Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 3, No. 2, May 2012), Pages: 35-44
www.jacr.iausari.ac.ir

A Probabilistic Bayesian Classifier Approach for Breast Cancer Diagnosis and Prognosis

Alireza Sadeghi Hesar
Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
Alireza.sadeghi89@yahoo.com

Received: 2012/04/17; Accepted: 2012/04/20

Abstract

Basically, medical diagnosis problems are the most effective component of treatment policies. Recently, significant advances have been formed in medical diagnosis fields using data mining techniques. Data mining or Knowledge Discovery is searching large databases to discover patterns and evaluate the probability of next occurrences. In this paper, Bayesian Classifier is used as a Non-linear data mining tool to determine the seriousness of breast cancer. The recorded observations of the Fine Needle Aspiration (FNA) tests that are obtained at the University of Wisconsin are considered as experimental data set in this research. The Tabu search algorithm for structural learning of bayesian classifier and Genie simulator for parametric learning of bayesian classifier were used. Finally, the obtained results by the proposed model were compared with actual results. The Comparison process indicates that seriousness of the disease in 86.18% of cases are guessed very close to the actual values by proposed model.

Keywords: *Breast Cancer, Data Mining, Bayesian Classifier, Genie Simulator*

1. Introduction

Advances in information technology have made progresses in other fields such as medical sciences. Today, data mining techniques are applied to develop physician's assistant systems. Basically, Data mining is evaluating the features of a dataset and assigning them to a set of predefined groups. This process is the most common data mining capabilities. Data mining can also produce a new pattern or a model based on historical and statistical databases. Then this extracted model can be used to classify new data sets. Generally, data mining consists of the following steps:

- Data Cleaning: this process removes irrelevant data from data set.
- Data Consolidation: at this step, selected data is converted into appropriate forms.
- Patterns Extraction: major task at this step is discovery and extraction of common features.
- Patterns Evaluation: it is critical step in which extracted patterns are compared with predetermined requirements.
- Classification: Finally, this step use obtained knowledge for new data classification.

Today, Bayesian Probabilistic Classifiers (BPCs) are very popular data mining techniques to encode relationships among a set of variables under uncertainty. The uncertainty is a common feature of bayesian modeling and medical diagnosis problems. So effectively a BPC can be used for detecting patterns in medical databases. Bayesian Probabilistic Classifier is a certain type of Bayesian Network that with the help of machine learning algorithms performs classification process on huge database with high accuracy.

The breast cancer is one of the most common diseases in women. The main reason for this lesion is the abnormal growth of a cell in breasts or nipples. Generally, the carcinogenic agents are divided into two major categories: Environmental factors and genetic factors. About 88% of breast cancers are due to genetic abnormalities. So people cannot avoid from this fatal happening. The most important treatment for breast cancer is radiotherapy classic methods. But awareness of the seriousness is very critical to use this method. Briefly, this paper propose a probabilistic classifier based on bayesian analysis to determine the likelihood of malignancy or benign for breast cancer. Figure 1 presents major procedures of project.

