Comparative analysis on bayesian classification for breast cancer problem

Wan Nor Liyana Wan Hassan Ibeni, Mohd Zaki Mohd Salikon, Aida Mustapha, Saiful Adli Daud, Mohd Najib Mohd Salleh
Faculty Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

ABSTRACT
The problem of imbalanced class distribution or small datasets is quite frequent in certain fields especially in medical domain. However, the classical Naive Bayes approach in dealing with uncertainties within medical datasets face with the difficulties in selecting prior distributions, whereby parameter estimation such as the maximum likelihood estimation (MLE) and maximum a posteriori (MAP) often hurt the accuracy of predictions. This paper presents the full Bayesian approach to assess the predictive distribution of all classes using three classifiers; naïve bayes (NB), bayesian networks (BN), and tree augmented naïve bayes (TAN) with three datasets; Breast cancer, breast cancer wisconsin, and breast tissue dataset. Next, the prediction accuracies of bayesian approaches are also compared with three standard machine learning algorithms from the literature; K-nearest neighbor (K-NN), support vector machine (SVM), and decision tree (DT). The results showed that the best performance was the bayesian networks (BN) algorithm with accuracy of 97.281%. The results are hoped to provide as base comparison for further research on breast cancer detection. All experiments are conducted in WEKA data mining tool.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

1. INTRODUCTION
Breast cancer is the top number one cancer in Malaysia. The statistic shows that 1.6 million new cases are diagnosed worldwide and 2,015,560 women will die of breast cancer every year. This information is sourced from the World Health Organization (WHO) and National Cancer of Malaysia [1]. Statistics from Malaysian Study on Cancer Survival (MySCAN) published by Ministry of Health in September 2018 has shown a total of 17009 patient suffering from breast cancer with 7372 death reported. Age group of 45-54 years old (34.9%) make up the largest patient and 75 years old and above (3.9%) is the smallest group of patient.

According to the National Breast Cancer Foundation [2], there are two type of tumors: malignant (cancerous) and benign (non-cancerous). The cancerous tumors aggressively invade and damage surrounding tissues. The malignant tumors have three grades to differentiate, where the lowest grade of malignant tumor is well differentiated while the highest grade is poorly differentiated. High grade tumors highly resemble healthy cells and have higher tendency to be aggressive. There is no exact symptom can accurately diagnose the cancer but breast cancer has symptom like swelling at some part or the entire breast, dimpling or skin irritation, breast or nipple pain, thickening or retraction of nipple, nipple discharge and lump in the underarm.
area. Breast cancer on male is rare but born as a male is not an exception to the risk of having breast cancer. The risk of cancer is about 1 in 1000.

Breast cancer research falls under the category of medical, which as in other fields, use data mining to analyze past experiences and identify trends and solutions to the present situations [3]. Data mining is well-known analytical methodology to extract such invaluable information and is especially efficient to work with large volume of medical data [4]. The methodology varies from predictive models that enable classification and prediction as well as clustering models that for discovering groups or patterns from data [5]. Salama et al. [6] compared five different classifiers using three different breast cancer datasets. The chosen classifiers were Naïve Bayes (NB), Multi-Layer Perception (MLP), Decision Tree (J48), Instance-based for K-Nearest Neighbor (IBK), and Sequential Minimal Optimization (SMO). Along the line, Aruna et al. [7] investigated the performance of different classification algorithms similar to [6], which were Naïve Bayes and Decision Tree J4 as well as new algorithms such as the Support Vector Machines, Radial Basis Neural Networks, and and simple Classification and Regression Trees (CART). Meanwhile, Bashir et al. [8] proposed a feature selection component to classification problem while maintaining conventional classification algorithms such as the Decision Tree, Bayesian algorithms, Rule-based algorithms, Neural Networks, Support Vector Machines, Associative classification, Distance-based methods, and Genetic algorithms.

