Bit-Tech

Vol.2, No.1, Oktober 2019
Available online at: http://jurnal.kdi.or.id/index.php/bt

Optimization of Application of Genetic Algorithm Using C4.5 Method to Predict Breast Cancer Disease

Hartana Wijaya

Buddhi Dharma University
Jalan Imam Bonjol No. 41, Tangerang, Indonesia
hartana.wijaya@ubd.ac.id

Article history:

Received 16 September 2019;
Revised 3 October 2019;
Accepted 10 October 2019;
Available online 30 October 2019

Keywords:

Breast cancer
Optimization
Accuracy
C4.5
Genetic algorithms
PSO

Abstract

Cancer is a big challenge for humanity. Cancer can affect various parts of the body. This deadly disease can be found in humans of all ages. However, the risk of cancer increases with age. Breast cancer is the most common cancer among women, and is the biggest cause of death for women. Then there are problems in the detection of breast cancer, causing patients to experience unnecessary treatment and huge costs. In a similar study, there were several methods used but there were problems due to the shape of nonlinear cancer cells. The C4.5 method can solve this problem, but C4.5 is weak in terms of determining parameter values, so it needs to be optimized. Genetic Algorithm is one of the good optimization methods, therefore the parameter values of C4.5 will be optimized using Genetic Algorithms to get the best parameter values. The results of this study are that C4.5 Algorithm based on genetic algorithm optimization has a higher accuracy value (96%) than only using the C4.5 algorithm (94.99%) and which is optimized with the PSO algorithm (95.71%). This is evident from the increase in the value of accuracy of 1.01% for the C4.5 algorithm model that has been optimized with genetic algorithms. So it can be concluded that the application of genetic algorithm optimization techniques can increase the value of accuracy in the C4.5 algorithm.

I. Introduction

Cancer is a big challenge for humanity. Cancer can affect various parts of the body. This deadly disease can be found in humans of all ages. However, the risk of cancer increases with age. Breast cancer is the most common cancer among women [3]. Information about tumors from certain examinations and diagnostic tests collected determines how extensive the cancer is. Stage of cancer is one of the most important factors in choosing treatment options, and uses Tumor, Nodes and Metastasis (TNM) systems.

Breast cancer is the second leading cause of death for women in the United States, and is the leading cause of death for women aged 40-59 years [2]. Although most are experienced by women, breast cancer can also occur in men. In the United States, from 40,600 breast cancer deaths in 2001, 400 were male [8]. By knowing the malignancy of cancer, treatment measures can be done better and the death rate can be reduced.

Over the years the use of various prediction/classification models in the medical domain has been strengthened largely due to better effectiveness and predictive abilities [1]. Because predicting the results of an accurate disease will help the doctor in making the right decisions, thus avoiding patients from unnecessary treatment and high costs.

Because breast cancer is one of the most common causes of female deaths throughout the world. So breast cancer needs to be accurately predicted, whether it is benign or malignant, so that appropriate medical action can be taken [11]. Several studies have also been conducted.

The C4.5 algorithm is the simplest, most easily implemented classification. But there are still difficulties in handling high dimensional data. In several studies the C4.5 algorithm method is considered to outperform Naïve Bayes. However, to determine attributes in predictive accuracy is still considered lacking. Searching for the optimal subset will be very expensive especially when attributes increase with the amount of data available. Sometimes it's not feasible to use. Therefore, a more accurate method is needed to optimize these predictions.

Genetic Algorithm is one of the reliable optimization methods so that it can be used to determine the optimal control parameters for a particular process. By applying the C4.5 method and hybrid genetics algorithms are expected

ISSN 2622-2728 (online) 2622-271X (print)
doi:

to accelerate the process and get the appropriate control parameter values, because it is proposed where C4.5 is a classifier and combined with Genetic Algorithms to solve problems such as breast cancer diagnosis.

