that the predictor rarely misclassifies. In what follows we consider the four different methods to assess the parameters listed above and present the calculated results in the form of tables.

**Using Cross validations**

Table 3 – Cross Validation Confusion Matrix details

| | **ZeroR** | **OneR** | **JRip** | **PART** | **Decision Table** |
|---|---|---|---|---|---|
| **True Positive** | 201 | 166 | 172 | 181 | 186 |
| **True Negative** | 0 | 22 | 31 | 23 | 24 |
| **False Positive** | 0 | 35 | 29 | 20 | 15 |
| **False Negative** | 85 | 63 | 54 | 62 | 61 |

From Table 3 it is observed that OneR has the lowest number of correctly classified instances. This is expected as OneR can be applied only when the instances are few and for large number of attributes or instances the efficiency is poor. The FP value of OneR is the highest for all the algorithms. It means that there are 35 cases when a recurrence event is incorrectly classified as no-recurrent event. The ZeroR algorithm has a 0 value for FP and the highest value of 85 for FN. This also causes a dangerous misclassification since a no-recurrence event is incorrectly classified as a recurrence event. Using Table 3 we get the results of the following table.

| | **ZeroR** | **OneR** | **JRip** | **PART** | **Decision Table** |
|---|---|---|---|---|---|
| **Time taken to build model (in secs)** | 0 | 0 | 0.06 | 0.13 | 0.07 |
| **Correctly classified instances** | 201 | 188 | 203 | 204 | 210 |
| **Incorrectly classified instances** | 85 | 98 | 83 | 82 | 76 |
| **Accuracy** | 0.70 | 0.66 | 0.71 | 0.68 | 0.70 |
| **Sensitivity** | 0.70 | 0.72 | 0.76 | 0.74 | 0.75 |
| **Specificity** | * | 0.39 | 0.52 | 0.44 | 0.45 |
| **Positive Predictive Value (PPV)** | 1.00 | 0.85 | 0.86 | 0.86 | 0.87 |
| **Negative Predictive Value (NPV)** | 0.00 | 0.26 | 0.36 | 0.27 | 0.28 |
| **Kappa Statistic** | 0.00 | 0.09 | 0.24 | 0.19 | 0.24 |

* - cannot be calculated as denominator is zero

Table 4 – Cross validation Analysis

From Table 4 we can note that the time taken by PART to build the model is the highest and the time taken by ZeroR and OneR are the least. The ratio between the number of correctly classified and incorrectly classified instances is approximately 2:1. JRip has the highest accuracy with 71% followed by ZeroR and Decision Table. The Kappa statistic result shows that there is no complete agreement with the true class. All the algorithms have nearly 70% sensitivity. Sensitivity is nearly two-thirds. The test is able to detect two-thirds of the people with the correct cancer and misses one-third of the people. They differ in the values of the specificity with JRip having the highest specificity of 52%. In other words, 31 persons out of 60 persons with negative results are truly negative and 29 individuals test positive for a type of cancer disease which they do not have. ZeroR has the highest PPV which mean that it correctly classifies the class as recurrent class. JRip has the highest NPV of 36%. ZeroR has a 0 NPV which means that the algorithm has the highest probability of misclassification. The value of kappa statistic is very low ranging from 0 to 0.24 for all algorithms which indicates chance agreement.

**Percentage Split – 66%**

| | ZeroR | OneR | JRip | PART | Decision Table |
|---|---|---|---|---|---|
| **True Positive** | 64 | 57 | 62 | 53 | 60 |
| **True Negative** | 0 | 9 | 2 | 14 | 6 |
| **False Positive** | 0 | 7 | 2 | 11 | 4 |
| **False Negative** | 33 | 24 | 31 | 19 | 27 |

Table 5 – Percentage Split Confusion Matrix details

From Table 5 we can see that ZeroR and JRip have the lowest number of correctly classified instances. The FP value of PART is 11 and is the highest for all the algorithms. This causes a dangerous misclassification since a recurrence event is incorrectly classified as no-recurrent event. The ZeroR algorithm has a 0 value for FP and the highest value of 33 for FN. This leads to a dangerous misclassification since a no-recurrence event is incorrectly classified as a recurrence event. Using Table 5 we get the results of the following table.

