DIFFERENT MACHINE LEARNING ALGORITHMS FOR BREAST CANCER DIAGNOSIS.

Adel Aloraini

Computer Science department, Qassim University- Saudi Arabia.
(a.oraini@qu.edu.sa)

ABSTRACT

Breast cancer affects many people at the present time. The factors that cause this disease are many and cannot be easily determined. Additionally, the diagnosis process which determines whether the cancer is benign or malignant also requires a great deal of effort from a doctors and physicians . When several tests are involved in the diagnosis of breast cancer, such as clump thickness, uniformity of cell size, uniformity of cell shape,…etc, the ultimate result may be difficult to obtain, even for medical experts. This has given a rise in the last few years to the use of machine learning and Artificial Intelligence in general as diagnostic tools. We aimed from this study to compare different classification learning algorithms significantly to predict a benign from malignant cancer in Wisconsin breast cancer dataset. We used the Wisconsin breast cancer dataset to compare five different learning algorithms , Bayesian Network, Naïve Bayes, Decision trees J4.8 , ADTree, and Multi-layer Neural Network along with t-test for the best algorithm in terms of prediction accuracy. The experiment, has shown that Bayesian Network is significantly better than the other algorithms.

KEYWORDS:

reast cancer , Bayesian Network , K2 algorithm.

1. INTRODUCTION

The diseases which cost so many lives, diagnostic computer-based applications are used widely. Robotics are playing a very important role in operating rooms. Also, the expert systems are presented in the intensive treatment rooms. In turn, using another aspect of Artificial Intelligence for breast cancer diagnosis is not unworthy. It is reported that breast cancer disease is the second most common cancer that affects women, and was the prevalent cancer in the world by the year of 2002[21]. Macmillan Cancer Support in London reports that, in the UK, breast cancer affects a significant number of Arab women and a small number of Arab men. This cancer is a very common type of cancer among women and the second highest cause of cancer death. In the United States, about one in eight women over their lifetime has a risk of developing breast cancer [1]. Breast cancer begins with the uncontrolled division of one cell inside the breast and results in a visible mass, called a tumour. The tumour can be either benign or malignant. The accurate diagnosis in determining whether the tumour is benign or malignant can result in saving lives. Therefore, the need for precise classification within the clinic is a cause of great concern for specialists and doctors. This importance of Artificial Inelligence has been motivated for the last 25 years, when scientists began to realise the complexity of taking certain decisions to treat particular diseases. The use of machine learning and data mining as tools in medical diagnosis becomes very effective and one of the critical diseases in medicine where the classification task plays a very essential role is the diagnosis of breast cancer. Therefore, machine learning

DOI : 10.5121/ijaia.2012.3603

techniques can help doctors to make an accurate diagnosis for breast cancer and make the correct classification of being benign or malignant tumour. There is no doubt that evaluation of data taken from the patient and decisions of doctors and specialists are the most important factors in the diagnosis, but expert systems and artificial intelligence techniques such as machine learning for classification tasks, also help doctors and specialists in a great deal

We aim in this paper to investigate different machine learning techniques. We will use several algorithms and apply them on *Wisconsin breast cancer dataset*. We will focus on five machine learning techniques; Bayesian Network, Naïve Bayes , ADTree, J48, and Multilayer Neural Network(back-propagation). We will primarily study these various algorithms and analyse their results.

The following sections give a brief introduction for the algorithms used in the experiment , and then discussion about the result and the analysis. Finally, we frame our future work in the conclusion section.

2. BAYESIAN NETWORK

Bayesian Networks are graphical structures that give a chance to represent and reason about an uncertain domain. Bayesian Networks can be viewed as the merger between the probability theory and the graphical Theory. Bayesian Networks describe the probability distribution governing a set of variables by specifying a set of conditional independence assumptions along with a set of conditional probabilities. The nodes in the network represent variables(either discrete or continuous), and arcs represent the dependencies between these variables. The important concept in Bayesian Network is the conditional independence(Figure 1).