II. Related Works/Literature Review

Data Mining

Data mining is an activity to extract to get important information that is implicit and previously unknown, from a data. Data mining is defined as the process of finding patterns in data. This process is automatic or (usually) semi-automatic [12]. Data mining is a process of discovering new patterns and trends by sorting through a lot of data stored in the repository, using pattern penalty technology and statistical and mathematical techniques [7].

Data Mining Classification Algorithm

Classification is the process of finding a model (function) that describes and distinguishes a data class or concept that aims to be used to predict the class of objects whose class labels are unknown [4]. Data classification consists of 2 steps process. The first is learning (training phase), where the classification algorithm is made to analyze training data and then represented in the form of a classification rule. The second process is classification, where test data is used to estimate accuracy from the classification rule [4].

Optimization Algorithm

Optimization is about finding optimal parameter values for an object or system that minimizes goals (costs) and functions [9]. In optimization, we are given a function, known as the objective function. The aim is to minimize or maximize the value of the objective function by adjusting various parameters. Each parameter combination marks a solution that may be good or bad, depending on the value of the objective function. Soft-computing techniques produce the best set of parameters that give the best values of the objective function given the constraints of time.

C4.5 Algorithm

C4.5 is part of an algorithm for classification in machine learning and data mining. C4.5 is an algorithm that is suitable for classification problems in machine learning and data mining [10]. C4.5 maps the attributes of the class so that they can be used to find predictions about data that has not yet appeared.

Genetic Algorithms

Genetic algorithms are global optimization algorithms derived from evolutionary ideas and inspiration [6]. In essence, it is a direct search method that does not depend on concrete problems and GA has gained wide application in image processing, biological science, neural networks, pattern recognition, machine learning and the like.

Particle Swarm Optimization (PSO) Algorithm

Particle Swarm Optimization is a population-based optimization technique developed by Eberhart and Kennedy in 1995. The PSO simulates the behavior of a group of birds looking for food. This behavior is described as follows:

A group of birds is looking for a piece of food in an area. All birds in the group did not know where the food was located, in the process of searching for food some birds suddenly separated from the herd and formed new flocks and returned to groups. The process of grouping birds aims to maintain the optimum distance between food with the bird and other birds. The most effective way to find these pieces of food is to follow the bird closest to the food.

The same behavior is also described in other groups of animals such as fish. In groups, flocks of fish experience competition in dividing food, but the herd also becomes easier to find new food in areas where food distribution is unknown [5]. The method is adopted in the PSO algorithm itself, which is to repeat it to optimize a problem by providing quality values. That way in large data can be known the best position in the calculation of data processing.

III. Methods

Types of Research

In this study using secondary datasets obtained from the UCI Machine Learning Repository made by Dr. WIlliam H. Wolberg from the University of Wisconsin Hospital for a total of 699 points. With the number of attributes there are 10 attributes consisting of:

1.	Sample code number	: id number
2.	Clump Thickness	: 1 - 10
3.	Uniformity of Cell Size	: 1 - 10
4.	Uniformity of Cell Shape	: 1 - 10

5.	Marginal Adhesion	: 1 - 10
6.	Single Epithelial Cell Size	: 1 - 10
7.	Bare Nuclei	: 1 - 10
8.	Bland Chromatin	: 1 - 10
9.	Normal Nucleoli	: 1 - 10
10.	Mitoses	: 1 - 10
11.	Class	: (2 for benign, 4 for malignant)

Thickness Clumps of benign cells tend to be grouped in monolayers, while cancer cells are often grouped in multilayers. While the diversity of cell size/shape of cancer cells tends to vary in size and shape. That is why these parameters play a role in determining whether or not cancer cells.

In marginal adhesion normal cells tend to stay together, where cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy. In single cell size epithelium size is related to the uniformity mentioned above. Significant epithelial cells may be enlarged by malignant cells.

Inti Bare is a term used for cores that are not surrounded by cytoplasm (the rest of the cell). They are usually seen in benign tumors. Whereas Chromatin Bland describes the "texture" of the uniformity of the nucleus seen in benign cells. In chromatin cancer cells tend to be rough. The normal nucleolus is a small structure seen in the nucleus.