| | ZeroR | OneR | JRip | PART | Decision Table |
|---|---|---|---|---|---|
| **Time taken to build model (in secs)** | 0.00 | 0.00 | 0.02 | 0.06 | 0.14 |
| **Correctly classified instances** | 64 | 66 | 64 | 67 | 66 |
| **Incorrectly classified instances** | 33 | 31 | 33 | 30 | 31 |
| **Accuracy** | 0.66 | 0.68 | 0.66 | 0.69 | 0.68 |
| **Sensitivity** | 0.66 | 0.70 | 0.67 | 0.74 | 0.69 |
| **Specificity** | * | 0.56 | 0.50 | 0.56 | 0.60 |
| **Positive Predictive Value (PPV)** | 1.00 | 0.89 | 0.97 | 0.83 | 0.94 |
| **Negative Predictive Value (NPV)** | 0.00 | 0.27 | 0.06 | 0.42 | 0.18 |
| **Kappa Statistic** | 0.00 | 0.18 | 0.03 | 0.26 | 0.14 |

* - cannot be calculated as denominator is zero

Table 6 – Percentage Split analysis

From Table 6 we can ascertain that the time taken by Decision Table to build a model is the highest. The time taken by ZeroR and OneR is the least. The ratio between the number of correctly classified and incorrectly classified instances is approximately 2:1. Since the method uses only 1/3$^{rd}$ of the instances the total number of instances is reduced to 97. PART has the highest accuracy with 69%. PART has approximately 74% sensitivity. All the algorithms have nearly 70% specificity except ZeroR. Though ZeroR has 100% PPV it has 0% NPV which means that the algorithm has the highest probability of misclassification. PART has the highest NPV. The value of kappa statistic is very low ranging from 0 to 0.26 for all algorithms which indicates chance agreement.

**Using Training set**

| | ZeroR | OneR | JRip | PART | Decision Table |
|---|---|---|---|---|---|
| **True Positive** | 201 | 189 | 190 | 184 | 196 |
| **True Negative** | 0 | 19 | 30 | 45 | 29 |
| **False Positive** | 0 | 12 | 11 | 17 | 5 |
| **False Negative** | 85 | 66 | 55 | 40 | 56 |

Table 7 – Training Set Confusion Matrix details

From Table 7 we can see that ZeroR and JRip have the lowest number of correctly classified instances. The FP value of PART is 17 and is the highest for all the algorithms. This causes a dangerous misclassification since a recurrence event is incorrectly classified as no-recurrent event. The ZeroR algorithm has a 0 value for FP and the highest value of 85 for FN. This leads to a dangerous misclassification since a no-recurrence event is incorrectly classified as a recurrence event. Using Table 7 we get the results of the following table.

| | **ZeroR** | **OneR** | **JRip** | **PART** | **Decision Table** |
|---|---|---|---|---|---|
| **Time taken to build model (in secs)** | 0.00 | 0.00 | 0.01 | 0.01 | 0.07 |
| **Correctly classified instances** | 201 | 208 | 220 | 229 | 225 |
| **Incorrectly classified instances** | 85 | 78 | 66 | 57 | 61 |
| **Accuracy** | 0.70 | 0.74 | 0.77 | 0.80 | 0.79 |
| **Sensitivity** | 0.70 | 0.75 | 0.78 | 0.82 | 0.78 |
| **Specificity** | * | 0.61 | 0.73 | 0.73 | 0.85 |
| **Positive Predictive Value (PPV)** | 1.00 | 0.94 | 0.95 | 0.92 | 0.98 |
| **Negative Predictive Value (NPV)** | 0.00 | 0.22 | 0.35 | 0.53 | 0.34 |
| **Kappa Statistic** | 0.00 | 0.20 | 0.35 | 0.48 | 0.38 |

* - cannot be calculated as denominator is zero

Table 8 – Training Set analysis

From Table 8 we can observe that the time taken by Decision Table to build a model is the highest. The time taken by ZeroR and OneR is the least. The ratio between the number of correctly classified and incorrectly classified instances is approximately 3:1. PART has the highest accuracy of 80%. All the algorithms have more than 70% sensitivity with PART having the highest at 82%. All algorithms have above 60% specificity except ZeroR. Decision Table has the highest specificity with 85%. Though ZeroR has 100% PPV it has 0% NPV which means that the algorithm has the highest probability of misclassification. PART has the highest NPV. The Kappa statistic result shows that there is no complete agreement with the true class but since PART has approximately 0.5 it means that it is nearing to perfect agreement.