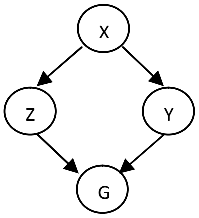


Figure.1 A simple structure of Bayesian Network.

The structure in Figure.1 shows that the probability of G is conditionally independent from X, given the probability of Z, Y and the relationship can be given in the following equation:

$$P(G|X,Z,Y)= P(G|Z,Y)$$

Each node is asserted to be conditionally independent of its non-descendants , given its immediate parents. This in fact, reduces the complexity of learning the target function. The joint probability distribution of all nodes in the network can be describe as follows :

i=1

$$P(y_1, y_2, y_3, \dots, y_n) = \prod_{i=1}^{n} P(\, y_i | \text{Parents}(y_i))$$

3. NAÏVE BAYES

Naïve Bayes are applied into learning tasks where each instance x is described by a conjunction of attribute values and where the target function ***f(x)*** can take on any value from some finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values$< a_1, a_2, a_3, \dots, a_n >$. The learner is asked to predict the target value for the new instance. The Bayesian approach used to classify the new instance is to assign the most probable target value, given the attribute values $< a_1, a_2, a_3, \dots, a_n >$. that describe the instance.

$$V_{NB} = \text{argmax} \prod_{i=1}^{n} P(a_i|V_j)P(V_j)$$

Naïve Bays can be considered as a special case of Bayesian Network(Figure.2).

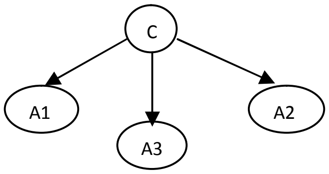


Figure.2 Naïve Bayes as a special case of Bayesian Network.

In Figure.2, each variable(A*i*) is conditionally independent from other variables, given its class (C). However, the assumption in the Naïve Bayes has more constraining than the global assumption in the Bayesian Network, where in Bayesian Network the variables are governed by specifying a set of conditional independence assumptions along with a set of conditional probabilities.

4. NEURAL NETWORK-BASED ALGORITHMS (NN)

A neural network can be defined as a model of reasoning based on the human brain[15]. Mainly, our brain consists of an intensive set of nerve cells connected to each other. The brain in humans contains about 10 billion and 60 trillion connections (synapses). When the brain uses multiple neurons in parallel, it can accomplish jobs much faster than the fastest computers nowadays.

Typically, the neuron consist of soma which represents the cell body, a number of fibers called dendrites, an axon which is a single long fiber , and synapses that represents connections between cells(Figure.3).

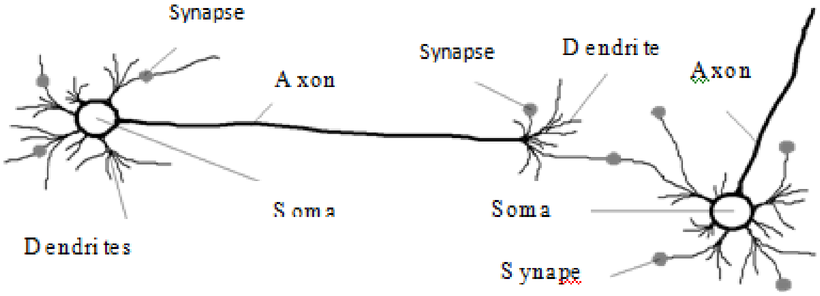


Fig.3. Biological Neural Network[15].

Similarly, an artificial Intelligence (ANN) is an interconnected group of artificial neurons that are based on a mathematical model to process information through neurons. ANN is sometimes called an adaptive system because it has the ability to change its structure based on the information being processed in the network. ANN's structure has input layers, hidden layers, and output layers (Figure.4).

