

AN INTEGRATION OF BAYESIAN BELIEF NETWORKS
IN MULTI-CRITERIA DECISION ANALYSIS:
AN APPLICATION IN MEDICAL DECISION MAKING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

DİLAN KARATEPE

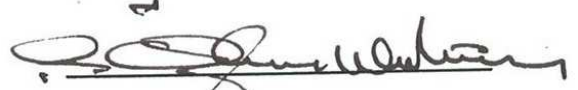
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

SEPTEMBER 2007

Title of the Thesis : **An Integration of Bayesian Belief Networks in Multi-criteria Decision Analysis: An Application in Medical Decision Making**

Submitted by **Dilan Karatepe**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University



Prof.Dr. Ç. Özhan ULUATAM
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.



Prof.Dr. Levent KANDİLLER
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.



Prof.Dr. Fetih YILDIRIM
Supervisor

Examination Date : 28.09.2007

Examining Committee Members

Yrd.Doc.Dr. Suat Kasap (Hacettepe University)



Prof. Dr. Fetih Yıldırım (Çankaya University)



Dr. Özlem Türker Bayrak (Çankaya University)



STATEMENT OF NON PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Dilan KARATEPE

Signature :



Date :

16.11.2007

ABSTRACT

AN INTEGRATION OF BAYESIAN BELIEF NETWORKS
IN MULTI-CRITERIA DECISION ANALYSIS:
AN APPLICATION IN MEDICAL DECISION MAKING

Karatepe, Dilan

M.Sc., Department of Industrial Engineering

Supervisor : Prof. Dr. Fetih Yıldırım

September 2007, 90 pages

This thesis analyzes the Medical counseling problem regarding breast cancer disease by using Bayesian Belief Network structure to visualize the conditional dependencies between variables and their impact on each other. Analytical Hierarchy Process (AHP) method is utilized to approximate the conditional probabilities of the diagnostic Fine Needle Aspiration (FNA) test outcome. The accuracy of the results obtained from surgical biopsy are compared to the actual outcomes and evaluated in terms of consistency and error magnitude.

Keywords: Bayesian Belief Networks, Multi-criteria Decision Analysis, AHP, Medical Decision Making, Fine Needle Aspiration (FNA)

ÖZ

ÇOK ÖLÇÜTLÜ KARAR ANALİZİNİN BAYES AĞLARI İÇİNDE KULLANIMI: TIBBİ KARAR VERME SÜREÇLERİNDE BİR UYGULAMA

Karatepe, Dilan

Yükseklisans, Endüstri Mühendisliği Anabilim Dalı

Tez Yöneticisi: Prof. Dr. Fetih Yıldırım

September 2007, 90 sayfa

Bu çalışma, meme kanseri hastalığına yönelik tıbbi danışmanlık problemini, değişkenler arasındaki koşullu olasılıkları ve birbirleri üzerindeki koşullu bağımlılıklarını görünür hale getirmek amacıyla Bayes Ağları kullanılarak incelemiştir. Analitik Hiyerarşi Süreci metodundan, İnce İğne Aspirasyonu tanı testine ait koşullu olasılıkları hesaplamada faydalanılmıştır. Sonuçların doğruluğu biyopsi ameliyatından elde edilen gerçek sonuçlarla karşılaştırılmış ve tutarlılık ve hatanın büyüklüğü yönünden değerlendirilmiştir.

Anahtar Kelimeler: Bayes Ağları, Çok-Kriterli Karar Analizi, Analitik Hiyerarşi Süreci , Tıbbi Karar Verme, İnce İğne Aspirasyonu

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. Fetih YILDIRIM for his guidance, criticism, encouragements and insight throughout the research.

I would also like to thank the thesis jury Assistant Professor Suat KASAP and Dr. Özlem TÜRKER BAYRAK for reading and evaluating the thesis.

I would like to express my thanks to Professor Şevket RUACAN and his assistant Dr. Sevgen ÖNDER for their help during interpretation of the data.

The guidance of Assistant Professor Selim AKSOY for the data set and the technical assistance of Ms. Miray Aslan are gratefully acknowledged.

TABLE OF CONTENTS

STATEMENT OF NON PLAGIARISM.....	iii
ABSTRACT.....	iv
ÖZ.....	v
ACKNOWLEDGMENTS.....	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	ix
LIST OF TABLES	x
CHAPTERS:	
1. INTRODUCTION	11
2. BAYESIAN APPROACH.....	13
2.2 Some Advantages of Bayesian Methods.....	15
2.3 Basics of Bayesian Analysis.....	17
3. MEDICAL DECISION MAKING	20
3.1 Medical Diagnosis	20
3.2 Estimating Test Accuracy Parameters	21
3.2.1 Sensitivity and Specificity	21
3.2.2 Positive and Negative Predictive Values	22
4. BAYESIAN BELIEF NETWORKS	24
4.1 Definition.....	24
4.2 Introduction to Inference in Bayesian Networks	28
4.3 Advanced Inference for Bayesian Networks	29
4.3.1 Variable Elimination.....	30
4.3.2 Junction Tree Algorithm	31
4.3.2 Pearl’s Message Passing Algorithm	33
5. MULTI-CRITERIA DECISION ANALYSIS	45
5.1 Literature Review	45
5.2 An Overview of MCDA Methods.....	47
5.3 Definitions	48
5.4 Definition of Analytical Hierarchy Method (AHP).....	49
5.5 An Example of AHP	51

6. INTEGRATING BAYESIAN BELIEF NETWORKS AND ANALYTICAL HIERARCHY PROCESS	53
7. AN APPLICATION IN MEDICAL DIAGNOSIS.....	55
7.1 Information about the Application Data	55
7.1.1. Wisconsin Prognostic Breast Cancer (WPBC) Dataset (Set 1).....	57
7.1.2 Wisconsin Diagnostic Breast Cancer (WDBC) Dataset (Set 2)	58
7.2 Analysis of the Data	59
7.2.1 Outline of the Analysis:.....	60
7.2.2 Analysis Procedure:	61
7.2.3 Analysis 1: Finding causal dependencies between attributes.....	62
7.2.4 Analysis 2: Construction of the BBN model	63
7.3 Entering evidence into the network.....	66
7.4 Checking The Accuracy Of The AHP Part Of The Model	67
7.5 Checking the Accuracy of the BBN model	68
7.6 Comment on Results	69
8. CONCLUSIONS.....	70
REFERENCES	R1
APPENDICES	
A: Conditional and Marginal Probability Tables of Hypothetical Problem	A1
B: Evidence Propagation Using Pearl’s Message Passing Algorithm.....	A3
C: Tables and Figures that are not included in Chapter 7	A9

LIST OF TABLES

Table 3.1: An example showing the effect of disease prevalence on the predictive values.....	22
Table 3.2: Calculation of Sensitivity and Specificity	22
Table 4.1: Marginal probabilities of the nodes in the hypothetical problem after evidence 1 and evidence 2 are introduced	43
Table 5.1: Saaty's relative importance scale.....	50
Table 5.2: Random Consistency Index (RI).....	52
Table-7.1: Classification of the correlation coefficient values	62
Table-7.2: Correlation matrix of attributes in Set 2.....	A8
Table-7.3: Correlation matrix of attributes in Set 2 (summarized)	A8
Table-7.4: Correlation matrix of attributes in Set 2 (converted into words).....	A9
Table-7.5: AHP priority matrix for parents of FNA node.....	65
Table-7.6: Priorities of size, shape, texture and smoothness nodes.....	65
Table-7.7: Priorities of concavity and symmetry nodes	65
Table-7.8: Contingency Table	68
Table-7.9: Effect of evidences on prior probabilities of the nodes	69

LIST OF FIGURES

Figure-4.1: Sample factored joint distributions	29
Figure 4.2: An example of a small Bayesian Network: The Asia Network.....	29
Figure 4.3: Asia Network after moralization.....	31
Figure 4.4: The Asia network, triangulated	32
Figure 4.5: The junction-tree for the Asia network	32
Figure 4.6: Equilibrium in the network after evidence introduction	35
Figure 4.7: Downward propagation	35
Figure 4.8: Downward propagation after initiation of variable X to x_1	36
Figure 4.9: (a) A directed cyclic graph that is not singly connected (b) A singly connected network that is not a tree	39
Figure 4.10: BBN of the hypothetical problem generated in MSBN	42
Figure 4.11: Prior and posterior probabilities of the nodes	43
Figure 4.12: Graph of hypothetical example showing prior probabilities and posterior probabilities of the sink node.....	44
Figure 5.1: MCDM Hierarchy	49
Figure-7.1: A magnified image of a breast FNA.....	56
Figure-7.2: Scatter plot identifying the relations between attributes	A9
Figure-7.3: Scatter plot matrix identifying the difference of the attribute values (Set 2) in terms of their type of diagnosis	A10
Figure-7.4: Three dimensional scatter plot showing the dependency of compactness on area and perimeter.....	A10
Figure-7.5: Dependency between concave points and concavity; fractal dimension and symmetry.....	A10
Figure-7.6: BBN of FNA study.....	63
Figure-7.7: CPTs of BBN model.....	A11
Figure-7.8: Marginal probabilities of the nodes after the introduction of evidence 1	66
Figure-7.9: Marginal probabilities of the nodes after the introduction of evidence 2	67
Figure-7.10: Effect of evidences on prior probabilities of the nodes.....	69

CHAPTER 1

1. INTRODUCTION

The complexity of the decisional problems that arise in Medical Sciences requires greater specialization of individuals in this new context and needs the use of approaches more open and flexible than traditional decision-making techniques.

