Empirical Study on Data Mining Algorithms related to Breast Cancer

Jerrin James¹ and Sarvanakumar K.²

(¹PG Scholar, Department of Computer Science, Christ University, Hosur Road, Bengaluru, Karnataka)
(²Associate Professor, Department of Computer Science, Christ University, Hosur Road, Bengaluru, Karnataka)

Abstract: Data Mining is one of the major research areas that is used in the medical field. Research work in this area is aimed to make prediction of diseases more accurate and efficient. Breast Cancer is one of the major domain of data mining where big data play an important role. There are many algorithms available in data mining to analyze breast cancer. The algorithms used are J48 decision tree algorithm and Naïve Bayes algorithm. Through the research work, the result will make it easy for selecting the appropriate algorithm and the way for which algorithm gives better efficiency. Weka is the tool used to compare the two algorithms. The technology advancement has made research more precise.

Keywords: Decision Tree, Naïve Bayes, J48, Breast Cancer

I. INTRODUCTION

Through data mining, a process of finding essential relation between the data is done and then that is used to find the relation for future advancement and research. Trying to create a prediction models for future cases of breast cancer. Making a prediction[1] has always been a challenging task as it is not easy to get and create a relation between the store data, thus through data mining the task of prediction has been controlled to a considerable level.

Breast cancer[2] is a disease where malignant cells form in the tissues of the breast. It is the second leading cause of death among women. The leading risk factor for breast cancer is simply being a woman. Breast cancer does occur in men and the disease is 100 times more common in women than in men and women are at 200 times the risk of developing the disease compared to risk in men. Most breast cancer – about 85 percent – occurs in women who have no family history of breast cancer.

Need of the study:

The requirement for the paper is to help fellow researchers for future understanding of what algorithms should preferred in the two which make the research work faster and easier. Thus with the help of the work selection for better algorithm can be considered.

II. LITERATURE SURVEY

Breast cancer is the second most cause cancer found in the world today. The only way to cure breast cancer is to predict it in the early stage itself. The prediction[3] of breast cancer has been a research problem for many researchers. Since the early dates of the related research, much advancement is seen in several fields related to breast cancer, thanks to innovative biomedical technologies, better and proper explanatory prognostic factors are being measured and recorded. The advancement of technology have given an era of cheaper software and hardware to store data properly of any amount. Creating a framework for the research makes it easy for research scholars to run their data and enhance the efficiency. Thus graphical user interface makes working more interactive and easy. Segmentation is an important method to extract information from complex medical images and has wide application in medical field. The main objective of the image segmentation is to separate or partition an image into mutually exclusive and exhausted regions so that each Region of Interest (ROI) is spatially contiguous and the pixels within the region are homogeneous with respect to a predefined criterion. Segmentation is done to decompose or separate an image into meaningful parts for further analysis, resulting in a higher level presentation or representation of the image pixels like the front objects and the background. It is often seen in medical images, such as X-ray radiography/tomography and MRI (Magnetic Resonance Image). For example, the intensity in homogeneity in MRI often appears as intensity differs across the image represented, which generates from Radio Frequency (RF) coils or acquisition sequences. Thus the resultant intensities of the similar tissue vary with the locations or areas in the image. The noise (the disturbance) in MR images is Rician distributed and can affect the performances of Classification methods. Brain Image Segmentation based on Self-Organizing Map Network Magnetic Resonance (SOMNMR) Imaging is an advanced medical imaging method providing rich information about the human soft tissue anatomy. The aim of MRI segmentation or separation is to accurately identifying the principal tissue structures in the image volumes. A new unsupervised MR image segmentation technique based on self-organizing feature. Image segmentation is typically used to locate object and boundaries in image. The result of image segmentation is a set of regions that collectively cover the full image, or a set of outline extracted from the image.
In this modern state of medical development and research, breast cancer is one of the main research areas where prediction and finding any relation amongst the data is very important. Therefore, there are many algorithms to overcome this problem but to get the best and most suitable algorithm is actually a difficult situation. With this research, two algorithms are compared to find the more efficient one.

III. ALGORITHM

Decision Tree:

A decision tree \(^4\) is a flowchart-like structure in which each internal node represents a condition on an attribute which can be true or false, each branch represents the outcome of the test and each leaf node represents a class label. The paths from root to leaf represent classification rules.

In decision making or analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values or result of competing alternatives are calculated.

In a decision tree there are types of nodes:
- Decision nodes - depicted by squares
- Chance nodes - depicted by circles
- End nodes - depicted by triangles

Decision tree algorithm is commonly implemented or used in operation research and operation management. If in process decisions have to be taken online without revoking under incomplete knowledge, a decision tree should be paralleled by a model of probability as a best and proper choice model or an algorithm of online selection model. Another use of decision tree is as a means for calculating conditional probabilities in a descriptive manner.

J48 Decision Tree:

J48\(^5\) is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA\(^6\) data-mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms, the classification is performed recursively until every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Naïve Bayes:

Naïve Bayes\(^6\) is a simple procedure for constructing classifiers: models that assign class labels to instances of a problem, represented as vectors of feature values, where from a finite set the class label are drawn. It is not a single algorithm alone for training such classifiers, but a family or group of algorithms based on a common principle which states that all naive Bayes classifiers considers that the value of a particular feature is not dependent of the value of any other feature, given the class variable. For example, a fruit might be assumed to be an apple if it is red, round, and about 10 cm in diameter. A naïve Bayes classifier considers each of these features to contribute solo to the probability that this fruit is an apple, regardless of any possible co-relations between the colors or roundness or diameter features.