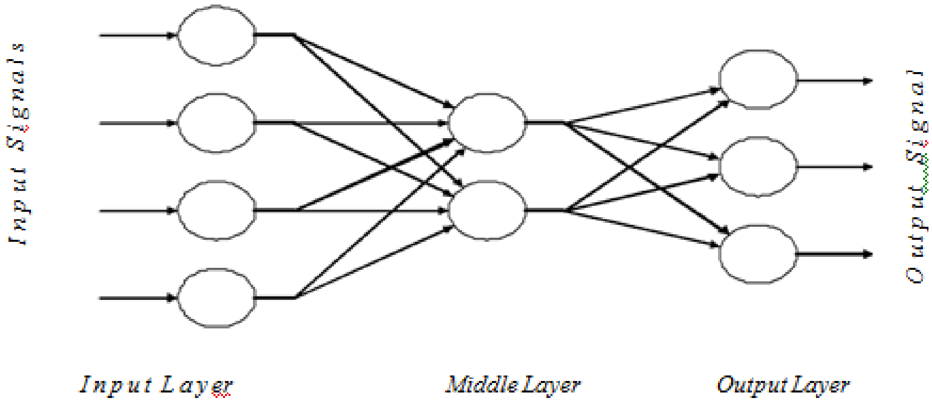


Fig.4. Artificial Biological Neural Network [15].

The figure above shows that the neurons are connected by links, and each link has a numerical weight associated with it. The weights express the strength of each neuron input and the learning is achieved by adjusting the weights through the links.

4.1.The perceptron

The perceptron is considered as the simplest form of neural networks, and is based on the McCulloch and Pitts neuron model [15]. It has just a single neuron with adjustable synaptic weights and a hard limiter (Figure.5).

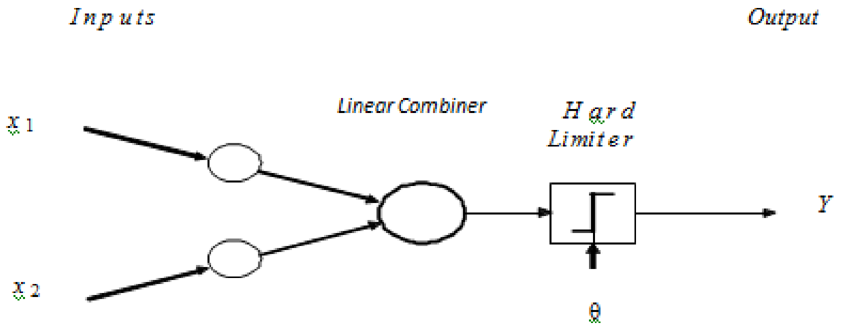


Fig.5. Single-layer two-input perceptron[15].

4.2. Multilayer neural network

A multilayer neural network is a feed-forward neural network (Figure.6). It has an input layer, one or more hidden layers, and output layer. The most popular algorithm is back-propagation which was first proposed in 1969 [15].

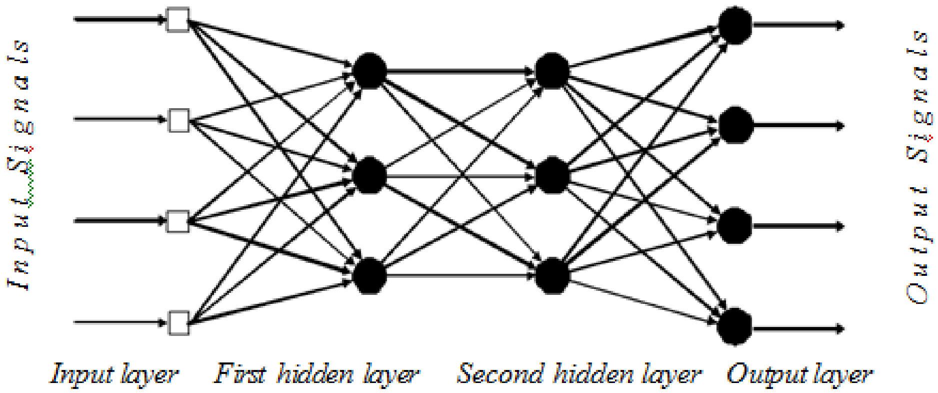


Fig6. Multilayer perception with two hidden layers[15].

