Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

Breast Cancer Detection using Data Mining Classification Methods

Ni Wayan Parwati Septiani[1)], Rayung Wulan[2)], Mei Lestari[3)]
Teknik Informatika Universitas Indraprasta PGRI

Wayan.parwati@gmail.com

Abstract. Breast cancer can be happen to anyone, man and women. Breast cancer is an uncontrollable growth of breast cells. These cells form a tumor. There are two types of tumor, cancerous (malignant) and non cancerous (benign). IARC fact sheet shows that breast cancer is the second causes of cancer death in more developed regions (198,000 deaths, 15.4%) after lung cancer. Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal Adhesion, Single Ephitelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses are attributes that been used to diagnose breast cancer. C4.5, Naive Bayes and k-Nearest Neighbor are data mining classification method that commonly used to detect desease. This paper presents comparison of data mining classification methods to detect breast cancer.

Keywords: data mining, classification methods, breast cancer, accuracy

1. Introduction

Data mining and data mining analysis is a process to form pattern and descriptive models, understandable, and predictive from a big size of data[1]. there are methods in data mining, Estimation Methods, Prediction/Forecasting Methods, Classification Methods, Clustering Methods and Association Methods. C4.5, Naive Bayes and k-Nearest Neighbor are Data mining Classification Methods.

Breast cancer is the most common cancer that happen in women. IARC fact sheet shows that 25% of 1.67 million cancer case in 2012. Therefore it is important to diagnose cancer in early phase. This paper apply C4.5, Naive bayes and k-NN algorithm in detecting breast cancer (malignant).

Tabel 1. Estimated Incidence, mortality and prevalence worldwide in 2012

	cases (thausands)	**Death (Thausands)**	**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

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

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
China	187	48	697
India	145	70	397
European Union (EU-28)	362	92	1444

2. Applied Models for Breast Cancer Detection

2.1. C4.5 Algorithm

C4.5 is the most used classification algorithm to built decision tree. C4.5 algorithm is a supervised learning algorithm, it requires training set data and each data has input object and desired output value (class). Decision trees are simply responding to a problem of discrimination, is one of the few methods that can be presented quickly enough to a non specialist without getting lost to understand mathematical formulations, and the most powerful and preferred method in machine learning is C4.5 algorithm[2].

C4.5 Algorithm has improve ID3 behaviors as follows[2]:

- A possibility to use continuous data
- Using unknown (missing values)
- Ability to use attributes with different weights
- Pruning the tree after created (Pessimistic prediction error and subtree raising)

The following steps are C4.5 algorithm

a. Calculate gain of each attribute. Determine attribute with the highest gain value as a root. Formulation of gain computation is

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

Where,

$$Entropy(S) = \sum_{i=1}^{n} -pi \times \log_2 pi$$

Where,
S : set of cases
A: attribute
N: number of partition
$|S_i|$: number of cases in i partition
$|S|$: number of cases in set of S cases
P_i: Proportion S_i to S

b. Create a decision node that splits on root node and become children of nodes
c. Repeat the steps for children nodes.

2.2. Naive Bayes Algorithm

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

Naive Bayes is one of data mining classification methods, it is a simple probabilistic-based prediction techniques that are based on the application of Bayes theorem with the assumption of strong independency[3].

Data analysis on bayesian algorithm are as follows [4]:

1. Set the full probability model, the joint distribution of number of data that can be observed nor hidden. Made a model that is consistent with the science of the underlying scientific issues and data collection process.
2. Calculate and interpret posterior distribution
3. Evaluate the suitability of the model, and the implications of the posterior distribution results.

Bayesian Theorem formulation is as follows:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Where:

P(H|E) : The probability of hypothesis H given the evidence E

P(E|H) : The probability the evidence E given the hypothesis H it is also referred as the likelihood

P(E) : The probability of the evidence irrespective of knowledge about H. since H can be either true or false

$P(E) = P(E|H) * P(H) + P(E|not\ H) * P(not\ H)$

2.3. K-Nearest Neighbor Algorithm

The aim of k-Nearest Neighbor (k-NN) Algorithm is to classify the new object based on attributes and training data. This classification is memory based model, it classify the new object with the most frequent label on training data (majority voting). The following are steps of k-NN algorithm:

1. Determine k number of nearest neighbor
2. Calculate similarity distance between new object and training data
3. Sort the distance and determine k-minimum distance
4. Label the new object with the most frequent label on training data

3. Data Analysis and Implementation

This paper presents comparison of data mining classification methods, C4.5 algorithm, Naive Bayes and k-NN. The main process of this experiment are as follows:

1. Data prepocessing, data that been used in this paper are breast cancer dataset from public repository UCI data repository. There are 134 testing data and 536 training data classified as benign and malignant. There are 9 attributes, *Clump Thickness, Uniformity of cell size, Uniformity of cell shape, Marginal Adhesion, Single Ephithelial cell size, Bare nuclei, Bland Chromatin, Normal Nucleoli a nd Mitoses.*
2. Methods, there are three algorithm to be compared C4.5, naive bayes and k-NN. This methods are compared based on accuracy.
3. Output, results (outputs) can be a models, formulations or functions and decision tree.
4. Evaluation, evaluation for classification is confusion matrix that often used to describe performance of classification models on a set of testing data.

Conceptual framework of research are shown as follow:

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

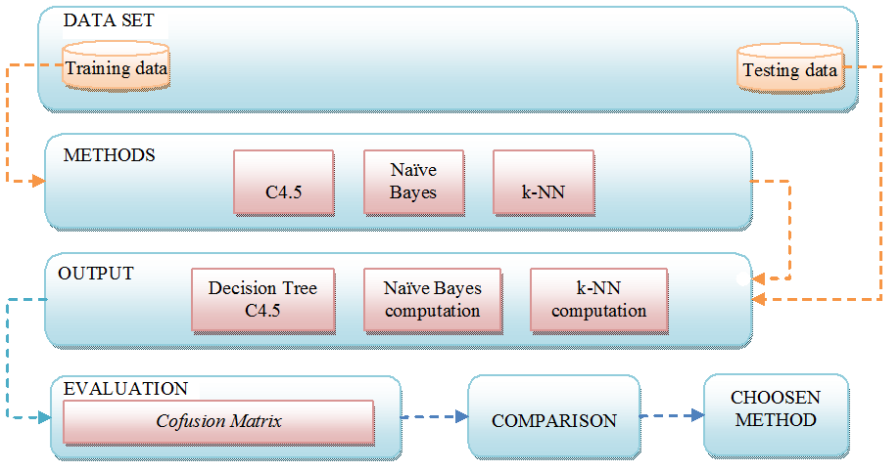

Figure 1. Conceptual Framework

3.1. C4.5 Application

To built a decision tree, calculate entropy of each class (malignant and benign) using training data set.

$$Entropy(Total) = (-\frac{332}{536}\log_2(\frac{332}{536})) + (-\frac{204}{536}\log_2(\frac{204}{536}))$$

$$Entropy(Total) = 0{,}95846584$$

To have gain of each attribute, first step is calculate attribute's entropy of each cases. For the attribute clump thickness there are 10 cases 1, 2, 3, …10.

Entropy of clump thickness case of value 1

$$Entropy = (-\frac{number\ of\ malignant}{total\ cases}\log_2(\frac{number\ of\ malignant}{total\ cases})) + (-\frac{number\ of\ benign}{total\ cases}\log_2(\frac{number\ of\ benign}{total\ cases}))$$

$$Entropy = (-\frac{3}{110}\log_2(\frac{3}{110})) + (-\frac{107}{110}\log_2(\frac{107}{110})) = 0{,}18502$$

Table 1. Gain of clump thickness

Attribute		Number of cases(S)	*Malignant* (Si)	*Benign* (Si)	*Entropy*	*Gain*
Total		536	204	332	0,95846	
Clump Thickness						
	1	110	3	107	0,18052	
	2	40	3	37	0,38431	
	3	76	11	65	0,59651	
	4	57	7	50	0,53738	
	5	89	33	56	0,95127	
	6	30	17	13	0,98714	
	7	20	20	0	0	
	8	39	35	4	0,47707	
	9	13	13	0	0	
	10	62	62	0	0	
						0,50309177

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

Table 2. Gain 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 Ephithelial cell size	0,5076148
Bare Nuclei	0,6231893
Bland Chromatin	0,4245069
Normal Nucleoli	0,5033751
Mitoses	0,2248786

Table 2 shows that attribute uniformity of cell size has the highest value 0,705120126, therefore uniformity of cell size become root node for the decission tree. Next step recalculate entropy and gain to find children node. Repeat steps to have children nodes.