The graphical model techniques that have been developed over the recent years try to help human-minds to understand the world around us, i.e stating the relations between variables, collecting data and processing it to have reasonable results. In order to make inferential statements or decisions concerning the parameters of a given process or population, information is obtained from the process or population. In the light of this sample information, Bayes' Theorem is a tool to revise the probabilities. As another new decision tool, Multi-criteria decision methods enable the consensus in a case of multi-actor decision-making and multiple solution alternatives.

For the cases where both statistical data and expert judgments are present, there is a need for an integrative model that combines the expert's beliefs with the historical data and gives accurate results with acceptable errors. The Fine Needle Aspirate diagnostic test is an example of such a case. For breast cancer tumors, the Fine Needle Aspirate test provides evidence about the tumor's status (benign or malignant) and directly effects the doctor's surgical biopsy decision. However, the researches have shown that the 80 percent of the patients are directed to the surgical biopsy although their tumor is benign. This comes from the fact that there is no standardization for the evaluation of FNA test outcomes with accurate results. FNA image has both qualitative and quantitative (measurable) parameters. The evaluation is realized by the cyto-pathologist generally without interpretation of measurable data.

In this study, we aimed to model the factors of FNA test such that the model outcome has the minimum error compared to the real world. The thesis is divided into two parts. First part provides a concise overview of important definitions, principles and techniques focusing on inference, graphical models and decision-making. First a brief introduction of Bayesian approach to probability and its usage in diagnostic test evaluation are given. The applications appeared in literature are mentioned which use Bayesian probability and multi-criteria decision methods together. Second part gives a detailed application on breast cancer tumors. In this part of the thesis study, the interdependencies between the features of a visual interpretation test and its outcome are found in terms of probabilities as a result of the AHP study, and then they are used as the conditional probabilities to construct the Bayesian Belief Network model for the Fine Needle Aspiration test, which gives the cyto-pathologist an idea about the breast tumors behavior. The accuracy of the generated model is measured in terms of consistency index of the AHP pair-wise comparison matrices and the error between the model result and actual outcome. The compliance of the model is evaluated both for the diagnostic data and prognostic data sets. This integrated approach enabled us to assess the interdependencies between uncertain variables and express them in terms of conditional probabilities.

The methodology proposed in this study is important due to the fact that it combines the expert knowledge with the statistical data by utilizing expert judgment in conditional probability generation process. The conditional probabilities are approximated with an integrated AHP approach. Additionally, this study succeeded to generate an approximate model of Fine Needle Aspiration with an acceptable error and inconsistency, which would help the pathologists for making the hard surgical biopsy decision.

CHAPTER 2

2. BAYESIAN APPROACH

In the Bayesian approach to statistics, all available information is utilized in order to reduce the amount of uncertainty that is present in the decision-making problem. As new information is obtained, it is combined with any previous information to form the basis for statistical procedures. The mechanism used to combine the new information with the previously available information is known as *Bayesian approach to probability*. Bayes' theorem involves the use of probabilities to generate beliefs. The content of this chapter consists of the basics of conditional probability and Bayesian approach as well as the differences between frequentists and Bayesian approach to probability.

2.1 The Conditional Nature of Probability

An important concept in the theory of probability is that of *conditional probability*, which is interested in the probability that one event will occur, given that a particular second event has occurred or will occur.

For instance, one might be interested in the probability that;

- sales of a certain firm's products will go down, given that a rival firm introduces a competing product
- the price of a certain stock will go up, given that taxes remain the same
- breast cancer develops on a woman by age 40 given that she carries a deleterious copy of breast cancer susceptibility gene.

In a sense, all probabilities are conditional upon some assumptions, upon the details of an experiment, upon some action or upon numerous similar factors. For example, if one says that the probability of rain is $1/3$, we should write;

$$P(\text{rain} \mid \text{current atmospheric conditions}) = 1/3$$

Of course, the notion of conditional probability is not restricted to the case in which there is only one “given” event. For example, the probability that a person is involved in an automobile accident in a one-year period, we might be interested in

$$P(\text{accident} \mid \text{age of the person})$$

or

$$P(\text{accident} \mid \text{occupation of the person})$$

or

$$P(\text{accident} \mid \text{number of miles the person drives per year})$$

then the conditional probability of the event becomes:

$$P(\text{accident} \mid \text{age of the person, occupation of the person, and number of miles the person drives per year})$$

Statistical methods, such as significance test and confidence intervals which can be interpreted in terms of the frequency of certain outcomes occurring in hypothetical repeated realizations of the same experimental situations. Under this approach, parameters of interest are obtained from the observed data. Concepts such as hypothesis testing, power calculation, p values, random forest, bootstrap and cross-validation are commonly used by classical statisticians (e.g., frequentists).

Frequentists view data as random variables while Bayesians view the unknown parameters as unknown parameters as random because those quantities are unknowable and are the quantities about which belief statements are needed. (Harrel, and Shieh, 2001).

As it is stated by Draper (2006); two main ways to think about the meaning of probability have been developed:

- *the frequentist (or relative frequency) approach*, in which attention is restricted to the phenomena that are *repeatable* under *identical conditions* (with each repetition logically *independent* of the others) and the probability $P(A)$ of an *event A* is regarded as the long-run relative frequency with which A would occur in the repetitions; and
- *the Bayesian approach*, in which A can be any (true/false) propositions you want (in other words, in this approach attention need not be restricted to

repeatable phenomena) and $P(A)$ is a numerical measure of the *weight of evidence* in favor of the statement that A is true.

Bayesian statisticians, think of the parameters as random variables, from a prior belief of regarding these parameters, and the use of the observed data to update their belief in the posterior distribution. The Bayesian posterior distribution offers a lot more flexibility in the type of evidence one can report and produce results more transparent to interpret. Bayesian applications are especially useful in studies involving multiple endpoints, which are common in clinical studies. (Harrel and Shieh, 2001)

2.2 Some Advantages of Bayesian Methods

The advantages of the Bayesian Methods can be summarized as follows:

In contrast to frequentist methods, Bayesian methods answer the right questions and agree with natural common sense. That is, they give explicit probability distributions for both parameters and future outcomes and revise these probabilities as new evidence becomes available. All probabilities are appropriately conditioned on the observed data, and users can find any probabilities of interest. Examples are the probability that one treatment effect is greater than another, the probabilities of various side effects from a given treatment, or the probability that a particular patient will survive a surgical procedure. (Winkler, 2001)

Important basic principles are consistently followed by Bayesian procedures. Most notable among these is the likelihood principle. In conditioning on the observed data, Bayesian methods ignore the likelihoods of any possible past outcomes that might have but did not occur. Only the likelihoods associated with the outcomes that actually occurred are used. Any probability manipulations, such as the determination of posterior and predictive probabilities, follow the usual rules of probability theory. (Winkler, 2001)

The output of Bayesian methods is ideal for decision-making and therefore for healthcare decision making. Posterior and predictive probabilities represent the uncertainties of interest to decision makers and can be used in calculating any expected values of interest such as expected payoffs, expected losses, or expected utilities. Prior probabilities play an important role in pre-posterior decisions, which are decisions about whether to gather information, how to gather it, and how much

information to gather. Such decisions are very important in settings such as the testing of new drugs or medical procedures (Raiffa and Schlaifer, 1961). Work in the 1950s and early 1960s leading to the recent increase in interest in Bayesian methods was motivated in large part by decision-making considerations.