Learning in multilayer network follows the same scheme as for a perceptron. A training data is presented to the network, and then the network computes its output pattern. If there is an error (difference between actual and desired output patterns), the weights in this case are adjusted to reduce the error by propagate the error backward through the output layer and hidden layers and then repeat the process again until we reach a certain criteria(usually the sum of squared error >= sufficient small number).

5.DECISION TREES ALGORITHMS

The Decision Trees are among the most popular inductive inference algorithms. The learned function in the decision trees is represented by a tree. Also, after the decision tree is constructed the IF…Then statements can be inferred easily.

5.1. ID3

The ID3 technique to build decision trees(with just nominal attributes) is based on the information theory and attempts to minimize the expected number of comparisons. The concept used to quantify information is *entropy*. Entropy is used to measure the amount of uncertainty , surprise or randomness in a set of data. Therefore, the ID3 finds the attribute that has the highest gain information or the lowest Entropy. Certainly, when all data in a set belongs to a single class, there is no uncertainty. In this case, the entropy is zero.

$$\text{Entropy}(S)= - p+ \log2(p+) - p- \log2(p-).$$

The goal of decision trees is to iteratively partition the given dataset into subsets, where all elements in each final subset belong to the same class[6].

5.2. C4.5

The decision tree algorithm C4.5 was designed by Quinlan [19]. It is designed to improve ID3 algorithm in different ways such as:

- Nominal and numeric data: the C4.5 can be used to classify either nominal or numeric attributes.

- Massing data: When the decision tree is built, missing data is simply ignored. To classify a record with a missing attribute value, the value for that item can be predicted based on what is known about the attribute values for the other records.

5.3. Alternating Decision Tree (ADTree)

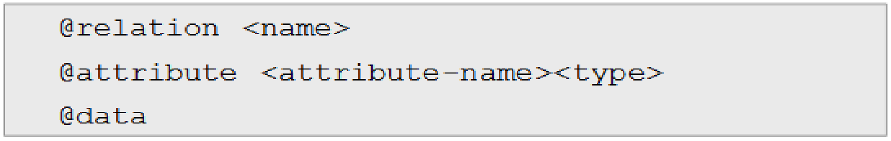
The ADTree is considered as another semantic for representing decision trees [9]. In the ADTree, each decision node is replaced by two nodes: a prediction node (represented by an ellipse), and a splitter node (represented by a rectangle). The decision tree in ADtree algorithm is identical while the prediction node is associated with a real valued number. As it is stated in the decision tree, an instance is mapped into a path along with the tree from the root to one of the leaves. However, unlike decision trees, the classification in ADTree that is associated with the path is not the label of the leaf. Instead, it is the sign of the sum of the prediction along the path.

6.THE EXPERIMENT WITH WEKA TOOL

The Weka tool is a collection of machine learning algorithms, and data preprocessing tools [24]. It provides a complete implementation for data sets being used. Weka tool has a variety of methods for transforming data sets, such as the algorithms for discretisation and filtering. Generally, the Weka tool has the methods for all standard machine learning algorithms (regression, classification, clustering, association rules, and attribute selection, etc) [24]. The data sets in Weka take the form of a single relation table in the ARFF format (Attribute- Relation File Format). The Weka tool has the ability to prepare the input data, evaluate learning methods statistically, and visualizing the input data with its result (output) during learning process.

6.1. Preparing datasets in Weka

The standard way of representing datasets in Weka tool is called an ARFF (Figure.7).



```
@relation <name>
@attribute <attribute-name><type>
@data
```

Fig7. ARFF structure.

6.1.2 ARRF Structure

6.1.2.1 Header :

The header section of an ARFF file is very simple and merely defines the name of the dataset along with set of attributes and their associated types [20]. The following tags are considered when the file is constructed:

6.1.2.1.1 @ Relation :

The relation<name> tag is declared at the beginning of the file, where <name> is referred to the relation name you intend to use.

6.1.2.1.2 . @ Attribute :

Attributes are defined as ***@attribute <attribute-name><type>***, and can take either numeric, nominal, or string values.