Following previous research, this work used breast cancer datasets obtained from UCI Machine Learning Repository from [9], which are Breast Cancer [10], Breast Cancer Wisconsin [11], and Breast Tissue [12]. This paper presents comparative analysis on the performance of three classification algorithms, which are the Naïve Bayes (NB) [13], Bayesian Networks (BN), and Tree Augmented Naïve Bayes (TAN) [14] as well as with algorithms from the existing literature such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Trees (DT) based on breast cancer dataset. Following [15], this paper will be using accuracy, precision, recall and F-measure for evaluation [16]. These standard measures have significantly higher correlation with human judgments and an intuitive interpretation. NB classifiers improved accuracy with high speed [17-19]. The results showed that the NB classifier provides better performances [20-21]. The accuracy, sensitivity and specificity up to 90% and above [22]. Dian E. R., Nurizal D. P. & Machsus Machsus [23] is promising and able to enhance the prediction on Breast Cancer Wisconsin data.

The rest of the paper is organized as follows. Section 2 describes the research methodology and the proposed classification approach on dataset for finding the best performance of algorithm. Section 3 shows the experimental results and finally Section 4 concludes the work and highlights a direction for future research.

2. RESEARCH METHOD

This section presents a data mining approach for a classification problem. A data mining process that describe about data mining approach to tackle the problem. This methodology is implement to get the best result from the classification experiment. In this, research using Knowledge Discovery in Database (KDD) from [14] as model methodology. There are important phases that have to get the best result at the research. There are five phase in KDD which are Selection, Preprocessing, Transformation, Classification and Evaluation. Figure 1 shows the KDD model. Based on Figure 1, KDD model is an iterative process when evaluation measures can be enhanced. The detail of each phase is discussed in the next sub-sections.

![Figure 1. KDD model [14]](image-url)
2.1. Dataset selection

This research used three breast cancer dataset was obtained from UCI machine learning repository from Brown [9], which are breast cancer (WDBC), Wisconsin breast cancer (WBC), and breast tissue. The Breast Cancer Wisconsin dataset contains 699 instances with two classes that are benign and malignant. The benign class contains 458 instances while the malignant contains 241 instances. Breast cancer contains 10 attributes; which are sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. The excerpt of the dataset is shown in Figure 2.

![Figure 2. Breast cancer wisconsin (BCW) dataset](image)

The breast cancer dataset contains 286 instances with two classes that are recurrence-events and no-recurrence-events. The recurrence-events class contains 85 instances while the no-recurrence-events contains 201 instances. This dataset contains nine attributes, which include age, menopause, tumor-size, inv-nodes, node caps, deg-malig, breast, breast-quad, and irradiat. The excerpt of the dataset is shown in Figure 3.

![Figure 3. Breast cancer (BC) dataset](image)

The Breast Tissue dataset contains 106 instances with six classes that are Carcinoma (CAR), Fibro-adenoma (FAD), Mastopathy (MAS), Glandular (GLA), Connective (CON), and Adipose (ADI). The class Carcinoma (CAR) contains 21 instances, Fibro-adenoma (FAD) contains 15 instances, Mastopathy (MAS) contains 18 instances, Glandular (GLA) contains 16 instances, Connective (CON) contains 14 instances, and Adipose (ADI) contains 22 instances. Breast Tissue contains nine attributes that include
impedivity (ohm) at zero frequency, phase angle at 500 KHz, high-frequency slope of phase angle, impedance distance between spectral ends, area under spectrum, area normalized by DA, maximum of the spectrum, distance between 10 and real part of the maximum frequency point, and length of the spectral curve. The excerpt of the dataset is shown in Figure 4.

![Figure 4. Breast tissue (BT) dataset](image)

2.2. Preprocessing

The types of the breast cancer wisconsin dataset are categorized as asymmetric. It means they are class breast cancer dataset represented using into categories benign or malignant. The attributed value of breast cancer dataset not completely record. That have only one attribute have missing values which are Bare Nuclear that have 2% of missing value are denoted by ‘?’ . To handle the missing value in machine learning is one of the important things to get the best accuracy.

In this paper imputation is a way of handling missing value by replacing them with meaningful replacement values. Since ‘Bare Nuclei’ is a categorical attribute, this experiment has use to handling the missing values of the attribute. Based on the breast cancer dataset given the mode for the attribute ‘Bare Nuclei’ is ‘1-10’ since it has the maximum value which is 10. Figure 5 and Figure 6 show the before and after dealing the missing value.