**Using Testing data**

| | **ZeroR** | **OneR** | **JRip** | **PART** | **Decision Table** |
|---|---|---|---|---|---|
| **True Positive** | 201 | 186 | 189 | 184 | 196 |
| **True Negative** | 0 | 19 | 30 | 42 | 27 |
| **False Positive** | 0 | 15 | 12 | 17 | 5 |
| **False Negative** | 85 | 66 | 55 | 43 | 58 |

Table 9 – Testing Set Confusion Matrix details

The Table 9 illustrates that ZeroR and OneR have the lowest number of correctly classified instances. The FP value of PART is 17. This causes a dangerous misclassification since a recurrence event is incorrectly classified as no-recurrent event. The ZeroR algorithm has a 0 value for FP and the highest value of 85 for FN. This leads to a dangerous misclassification since a no-recurrence event is incorrectly classified as a recurrence event. All the algorithms have a considerable higher value of FN. Using Table 9 we get the results of the following table.

| | **ZeroR** | **OneR** | **JRip** | **PART** | **Decision Table** |
|---|---|---|---|---|---|
| **Time taken to build model (in secs)** | 0.00 | 0.00 | 0.57 | 0.02 | 0.06 |
| **Correctly classified instances** | 201 | 207 | 222 | 222 | 217 |
| **Incorrectly classified instances** | 85 | 79 | 64 | 64 | 69 |
| **Accuracy** | 0.70 | 0.72 | 0.77 | 0.79 | 0.78 |
| **Sensitivity** | 0.70 | 0.74 | 0.77 | 0.81 | 0.77 |
| **Specificity** | * | 0.56 | 0.71 | 0.71 | 0.84 |
| **Positive Predictive Value (PPV)** | 1.00 | 0.93 | 0.94 | 0.92 | 0.98 |
| **Negative Predictive Value (NPV)** | 0.00 | 0.22 | 0.35 | 0.49 | 0.32 |
| **Kappa Statistic** | 0.00 | 0.18 | 0.34 | 0.44 | 0.35 |

* - cannot be calculated as denominator is zero

Table 10 – Testing Set analysis

The testing data method uses all the instances in the testing set. From Table 10 we note that the time taken by JRip to build a model is the highest. The time taken by ZeroR and OneR is the least. The ratio between the number of correctly classified and incorrectly classified instances is around 3:1. PART has the highest accuracy with it being 79%. All algorithms have approximately 70% and above sensitivity with PART 81%. All algorithms have above 70% specificity except ZeroR and OneR. Decision Table has the highest specificity. Decision Table has the highest PPV with 98%. Though ZeroR has 100% PPV it has a 0 NPV which means that the algorithm has the highest probability of misclassification. PART has the highest NPV with 49%. The Kappa statistic result shows that there is no complete agreement with the true class. The maximum agreement towards perfect agreement is 0.44 using PART algorithm.

From the above tables it is found that for ZeroR algorithm specificity cannot be calculated. This means that the algorithm has the highest probability of misclassification. The algorithm also has the highest number of FN value for all the four methods. It is also found that the Kappa Statistic is the highest for the method of training data set. It is has a 50% agreement with the true class. PART has approximately 50% agreement with the true class. The number of correctly classified instances is considerably more than the number of incorrectly instances using all the methods for all algorithms except ZeroR. JRip, PART and Decision Table algorithms have a consistent sensitivity value using all the four models. JRip and PART have a nearly 90% sensitivity for the testing data model. The analysis of the experimental results leads to the understanding that the steps of algorithm, the models embedded in the algorithm and the statistical methods used in prediction process play an important role in the prediction of data.

The main aim of the second experiment is to predict by quantifying an identifier. To do this we have considered mode of the dataset with respect to each attribute. This is similar to OneR applied earlier.