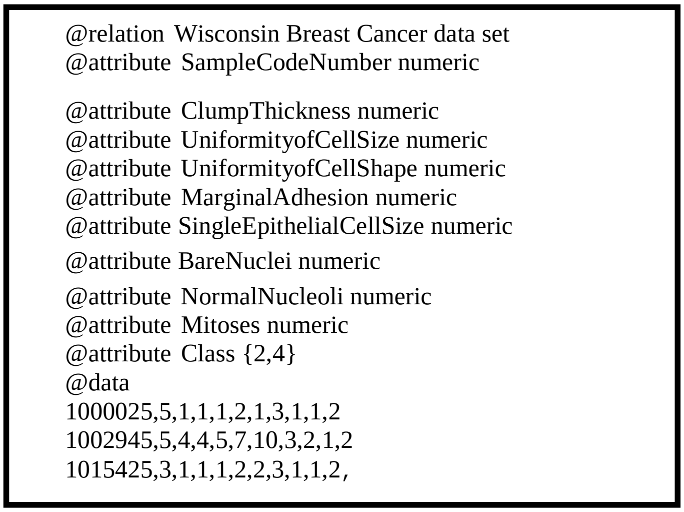
6.1.2.1.3 @Data [<data-format>].

The instances in @data tag are written one per line with values for each attribute in turn, separated by commas. If a value is missing, it is represented by a single question mark "?". The class attribute is always written as a last attribute.

6.2 The experiments with dataset

6.2.1 Wisconsin Breast Cancer data set

This dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr.William H. Wolberg, and formalized into ARFF file ,Figure 8.



```
@relation  Wisconsin Breast Cancer data set
@attribute  SampleCodeNumber numeric

@attribute  ClumpThickness numeric
@attribute  UniformityofCellSize numeric
@attribute  UniformityofCellShape numeric
@attribute  MarginalAdhesion numeric
@attribute SingleEpithelialCellSize numeric
@attribute BareNuclei numeric
@attribute  NormalNucleoli numeric
@attribute  Mitoses numeric
@attribute  Class {2,4}
@data
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2,
```

Fig.8. Part of Wisconsin breast cancer dataset in ARFF format.

Wisconsin Breast Cancer data set consists of (10) features. All are numeric type:

Clump Thickness, *Uniformity of Cell Size*, *Uniformity of Cell Shape*, *Marginal Adhesion*, *Single Epithelial Cell Size*, *Bare Nuclei*, *Bland Chromatin*, *Normal Nucleoli*, *Mitoses*. In addition, the class attribute has two values that the classification relies on: *Benign cancer* represented by (2) , and *Malignant cancer* represented by (4). The dataset has 699 instances [Benign (458 instances), Malignant (241 instances)]. The data set has 16 instances with missing data (unavailable data) represented by "?".