![Figure 5. Before dealing with missing values](image)

![Figure 6. After dealing with missing values](image)

2.3. Transformation

The breast cancer dataset is divided into training and testing sets based on 10-fold validation method where nine parts is for training the algorithm and the last one part for assessing the algorithm. This breast cancer dataset will be using accuracy estimation because this is effective measure of the performance of a classifier. Figure 7 shows the procedure of accuracy rate estimation adopted from Mitani and Hamamoto [24].

![Figure 7. Accuracy estimation](image)
Comparative analysis on bayesian classification for breast cancer... (Wan Nor Liyana Wan Hassan Ibeni)

2.4. Classification algorithms

In this work, there Bayesian classification algorithms explored, which are the Naïve Bayes (NB), Bayesian Networks (BN), Tree-Augmented Naïve Bayes (TAN).

Naïve Bayes (NB) algorithm is one type of probabilistic classifier based on the Bayes theorem. NB produces a probability that a given instance belongs to that class rather than prediction. One main advantage of NB is that it only requires small amount of training data because it is based on major assumption that an attribute value on a given feature is always independent from the values of other features. This assumption forms the basis of NB and is widely known as class conditional independence [25].

Bayesian Networks (BN) is a type of directed acyclic graph where the nodes represent domain variables and are connected with arcs that represent dependencies between the variables. BN is composed of the network structure and its conditional probabilities. Meanwhile, the Tree Augmented Naïve Bayes (TAN) incorporates some dependencies between the attributes by building a directed tree among the attribute variables. This means the n attributes will form a directed tree that represents the dependency relations between the attributes. This learning algorithm creates a TAN graph structure whereby a single class variable has no parents while other variables have the class as a parent or at most one other attribute as a parent.

Next, the results from Bayesian classification algorithms will be compared with K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) from the literature. K-Nearest Neighbor Algorithm (KNN) is an algorithm to classify the object based on the learning data that is closest to the object. There has two basic k-Nearest Neighbor Classification algorithm to be considered, (1) finding the k training instances that are closest to the unseen instances or (2) taking the most commonly accruing classification for these k instances. Using k is to reduce the effect of the presence of point noise. A number of neighbor making a decision is considered better than a single neighbor making a decision.

Support Vector Machine (SVM) is a classification technique used in the case of classification. SVM scans for the best hyper plane which serves as a separator of two classes in the input space, where the input data serve as a vector in an n-dimensional space. The hyper plane separation sets as margin to be as big as possible between both sets of data. The margin is calculated by constructing two parallel hyper-planes that separate between the two sets of data. In an obvious case, a good separation can be accomplished with a hyper-plane which has the largest distances to neighboring data points in both classes. This means the larger the low margin of error, the more generalized a classifier could be.

2.5. Evaluation metrics

The evaluation metrics used in the experiments are accuracy, precision, recall and F-measure. The equations respectively, where TP is the number of true positive, TN is the number of true negatives, FP is the number of false positive and FN is the number of false negatives.

a. Accuracy

Accuracy is total number of samples correctly classified to the total number of samples classified. The formula for calculating accuracy is shown in (1).
b. **Precision**

Precision is the number of samples is categorized positively classed correctly divided by total samples are classified as positive samples. The formula for calculating precision is shown in (2).

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]  

(2)

c. **Recall**

Recall is the number of samples is classified as positive divided by the total sample in the testing set positive category. The formula for calculating recall is shown in (3).

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]  

(3)

d. **F-Measure**

F-Measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. The formula for calculating F-Measure score is shown in (4).

\[
\text{F-Measure} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}
\]  

(4)