Bayesian methods force careful thought. They require more inputs than are needed in frequentist procedures, and this generally encourages more careful thought about the model. This is not to say that many frequentist analyses are not done with careful modeling and thought, but as noted above, frequentist procedures are easier to apply mindlessly. Bayesian analyses often heighten awareness of some modeling issues. (Savage, 1954)

In general, Bayesian analyses are more thorough and more transparent. There are fewer formal inputs to a frequentist analysis, which leaves greater leeway for modeling choices that are not always explicitly discussed. The need to specify explicitly the inputs to a Bayesian analysis makes the analysis and any assumptions more transparent to observers and to decision makers for whom the analysis is relevant. (Schlaifer, 1959)

Bayesian methods allow for the formal incorporation of relevant information other than the data immediately at hand. Some actually view this as a weakness of the Bayesian approach, but in important real-world problems it is important to draw on any information that may be available pertaining to the question of interest. Excluding available information as the frequentist approach does is just as much a subjective judgment as including it explicitly and is less defensible. Lilford and Braunholtz (1996) state, “Health issues are now much more complex and the amount of disparate evidence that impacts on belief has increased. Only the Bayesian approach can do justice to all this information and provide the probabilistic basis for action.”

Bayesian techniques lend themselves better to situations with messy data sets and with multiple data sets. Prior information, as expressed through the prior distribution, can serve to identify such models and enable us to differentiate between these different combinations (Bernardo and Smith, 1994).

2.3 Basics of Bayesian Analysis

Three basic components in the Bayesian analysis are the prior distribution, likelihood function, and posterior distribution. The prior distribution describes analysts' belief *a priori*; the likelihood function captures how data modify the prior knowledge; and the posterior distribution synthesizes both prior and likelihood information. The Bayesian approach treats the parameters of interest as random variables, uses the entire posterior distribution to quantify the evidence, and reports evidence in a probabilistic manner.

Frank P. Ramsey in *The Foundations of Mathematics* (1931) first used subjective belief as a way of interpreting probability. Ramsey proposed this interpretation as a complement to the frequency interpretation of probability, which was more established and accepted at the time. The statistician Bruno de Finetti in 1937 adopted Ramsey's view as an alternative to the frequency interpretation of probability. L.J. Savage expanded the idea in *The Foundations of Statistics* (1954).

Formal attempts have been made to define and apply the intuitive notion of a "degree of belief". The most common application is based on betting: a degree of belief is reflected in the odds and stakes that the subject is willing to bet on the proposition at hand.

General form of Bayes' theorem can be stated as Equation-2.1 below. If the J events A_1, A_2, \dots, A_J are mutually exclusive (i.e, no two of the events can both occur; if one of them occurs, none of the other $J-1$ can occur) and collectively exclusive (i.e, one of the events *must* occur; the J events exhaust all of the possible results), and B is another event (the "given" event), then;

$$P(A_j | B) = \frac{P(B | A_j)P(A_j)}{P(B | A_1)P(A_1) + P(B | A_2)P(A_2) + \dots + P(B | A_J)P(A_J)} \quad (2.1)$$

Two distinct features of the Bayesian approach are the use of the prior distribution and the way θ , the parameter (or parameters) of interest, is conceptualized. The likelihood function serves as an intersection between frequentists and Bayesian statisticians; it is a common element in both approaches. Frequentists treat θ as a fixed unknown value and use the observed data to estimate θ . However, Bayesian statisticians treat θ as an unknown random variable. They have a prior belief

regarding θ , update their prior belief with the data observed, and summarize this information in the posterior distribution.

$$\Pr(\theta \mid \text{data}) \propto \Pr(\text{data} \mid \theta) \Pr(\theta) \quad (2.2)$$

The posterior distribution obtained from the Equation 2.2 contains, from left to right, information on both the prior belief of analysts and the likelihood for the observed data, obtained from the experiment or study.

An example for Bayesian Approach used in clinical studies is the GUSTO Study (Global Utilization of Streptokinase and t-PA for Occluded Coronary Arteries), which analysis the short-acting Nifedipine and the dose-related mortality of Nifedipine. 53 patients were enrolled in GUSTO. The purpose of this study was to evaluate if either drug alone, or a combination of both medications, is more effective in the treatment of heart attacks.

The application of Bayesian Analysis can be summarized in five steps as follows:

Step 1 involves determining the loss, called Bayes loss, if the wrong recommendation is made, which could be represented in clinical terms (e.g., lives lost, years of life lost, or quality-adjusted years of life lost), in economic terms (direct or total costs saved), or in combination of both (cost-effectiveness or net health benefits (Stinnett and Mullahy,1998)). Many trials directly measure the outcome of interest for deciding a policy. However, in some instances, the trial may have to rely on a surrogate measure. For example, for trials of drugs to treat hepatitis C, the main outcome measure is viral load, because it would be exceedingly difficult to keep patients in trials lasting many years to show an effect on liver cirrhosis, and thus mortality. In either case, it can be transformed mathematically into a relevant measure of risk relevant to the decision maker.

In **step 2**, the policy maker specifies the prior distribution function before the trial (the results of a Bayes analyses are likely to be viewed more credibly if the prior is defined before trial data are analyzed). For both standard and experimental treatment, the functions, denoted as $p_i(\theta_i)$, give the probability for each possible value of θ_i . If there exists an abundance of prior information (e.g., biological plausibility, epidemiologic studies, prior randomized studies), then the prior distribution function is likely to appear narrower than if there exists little prior information.

At **Step 3**, the statistician estimates the probabilities of observing the outcomes seen in the clinical trial, for any given level of parameter, θ . These represent the likelihood functions, $f_i(\mathbf{x}_i|\theta)$.

Bayes rule is used to estimate the posterior distribution of the θ 's (**step 4**) based on the prior distribution and the likelihood function. For example, in a clinical trial enrolling 200 patients, 100 for standard treatment and 100 for experimental treatment, the survival rate of patients became 45% and 65%, respectively.

The Bayes method permits the decision maker to be more explicit about the risk of making the wrong recommendation. For either the prior or posterior distribution, the expected number of lives lost because of making wrong decision due to uncertainty in the true survival rates can be computed, and is called Bayes risk (**step 5**). The risk of deciding to recommend using the experimental treatment, when not using it is less harmful, occurs when $\theta_1 < \theta_0$. Conversely, the risk of deciding not to recommend using the experimental treatment, when using it is less harmful, occurs when $\theta_1 > \theta_0$. (Hornberger, 2001).

CHAPTER 3

3. MEDICAL DECISION MAKING

Medical decision-making often involves making a diagnosis and selecting appropriate treatment. They help determine what prevention program to promote, what diagnose to make, what tests to order and what treatments to perform. Medical decisions are one of the most important factors determining the cost and quality of medical care. Since in medical decision-making, decisions in the light of information are converted into action; the alternatives, events and outcomes should be clearly defined to prevent wrong decisions. In this chapter, the usage of the test accuracy parameters which are calculated using contingency tables are explained in detail with an example.

3.1 Medical Diagnosis

While choosing one of many alternating test, the decision depends on several factors regarding the test:

- Accuracy of the test
- Availability
- Difficulty in performance
- and cost of the test

Another important consideration in making diagnostic decision is to weigh up how much additional information the test will add to what is already known. The accuracy of tests is reported in terms of their sensitivity, specificity and predictive values. The meaning of these terms should be clearly understood by the doctors not to lead any wrong conclusions. For example, in a district with low disease prevalence, some doctors grossly overestimate the disease probability from a screening test, when the patient has a positive test result. This is due to the fact that, the predictive values of the tests change according to the prevalence of the disease on the population, on which the test has been performed. An example for this situation is given in Section 3.2.2.

3.2 Estimating Test Accuracy Parameters

There are several characteristics that can be used to describe the quality and usefulness of a test. Accuracy is one of these characteristics that can be expressed through sensitivity and specificity, positive and negative predictive values. Each measure of accuracy should be used in combination with its complementary measure.

- Sensitivity complements specificity
- Positive predictive value complements negative predictive value

3.2.1 Sensitivity and Specificity

Sensitivity is the proportion of patients who were positive for the test among all patients with the disease. Specificity is the proportion of patients who were negative for the test among all patients without the disease.

Generally the sensitivity and specificity depend on the cut-off values and may have some trade-off. A more sensitive test may be less specific and a more specific test may be less sensitive, so the decision on what test to request is often not easy. The answer depends on the purpose of doing the test.

For example, a family physician has to decide to rule out the possibility of a treatable disease because the outcome is dangerous, that is, early detection of cervical cancer so that surgical intervention can be done immediately. If the purpose is making sure that the patient does not have cervical neoplasm, then a more sensitive test will be the right choice. Therefore, the physician may request a regular *Pap smear* test, which is more sensitive but not specific to *cervical neoplasm*. However, if the physician has to decide to only recommend treatment for those who really have the disease because the effect of treatment for a non-diseased patient can harm the patient physically, emotionally or financially. In such a case, a more specific test will be the right choice because a very specific such as cervical *biopsy* test is rarely positive in the absence of the disease (Espallardo, 2003).