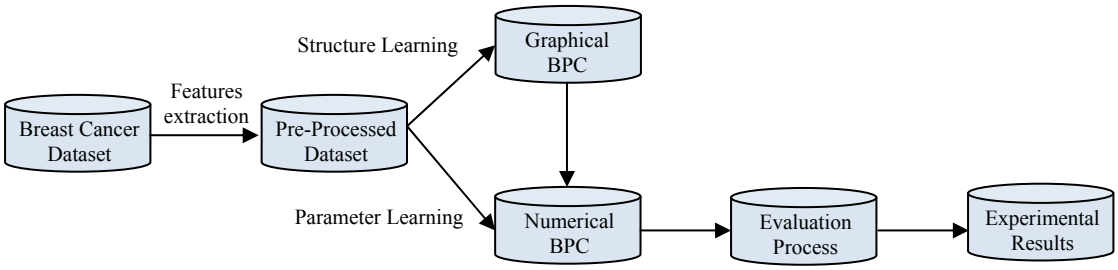

Figure 1. Main phases of project

So far, many models have been proposed in the field of breast cancer diagnosis using data mining and machine learning techniques such as decision trees, multiple linear regression (MLR), Artificial Neural Networks (ANNs) and discriminate analysis. Hussein A. Abbass in [1] developed a breast cancer diagnosis model based on an evolutionary artificial neural network (EANN) using the pareto differential evolution algorithm augmented with local search. Xiaobo Zhou et al. in [2] expanded an improved BBN using the linear regression technique to gene classification. The results show that proposed method can effectively identify important genes of breast tumors. In research of Cheng chen and Chou Hsu [3] Genetic Algorithms were proposed to achieve the best patterns in breast cancer data sets. Young U. Ryu et al. applied isotonic separation as a data classification/separation technique for breast cancer prediction in [4].Nicandro Cruz-Ramírezet al. in [5] evaluated bayesian classifiers performance for the diagnosis of breast cancer using two different real datasets. Susan M. Maskery et al in [6] simulated breast pathology using Bayesian Belief Networks (BBNs). In this study, 1631 different instances were considered. Wei-Chang Yeh et al. in [7] proposed a hybrid approach based on statistical methods and particle swarm optimization (PSO) for mining breast cancer patterns. Imran Kurt Omurlu et al in [8] compared bayesian analysis and Cox regression analysis in a cancer database that was obtained from 400patients between 1998 and 2007. F.K. Ahmad in [9] proposed bayesian computational analysis to

construct sequential genenet works from breast cancer databases. In this paper, effectiveness of four different geneson breast cancer were studied.
In the next section, the characteristics of experimental data set are briefly reviewed. The Bayesian Probabilistic Classifier will be introduced in the section 3. Section 4 describes the proposed method with full details. Finally, Section 5 summarizes the main results and conclusions.

2. Experimental Data Set

Availability of reliable data set for intelligent tools learning is most needed. Fortunately, there are several medical data sets that researchers can use them for benchmark and test processes. For example, Genetic STATLOG DNA (GSD) data set is frequently used for genetic cloning in intelligent systems learning or SPECT Heart data set that is used to simulate the artificial hearts. In this study, Wisconsin Diagnostic Breast Cancer (WDBC) is used for proposed model learning process. WBCD database is results of the Fine Needle Aspiration (FNA) test that has been achieved in medical and research Center of Wisconsin (MRCW) by Dr. William H. Wolberg in 1992-1995. Table 1 presents the features of WBCD.

Table1. Features of WBCD database

Attribute	Description
Data Set Characteristics:	Multivariate
Attribute Characteristics:	Real
Associated Tasks:	Classification
Number of Instances:	559
Number of Attributes:	32
Missing Values?	No
Area:	Life
Date Donated	1995/11/01
Number of Web Hits:	117638

3. Bayesian Probabilistic Classifiers

So far, many methods have been proposed for the uncertainty modeling. The proposed methods are often based on fuzzy logic and probability theory. Bayesian Probabilistic Classifiers are also based on probabilities and use graphical models to express dependencies between variables in field of study. Generally, BPCs are certain type of Bayesian Networks that are useful for cases in which the future situation depends on the current situation. So we can use them for reasoning and decision under uncertainty. In other words, each BPC as a computational structure, inferences the Joint Probability Distribution (JPD) of a set of related variables using Observational data. The BPCs use a Directed Acyclic Graph (DAG) to encode relationships among a set of variables. The nodes of DAG represent discrete variables and arcs represent Conditional in dependence of variables. The JPD computation in BPCs can be expressed mathematically using Equation 1:

$$P_r(Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^{n} P\left(\frac{Y_i}{\pi_i}\right) \quad (1)$$

Equation 1 shows that the joint probability distribution for node Y in DAG is product of the probability of each Y_i of node Y given the parents of Y_i .

Each node of DAG has a Conditional Probability Table (CPT) that presents probability of each state of node according to any combination of parent states. An example of a bayesian probabilistic analysis with complete CPTs is shown in Figure 2.