1. The prediction done on node-caps shows that highest frequency of cancer is in the class ‘no’. The maximum of instances classified is 201 and 21 incorrectly predicted.
2. The prediction done on inv-nodes shows that highest frequency of cancer is in the period ‘0-2’. The maximum of instances classified is 204 and 9 incorrectly predicted.
3. The prediction done on irradiat shows that highest frequency of cancer is in the class ‘no’. The maximum of instances classified is 210 and 8 incorrectly predicted.
4. The prediction done on menopause shows that highest frequency of cancer is in the class ‘prem_no’. The maximum of instances classified is 120 and 30 incorrectly predicted.
5. The prediction done on deg-malign shows that highest frequency of cancer is in the class ‘2’. The maximum of instances classified is 113 and 17 incorrectly predicted.
6. The prediction done on breast shows that highest frequency of cancer is in the region ‘left’. The maximum of instances classified is 107 and 45 incorrectly predicted.
7. The prediction done on age shows that highest frequency of cancer is in the period ‘40 – 49’. The maximum of instances predicted is 81 and 9 incorrectly predicted.
8. The prediction done on breast-quad shows that highest frequency of cancer is in the region ‘left_low’. The maximum of instances classified is 75 and 35 incorrectly predicted.
9. The prediction done on tumour size shows that highest frequency of cancer is in the period ‘25 – 29’. The maximum of instances classified is 25 and 29 incorrectly predicted.

The individual modes of each attribute are as follows

| Attributes | Node-caps | Inv-nodes | Irradiat | Menopause | Deg-malign | Breast | Age | Breast-quad | Tumour-size |
|---|---|---|---|---|---|---|---|---|---|
| Modal values | 201 | 204 | 210 | 120 | 113 | 107 | 81 | 78 | 25 |

Table 11 – Individual Quantifying Statistics

The second experiment predicts based on identifiers i.e. attributes that contribute most to the prediction. The experiment identifies the attribute that makes the most impact on the result of prediction and then using that checks the next attribute. This continues until a significant mode is present. The attributes are arranged in the order of maximum correctly classified instances. The table below gives the different mode values keeping a previous attribute as a constant in this experiment.

| Attributes | Irradiat | Inv-nodes | Node-caps | Menopause | Deg-malign | Breast | Age | Breast-quad | Tumour-size |
|---|---|---|---|---|---|---|---|---|---|
| Modal values | **218** | 218 | 218 | 218 | 218 | 218 | 218 | 218 | 218 |
| | 218 | **183** | 183 | 183 | 183 | 183 | 183 | 183 | 183 |
| | 218 | 183 | **177** | 177 | 177 | 177 | 177 | 177 | 177 |
| | 218 | 183 | 177 | **90** | 90 | 90 | 90 | 90 | 90 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 218 | 183 | 177 | 90 | **NA** | | | | |
| | 218 | 183 | 177 | 90 | | **NA** | | | |
| | 218 | 183 | 177 | 90 | | | **NA** | | |
| | 218 | 183 | 177 | 90 | | | | **NA** | |
| | 218 | 183 | 177 | 90 | | | | | **NA** |

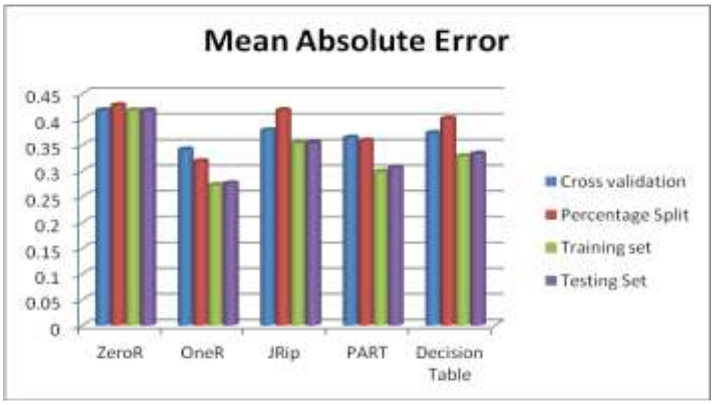
Table 12 – Quantifying Statistics

From Table 12 it is found that the attributes of irradiat, inv-nodes, node-caps and menopause helps in the prediction of new class. A rule is identified from the above table that helps in the prediction of the class.

```
If there is no irradiation and
   If inv-nodes between 0 and 2
      If there is no node caps and
         If menopause is between 40 and 49
            then class is identified
```