3.2.2 Positive and Negative Predictive Values

A positive predictive value is the proportion of patients with the disease among all patients who were positive for the test.

A negative predictive value is the proportion of patients who do not have the disease among those patients who were negative for the test.

It gives us the probability of the presence or absence of the disease if the test is positive or negative, respectively.

Predictive values are affected by the prevalence of the disease. For example, A test with 90% sensitivity and 80% specificity in a population that has 30% prevalence of the disease will have a positive predictive value of 66% and a negative predictive value 95%. However, if the same test is applied to an area where the prevalence of a disease is 10%, the positive predictive value becomes 33% and the negative predictive value becomes 99%. See Table 3.1

Table 3.1: An example showing the effect of disease prevalence on the predictive values

	High Prevalence				Low Prevalence		
	Disease	No disease	Total		Disease	No disease	Total
<i>Positive</i>	27	14	41	<i>Positive</i>	9	18	27
<i>Negative</i>	3	56	59	<i>Negative</i>	1	72	73
Total	30	70	100	Total	10	90	100

Table 3.2: Calculation of sensitivity and specificity

	Disease	No disease
<i>Positive</i>	True Positive (TP)	False Positive (FP)
<i>Negative</i>	False Negative (FN)	True Negative (TN)

Using Table 3.2, the prevalence, sensitivity, specificity, positive predictive value and negative predictive value parameters are calculated as follows:

$$\text{Prevalence} = (TP+FN)/(TP+FP+FN+TN)$$

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(FP+TN)$$

Positive Predictive Value = $TP / (TP + FP)$

Negative Predictive Value = $TN / (FN + TN)$

Using these formula for the values given in Table 3.1,

In high prevalence case,

sensitivity= 90%

specificity = 80%

prevalence =30%

Positive Predictive Value = 66%

Negative Predictive Value = 95%

In low prevalence case,

sensitivity= 90%

specificity = 80%

prevalence=10%

Positive Predictive Value = 33%

Negative Predictive Value = 99%

Thus, the prevalence of the disease has crucial effect on the test performance. A diagnostic test that is applied in high prevalence area will have higher positive predictive values when applied to a low prevalence area. The negative predictive values, however, hardly change in such a case. Therefore, the doctor should take this fact into account in order not to make wrong comments about patient's situation.

CHAPTER 4

4. BAYESIAN BELIEF NETWORKS

Bayesian Belief Networks (BBN), which are used to model the uncertainty in a domain able to make reasoning under uncertainty, has grown rapidly over the last few years. A BBN provides a model representation for the joint distribution of a set of variables in terms of conditional and prior probabilities, in which the orientations of the arrows represent influence or causality relations that are observed from data or expert opinion. In this chapter, the idea behind the Bayesian Belief Network propagation is identified, and then an hypothetical medical problem is generated and solved using Pearl's algorithm.

4.1 Definition

Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering – uncertainty and complexity – and in particular they are playing an increasingly important role in design and analysis of machine learning algorithms. Fundamental to the idea of a graphical model is the notion of modularity – a complex system is built by combining simpler parts. Probability theory provides the bond whereby the parts are combined, insuring that the system as a whole is consistent, and providing ways to interface models to data (Jordan, 1999).

A Bayesian Belief Network (Bayes Nets, Bayesian Networks) is a graphical model that encodes probabilistic relationships among variables of interest. Heckerman (2004) stated the advantages of Bayes Nets for data analysis when used in conjunction with statistical techniques:

- Bayesian Networks can handle incomplete data sets. When one of the inputs is not observed, however, most models will produce an inaccurate prediction, because they do not encode the correlation between the input variables. When we use Bayesian networks constructed from statistical data, missing data is processed by EM algorithm.
- Bayesian Networks allow one to learn about causal relationships. Learning about causal relationships are important for at least two reasons. The process is useful when we are trying to gain understanding about a problem domain. In addition, knowledge of causal relationships allows us to make predictions in the presence of interventions. For example, a marketing analyst may want to know whether or not it is worthwhile to increase exposure of a particular advertisement in order to increase the sales of a product. To answer this question, the analyst can determine
 - Whether or not the advertisement is a cause for increased sales, and to what degree. The use of Bayesian Networks helps to answer such questions even when no experiment about the effects of increased exposure is available.
- Bayesian Networks in conjunction with Bayesian statistical techniques facilitate the combination of domain knowledge and data. Anyone who has performed a real-world analysis knows the importance of prior or domain knowledge, especially when data is scarce or expensive. The fact that some commercial systems (i.e., expert systems) can be built from prior knowledge alone is a testament to the power of prior knowledge. Bayesian networks have a causal semantics that makes the encoding of causal prior knowledge particularly straightforward. In addition, Bayesian networks encode the strength of causal relationships with probabilities.

Consequently, prior knowledge and data can be combined with well-suited techniques from Bayesian statistics.

Belief networks are capable of representing the probabilities over any discrete sample space. The probability of any sample point in that space can be computed from the probabilities in the belief network. The key feature of belief networks is their explicit representation of the conditional independence and dependence among events.

Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over fitting of data. There is no need to hold out some of the available data for testing. Using the Bayesian approach, models can be “smoothed” in such a way that all available data can be used for training.

Bayesian probability of an event x is a person’s *degree of belief* in that event. Whereas a classical probability is a physical property of the world (e.g., the probability that a coin will land heads), a Bayesian probability is a probability of the person who assigns the probability (e.g., your degree of belief that the coin will land heads).

One important difference between physical probability and personal probability is that, to measure the latter, we do not need repeated trials. For example, if we imagine the repeated tosses of a sugar cube onto a wet surface, every time the cube is tossed, its dimensions will change slightly. Thus, although the classical statistician has a hard time measuring the probability that the cube will land with a particular face up, the Bayesian simply restricts his or her attention to the next toss, and assigns a probability.

Heckerman (1996) listed the advantages of graphical models when used in conjunction with statistical techniques. Since the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. A Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Since the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over-fitting of data.

Due to the advances in Bayesian algorithms, BBNs have been employed to support decision making in a variety of domains, that include;

- medical diagnosis and advice
- software safety assessment
- system reliability prediction
- system fault analysis
- PC software user-assistance and trouble shooting

- Financial risk and return analysis
- Weapons scheduling
- Agriculture related applications.

Bayesian networks have become increasingly popular in biomedicine and health-care for handling the uncertain knowledge involved in establishing diagnoses of diseases, in selecting optimal treatment alternatives, and predicting treatment outcome in various areas. Bayesian networks are also increasingly developed in areas of health-care that are not directly related to the management of disease in individual patients. Some examples of the use of Bayesian networks in medicine are clinical epidemiology for the construction of disease models and within bioinformatics for the interpretation of micro-array gene expression data.

Bayesian networks have been applied to problems in medical diagnosis (Heckerman 1990; Franklin et al., 1989), map learning (Dean, 1990), language understanding (Charniak and Goldman, 1989; Goldman, 1990), vision (Levitt, Mullin, and Binford 1989), heuristic search (Hansson and Mayer, 1989).

To construct a Bayesian Belief Network, the graph structure or net topology must first be determined. This involves deciding upon the uncertain variables of interest, which provide the nodes and attributing values to these variables, if discrete in nature. This latter activity provides node states. The conditional relationships between the nodes must then be identified and these provide arcs. Finally, the BBN topology must be quantified or populated with node probability values, which describe the nature of uncertainty in the domain (Rajabally et al., 2004).

An independence assumption is made with BBNs; x_i , given its parents π_i , is independent of any other variables except its descendents. Equation 4.1 gives the joint probability distribution of $X = (x_1, \dots, x_n)$, which can be factored out as a product of the conditional distributions in the network:

$$P(X) = \prod_{i=1}^n P(x_i | \pi_i) \quad (4.1)$$

The uncertainty of the interdependence of the variables is represented locally by the conditional probability table (CPT). $P(x_i | \pi_i)$ associated with each node x_i , where π_i is the parent set of x_i .

Following definitions are related to dependency between the nodes in Bayes Nets.

Definition 4.1: (Conditional Independence) Two variables A and C are conditional independent given variable B if the Equation 4.2 holds:

$$P(A | B) = P(A | B, C) \quad (4.2)$$

Definition 4.2: (Marginal Independence) A variable A and a variable C are marginally independent if the Equation 4.3 holds:

$$P(A, C) = P(A) P(C) \quad (4.3)$$

Definition 4.3: (d-Separation) Two distinct variables A and B are d -separated in a causal network if, for all paths between A and B there is a variable V such that the connection is either serial or diverging and variable V is instantiated or the connection is converging and neither V or any of its descendants are instantiated (have received evidence).