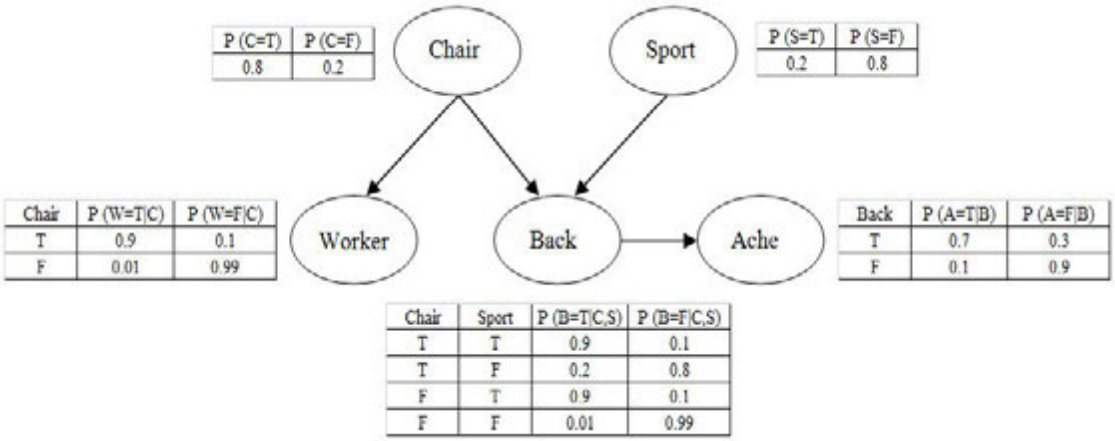

Figure 2. Inference method of bayesian

The problem of BPC learning can be divided into two different related phases:
1) learning the structure process or graphical modeling that focuses on DAG construction 2) learning the parameter process or numerical modeling that focuses on values of CPTs computation. But two different methods can be used for structure learning: constraint based methods and score based methods. Constraint based approach relies on conditional independences and dependences between variables. The constraint based method is more accurate consistently but it is only well-applicable when infinite or huge data collection is used. But, score based methods apply heuristic search algorithms to find the best structure because finding the structure of BPCs is a NP-Complete problem.

4. Proposed Method

Selecting an appropriate algorithm is the most difficult step in BPC structure construction. Usually, researchers rely on the results of previous studies and apply proposed algorithms. But in this paper, for the most reliable method discovery, several algorithms were evaluated and each of them independently, was imposed on the WDBC database. Four used heuristic algorithms are:

- Greedy Thick Thinning
- Tabu Search
- K2 Search
- Iterated Hill Climbing

The Computations were done on a computer with a 2.6 GHz Core 2 CPU, and 2 GB of RAM memory. Figure 3, 4 provide comparison of all the algorithms based on

construction time of the structure and classification accuracy metrics. As a result, from the point of classification accuracy, the Tabu search is the best algorithm and K2 is less time consuming compared to other algorithms.