In normal cells the nucleolus is usually very small if visible. In nucleoli cancer cells become more prominent, and sometimes there are more of them. Finally, mitosis is the core part plus cytokines and produces two daughter cells that are identical during prophase. This is the process by which cells divide and repeat. Pathologists can determine the level of cancer by calculating the number of mitoses.

Research Steps

The research method in this study is experimental research with research stages as follows:

1. Data Collection
   This research begins with data collection. It is a dataset that is used by similar researchers.
2. Initial Data Processing
   The dataset to be used is processed first.
3. The Proposed Model/Method
   The model/method proposed by the researcher is the C4.5 method which is optimized using genetic algorithms.
4. Experiments and Testing Models
   The dataset that will be used after processing is tested against the proposed model.
5. Evaluation and Validation of Results
   After testing the dataset, the results will appear in the form of accuracy values. Which is then analyzed and evaluated, only after being evaluated can conclusions be drawn from the results of this study.

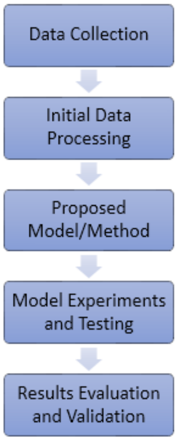

Data Collection
Initial Data Processing
Proposed Model/Method
Model Experiments and Testing
Results Evaluation and Validation

Figure 1. Research Steps

Proposed Model/Method

The method proposed in this study is processing datasets to optimize parameter values and measure the level of accuracy compared to models that are not optimized. Thus, which parameter is best known. This classification data will be tested for accuracy using 10 folds of x-validations and ROC Curve. As seen in figure 2 as below.

a. Selection of datasets for training and datasets for testing are conducted.
b. By using 10 fold cross validation, the dataset will be separated into training data (10%) and testing data (90%) and repeated 10 times.
c. The next step is to determine the parameters that can be determined by yourself to get maximum results. Then optimization is done using genetic algorithms.
d. The model formed will be directly tested with the testing dataset formed, and the accuracy of the model will be averaged.
e. Comparing the results of accuracy and performance results between the C4.5 model using genetic algorithms with the C4.5 model without genetic algorithms. The results will be compared according to the accuracy values obtained from the confusion matrix and the AUC values obtained from ROC Curve.

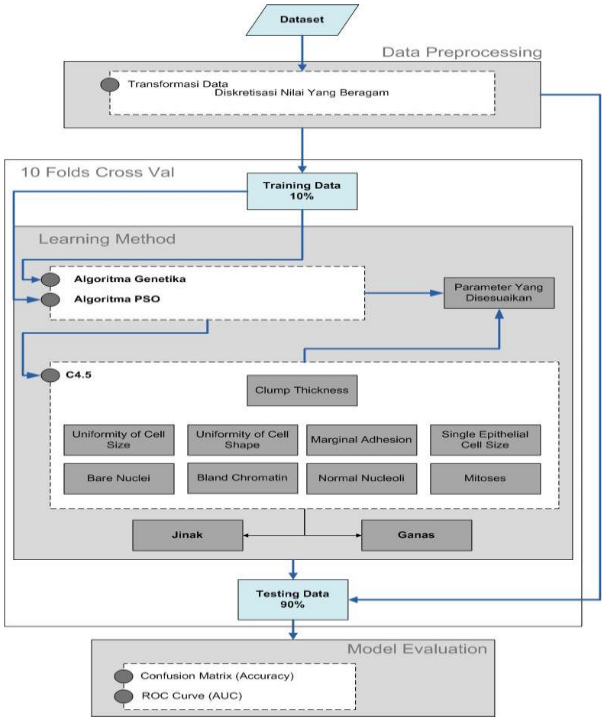


Figure 2. Proposed Model/Method

Model Experiments and Testing

Model testing uses 10 fold cross validation which will randomly take 10% of the training data as testing data, this process is repeated 10 times and the results of model testing are in the form of accuracy, precision, and recall averaged. This testing process is carried out with rapid miner in the building block used for prediction.