4.2 Introduction to Inference in Bayesian Networks

Although it has been proven that inference in BBNs with general directed a-cyclic graph (DAG) structure is NP -hard (deterministic non-polynomial time hard) (Cooper, 1990), probabilistic inference algorithms that are more efficient than the brute force use of gigantic joint probability table have been developed by exploring the interdependency captured by the network structure (Watthayu and Peng, 2004). Most important among them are algorithms for computing posterior probabilities $P(x_i | e)$, where e denotes evidence, the observed values for some variables. These class of algorithms include “belief propagation” by Judea Pearl (1988) and “Junction Tree” by Shafer (1996) and Jensen (1995). Belief Propagation and Junction Tree are methods of *exact solution*. For extremely large BBNs, various statistical sampling techniques (e.g. Markov Chain Monte Carlo (MCMC) sampling) are used which find approximate solutions (Castillo et al., 1997).

A graphical model specifies a complete joint probability distribution (JPD) over all the variables. Given the JPD, we can answer all possible inference queries by marginalization (summing out over irrelevant variables).

Some examples are given in Figure-4.1 below showing the application of Bayes Rule in directed graphs. Their joint probability representations are as follows:

- (a) $P(x_1, x_2, x_3) = P(x_3 | x_1) P(x_2 | x_1) p(x_1)$
- (b) $P(x_1, x_2, x_3) = P(x_3 | x_1) P(x_1 | x_2) p(x_2)$
- (c) $P(x_1, x_2, x_3) = P(x_3 | x_2, x_1) P(x_2) p(x_1)$
- (d) $P(x_1, x_2, x_3) = P(x_2 | x_3, x_1) P(x_3 | x_1) p(x_1)$
- (e) $P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_6 | x_5) P(x_5 | x_3, x_2) P(x_4 | x_2, x_1) P(x_3 | x_1) P(x_2 | x_1) P(x_1)$

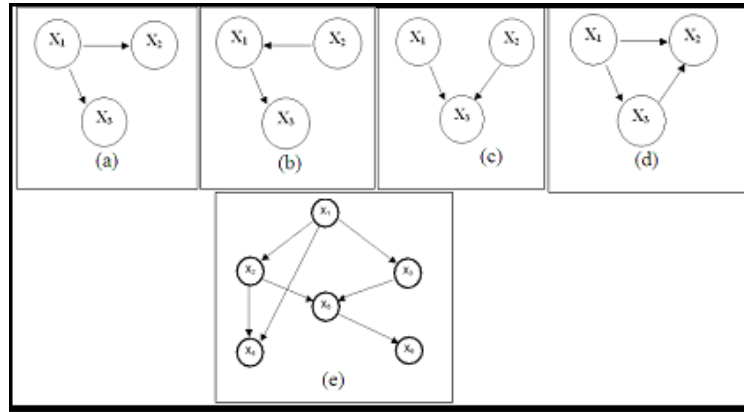


Figure-4.1: Sample factored joint distributions

Figure-4.2 is an example of a small Bayesian Network. (Lauritzen and Spiegelhalter, 1988). The initial conditional probabilities of each node are given for each state of parent nodes.

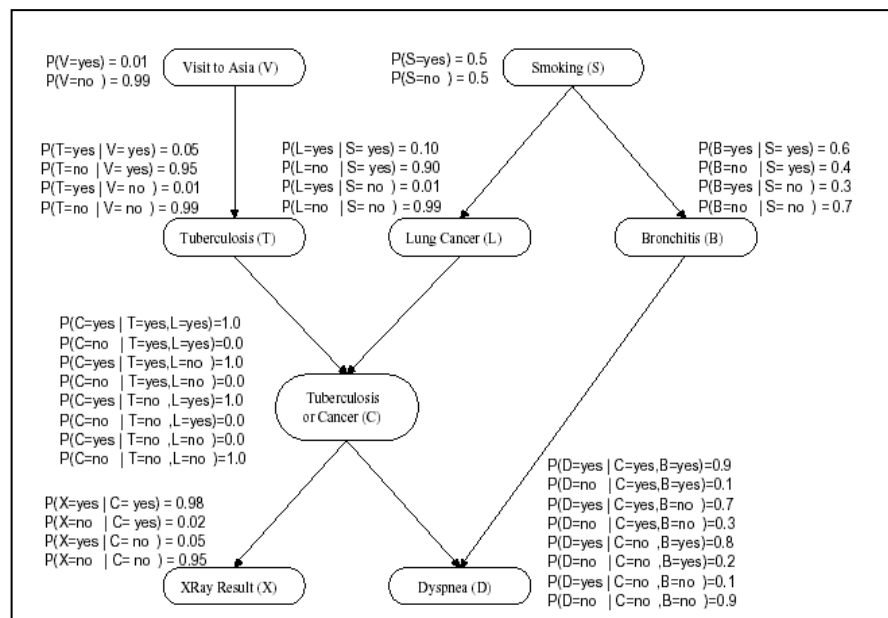


Figure 4.2: An example of a small Bayesian network: The Asia network

4.3 Advanced Inference for Bayesian Networks

Inference in Bayes Nets can be performed either exactly or approximately. Variable Elimination, Pearl's Message Passing and Junction Tree algorithms provide exact inference, whereas Markov Chain Monte Carlo methods provide approximate solution for the node probabilities after evidence.

4.3.1 Variable Elimination

Variable elimination (VE) is a query-based algorithm that formalizes the process used to derive the marginal probability of X -node from the Asia network in Figure 4.3. VE computes a distribution over its query variables by marginalizing other variables from the joint probability one by one.

Variable elimination begins by creating a pool of distributions, which initially contains the CPTs of the Bayesian network. A variable to be marginalized is selected, and all distributions defined over that variable are removed from the pool.

These distributions are multiplied into a single distribution, and the selected variable is marginalized from the resulting distribution. This distribution is then placed in the pool, and the process is repeated, until all non-query variables have been marginalized. The remaining distributions in the pool are combined using multiplication, and the resulting distribution is normalized, giving us the posterior probability over the query variables.

The complexity of the algorithm is $O(n \exp(w_p))$, where n is the number of variables in the Bayesian network, is the ordering in which the variables are eliminated, and w_p is the *induced width* of the variable ordering, which is equivalent to the number of variables in the largest intermediate distribution. The induced width is a function of the variable ordering. Finding an optimal elimination ordering is a hard problem, but with several heuristics giving good approximations to the optimal (Kjaerulff, 1990), (Rose, 1974). The primary advantages of the *VE algorithm* are its simplicity and its dynamic nature. The algorithm is very straightforward to implement, and no recompilation takes place, allowing the algorithm to exploit barren and d-separated variables at runtime. The main disadvantage to VE is that it requires k runs to compute the individual posterior for k variables. Much of the work is repeated for each computation, something that other methods are able to avoid.

There exist several variants to the VE algorithm. *Bucket Elimination* (Dechter, 1996) places the distributions into separate pools (or buckets) according to the domains of the distributions, thus eliminating the need to search for distributions defined over a particular variable when marginalizing. *Mini-buckets* (Dechter, 1997) is an algorithm that computes an approximation to the posterior of the query variables in less time and space than VE. In the mini-bucket algorithm, the set of distributions

of a bucket is partitioned into smaller buckets, and each smaller bucket is processed the same as a standard bucket in Bucket Elimination. This further partitioning typically creates smaller intermediate distributions, which reduces the time and space requirements of the algorithm, at the expense of an exact answer.

4.3.2 Junction Tree Algorithm

Junction-tree propagation (JTP) is a batch update technique that pre-compiles the Bayesian network into a junction-tree. ((Jensen et al., 1990), (Jensen,1996), (Lauritzen and Spiegelhalter, 1988)), Computing over the junction tree allows the posterior probability of each variable to be computed simultaneously and efficiently.

A **junction-tree** is an undirected, acyclic graph derived from the Bayesian network. Each node in the junction-tree, called a cluster, is a subset of the variables from the Bayesian network. The JTP algorithm calculates a joint probability distribution over each cluster in the junction-tree. Once JTP completes, the posterior probability of a variable can be obtained from any cluster containing that variable by marginalizing out all other variables in that cluster, and normalizing the resulting distribution. The clusters of a junction-tree are identified after the Bayesian network is moralized and triangulated. To moralize the Bayesian network, the parents of each variable are married (an edge is placed between any two variables that share a common child and do not already have an edge between them), and the direction of all links are dropped (Figure 4.3). Triangulating a graph ensures that any cycles of length greater than 3 have a chord intersecting them (Figure 4.4). Triangulating a graph is typically done through an elimination procedure, where one variable is eliminated from the graph, and edges are added between the remaining neighbors of the eliminated variable. The triangulated graph is the original graph with these new added edges.

