Brief review of classification algorithm on breast cancer

Ni Wayan Parwati Septiani*, Mei Lestari and Rayung Wulan
Universitas Indraprasta PGRI

*ni_wayanparwati@unindra.ac.id

Abstract. Breast cancer is a malignant neoplasm disease which is an abnormal growth of breast tissue that is different from surrounding tissue. It became the second most case of death that caused by cancer after lung cancer. Several studies have been done to diagnose patients of breast cancer using various type of algorithm. This paper presents a brief review of breast cancer detection using datamining classification algorithm. C4.5, Naive Bayes and K-Nearest Neighbour algorithm are applied on breast cancer data set. Outputs are decision tree, models and computations. This study aims to evaluate performance of those three-data mining classification algorithm.

1. Introduction
Breast cancer starts when cells in the breast begin to grow out of control, and usually form a tumor that can be seen on an X-ray or form as a lump. It is important to understand that most of breast lump is not cancer (malignant), they are benign that has abnormal growth, but they do not spread outside of the breast and not a life threatening tumor [1]. Typically, young woman present with more benign pathologies; but malignancies can also occur in young women. In all women aged 40 and over presenting with abnormalities of the breast, a primary breast cancer should be ruled out because it is the leading cancer among women in developed country [2]. According to GLOBOCAN data, International Agency for Research in year 2012, breast cancer is most common disease that diagnosed on women [3].

Table 1. Estimated Incident, Death and Prevalence of the World in 2012

| Region                        | Cases (thousands) | Death (thousands) | 5-year Prev |
|-------------------------------|-------------------|-------------------|-------------|
| World                         | 1671              | 522               | 6232        |
| More Developed Region         | 788               | 198               | 3201        |
| Less Developed Region         | 883               | 324               | 3032        |
| WHO Africa Region (AFRO)      | 100               | 49                | 318         |
| WHO Americas Region (PAHO)    | 408               | 92                | 1618        |
| WHO East Mediterranean Region (EMRO) | 99     | 42                | 348         |
| WHO Europe Region (EURO)      | 494               | 143               | 1936        |
| South East Asia Region (SEARO) | 240              | 110               | 735         |
| WHO Western Pacific Region (WPRO) | 330         | 86                | 1276        |
| IRC membership (24 countries) | 935               | 257               | 3591        |
| United States of America      | 233               | 44                | 971         |

Source: GLOBOCAN, IARC 2012
Data mining is a process of forming a new descriptive, understandable and predicted pattern from a large scale of data [4]–[6]. Recognizing data from other databases using models that has been build using certain algorithm is one of data mining techniques. Data mining is a multidisciplinary field that combine data statistics, machine learning, artificial intelligence, and database technology. There are top 10 data mining algorithm identified by IEEE International Conference on Data Mining (ICDM) in December 2006, i.e. C4.5, K-means, SVM, apriori, EM, PageRank, AdaBoost, k-Nearest Neighbor, Naïve Bayes, CART [7].

C4.5 Algorithm is the most common used data mining classification algorithm. It is use to form a decision tree using the gain value. These algorithm is easy to implement and to understand. C4.5 Algorithm is development of ID3 algorithm that allow to use continuous data, allow to use unknown missing value, able to use attributes with different weight and allow to prune the tree after being created [8]. There are three types of nodes in decision tree, root node, branch node and leaf node. The following formula are gain and entropy computation.

\[
\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)
\]

\[
\text{Entropy}(S) = \sum_{i=1}^{n} -P_i \times \log_2 P_i
\]

Where,
S is set of cases; A is attribute; n is number of values of attribute A in S; S is information entropy of S
S_i is number of cases in i partition; P_i is frequency of class i in S.

Naïve bayes is a simple probabilistic-based prediction technique based on the application of Bayes theorem with strong assumptions of independence (naive) [9]. Here is Bayes theorem formula:

\[
P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}
\]

Where,
P(C|X) is the posterior probability of target class given predictor attribute,
P(X|C) is likelihood which probability of predictor given class,
P(C), Prior probability of class,
P(X), prior probability of predictor

Predicted outcome of naïve Bayes algorithm is a class with the highest posterior probability. To find the posterior probability, by first is constructing a frequency table of each attribute against the target (classes), then transforming frequency table into likelihood table.

k-Nearest Neighbour, is an algorithm that stores all available cases and classifies new cases by distance function. It represents classifying algorithm based on closest k object in its neighbourhood [10]. A new cases labelled with the most frequent label (majority voting). In general, a large k-value as it eliminate overall noise, but there is no guarantee as it built boundary between classification and make it blurred.

2. Methodology
Main procedure of this study includes preparing data set, applying models (c4.5, naïve Bayes, k-NN algorithm), showing the output, evaluating performances and implementing selected algorithm into a system. As shown in Figure 1.
This paper presents comparison of C4.5, naïve Bayes and k-NN algorithms to detect breast cancer. By first determine data that used. There are 134 testing data and 536 training data that derived from UCI public repository [11]. Classifications of breast cancer are benign (non-cancerous) and malignant (cancerous). Next step is to process training data using C4.5, naïve Bayes and k-NN algorithm. The outcome can be computation, models, decision tree. Last, use testing data to evaluate the output models. By using confusion matrix to describe performance of those three algorithms.