In conducting this research, experiments and the process of testing the proposed model are needed. The experimental process and testing the model using a part of the existing dataset. All datasets are then tested by the method proposed in the Rapid Miner 5 application. Following are the models implemented in the Rapid Miner 5 application, i.e.:

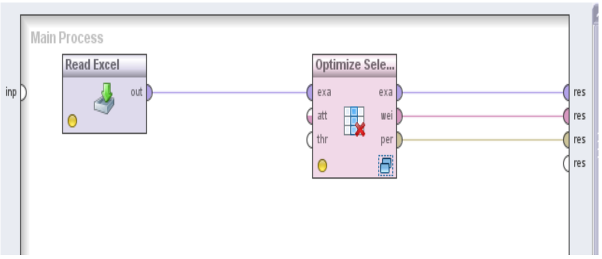


Figure 3. Model Proposed in Rapid Mider 5

In Figure 3 is the relationship between the proposed model, namely read excel from the dataset. Then connecting between the two processing processes in terms of validation, this is done to estimate the performance of the operator. After that, the results of the training validation are connected to the optimize selection example as the set in section. The results of this section, the weight attribute is used as result2, and the performance of optimize selection is connected to result 3.

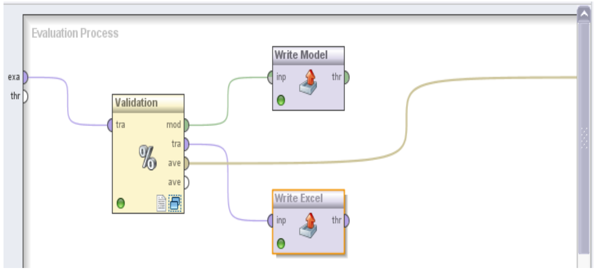


Figure 4. Cross Validation on Rapid Miner 5

In figure 4 is part of the evaluation process which is to produce a model in the form of a decision tree and example set in an excel spreadsheet.

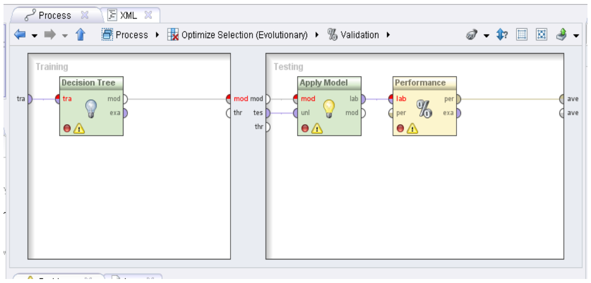
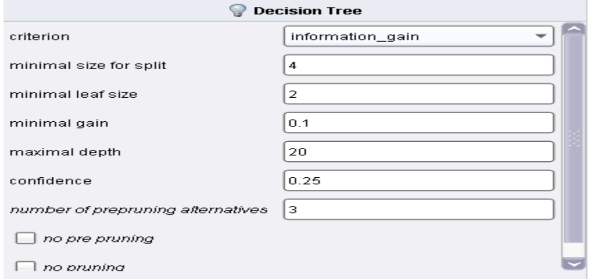


Figure 5. Methods for Cross Validation on Rapid Miner 5

The dataset is also trained in Figure 5 which uses a decision tree to produce the performance of the algortma used. Where information gain is filled with parameters such as figure 6 below:



Decision Tree

criterion	information_gain
minimal size for split	4
minimal leaf size	2
minimal gain	0.1
maximal depth	20
confidence	0.25
number of prepruning alternatives	3

□ *no pre pruning*

□ *no pruning*

Figure 6. Parameters for The Decision Tree Process

Results Evaluation and Validation

Evaluation of the model formed will be done by measuring accuracy and area under curve. Accuracy is measured using confusion matrix, and the AUC value will be measured using ROC Curve. Confusion matrix will describe the results of accuracy starting from the correct positive predictions, false positive predictions, correct negative predictions, and false negative predictions. So that the model formed can be directly tested with data randomly separated by 10 fold cross validation.