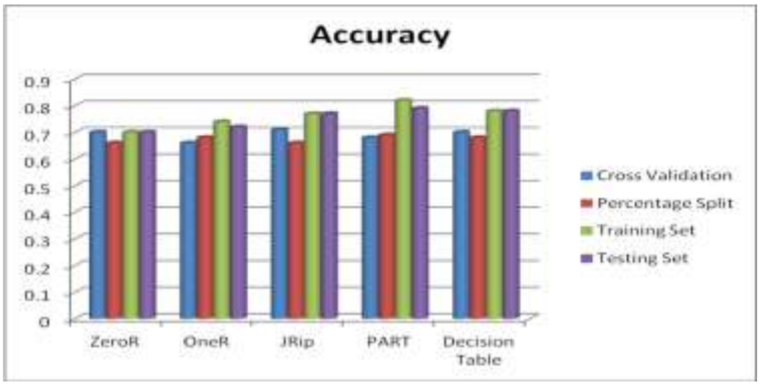
## VI. Results and Discussions

This paper deals with finding efficiency of an algorithm given a large dataset. We have considered five algorithms namely Decision Tree, PART, OneR, JRip and ZeroR and four different predictive methods of cross validation, percentage split, testing data and training data. We first considered the mean absolute error for each of the algorithms in each method. Further we have considered the confusion matrix for the calculation. The measures on which the prediction algorithms are analyzed are (i) number of correctly classified and incorrectly classified instances, (ii) accuracy, (iii) sensitivity, (iv) specificity, (v) positive predictive value (PPV), (vi) Negative predictive value (NPV). In what follows we represent the calculated results in the form of graphs.



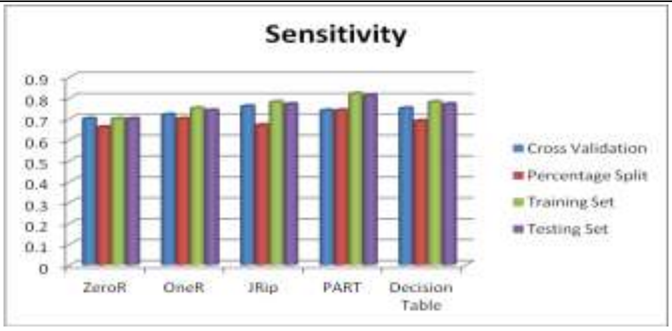
Graph 1 – MAE Statistics

The analysis of Graph 1 shows that ZeroR has a consistently high mean absolute error of 0.4. For prediction it is always preferrable to have a low value for the mean absolute error. It is difficult to determine a better predictive algorithm based on only the MAE statistics. We consider the other measures to determine the best predictive algorithm.



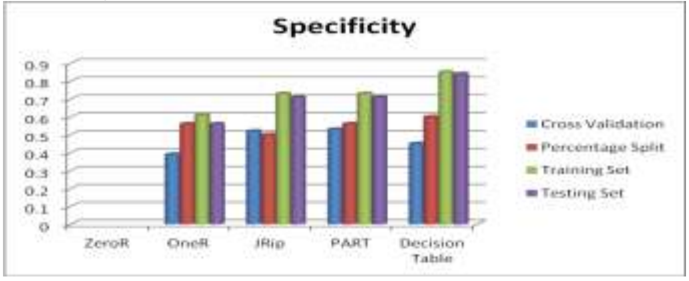
Graph 2- Accuracy Statistics

The analysis of Graph 2 shows that for all the algorithms the accuracy is nearly 70% for the methods of Training Set and Testing Set.



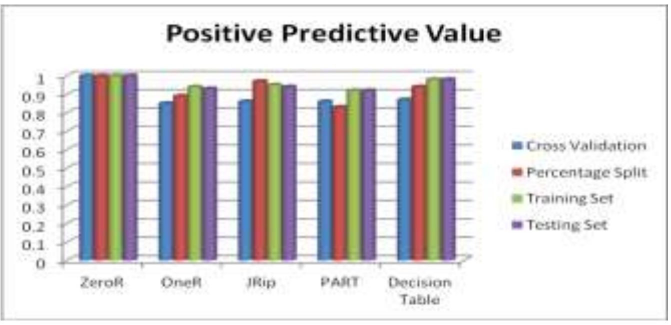
Graph 3- Sensistiviy Statistics

The analysis of Graph 3 shows that PART has a good sensitivity rate with respect to the the method of testing set. Overall the sensitivity of PART is above 70%. Decison Tree and JRip have nearly the same performance in terms of sensitivity in all the methods.