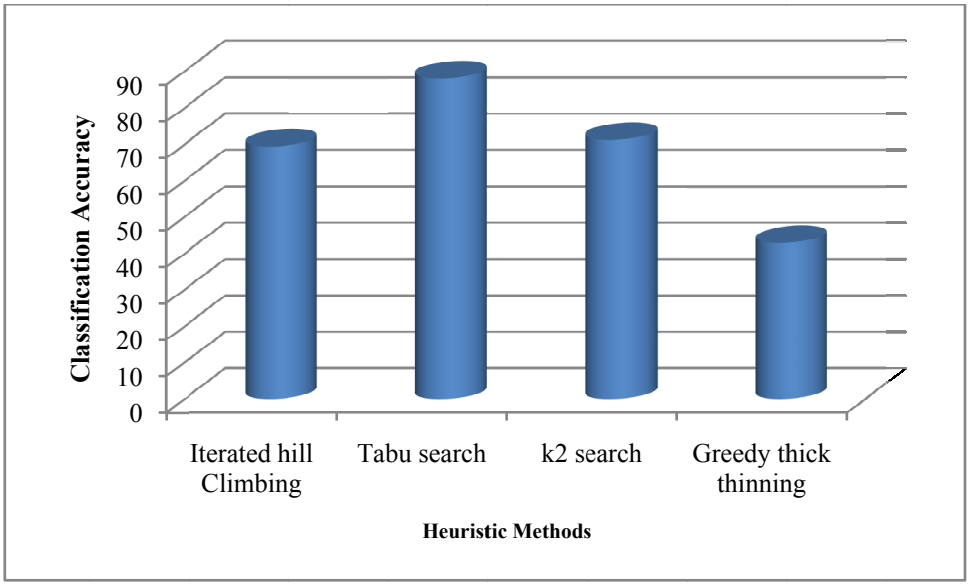

Figure 3. Classification accuracy of methods on WDBC database

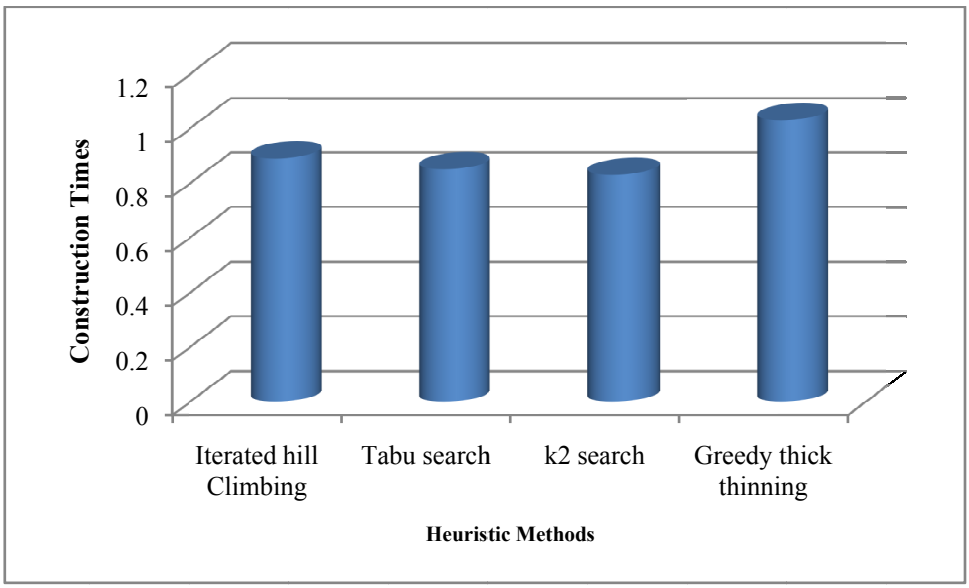

Figure 4. Classification accuracy of methods on WDBC database

4.1 Variables Extraction

Breast cancer is one of the most common cancers among women. The main symptom is the observed Mass or tumor in the breasts or nipples. But, there are cases of cancer that no mass has been observed. Now, FNA test method is used for cancer diagnosis usually. FNA test is very high cost to ordinary people but no surgery, no bleeding and also high speed are considered as important advantages of this method. WBCD database that was

described in Section II is results of the FNA test that has been achieved in the Medical and research Center of Wisconsin (MRCW). All considered Factors in these experiments are discrete variables that values between 1 and 9 are allocated to them. Table 2 presents effective factors of FNA test. These factors are considered as main variables.

Table 2. Effective factors of FNA test (Features Extraction Process)

No.	Variable
v1	Clump Thickness
v2	Uniformity Of Cell Size
v3	Uniformity Of Cell Shape
v4	Marginal Adhesion
v5	Single Epithelial Cell Size
v6	Bare Nuclei
v7	Bland Chromatin
v8	Normal Nuclei
v9	Mitosis

4.2 Graphical BPC Construction

According to results of previous sections, Tabu search algorithm is a good option for our study. Pseudo-code of algorithm is shown as follows:

```
Tabu Search Pseudo-Code
1:   s ← s0
2:   sBest ← s
3:   List ← null
4:   while (not stoppingCondition())
5:      candidateList ← null
6:      for(sCandidate in sNeighborhood)
7:           if(not containsTabuElements(sCandidate, List))
8:                candidateList ← candidateList + sCandidate
9:           end
10:     end
11:     sCandidate ← LocateBestCandidate(candidateList)
12:     if(fitness(sCandidate) > fitness(sBest))
13:          sBest ← sCandidate
14:          tabuList ← featureDifferences(sCandidate, sBest)
15:          while(size(List) > maxTabuListSize)
16:               ExpireFeatures(List)
17:          end
18:     end
19:  end
20:  return(sBest)
```