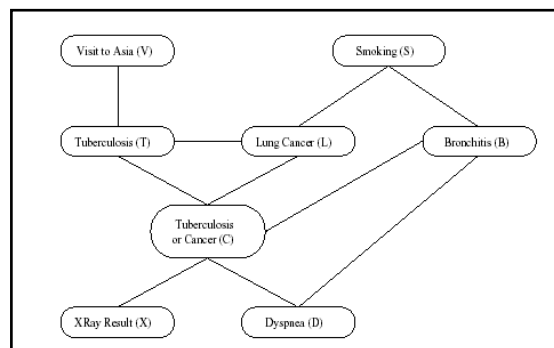


Figure 4.3: Asia Network after moralization. (Marriage between T and L, and between C and B, also the direction of the arcs has been dropped)

Definition 4.1: (Cluster) A cluster Q_i corresponds to a set of variables V_i where V_i can consist of an arbitrary subset of nodes from a BBN.

Definition 4.2: (Junction Trees) A cluster tree is a junction tree (JT) if for every pair of clusters Q_i and Q_j , $Q_i \cap Q_j$ is contained in every cluster on the path ρ between Q_i and Q_j . This is called the *running intersection property*.

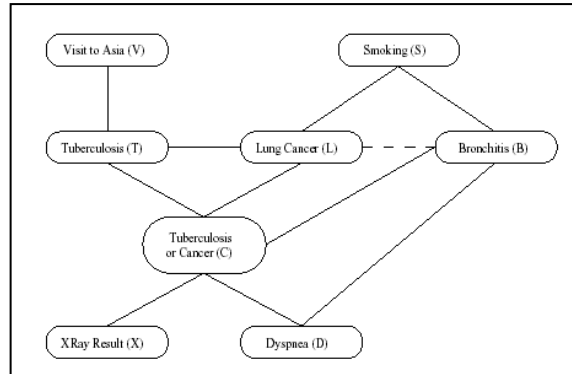


Figure 4.4: The Asia network, triangulated

Each maximal clique in the moralized, triangulated graph contains the variables for a cluster in the junction-tree. Once the clusters of the graph have been identified, the junction-tree can be constructed. A vertex is created for each cluster, and edges are added between the vertices such that 1) the graph is connected, with no loops and 2) the running intersection property is maintained.

If a junction-tree maintains the running intersection property, then if two clusters share a common variable, then all clusters along the path between those two clusters contain the variable as well. Each edge in the junction tree is also labeled with a variable set, known as its separator set. The separator set is just the intersection of the clusters that the edge connects.

Figure 4.5 given below shows a junction-tree for the Asia network.

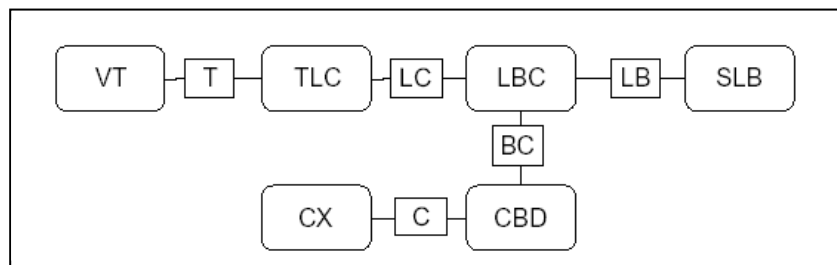


Figure 4.5: The junction-tree for the Asia network. Clusters are shown as rectangles with rounded corners, and separator sets are shown as rectangles with square corners.

Inference in a junction tree proceeds with each node passing messages to each other. These messages take the form of a distribution. One message is passed from each cluster to each of its neighbors. These messages are combined into a final distribution at each node, and the posterior probability for a variable at a cluster can be obtained by marginalizing away all other variables in the cluster. The complexity of inference in junction-trees is $O(n \exp(w))$, where w is the size of the largest clique. The clique sizes depend on the triangulation of the Bayesian network, which in turn depends on the variable ordering used to determine fill-edges. Finding the optimal variable elimination ordering for this problem is NP-hard [Kjaerulff, 1990]. In fact, the problem of finding an optimal variable ordering is the same for both Variable Elimination (VE) method and JTP, so the same heuristics can be applied.

The primary advantage of JTP is that it calculates the individual posterior of each variable simultaneously. That is, after the completion of the algorithm, the posterior of any variable is available from the distribution of any cluster containing that variable. One disadvantage of a junction-tree is that its space requirement is exponential on the size of its largest clique. As well, because the junction-tree is a precompiled structure, it is more difficult to take advantage of barren variables and d-separation.

4.3.2 Pearl's Message Passing Algorithm

Pearl (1988) developed a message-passing algorithm for inference in Bayesian networks. Pearl's message passing algorithm computes the posterior probability distribution of each variable in a Bayesian network in a single run ((Dechter, 1997), (Dechter and Rish, 2003)). The algorithm only computes correct probabilities for a singly-connected network. Hence, the algorithm is typically used in conjunction with a conditioning algorithm that renders the network singly-connected.

During the message passing algorithm, a variable in the Bayesian network becomes a processing unit. The variable receives messages from its neighboring nodes. These messages are in the form of a distribution, representing information from another part of the network. A variable uses these messages to calculate the posterior probability distribution over itself, as well as to calculate messages to send to its neighbors. The message sent to a neighboring variable is a summary of all information received from all other neighbors. The algorithm terminates when all messages have been sent.

The number of messages sent during the message passing algorithm is $2e$, where e is the number of arcs in the network (since a node sends and receives one message from each neighbor). Calculating messages to be sent to parent variables takes $O(\exp(f))$ time, where f is the size of the largest family (calculating a message to be sent to a child can be done in time linear on the size of the variable, once the posterior probability has been calculated, and therefore does not contribute to the complexity). Calculating the posterior probability of a variable from the messages also takes $O(\exp(f))$ time. Hence, the overall time for the algorithm is $O((n + e) \exp(f))$. The space required by the algorithm is $O(n \exp(f))$, for CPT storage (the messages passed are linear in the domain size of the variables, and therefore do not contribute to the space complexity).

The advantage of Pearl's algorithm is its low resource requirements: in terms of complexity, it is among the fastest and smallest inference algorithms for Bayesian networks to date. The algorithm calculates posterior probability distributions for each variable simultaneously, as opposed to a single distribution as in VE. Also, because each variable processes independently, much of the computation can be done in parallel. However, the algorithm works only for singly-connected networks, which in practice occurs infrequently.

Pearl conjectured that running the message passing algorithm in a multiply-connected network (containing undirected loops) might stabilize to an equilibrium, even though the posteriors at equilibrium may not be representative of the real posteriors. Murphy et al. (1999) explored this idea on general probabilistic networks, attempting to ascertain empirically if message-passing was a reasonable approximation approach on "loopy" networks. The results showed that when convergence occurred, the approximations were quite good, outperforming other standard approximation methods given a similar amount of running time. However, the algorithm would exhibit oscillatory behavior over certain networks, and never converge.

In Pearl's Algorithm, the impact of each new piece of evidence is viewed as a perturbation that propagates through the network via message-passing between neighboring variables. The example shown in Figure 4.6 requires five time periods to reach equilibrium after the introduction of data.

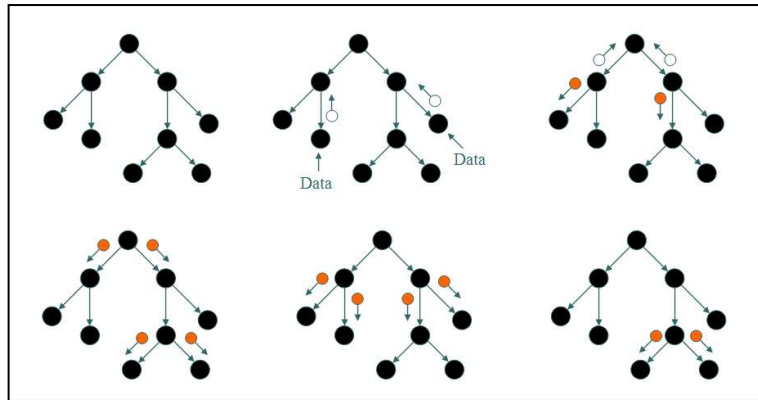


Figure 4.6: Equilibrium in the network after evidence introduction

Examples given in Figure 4.7 given below show the types of propagations in belief networks (Neapolitan,2004).