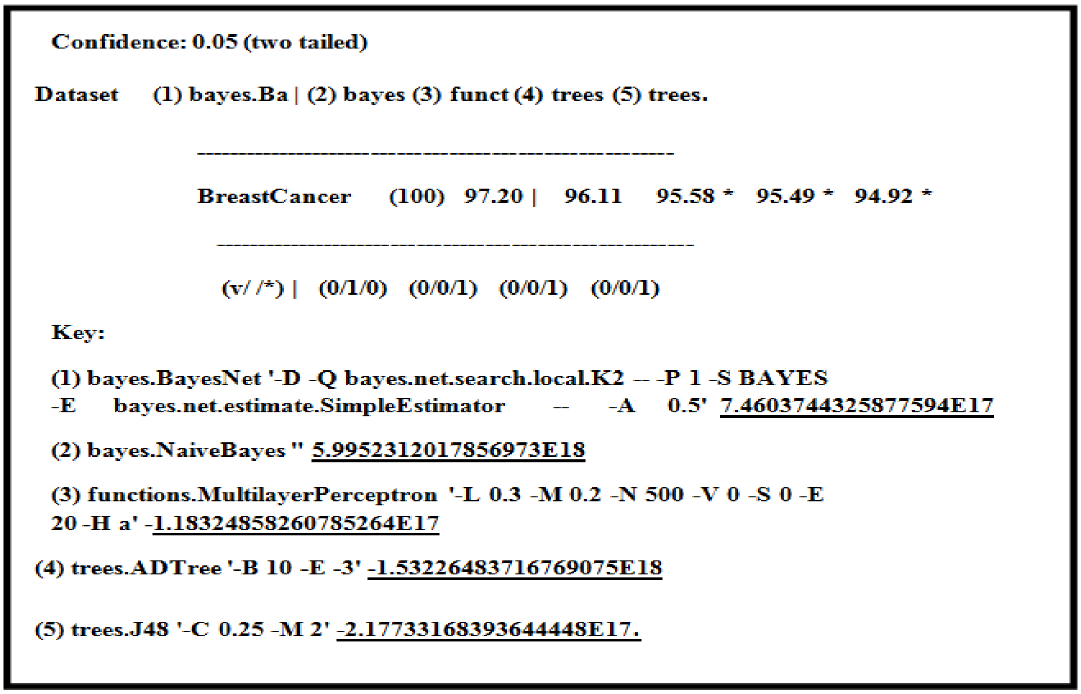
7. THE RESULT AND THE ANALYSIS

We aimed from this study to compare different learning algorithms to classify a benign cancer from malignant cancer in Wisconsin breast cancer dataset . One way to do the comparison is to estimate the error using any estimation procedure such as cross-validation for each learning algorithm and choose the algorithm whose estimate is smaller. This is still reasonable, if one algorithm has a lower estimated error than others in one dataset, then this algorithm can be consider as the model for this dataset. However, the difference between algorithms in this way might not be the significant difference. It is important to estimate the difference in errors significantly. The significant test called t-test(Student’s t test) can be used to determine whether one learning algorithm outperforms another on a particular learning task.

7.1 The K-fold cross-validated paired t test

This t-test is the most popular test among machine learning researchers and the one we used in Weka tool. The dataset is divided into k disjoint sets of equal size T1,T2,T3,…,Tn [Dietterich,1996] where the size is at least 30. It then trains and tests the learning algorithm k times. In each time we use Ti as a test set and the union of the other , where Tj,J≠ I, as training sets.

We began by specify the significant level (confidence equals to 0.05). The experiment as shown In figure 9, clarifies that the Bayesian Network is significantly better among the other algorithms. It is notable that the difference between Bayesian Network and Naïve Bayes is not significantly and this is acceptable since Naïve Bayes is just a special case of Bayesian network, where the constraints on the assumptions in Bayesian Network are more global than Naïve Bayes(see section 2.1).



```
  Confidence: 0.05 (two tailed)

Dataset    (1) bayes.Ba | (2) bayes (3) funct (4) trees (5) trees.

                  ------------------------------------------------------

                  BreastCancer    (100)  97.20 |   96.11    95.58 *   95.49 *   94.92 *

                    ------------------------------------------------------

                     (v/ /*) |   (0/1/0)   (0/0/1)   (0/0/1)   (0/0/1)

  Key:

  (1) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES
  -E     bayes.net.estimate.SimpleEstimator      --     -A     0.5' 7.4603744325877594E17

  (2) bayes.NaiveBayes '' 5.9952312017856973E18

  (3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E
  20 -H a' -1.18324858260785264E17

(4) trees.ADTree '-B 10 -E -3' -1.53226483716769075E18

(5) trees.J48 '-C 0.25 -M 2' -2.17733168393644448E17.
```

Fig.9. The result of the experiment .

The asterisks in Figure.9 means that the result is worse than the baseline (Bayesian Network in this case). For example, the function multilayer neural network is significantly worse than Bayesian Network. Figure.10 shows the Bayesian network that has been learned .