The nodes of proposed bayesian classifier are the variables shown in Table 2 that are responsible to record observations. 9 different modes are defined for each node. The arcs represent effective relations between two nodes in the network. In real, the states of node in FNA test, help to estimate the states of other nodes that are connected to it.

After the number of nodes and states and arrangement of the all nodes was determined, attributed values to each variable in table 2 are analyzed by Tabu search. The proposed structure for bayesian classifier that is generated by Tabu search algorithm is shown in Figure 5. This structure learned by WBCD database. The learned bayesian classifier, divides each instance into Benign class and Malignant class.

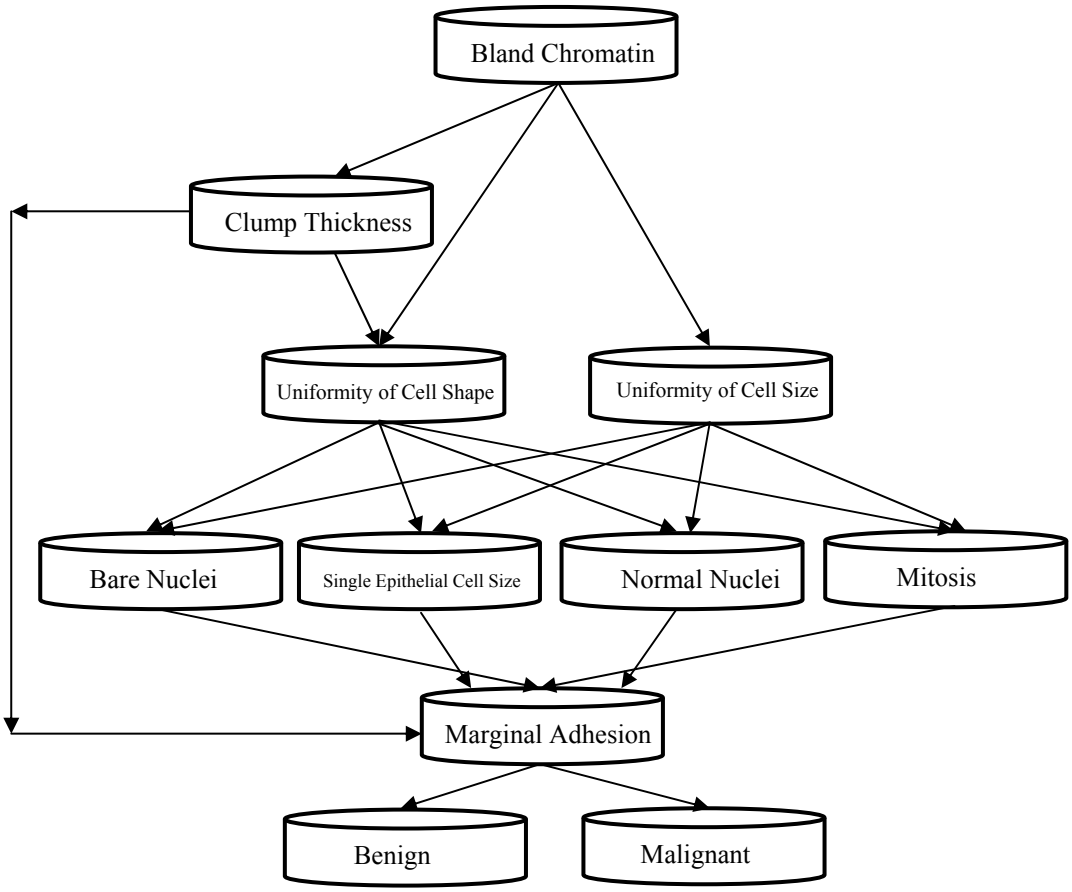

Figure 5. Generated classifier by Tabu search

4.3 Numerical BPC Construction

Maximum Likelihood Estimation or MLE is the most common method for parameter learning. Usually, developed softwares to bayesian construction such as BNT[1], Genie, Netica and Agena Risk apply MLE method. The definition of main likelihood function $L(\varphi: D)$ is equal to probability of independent instances D with parameters φ and is calculated as follows:

1. Bayesian Network Tool Box of Matlab Software

$$L\,(\varphi{:}\,D)\;=\;\prod_{m=1}^{M}P\,(x[m]|\varphi)$$

$$L\,(\varphi{:}\,D)\;=\;\sum_{m=1}^{M}log\;P\,(x[m]|\varphi)$$

$$\hat{\varphi}\;=arg\;max\;L\,(\varphi{:}\,D)$$

$$=\prod_{m=1}^{M}\prod_{i=1}^{N}P\,(x_i[m]|Pa_i[m]{:}\,\theta_{x_i|Pa_i})$$

$$=\prod_{i=1}^{N}[\prod_{m=1}^{M}P\,(x_i[m]|Pa_i[m]{:}\,\theta_{x_i|Pa_i})]$$

$$=\prod_{i=1}^{N}L_i\;(\theta_{x_i|Pa_i}\;:\;D)$$

$$L_i\;(\theta_{x_i|Pa_i}\;:\;D)\;=\;\prod_{m=1}^{M}P\,(x_i[m]|Pa_i[m]\;:\;\theta_{x_i|Pa_i}$$

For example, the output of Genie simulator for produced classifier by Tabu search is shown in figure 6. Also, figure 6 presents the produced values of CPTs for one of real instances. The computed probability of malignant and benign is 14% and 86%.Comparison process of the actual values with predicted values by Genie simulator for the first 150 patient of WBCD is evident in scatter charts of figure 7,8,9.