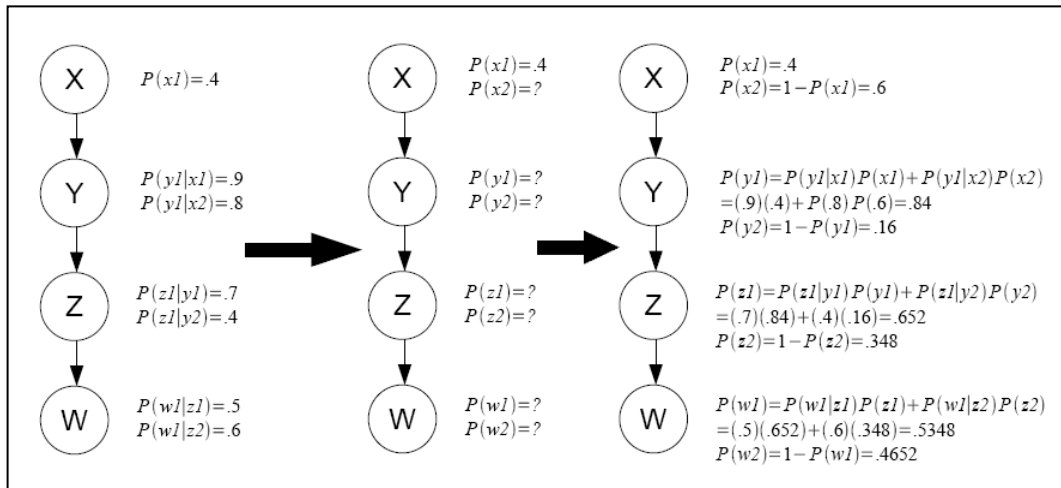


Figure 4.7: Downward propagation

Down propagation:

For example, if variable X is instantiated to x_1 , then

$$\begin{aligned}
 P(y_1) &= P(y_1|x_1) = 0.9 \\
 P(z_1|x_1) &= P(z_1|y_1, x_1)P(y_1|x_1) + P(z_1|y_2, x_1)P(y_2|x_1) \\
 &= P(z_1|y_1)P(y_1|x_1) + P(z_1|y_2)P(y_2|x_1) \\
 &= (0.7)(0.9) + (0.4)(0.1) = 0.67 \\
 P(w_1|x_1) &= P(w_1|z_1, x_1)P(z_1|x_1) + P(w_1|z_2, x_1)P(z_2|x_1) \\
 &= P(w_1|z_1)P(z_1|x_1) + P(w_1|z_2)P(z_2|x_1)
 \end{aligned}$$

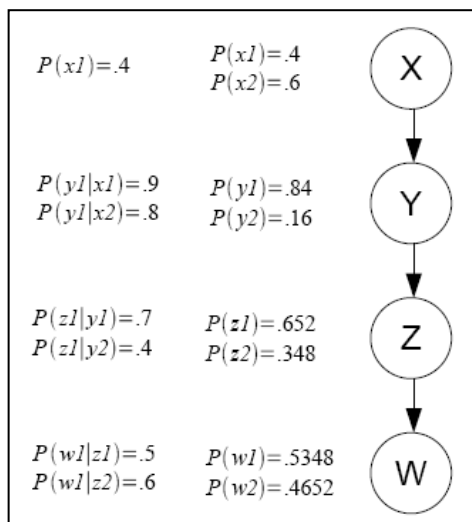


Figure 4.8: Downward propagation after initiation of variable X to x_1

Upward propagation:

For example, if variable W is instantiated to w_1 ,

$$P(z_1|w_1) = \frac{P(w_1|z_1)}{P(w_1)} = \frac{(0.5)(0.652)}{(0.5348)} = 0.6096$$

$$P(w_1|y_1) = P(w_1|z_1, y_1)P(z_1|y_1) + P(w_1|z_2, y_1)P(z_2|y_1) = P(w_1|z_1)P(z_1|y_1) + P(w_1|z_2)P(z_2|y_1) = (0.5)(0.7) + (0.6)(0.3) = 0.53$$

$$P(y_1|w_1) = \frac{P(w_1|y_1)P(y_1)}{P(w_1)} = \frac{P(w_1|y_1)(0.84)}{(0.5348)} = \frac{(0.53)(0.84)}{0.5348} = 0.832$$

$$P(w_1|x_1) = P(w_1|y_1)P(y_1|x_1) + P(w_1|y_2)P(y_2|x_1) = P(w_1|y_1)(0.9) + P(w_1|y_2)(0.1) = [P(w_1|z_1, y_1)P(z_1|y_1) + P(w_1|z_2, y_1)P(z_2|y_1)](0.9) + [P(w_1|z_1, y_2)P(z_1|y_2) + P(w_1|z_2, y_2)P(z_2|y_2)](0.1) = [(0.5)(0.7) + (0.6)(0.3)](0.9) + [(0.5)(0.4) + (0.6)(0.6)](0.1) = 0.533$$

$$P(x_1|w_1) = \frac{P(w_1|x_1)P(x_1)}{P(w_1)} = \frac{(0.533)(0.4)}{(0.5348)} = 0.398654$$

4.3.2.1 Details of Pearl’s Message Passing Algorithm

Given a set a of values of a set A of instantiated variables, the algorithm determines $P(x|a)$ for all values x of each variables X in the network. It accomplishes this by initiating messages from each instantiated variable to its neighbors. These neighbors pass messages to their neighbors. The updating does not depend on the

order in which the messages are initiated, which means the evidence can arrive in any order.

Inference in rooted trees

A rooted tree is a directed acyclic graph, where the root has no parent, every other node has precisely one parent, and every node is a descendent of the root.

The algorithm for rooted trees is based on Theorem-1 (Neapolitan, 2004).

Theorem-1: Let a be a set of values of a subset A of V , for each variable X , the λ messages, λ values, π messages and π values are defined as follows:

1. Define λ messages:

For each child Y of X , for all values of x ,

$$\lambda_Y(x) \equiv \sum_y P(y|x)\lambda(y)$$

2. Define λ values:

If $X \in A$ and X 's value is \hat{x} ,

$$\lambda(\hat{x}) \equiv 1$$

$$\lambda(x) \equiv 0 \quad \text{for } x \neq \hat{x}.$$

If $X \notin A$ and X is a *leaf*, for all values of x ,

$$\lambda(x) \equiv 1$$

If $X \notin A$ and X is a *nonleaf*, for all values of x ,

$$\lambda(x) \equiv \prod_{U \in CH_X} \lambda_U(x), \quad \text{where } CH_X \text{ denotes the set of children of } X.$$

3. Define π messages:

If Z is the parent of X , then for all values of z ,

$$\pi_X(z) \equiv \pi(z) \prod_{U \in CH_Z - \{X\}} \lambda_U(z)$$

4. Define π values:

If $X \in A$ and X 's value is \hat{x} ,

$$\pi(\hat{x}) \equiv 1$$

$$\pi(x) \equiv 0 \quad \text{for } x \neq \hat{x}.$$

If $X \notin A$ and X is the *root*, for all values of x ,

$$\pi(x) \equiv P(x)$$

If $X \notin A$, X is not the *root*, and Z is the parent of X , for all values of x ,

$$\pi(x) \equiv \sum_z P(x|z)\pi_X(z)$$

5. Given the above definitions, for each variable X , we have for all values of x ,

$$P(x|a) = \alpha \lambda(x) \pi(x), \quad \text{where } \alpha \text{ is a normalizing constant.}$$

Pearl presented an algorithm based on Theorem-1. The steps of the algorithm include initialization and updating processes. Updating process includes sending λ messages to parents and π messages to child nodes.