### 3. Result and Discussion

#### 3.1. C4.5 Algorithm

First step of C4.5 algorithm is to compute number of malignant (cancerous) cases and number of benign cases (non-cancerous). Then compute the entropy of each classes using the training data.

\[
\text{Entropy}(\text{Total}) = -\left(\frac{332}{536} \log_2\left(\frac{332}{536}\right)\right) + \left(-\frac{204}{536} \log_2\left(\frac{204}{536}\right)\right)
\]

\[
\text{Entropy}(\text{Total}) = 0.958462584
\]

To find gain of each attribute, by first calculate the entropy of each attribute.

| Attribute               | Gain    |
|-------------------------|---------|
| Clump thickness         | 0.5030918 |
| Uniformity of cell size | 0.7051201 |
| Uniformity of cell shape| 0.6854048 |
| Marginal Adhesion       | 0.4434172 |
| Single Epithelial cell size | 0.5076148 |
| Bare nuclei             | 0.6231893 |
| Bland chromatin         | 0.4245068 |
| Normal nucleoli         | 0.5033751 |
| Mitoses                 | 0.2248785 |

Table 2 shows the uniformity of cell size has the highest value 0.7051201, hence the uniformity of cell size become the root of decision tree. Then recalculate gain and entropy of each attribute to form a complete decision tree and rules for breast cancer detection.

#### 3.2. Naïve Bayes Algorithm

The computation using naïve Bayes algorithm, by first is to find the prior probability for classes. P(malignant) is 204/536=0.380597, and P(benign) is 332/536=0.61403. Then, constructing frequency table of each attribute against the target. Table 3. Shows frequency table of clump thickness. Table 4 shows the likelihood of clump thickness.
Table 3. Frequency table of clump thickness

| Frequency Table | Malignant | Benign |
|----------------|-----------|--------|
| 1              | 3         | 107    |
| 2              | 3         | 37     |
| 3              | 11        | 65     |
| 4              | 7         | 50     |
| 5              | 33        | 56     |
| 6              | 17        | 13     |
| 7              | 20        | 0      |
| 8              | 35        | 4      |
| 9              | 13        | 0      |
| 10             | 62        | 0      |

Table 4. likelihood table of clump thickness

| Likelihood Table | Malignant | Benign |
|------------------|-----------|--------|
| 1                | 0,027273 | 0,972727 |
| 2                | 0,075 | 0,925 |
| 3                | 0,144737 | 0,855263 |
| 4                | 0,122807 | 0,877193 |
| 5                | 0,379787 | 0,629213 |
| 6                | 0,566667 | 0,433333 |
| 7                | 1 | 0 |
| 8                | 0,897436 | 0,102564 |
| 9                | 1 | 0 |
| 10               | 1 | 0 |

Following table is a new data X to be predicted whether it is malignant or benign

Table 5. Data X

| Data X | Atribut | Value | P(X|Ci) |
|--------|---------|-------|-------|
|        | Clump Thickness | 4 | 0,122807 | 0,877193 |
|        | Uniformity of cell size | 4 | 0,818182 | 0,181818 |
|        | Uniformity of cell shape | 2 | 0,136364 | 0,863636 |
|        | Marginal Adhesion | 1 | 0,094276 | 0,905724 |
|        | Single Ephithelial cell Size | 2 | 0,076087 | 0,923913 |
|        | Bare Nuclei | 5 | 0,666667 | 0,333333 |
|        | Bland Chromatin | 2 | 0,053571 | 0,946429 |
|        | Normal Nucleoli | 1 | 0,105919 | 0,894081 |
|        | Mitoses | 2 | 0,814815 | 0,185185 |

From table 5, we can compute the likelihood of malignant is 3.0294E-07, likelihood of benign is 0.006021, P(malignant|x) = 1.15298E-07, and P(benign|x) = 0.0003729. as the computation shows that P(benign|x) greater than P(malignant|x), new data on table 6 labeled as benign (non cancerous).

3.3. K-Nearest Neighbour Algorithm
k-Nearest Neighbour classifying a case by majority vote of its neighbour, with the case being assign to the class that most common amongst its nearest neighbour measured by distance function. In this paper we use 9 for the k value. Table 6 is a sample of breast cancer testing data.