3. **RESULTS AND DISCUSSION**

The performance result for the breast cancer dataset, which by three algorithms is naïve bayes (NB), bayesian networks (BN), and tree augmented naïve bayes (TAN) has been implemented by using Waikato Environment for knowledge analysis (WEKA) software. The performance result can trace the breast cancer benign or malignant is correctly or incorrectly classified. The accuracy of an algorithm can be obtained by comparing the accuracy of result related work and this using new algorithm in this research. Table 1 shows the accuracy and precision percentage for breast cancer dataset in different training and testing environment.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-Measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>94.992</td>
<td>96.943</td>
<td>95.483</td>
<td>96.207</td>
</tr>
<tr>
<td>SVM</td>
<td>96.852</td>
<td>97.161</td>
<td>98.017</td>
<td>98.591</td>
</tr>
<tr>
<td>DT(J48)</td>
<td>94.992</td>
<td>95.633</td>
<td>96.688</td>
<td>97.167</td>
</tr>
<tr>
<td>BAN</td>
<td>97.281</td>
<td>96.806</td>
<td>99.325</td>
<td>97.895</td>
</tr>
<tr>
<td>NB</td>
<td>95.994</td>
<td>95.196</td>
<td>98.642</td>
<td>96.888</td>
</tr>
<tr>
<td>TAN</td>
<td>96.280</td>
<td>95.851</td>
<td>98.430</td>
<td>97.123</td>
</tr>
</tbody>
</table>

The results showed that BN algorithm has the higher classification accuracy based on breast cancer dataset. BN has the highest value of accuracy, which is 97.281% while NB is only achieved 95.994% and TAN achieved 96.280%. The comparison the result of the related paper using other algorithms which is KNN only achieved 94.992%, SVM 96.852% and DT 94.992% of accuracy. This proven the BN algorithm is the best classification using breast cancer dataset.

Precision is the predictive value for a class label of whether positive or negative depending on the class it is calculated for. This is essentially the predictive power of the classification algorithm. Figure 8 shows precision of BN is 96.506%, higher than NB 95.994% and TAN 95.851%.

Next, recall is the fraction of relevant instances that are retrieved. A high recall value indicates that the classification algorithm is able to return most of the relevant results. Figure 9 shows BAN is higher recall, which is 99.325% as compared to the other algorithm NB with 98.642% and TAN with 98.430%.
Finally, F-measure is a measure of accuracy that considers both the precision and the recall rate of the test and computes a composite score. A good score favors the algorithms with higher sensitivity and challenges algorithms with higher specificity. Based on the considerations, we can conclude that BN is preferable to NB and TAN. The F-measure for BAN is higher with 99.895% while NB 96.888% and TAN 97.123%. The results are shown in Figure 10.

To conclude the experiments, Table 2 shows the comparisons of the proposed Bayesian approach across all three datasets, which are breast cancer, breast cancer wisconsin, and breast tissue dataset and the result for breast cancer datasets with three classification algorithms.

Based on the table, BN algorithm has higher results than NB and TAN algorithm with the Breast Cancer and Breast Cancer Wisconsin dataset while than the second higher in the Breast Tissue dataset it because the attribute Breast tissue more to specification to tissue and that have many classes depend Breast Cancer and Breast Cancer Wisconsin have only two classes.
Figure 10. Comparison of F-measure

Table 2. Comparison across different datasets

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Naïve Bayes (NB)</th>
<th>Bayesian Networks (BN)</th>
<th>Tree Augmented Naïve Bayes (TAN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer Wisconsin</td>
<td>72.028%</td>
<td>72.377%</td>
<td>67.332%</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>95.994%</td>
<td>97.281%</td>
<td>96.280%</td>
</tr>
<tr>
<td>Breast Tissue</td>
<td>70.754%</td>
<td>66.037%</td>
<td>62.264%</td>
</tr>
</tbody>
</table>

4. CONCLUSION
Classification is an important technique of the data mining with applications in various fields. This paper presented a comparative experiment on different techniques evaluated on the breast cancer dataset. Six classifiers were compared based on accuracy to select the best result to be used in the classification task. The best result between three algorithm chose in this paper showed Bayesian Networks (BN) classification because has the highest accuracy which is 97.281%. Next, this paper also compared the performance of Bayesian algorithms based on different datasets, which are the Breast Cancer Wisconsin and Breast Tissue dataset to prove that BN algorithm has the best accuracy as compared to NB and TAN algorithms. In the present study few issue like high dimensionality, scalability and accuracy are to be considered for further research along with other algorithms not currently available in WEKA environment.

ACKNOWLEDGEMENTS
This research is supported by Universiti Tun Hussein Onn Malaysia.

REFERENCES
Comparative analysis on bayesian classification for breast cancer... (Wan Nor Liyana Wan Hassan Ibeni)