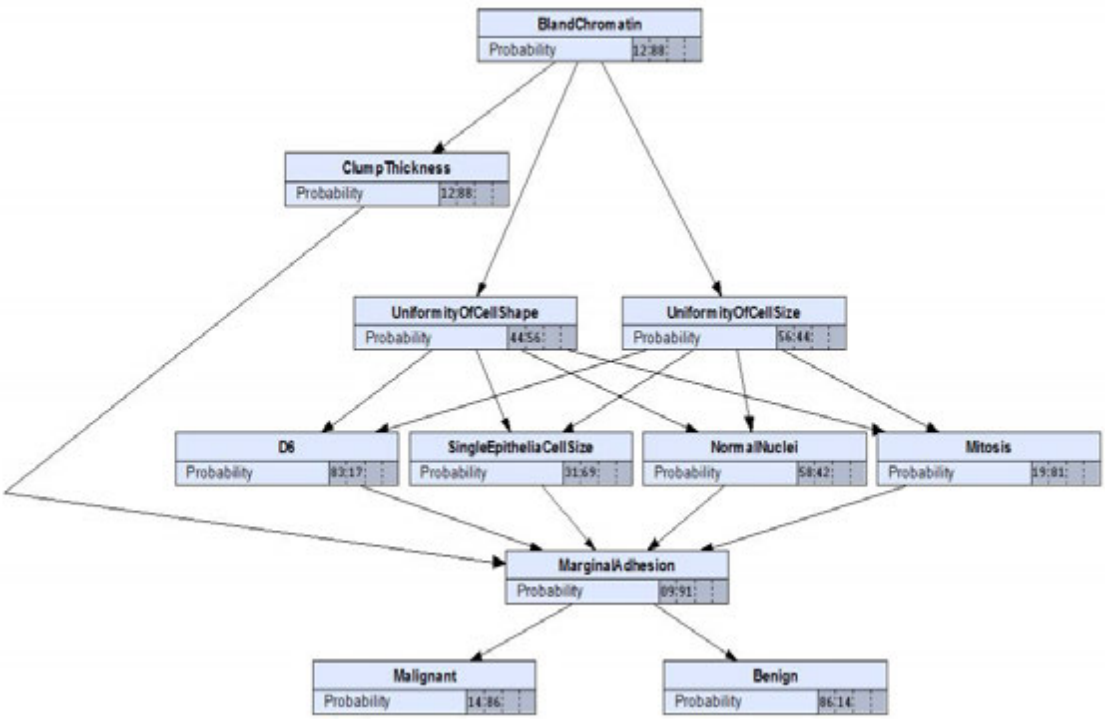

Figure 6. Output of genie simulator

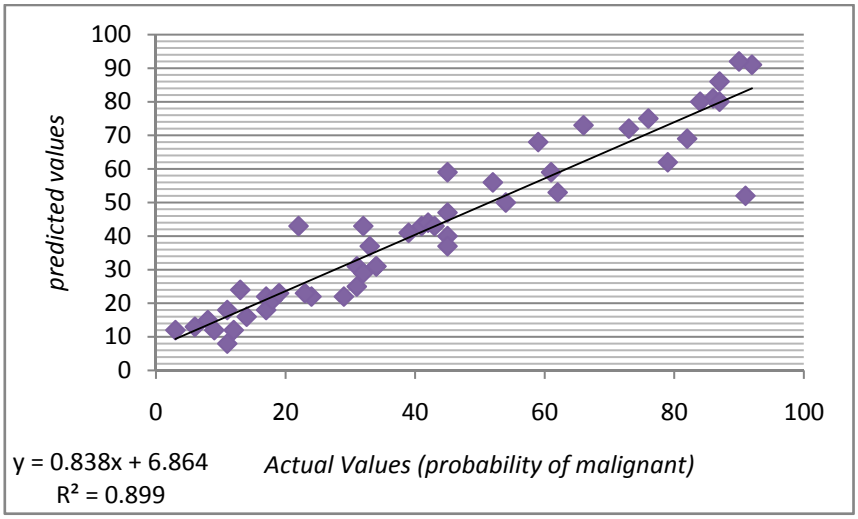

Figure 7. Comparison process of the actual values with predicted values – first 50 instances

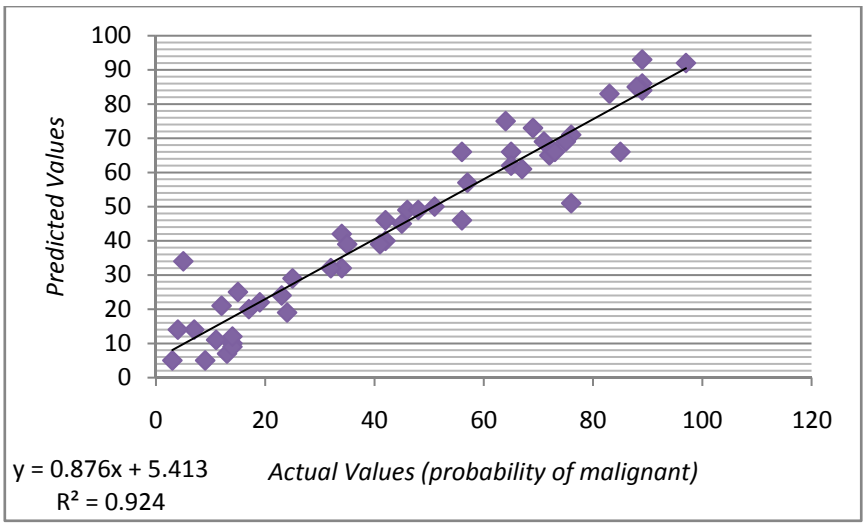

Figure 8. Comparison process of the actual values with predicted values – second 50 instances