Naïve Bayes classifiers are highly scalable, requiring a number of attributes linear in the number of variables (features/predictors) in a learning problem. Maximum-similarity training can be done by evaluating a closed-form expression, which takes linear time, rather than by costly iterative approximation as used for many other types of classifiers.

IV. METHODOLOGY

Weka Breast Cancer dataset is being used as the dataset for the process. Which is an already pre-processed dataset and where the Remove Missing Value with mean function (Replace Missing Value With User Constrant) of the Weka Tool removes the missing data. This function put the mean value of the particular column in missing cell of the dataset. Both the algorithms are being executed and the efficiency is taken for both algorithms.

The details of the dataset are as follows:

Relation:
The research being done on which domain is specified here. It gives a brief description about the research.

Instance:
It relates to the number of sample being used to check or analyse given in the dataset. For example, here consider one patient as one instance of the given domain. Every patient is taken as an instance of the domain.

Attributes:
Attributes relates to the characteristic of the given sample data. Listed below are the major 10 parameter for the algorithm. There are many attributes but these 10 constitute maximum to the result.

- **Age**: Patients’s age
- **Menopause**: Ceasing of menstruation
- **Tumour-size**: Size of the cancer tumour
- **Inv-nodes**: Lymph nodes that contain metastatic breast cancer visible on histological examination
- **Node-caps**: Capsule of the lymph nodes breasts
- **Breast-Quad**: Breast divided into four quadrant with nipple as center
- **Irradiat**: Irradiation is an X-ray to kill cancer cell
- **Class**: The class of the classifier

The Output and the execution of the algorithm show different term’s which relates to the result. The terms are as follows

- **Correctly Classified Instances**: The scenario where the instance of testing data correctly matches with the training data.
- **Incorrectly Classified Instances**: The scenario where the instance of testing data does not match with the training data.
- **Kappa Statistics**: Studies that measure the agreement between two or more observers should include a statistic.
- **Mean Absolute Error**: Quantity used to measure how close forecasts or predictions are to the eventual outcomes.
- **Root Mean Squared Error**: Measure of the differences between values predicted by a model or an estimator and the values actually observed.
- **Relative Absolute Error**: Magnitude of the difference between the exact value and the approximation value.
- **Relative Squared Error**: What it would have been if a simple predictor had been used.

**VI. RESULTS**

The two algorithms are executed using the weka tool and the method of Cross Validation fold also called as K-fold cross validation. In k-fold cross-validation, the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining K−1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

**Algorithm-J48**

<table>
<thead>
<tr>
<th>Relation:</th>
<th>breast-cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances:</td>
<td>286</td>
</tr>
<tr>
<td>Attributes:</td>
<td>10</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Menopause</td>
<td></td>
</tr>
<tr>
<td>Tumour-size</td>
<td></td>
</tr>
<tr>
<td>Inv-nodes</td>
<td></td>
</tr>
<tr>
<td>Node-caps</td>
<td></td>
</tr>
<tr>
<td>Breast-Quad</td>
<td></td>
</tr>
<tr>
<td>Irradiat</td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td></td>
</tr>
</tbody>
</table>

| Correctly Classified Instances | 216 | 75.5265 % |
| Incorrectly Classified Instances | 70  | 24.4735 % |
| Kappa statistics        | 0.2316 |
| Mean absolute error      | 0.6075 |
| Root mean squared error  | 0.4324 |
| Relative absolute error  | 17.8855 % |
| Root relative squared error | 94.4693 % |
| Total Number of Instances | 286 |
In J48 algorithm, it is seen that the number of correctly classified instance are 216 and the efficiency percentage is 75.52%, which is more than the Naïve Bayes algorithm. Therefore, J48 algorithm provides more precise result, which is more efficient and more reliable.

**Naïve Bayes Algorithm:**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>265</td>
<td>71.5783 %</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>81</td>
<td>28.4217 %</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.2657</td>
<td></td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.3372</td>
<td></td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.4694</td>
<td></td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>78.2606</td>
<td></td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>59.1872</td>
<td></td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>288</td>
<td></td>
</tr>
</tbody>
</table>

In the case Naïve Bayes algorithm, the efficiency is lower than J48 algorithm by around 6% which is high number when taking in consideration of medical field. More the efficiency more the algorithm will be preferred and accepted.

VII. CONCLUSION

Medical domain being the major area where research and technology advancement is very essential. Research in medical field has helped in eradicating many major diseases at their initial stage like cancer of different form and various other diseases as well. With the help of research the possibility of occurrence of diseases has been reduced and controlled. This has made people to take necessary action and steps against such possibility before its occurrence.

This research finds the better algorithm in the domain of breast cancer to get better efficient result. The two algorithms compared here are J48 decision tree algorithm and Naïve Bayes algorithm. In which it is very much evident that J48 decision tree algorithm provide better result with a difference of 8% which is indeed a major margin between the two algorithm. The result is more efficient in the case of J48. With the percentage split of 60% to 40% that is 60% of training data and 40% percentage of testing data. Therefore, with the help of the J48 decision tree algorithm we can get better result.

**Limitations:**

Only two algorithms are being implemented here, which only gives an option to select either one of them where in there are multiple algorithm available for different scenario.

**Future Enhancement:**

Maximum algorithms to be put into the paper which provide more option to the researchers. A tool can be implemented for the selection of algorithms which gives a simple interface for the user to select an algorithm and get the efficiency.

**Sources of Funding:**

The above mentioned case study is self-funded.

VIII. REFERENCES