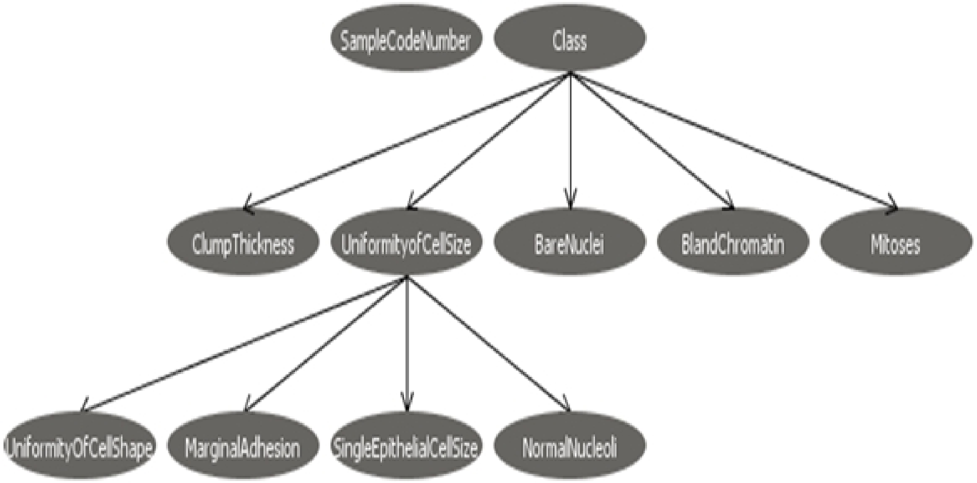


Fig.10 the Bayesian Network for Wisconsin breast cancer.

8. CONCLUSION AND FUTURE WORK

In this paper we have investigated five popular algorithms on one of the most important domains in medicine; namely breast cancer. We used Wisconsin breast cancer dataset that contains important risk factors. These factors usually are used to diagnose the breast cancer in labs and give a reliable result. We have shown that using machine learning can help in cancer detection. However, we also believe that the diagnosis of breast cancer can be more accurate if it is conducted in the genomic level. This will help in understanding for example the abnormality of genes and which are down-expressed / up-expressed. Starting from this motivation, we will work in the future in the genes-regularity networks using gene expression profiles for breast cancer. The consideration will be biased to Bayesian Networks and their interpretation to causal networks as it has not been investigated intensively in the literature.

9. REFERENCES

[1] [Bellaachia & Erhan, 2005] :Bellaachia, A . & Guven, E.(2005) Predicting Breast Cancer Survivability Using Data Mining Techniques[on line].s.n,Available from: http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datas ets/bellaachia.pdf[access 02 December 2005],

[2] [Bouckaert, 2007]: Bouckaert , R.(2007) Bayesian Network Classifiers in Weka,[on line].New Zealand. Available from : http://weka.sourceforge.net/manuals/weka.bn.pdf [accessed 2 July 2007].

[3] [Breastcancer,2007]. : Medical Experts(2007) Understanding Breast Cancer [online].USA : Sally Aman . Available from :http://www.breastcancer.org/symptoms/understand_bc/[accessed 26th July 2007].

[4] [Carlos & Sipper, 1999]: Andres , C. & Sipper , M.(1999) A fuzzy- genetic approach to breast cancer diagnosis. Artificial Intelligence in Medicine Journal , 7(2), pp131-155.

[5] [Coiera, 2003] : Coiera , E.(2003) Guide to Health Informatics . 2nd ed. London : Arnold.

[6] [Dunham, 2003]: Dunham , H.(2003) Data Mining : Introductory and Advanced Topics. U.S.A : Pearson Education ,Inc.

[7] [Fentiman,1998] : Fentiman , I.S.(1998) Detection and Treatment of Breast Cancer.2nd ed. London: Informa Health Care.

[8] [Fowler & Scott ,2000] : Fowler , M. & Scott, K.(2000) UML distilled (2nd ed.): a brief guide To the standard object modeling language. 2nd ed. Boston : Addison-Wesley Longman Publishing Co.