Graph 4 – Specificity Statistics

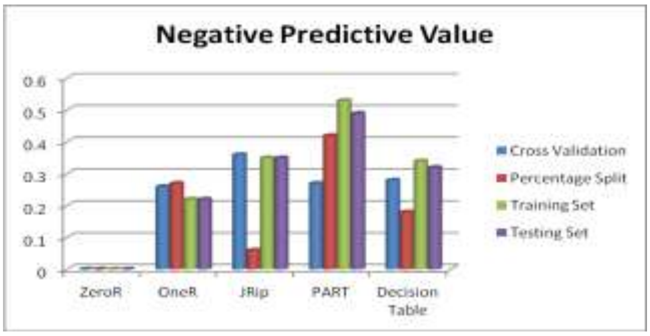
The analysis of Graph 4 shows that Decision Table has the best specificity rate for training data set and testing data set method. All the algorithms perform well in the specificity measure with respect to all methods except OneR which has a considerably lower measure for Cross Validation (around 30%) . ZeroR’s specificity measure cannot be calculated



.

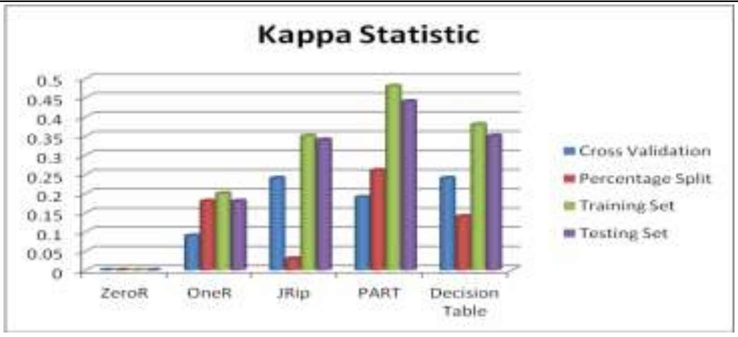
Graph 5 – PPV Statistics

The analysis of Graph 5 shows that all the algorithms have a consistently good PPV value of more than 0.8. ZeroR has the best PPV value of 1.



Graph 6 – NPV Statistics

The analysis of Graph 6 shows that PART has the best NPV value of all the algorithms with respect to the method of percentage split, training data set and testing data. ZeroR has a NPV value of 0 which shows that it has the highest probability of misclassification.



Graph 7 – Kappa Statistics

The analysis of Graph 7 shows that PART has a Kappa value ranging from 0.2 to nearly 0.5 for the methods. ZeroR has a Kappa value of 0 which shows that there is no complete agreement to the true class of prediction. Except OneR none of the algorithms have a consistent Kappa Statistic.

The analysis of the second experiment based on mode as an identifier predicts that attributes irradiat, inv-nodes, node-caps and menopause make the monotonically decreasing impact on the result of prediction which leads to deducing a classification rule.

## VII. Conclusion

From the analysis of the results we find that PART algorithm categorises the data to the true class better than the other algorithms. The method of cross validation yields the most consistent accurate result for all the different algorithms used for prediction. ZeroR algorithm results in 0% specificity and 0 NPV which means that this algorithm has the highest probability of misclassifications. In cross validation method Decision Tree algorithm has the best accuracy and specificity. PART algorithm has the best accuracy in percentage split method and in testing set method but not the best specificity. In training set method PART algorithm has the highest accuracy and specificity.

All though the results of the case study are presented by considering the parameters (i) number of correctly classified and incorrectly classified instances, (ii) accuracy, (iii) sensitivity, (iv) specificity, (v) positive predictive value (PPV), (vi) Negative predictive value (NPV), (vii) Kappa Statistic the numerical computations based on error reveals that for all the algorithms except ZeroR the mean absolute error is not uniform. The mean absolute error of ZeroR algorithm has a consistent mean absolute error which is around 0.45 which shows that the algorithm is not a good predictor. The second experiment enables us to deduce a classification rule.

From the above analysis we understand that the efficiency of a mining algorithm is found to be the function of many variables such as dataset consisting of huge historical information, the perspective of collected information, context in which predicted information will be used. The steps of algorithm, the models embedded in the algorithm and the statistical methods used in prediction process also play an important role. Prediction model is not a linear model to the present case study. We also note that reliable results can be produced if the mentioned points are carefully analyzed in the algorithm design.

## VIII. Acknowledgement

The authors would like to thank the college management for supporting to this research work.