With confusion matrix, the accuracy value of the model will be compared between the models formed with the C4.5 algorithm and the C4.5 algorithm that has been optimized. To see the quality of the model produced, the ROC curve will be created and the AUC value can be used as a measure to see the model formed.

IV. Results

Research Result

This study aims to find the best level of accuracy by comparing the level of accuracy produced by several optimization methods, namely Genetic Algorithm and PSO on C4.5 method in determining breast cancer prediction on the UCI Machine Learning Repository dataset.

The results of this study are the results of the processing of qualitative and quantitative data collected in accordance with the proposed model. Research will be conducted on all datasets.

Experimentation of Model Indicators

To get a good model in research, there are several indicators that will be adjusted to reach the maximum model. For making decision trees, the minimum gain indicator and pruning are adjusted to obtain a model with high accuracy. Measurement of accuracy and the value of the model formed will be based on using confusion matrix and ROC Curve. Table indicators and test results for decision trees can be seen in the table 1.

Table 1. Decision Tree Indicators

	No Pruning		**Pruning**	
Minimal Gain	ACC	AUC	ACC	AUC
0	94.99%	0.997	95.56%	0.991
0.1	**94.99%**	**0.997**	95.56%	0.991
0.2	95.42%	0.996	95.71%	0.994
0.3	95.28%	0.994	95.28%	0.991
0.4	94.57%	0.993	94.99%	0.985
0.5	94.99%	0.985	94.13%	0.982
0.6	94.85%	0.984	93.28%	0.979
0.7	85.84%	0.995	65.52%	1.000

Because testing is based on accuracy and AUC, the minimum value of the selected gain is 0.1 with no pruning (pruning and pre pruning). After obtaining an indicator that is suitable for making a decision tree, the model will be optimized to use Genetic Algorithm and PSO.

Model C4.5 Optimization with Genetic Algorithms

With Genetic Algorithms the data to be processed will be given weights to help attribute selection, and the C4.5 algorithm will be applied and calculated the level of accuracy. Genetic Algorithm in the Roullete Wheel technique. After all particles are counted, particles with the best accuracy value will be searched. The next loop, the other particles will randomly move towards the best particles. This process keeps repeating until the allowable repetition limit.

Genetic algorithm indicators are also adjusted to provide a large increase. From all studies, the increase in accuracy with attribute selection based on genetic algorithms in the study was relatively small. But if measured using ROC curve. The improvement obtained is quite good. Detail indicators can be seen in the table 2.

Tabel 2. Indicator for Genetic Algorithms

Population Size	**Max. Generations**	**ACC**	**AUC**
1	5	94.42%	0.914
2	5	94.00%	0.992
3	5	95.42%	0.998
4	5	95.56%	0.996
5	5	95.99%	0.997
1	4	93.70%	0.939
2	4	93.99%	0.990
3	4	95.14%	0.998

4	4	95.71%	0.997
5	4	95.71%	0.998
1	3	93.57%	0.924
2	3	94.84%	0.998
3	3	95.00%	0.997
4	**3**	**96.00%**	**0.997**
5	3	95.28%	0.998
1	2	93.42%	0.987
2	2	93.71%	0.997
3	2	95.00%	0.997
4	2	92.28%	0.996
5	2	95.28%	0.998
1	1	93.42%	0.987
2	1	93.13%	0.991
3	1	93.70%	0.954
4	1	95.28%	0.996
5	1	94.99%	0.998

The following is the weight of the best attributes and decision tree images of the models that have been optimized with Genetic Algorithms:

Tabel 3. Weight of the Best Attributes of Genetic Algorithms

Atribut	**Bobot**
Clump Thickness	1
Uniformity of Cell Size	1
Uniformity of Cell Shape	0
Marginal Adhesion	1
Single Epithelial Cell Size	1
Bare Nuclei	1
Bland Chromatin	0
Normal Nucleoli	0
Mitoses	0

Model C4.5 Optimization with PSO Algorithm

With PSO the data to be processed will be given weights to help improve the results of the calculation, giving this weight is given randomly by determining the minimum and maximum weights. After that each particle will have its own weight in the dataset, and the C4.5 algorithm will be applied and the level of accuracy calculated. After all particles are counted, particles with the best accuracy value will be searched. The next loop, the other particles will randomly move towards the best particles in order to find a better weight. This process keeps repeating until the allowable repetition limit.

After obtaining a suitable indicator to make a decision tree, the model will be tried to be optimized by using PSO. The PSO indicator is also adjusted to provide a large increase. Of all the studies, an increase in accuracy with PSO in the study was relatively small. But if measured using the ROC curve, the increase obtained is quite good. Detail indicators can be seen in the table 4.

Tabel 1. Indicator for PSO Algorithm

Inertia	**Global best**	**ACC**	**AUC**

0	0.1	95.56%	0.999
0.1	0.1	95.56%	0.999
0.2	0.2	95.56%	0.999
0.3	0.3	95.56%	0.999
0.4	0.4	**95.71%**	**0.998**
0.5	0.5	95.56%	0.999

After the weighting process with PSO, the attribute weights will be displayed, and the model that has been optimized will be compared with the model without optimization.

Tabel 2. Weight of the Best Attributes of the PSO Algorithm

Atribut	Bobot
Clump Thickness	0.995627459
Uniformity of Cell Size	0.548731844
Uniformity of Cell Shape	0
Marginal Adhesion	0.410208691
Single Epithelial Cell Size	0.783336492
Bare Nuclei	0.727799272
Bland Chromatin	1
Normal Nucleoli	0.283464633
Mitoses	0

Comparative Results

C4.5 algorithm can make a model that is accurate enough for the process of classification of breast cancer. With the existing dataset stopping as test data, the comparison between data on sick and healthy patients can remain comparable. The C4.5 algorithm model can produce a fairly accurate model that is equal to 94.99%. In order for these predictions to work more optimally, the model will be optimized using attribute selection with Genetic Algorithms and PSO.

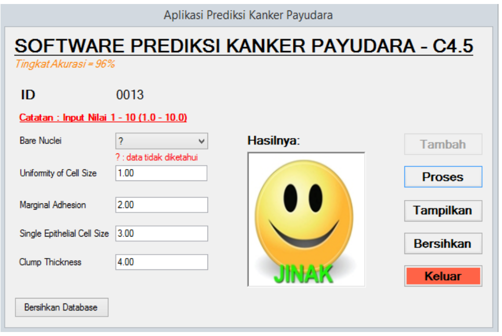
Genetic Algorithms and PSO will determine the attributes of data to be processed, attribute selection becomes effective if the data attribute is nominal, and the data held by the dataset consists mainly of numeric data. With the selection of these attributes, the accuracy of the model formed can be increased to 96.00%.

Table 6. Comparison of C4.5 Models before and after Optimization

	C4.5 algorithm	C4.5 + GA algorithm	C4.5 + PSO algorithm
Success tame predictions	443	445	445
Successful prediction of Ferocious	221	226	224
Model Accuracy	94.99%	**96.00%**	95.71%
AUC	0.997	0.997	0.998

Application of Selected Algorithms

Based on the results of evaluation and validation, it is known that the C4.5 algorithm optimized with Genetic Algorithms has the best accuracy and performance, so that the resulting rule is used as a rule for making interfaces that can facilitate predictions of breast cancer. The interface used in this study was made using Microsoft Visual Basic .Net 2010. Display for Graphical user interface (GUI) Application for Breast Cancer Prediction can be seen in Figure 7.