Table 6. breast cancer testing data

| Attribute          | Value |
|--------------------|-------|
| Clump Thickness    | 5     |
| Uniformity of Cell Size | 1     |
| Uniformity of cell Shape | 1     |
| Marginal Adhesion  | 1     |
| Single Ephithelial cell size | 2     |
| Bare Nuclei        | 1     |
| Bland Chromatin    | 3     |
| Normal Nucleoli    | 1     |
|                   |       |

Table 7. Breast cancer training data

| Atribut            | Value |
|--------------------|-------|
| Clump Thickness    | 5     |
| Uniformity of Cell Size | 1     |
| Uniformity of cell Shape | 1     |
| Marginal Adhesion  | 1     |
| Single Ephithelial cell size | 2     |
| Bare Nuclei        | 1     |
| Bland Chromatin    | 3     |

From table 7, we can compute the likelihood of malignant is 3.0294E-07, likelihood of benign is 0.006021, P(malignant|x) = 1.15298E-07, and P(benign|x) = 0.0003729. as the computation shows that P(benign|x) greater than P(malignant|x), new data on table 6 labeled as benign (non cancerous).
Mitoses 1
Class 2
Normal Nucleoli 1
Mitoses 1

Computation of distance function of testing data on table 6 and training data on table 7 is $(5-1)^2+(1-1)^2+(1-1)^2+(2-1)^2+(1-1)^2+(3-3)^2+(1-1)^2+(1-1)^2=16$. Next step is to sort distance value and set 9 minimum distance. As the most frequent label are benign hence, the new data labeled as benign.

3.4. Evaluation

To find performance of those three algorithm, we use confusion matrix. There are two class classification on breast cancer detection i.e. malignant and benign or called binary classification. After we use a dataset with known classes to build a model, then we use another dataset with known classes to evaluate the model. The evaluation process in general is comparing the predicted classes against the actual classes. Confusion matrix shows the number of correct and incorrect predictions. There are four basic terms on binary classification as follows [12]:
1. True Positives, cases in which predicted have disease and they do have disease,
2. True Negative, cases in which predicted do not have disease and they do not have disease,
3. False Positive, cases in which predicted have disease but they do not have disease,
4. False Negative, cases in which predicted do not have disease but they do have disease

Accuracy are the most common rates computed from a confusion matrix for a binary classifier, using the following formula:

$$\frac{(TP + TN)}{total}$$

Accuracy is measured by the area under ROC Curves (AUC). ROC curve is a commonly used graph that summarizes the performance of a classifier over all possible threshold [12].

Table 8. Accuracy of C4.5, Naïve Bayes and k-NN algorithm on breast cancer detection

|       | C4.5 | Naïve Bayes | k-NN |
|-------|------|-------------|------|
| Accuracy | 91.79% | 98.51% | 98.51% |
| AUC    | 0.928 | 1.000 | 0.999 |

Table 9 shows the accuracy of naïve bayes algorithm highest among the other algorithm 98.51%, hence naïve Bayes algorithm are applied to the system.

3.5. Implementation

The proposed system is implemented using java programming language which applied naïve Bayes classifier in it.
Figure 2. Interface of breast cancer detection system

// Pseudocode of proposed system
BEGIN
READ training data
PROCESSED training data
IF training data ready to use THEN
APPLIED naïve Bayes classifier on training data
INPUT New data to be classified
IF Button TEST is selected THEN
Classified new data
ELSE IF Button RESET is selected THEN
Reset Form
ELSE IF Button SAVE is selected THEN
Save classified data
END IF
ELSE
INPUT new data as training data
END

4. Conclusion
This paper presents comparison of three classifier algorithm, i.e. C4.5, Naïve Bayes and k-Nearest Neighbor to detect breast cancer (malignant). Computation shows that Naïve Bayes has the highest accuracy, hence the proposed system is built using naïve Bayes and java programming language. This system can be contributed in biomedical science.

Acknowledgement
The authors would like to thank Kemenristek DIKTI Kopertis Wilayah III for financial support of this research.

References
[1] American Cancer Society, 2016 Breast Cancer; What is breast cancer? Am. Cancer Soc. p. 1–127.
[2] Meisner A L W Houman Fekrazad M and Royce M E, 2008 Breast Disease: Benign and Malignant Med. Clin. North Am. 92, 5 p. 1115–1141.
[3] IARC, 2016 Globocan 2012 http://Www-Dep.Iarc.Fr/ GLOBOCAN, August p. 2012–2013.

[4] Larose D T, 2006 Data Mining Methods and Models.

[5] Hand D Hand D Mannila H Mannila H Smyth P and Smyth P, 2001 Principles of data mining 30.

[6] Larose D T. and Larose C D ., 2014 Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition.

[7] Wu X D et al., 2008 Top 10 algorithms in data mining Knowl. Inf. Syst. 14, 1 p. 1–37.

[8] HSSINA B MERBOUHA A EZZIKOURI H and ERRITALI M, 2014 A comparative study of decision tree ID3 and C4.5 Int. J. Adv. Comput. Sci. Appl. 4, 2.

[9] Gelman A Carlin J B Stern H S and Rubin D B, 2004 Bayesian Data Analysis.

[10] Bramer M, 2007 Principles of Data Mining.

[11] UCI, 2011, UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%2528Diagnostic%2529 .

[12] Markham K, 2014 Simple guide to confusion matrix terminology Data Sch. p. 1–9.