## References

[1] Avery, John. *Information Theory and Evolution.* USA: World Scientific Publishing Co. 2003.
[2] David Weatherall, Brian Greenwood, Heng Leng Chee and Prawase Wasi, *Science and Technology for Disease Control: Past, Present, and Future*, Disease Control Priorities in Developing Countries
[3] Leser, Ulf, et al. *Data integration in the Life Science.* USA: Springer. 2006.
[4] Paradis, Emmanuel. *Analysis of Phylogenetics and Evolution with R.*USA: Springer. 2011.
[5] Selzer, Paul. *Applied Bioinformatics: An Introduction.* USA: Springer. 2008.
[6] Rob, Peter, et al. *Database systems: design, implementation and management.* USA: Cengage Learning. 2009.
[7] *SDART.* Software Design and Research Technology Ltd., n.d. Web. 29 April 2012.
[8] Ramaswamy, Sridhar, et al. *Efficient Algorithms for Mining Outliers from Large Data Sets.* The Pennsylvania State University, 2010. Web. 29 April 2012.
[9] *Cornell University Library.* “A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.)” 2012. Web. 29 April 2012.
[10] Pandey, Hari. *Design Analysis and Algorithm.* New Delhi: University Science Press. 2008.
[11] Crawley, Michael. *Statistics: An Introduction Using R.* USA: John Wiley & Sons. 2011.
[12] Bertsimas, Dimitris, et al. “Algorithm Prediction of Health-Care Costs.” *Operations Research* 56. 6 (2008): 1382-1392.
[13] Rossiter. *An Introduction to Statistical Analysis.* Netherlands: International Institute of Geo Science and Earth Observation. 2006.
[14] Bartz-Beielstein, Thomas, et al. *Experimental Methods in Algorithm Design and Analysis.* USA: Springer. 2010.
[15] Bonate, Peter. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation.* USA: Springer. 2011.
[16] Yan, Xin, et al. *Linear Regression Analysis: Theory and Computing.* USA: World Scientific Publishing Co. 2009.

[17] Sen, Zekai. *Spatial Modeling Principles in Earth Sciences*. USA: Springer. 2009.

[18] *Technet*. Microsoft, n.d. Web. 29 April 2012.

[19] Anagnostopoulos. *Path Prediction through Data Mining*. ICPS, n.d. Web. 29 March 2012.

[20] Hof, Paul. *System Identification 2003*. UK: Elesevier Ltd. 2004.

[21] Djahantighi, Farhad et al. "An Effective Algorithm for Mining User behavior in Time-Periods." *European Journal of Scientific Research* 40. 10 (2010): 81-90.

[22] Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). "*The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains*." In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann.

[23] http://www.hakank.org/weka/BC.arff (25063)

## Appendix

1. ***Accuracy***

   *The rate of correct (incorrect) predictions made by the model over a data set (cf. coverage). Accuracy is usually estimated by using an independent test set that was not used at any time during the learning process. More complex accuracy estimation techniques, such as cross-validation and the bootstrap, are commonly used, especially with data sets containing a small number of instances.*

2. ***Sensitivity***

   *It measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).*

3. ***Specificity***

   *It measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition). A perfect predictor would be described as 100% sensitivity (i.e. predict all people from the sick group as sick) and 100% specificity (i.e. not predict anyone from the healthy group as sick), however theoretically any predictor will possess a minimum error.*

4. ***Positive Predictive Value(PPV)***

   *It is the proportion of positive test results that are true positives (such as correct diagnoses). It is a critical measure of the performance of a diagnostic method, as it reflects the probability that a positive test reflects the underlying condition being tested for. Its value does however depend on the prevalence of the outcome of interest, which may be unknown for a particular target population.*

5. ***Negative predictive value (NPV)***

   *It is a summary statistic used to describe the performance of a diagnostic testing procedure. It is defined as the proportion of subjects with a negative test result who are correctly diagnosed. A high NPV for a given test means that when the test yields a negative result, it is most likely correct in its assessment. In the familiar context of medical testing, a high NPV means that the test only rarely misclassifies a sick person as being healthy. Note that this says nothing about the tendency of the test to mistakenly classify a healthy person as being sick.*

6. ***Kappa Statistic***

   *The kappa statistic measures the agreement of prediction with the true class -- 1.0 signifies complete agreement. Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement.*