[9] [Freund & Mason, 1999]: Freund , Y. & Mason , L.(1999) The alternating decision tree learning algorithms[on line].s.n .Available from: www1.cs.columbia.edu/compbio/medusa/non_html_files/Freund_Atrees.pdf.[accessed1999]

[10] [Han & Kamber, 2001] : Han , J. & Kamber , M.(2001) Data Mining : Concepts and Techniques. San Francisco : Morgan Kaufmann. [11]- [Hesterberg et al., 1990] : Papalexopoulos , A. & Hesterberg , T.(1990) A regressionbased approach to short-term system load forecasting , Power Systems Journal 5(4),pp1535-1547.

[12] [Junfeng , 2002] : Junfeng, Qu (2002) An introduction to Data Mining Technology [on line] ,s.n. Available from : http://www.cs.uga.edu/~budak/courses/s02/nis/DMPres.ppt [accessed 08 Feb 2002].

[13] [Kiyan & Yildirim, 2004] : Kiyan, T. & Yildirim , T.(2004) Breast cancer diagnosis using statistical neural networks. Electrical & Electronics Engineering Journal, 4(2) pp1149-1153.

[14] [Korb & Nicholson, 2004] : Korb, K.& Nicholson , A.(2004) Bayesian Artificial Intelligence.USA : Chapman & Hall/CRC.

[15] [Michel, 2005] : Negnevitsky , M.(2005) Artificial Intelligence:A guide to Intelligent Systems. 2nd ed. Essex : Pearson Education Limited .

[16] [Nauck & Kruse, 1998] : Nauck, D. & Kruse, R (1998) Obtaining interpretable fuzzy classification rules from medical data . Artificial Intelligence in Medicine Journal , 16(no issue number), pp 149-169.

[17] [Norsys, 2007] : Norsys Software Corp(2007) Netica Tutorial [on line]. Canada: Norsys Software Corp . Available from : http://www.norsys.com/tutorials/netica/nt_toc_A.htm [last accessed 27July 2007].

[18] [Parkin et al., 2002] : Parkin, M., Bray , F., Ferlay, J . & Pisani , P. (2002) Global Cancer Statistics. CA Cancer J Clin Journal, 55[no issue number], pp.74-108.

[19] [Quinlan, 1993] :Quinlan, J.(1993) C4.5: programs for machine learning. San Francisco : Morgan Kaufmann Inc.

[20] [Roberts, 2005] : Roberts , A.(2005) Guide to Weka, [on line]. UK. Available from : http://www.andy-roberts.net/teaching/ai32/weka.pdf [accessed 1st March 2005].

[21] [Sakorafas et al., 2002] : Sakorafas , G . , Krespis , E . & Pavlakis , G .(2002) Risk estimation for breast cancer development; a clinical perspective . Surgical Oncology Journal, 10(no issue number),pp 183-192.

[22] [Seal et al., 2007]: Sahan, S ., Polat , K., Kodaz, H. & G, S .(2007)A new hybrid method based on fuzzy- artificial immune system and K-nn algorithm for breast cancer diagnosis . Computers in Biology and Medicine Journal, 37 (no issue number),pp.415-423.

[23] [Szolovits, 1982]: Szolovits, P.(1982)Artificial Intelligence in Medicine. USA : Westview.

[24] [Thodberg,1993] : Thodberg , H.(1993) Ace of Bayes Application of Neural Netwoks with pruning[on line] , s.n Available from : http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/959/ftp:zSzzSzarchive.cis.ohiotate.eduzSzpubzSzneuroprosezSzthodberg.acofbayes.pdf/thodberg93ace.pdf[accessed 19 May 1993].

[25] Witten & Frank, 2005] : Witten, I . & Frank , E.(2005) Data Mining: Practical Machine Learning Tools and Techniques. USA : Elsevier Inc.

[26] [Zhang, 2004] : Zhang , H.(2004) The Optimality of Naïve Bayes [on line].USA:aaai.Available from http://www.cs.unb.ca/profs/hzhang/FLAIRS04ZhangH.pdf[no date].