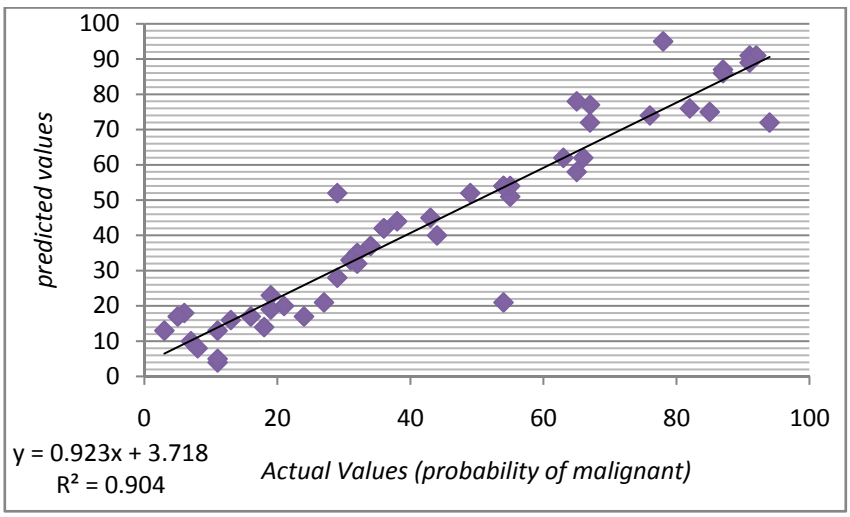

Figure 9. Comparison process of the actual values with predicted values – third 50 instances

5. Conclusion

Breast cancer is one of the main causes of death among women in developed countries. In recent decade, data mining techniques are applied to develop physician's assistant systems for cancer diagnosis. In this paper, a breast cancer diagnosis system was proposed using Bayesian Probabilistic Classifiers or BPCs. BPCs as a data mining tool encode relationships among a set of variables under uncertainty. In construction phase, several heuristic algorithms were evaluated on WDBC data base. Tabu search as a heuristic method was selected for graphical modeling of proposed BPC and Genie simulator (using Maximize Likelihood Estimation) was used to numerical modeling of BPN. The proposed model results for 150realpatient using scatter charts were shown in the final section. Probability of benign is not shown on the charts for more simplicity that is calculable from "100 - malignant" easily. On average observed difference between actual values and predicted values is about 13.83 %.As a result, BPC using its unique features (modeling under uncertainty) can extract common patterns from medical and clinical data bases. Generally, BPCs can be used in problems of modeling that very high accuracy, speed of construction and reliability are required such as medical diagnosis and security problems.

5. References

[1] H.A.Abbas, "An evolutionary artificial neural networks approach for breast cancer diagnosis", Journal of Artificial Intelligence in Medicine archive, Volume 25 Issue 3, July, 2002, Pages 265-281.

[2] X.Zhou, D.Huang, "Automatic 3D Facial Expression Recognition Based on a Bayesian Belief Net and a Statistical Facial Feature Model", Published in Proceeding ICPR '10 Proceedings of the 2010 20th International Conference on Pattern Recognition, Pages 3724-3727.

[3] C.Chen, Ch.Hsu,"A GAs based approach for mining breast cancer pattern", Published in Journal Expert Systems with Applications: An International Journal archive, Volume 30 Issue 4, May, 2006, Pages 674-681.

[4] [4]R.Young, Ch.R" Breast cancer prediction using the isotonic separation technique", *European Journal of Operational Research*, Volume 181, issue 2 (September 1, 2007), p. 842-854.

[5] [5]N. C-Ramírez, H. Gabriel"Diagnosis of breast cancer using Bayesian networks: A case study**",** Journal of Artificial Intelligence in Medicine archive, Volume 37 Issue 11, November, 2007 Pages 1553-1564.

[6] [6] S. Maskery, H.Hu" A Bayesian derived network of breast pathology co-occurrence", Journal of Biomedical Informatics archive,Volume 41 Issue 2, April, 2008,Pages 242-250.

[7] **[7]**W.Chang. Yeh, W.W.Chang "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method**",** *Expert Systems With Applications*, Volume 36, issue 4 (May, 2009), p. 8204-8211.

[8] [8]Ture, Mevlut; Tokatli, Fusun; Kurt Omurlu, Imran"The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data. ",*Expert Systems With Applications* vol. 36 issue 4 May, 2009. p. 8247-8254.

[9] [9] F.K. Ahmad "The inference of breast cancer metastasis through gene regulatory networks", 2012 Apr;45(2):350-62. Epub 2011 Dec 10.