3.2. Naive Bayes Application

For naive bayes computation, first step is to find the probability for each class P(H). Hypothesis that used in this paper are patient detected breast cancer (malignant) and patient not detected breast cancer (benign).

P(*malignant*) = 204 : 536 = 0,380597

P(*benign*) = 332 : 536 = 0,61403

Prior probability computation using naïve bayes on breast cancer data set as the following table. To determine new object for the new cases Posterior probability should be computed, as shown on table 3.

Table 3. Posterior Probability

Data X		P(X|Ci)	
Attribute	Value	Malignant	Benign
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

Total probability computation for each class

P(X|detected = *malignant*)

= 0.122807x 0,818182x 0,136364x 0,094276x 0,076087x 0,666667x 0,053571x 0,105919x 0,814815

= 3,0294E-07

P(X|detected = *benign*)

= 0,877193x 0,181818x 0,863636x 0,905724x 0,923913x 0,333333x 0,946429x 0,894081x 0,185185x 0,006021
= 0,006021
P(X|detected = *malignant*)P(*malignant*) = 3,0294E-07 x 0,380597
= 1,15298E-07
P(Xdetected = *benign*)P(*benign*) = 0,006021 x 0,619403
= 0.0003729

From the above computation resulted P(*benign*|X) higher than P(*malignant*|X), therefore this case labelled as benign (non cancerous).

3.3. k-NN Application

k-NN methods classified the new object by calculate distance of testing data and training data. In this paper number of neighbourhood that been used is 9. There are sample of training data on table 4 and sample of testing data (new object to be classified) on table 5.

Table 4. Sample of training data on breast cancer data set

Attribut	**Value**
Clump Thickness	1
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
Mitoses	1
Class	2

Table 5. Sample of testing data on breast cancer data set

Attribut	**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
Mitoses	1

Distance (similarity value) between testing data and training data is $(5-1)^2+(1-1)^2+(1-1)^2+(1-1)^2+(2-1)^2+(1-1)^2+(3-3)^2+(1-1)^2+(1-1)^2=16$

Table 6. Similarity (distance) value between testing data and training data

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhession	Single Ephithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	class	Distance
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0
1	1	1	1	2	1	3	1	1	Benign	0

Table 6 shows similarity (distance) value for 9 nearest neighbour. It shows that most frequent class is Benign, therefore the new object (table 5) are classified as benign (non cancerous).

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

4. Evaluation

Confusion matrix is method that commonly used to compute accuracy in data mining concept. A confusion matrix describe the performance of a classification model on a set of data for which the true values, labels or classes are known[5]. Area Under (a ROC) Curve (AUC) is a summary to measure accuracy of a quantitative diagnostic test. It is used in classification analysis in order to determine which of the used models predicts the classes best. The accuracy for three algorithm that been used to detect breast cancer are shown on table bellow.

Table 6. Accuracy and AUC value of breast cancer detection

	C4.5	Naïve Bayes	k-NN
Akurasi	91.79%	98.51%	98.51%
AUC	0,928	1,000	0,999

Naive bayes has the highest value from two other methods 98,51% for accuracy and 1,000 for AUC.

5. Conclusion

This paper presents the application of three classification methods on beast cancer data set. By using confusion matrix, shows that naive bayes and k-NN has the similar value for accuracy, that is 98,51%. For the AUC value, naive bayes has the highest value 1,000.

References

[1] Zaki. J. Mohammed & Wagner Meira JR. 2014. Data Mining And Analysis Fundamental Concepts and Algorithms. Cambridge University Press. USA.A reference

[2] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Eritali. A Comparative Study of Decission Tree ID3 and C4.5

[3] Prasetyo, E. 2012. Data Mining – Konsep dan Aplikasi Menggunakan Matlab. Andi Offset. Yogyakarta.

[4] Gelman, A., Carlin, J.B., Stern, H. S., Rubin, D. B. 2014. Bayesian Data Analysis. Taylor & Francis.

[5] Bramer, M., 2007. Principles of Data Mining. Springers.