Aplikasi Prediksi Kanker Payudara

SOFTWARE PREDIKSI KANKER PAYUDARA - C4.5

Tingkat Akurasi = 96%

ID 0013

Catatan : Input Nilai 1 - 10 (1.0 - 10.0)

Bare Nuclei ?

? : data tidak diketahui

Uniformity of Cell Size 1.00

Marginal Adhesion 2.00

Single Epithelial Cell Size 3.00

Clump Thickness 4.00

Hasilnya: JINAK

Figure 7. Graphical User Interface (GUI) Application for Breast Cancer Prediction Software

V. Discussion

From the results of the comparison above, it can be concluded that the results of this study show that the optimization of the C4.5 method with Genetic Algorithms can increase the value of accuracy compared to the use of the C4.5 method without being optimized and optimized with the PSO Algorithm. The increase obtained is 1.01%.

VI. Conclusions

From the research conducted, the determination of parameter values that have been optimized using Genetic Algorithms has been shown to increase the accuracy of predictions in breast cancer. The model formed by the C4.5 method based on Genetic Algorithms produces better accuracy than the C4.5 method without being optimized and C4.5 optimized with the PSO Algorithm. The results of this optimization are very important in determining which parameters are the best, resulting in high accuracy.

The increase can be seen from the increase of the accuracy value for the C4.5 algorithm model of 94.99%, after being optimized the C4.5 algorithm's accuracy based on Genetic Algorithms is 96% with an accuracy difference of 1.01%. For evaluation using ROC curve to produce an AUC (Area Under Curve) value for the C4.5 algorithm model produces the same value, namely 0.997. So that it can be concluded that the application of optimization techniques Genetic Algorithms can increase the value of accuracy in the C4.5 algorithm.

As for the model that is formed, it can later be developed or implemented into an application, so that it can help and facilitate health practitioners in diagnosing breast cancer, and the diagnosis results can be more accurate and reliable.

References

[1] Ali, Amna, and Ali Tufail. 2009. "A Survey of Prediction Models for Breast Cancer Survivability." Proceeding ICIS '09 Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human 1259-1262 .

[2] Calle, J. 2004. Breast cancer facts andfigures 2003—2004. American Cancer Society. p. 1—27.

[3] Dellen, Dursun, Glen WaLker, and Amit Kadam. 2004. "Predicting breast cancer survivability:a comparison of three data mining methods."

[4] Han, Jiawei, and Micheline Kamber. 2006. Data Mining: Concepts and Techniques. Second Edition. San Francisco: Elsevier Inc.

[5] Kennedy, J., & Eberhart, R. (1995, November-December). Particle Swarm Optimization. Proceedings of IEEE International Conference on Neural Networks , 1942-1948.

[6] K. Hornik, M. Stinchcombe, H. White. 1989. "Multilayer Feedforward Networks are Universal Approximators." *Journal Neural Networks Volume 2 Issue 5, 1989* 359-366.

[7] Larose, Daniel T. 2005. *Discovering Knowledge In Data : An Introduction to Data Mining.* Canada: John Wiley & Sons, Inc., Hoboken, New Jersey.

[8] Jerez-Aragone´s JM, Gomez-Ruiz JA, Ramos-Jimenez G. 2003. "A combined neural network and decision trees model for prognosis of breast cancer relapse."

[9] Shukla. 2010. *Real Life Application of Soft Computing.* Taylor and Francis Group, LLC.

[10] Wu, Xindong, and Vipin Kumar. 2009. *The Top Ten Algorithms in Data Mining.* Taylor & Francis Group, LLC.

[11] Yaoying, Huang, Li Wanggen, and Ye Jiao Xiao. 2011. "A Study of Genetic Neural Network as Classifiers and its application In Breast Cancer Diagnosis." *Journal Of Computer.*

[12] Witten, Ian H., Frank Eibe, and Mark A. Hall. 2011. *Data Mining : Practical Machine Learning Tools and Techniques. 3rd Edition.* United States: Elsevier

[13] Dataset:http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data