Initial_tree (Bayesian-network:(G,P) where $G=(V,E)$, set of variables: A , set of variable values: a)

```

A =  $\emptyset$ ;  $a = \emptyset$ ;
for (each  $X \in V$ )
  for (each variable  $x$  of  $X$ )
     $\lambda(x)=1$ ; //Compute  $\lambda$  values.
  for (the parent  $Z$  of  $X$ ) //Does nothing if  $X$  is a root.
    for (each value  $z$  of  $Z$ )
       $\lambda_x(z)=1$ ; //Compute  $\lambda$  messages.
for (each value  $r$  of the root  $R$ )
   $P(r|a)=P(r)$ ; //Compute  $P(r|a)$ .
   $\pi(r)=P(r)$ ; //Compute  $R$ 's  $\pi$  values.
  for (each child  $X$  of  $R$ )
    send_  $\pi$ _msg( $R,X$ );

```

Update_tree (Bayesian-network:(G,P) where $G=(V,E)$, set of variables: A , set of variable values: a , variable: V , variable value \hat{v})

```

A =  $A \cup \{V\}$ ;  $a = a \cup \{\hat{v}\}$ ; //Add  $V$  to  $A$ .
 $\lambda(\hat{v})=1$ ;  $\pi(\hat{v})=1$ ;  $P(\hat{v}|a)=1$ ; //Instantiate  $V$  to  $\hat{v}$ .
for (each value of  $v \neq \hat{v}$ )
   $\lambda(v)=0$ ;  $\pi(v)=0$ ;  $P(v|a)=0$ ;
if ( $V$  is not the root and  $V$ 's parent  $Z \in A$ )
  send_  $\lambda$ _msg( $V,Z$ );
for (each child  $X$  of  $V$  such that  $X \notin A$ )
  send_  $\pi$ _msg( $V,X$ );

```

send_ λ _msg(node Y , node X);

```

for (each value of  $x$ )
 $\lambda_Y(x) = \sum_y P(y|x) \lambda(y)$ ; //Y sends X a  $\lambda$  message.
 $\lambda(x) = \prod_{U \in CH_x} \lambda_U(x)$ ; //Compute X's  $\lambda$  values.
 $P(x|a) = a\lambda(x) \pi(x)$ ; // Compute  $P(x|a)$ 
normalize  $P(x|a)$ ;

```

if (X is not the root **and** X 's parent $Z \notin A$)

```

  send_  $\lambda$ _msg( $X,Z$ );
for (each child  $W$  of  $X$  such that  $W \neq Y$  and  $W \notin A$ )
  send_  $\pi$ _msg( $X,W$ );
send_  $\pi$ _msg(node  $Z$ , node  $X$ );
for (each value of  $z$ )

```

$$\pi_X(z) = \pi(z) \prod_{Y \in CH_z - \{X\}} \lambda_Y(z); \quad //Z \text{ sends } X \text{ a } \pi \text{ message.}$$

for (each value of x)

$$\pi(x) = \sum_z P(x|z) \pi_X(z); \quad // \text{Compute } X\text{'s } \pi \text{ values.}$$

$$P(x|a) = \alpha \lambda(x) \pi(x); \quad // \text{Compute } P(x|a)$$

normalize $P(x|a)$;

for (each child Y of X such that $Y \notin A$)

send_ π _msg(X, Y);

Inference in singly connected trees

A directed acyclic graph is called *singly connected* if there is at most one chain between any two nodes. Otherwise, it is called multiply connected.

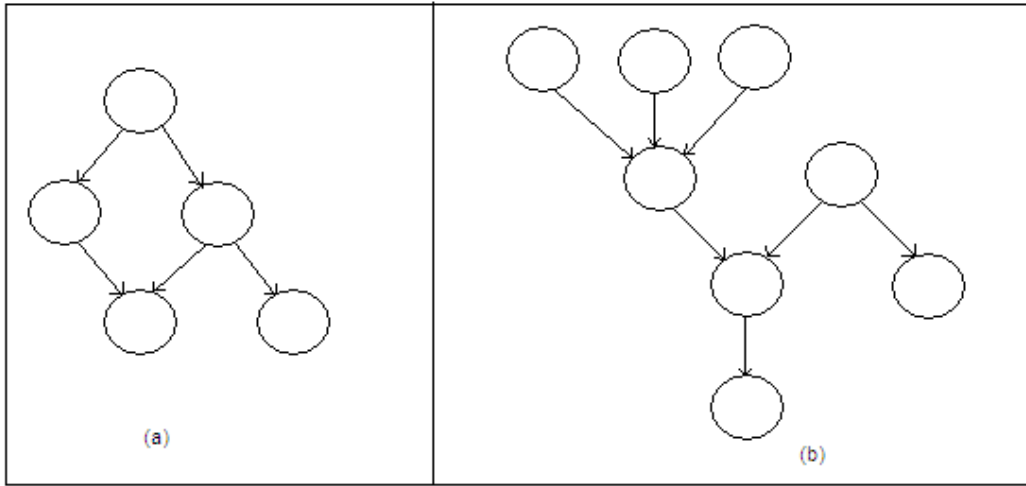


Figure 4.9: (a) A directed cyclic graph that is not singly connected (b) A singly connected network that is not a tree (Neapolitan,2004).

Theorem-2: Let (G,P) be a Bayesian network that is singly connected, where $G=(V,E)$, and let a be a set of values of a subset A of V , for each variable X , the λ messages, λ values, π messages and π values are defined as follows:

1. Define λ messages:

For each child Y of X , for all values of x ,

$$\lambda_Y(x) \equiv \sum_y \left[\sum_{w_1, w_2, \dots, w_k} \left(P(y|x, w_1, w_2, \dots, w_k) \prod_{i=1}^k \pi_Y(w_i) \right) \right] \lambda(y)$$

where W_1, W_2, \dots, W_k are the other parents of Y .

2. Define λ values:

If $X \in A$ and X 's value is \hat{x} ,

$$\lambda(x) \equiv 1$$

$$\lambda(x) \equiv 0 \quad \text{for } x \neq \hat{x}.$$

If $X \notin A$ and X is a *leaf*, for all values of x ,

$$\lambda(x) \equiv 1$$

If $X \notin A$ and X is a *nonleaf*, for all values of x ,

$$\lambda(x) \equiv \prod_{U \in CH_X} \lambda_U(x), \quad \text{where } CH_X \text{ is the set of all children of } X.$$

3. Define π messages:

Let Z be a parent of X , then for all values of z ,

$$\pi_X(z) \equiv \pi(z) \prod_{U \in CH_Z - \{X\}} \lambda_U(z)$$

4. Define π values:

If $X \in A$ and X 's value is \hat{x} ,

$$\pi(x) \equiv 1$$

$$\pi(x) \equiv 0 \quad \text{for } x \neq \hat{x}.$$

If $X \notin A$ and X is the *root*, for all values of x ,

$$\pi(x) \equiv P(x)$$

If $X \notin A$, X is a *nonroot*, and Z_1, Z_2, \dots, Z_j are the parents of X , for all values of x ,

$$\pi(x) \equiv \sum_{x_1, x_2, \dots, x_j} \left(P(x | z_1, z_2, \dots, z_j) \prod_{i=1}^j \pi_X(z_i) \right)$$

5. Given the above definitions, for each variable X , we have for all values of x ,

$$P(x | a) = a \lambda(x) \pi(x), \quad \text{where } a \text{ is a normalizing constant.}$$

Pearl presented an algorithm based on Theorem-2 (Neapolitan, 2004). The steps of the algorithm include initialization and updating processes. Updating process includes sending λ messages to parents and π messages to child nodes.

Initial_tree (Bayesian-network: (G, P) where $G=(V, E)$, set of variables: A , set of variable values: a)

$A = \emptyset; a = \emptyset;$

for (each $X \in V$)

for (each variable x of X)

$\lambda(x) = 1;$ //Compute λ values.

for (the parent Z of X)

 //Does nothing if X is a root.

for (each value z of Z)

$\lambda_X(z) = 1;$ //Compute λ messages.

for (each value r of the root R)

$P(r | a) = P(r);$ //Compute $P(r | a)$.

$\pi(r) = P(r);$ //Compute R 's π values.

for (each child X of R)

$send_ \pi_msg(R, X);$

Update_tree (Bayesian-network:(G,P) where $G=(V,E)$, set of variables: A , set of variable values: a , variable: V , variable value \hat{v})

$A = A \cup \{V\}; a = a \cup \{\hat{v}\};$ //Add V to A .

$\lambda(\hat{v})=1; \pi(\hat{v})=1; P(\hat{v} | a)=1;$ //Instantiate V to \hat{v} .

for (each value of $v \neq \hat{v}$)

$\lambda(v)=0; \pi(v)=0; P(v | a)=0;$

if (V is not the root and V 's parent $Z \in A$)

$send_ \lambda_msg(V,Z);$

for (each child X of V such that $X \notin A$)

$send_ \pi_msg(V,X);$

send_ λ _msg(node Y , node X);

for (each value of x)

$\lambda_Y(x) = \sum_y P(y|x) \lambda(y);$ // Y sends X a λ message.

$\lambda(x) = \prod_{U \in CH_x} \lambda_U(x);$ //Compute X 's λ values.

$P(x | a) = \alpha \lambda(x) \pi(x);$ // Compute $P(x | a)$

normalize $P(x | a);$

if (X is not the root **and** X 's parent $Z \notin A$)

$send_ \lambda_msg(X,Z);$

for (each child W of X such that $W \neq Y$ and $W \notin A$)

$send_ \pi_msg(X,W);$

send_ π _msg(node Z , node X);

for (each value of z)

$\pi_X(z) = \pi(z) \prod_{Y \in CH_Z - \{X\}} \lambda_Y(z);$ // Z sends X a π message.

if ($X \notin A$) {

for (each value of x) {

$\pi(x) = \sum_{z_1, z_2, \dots, z_j} P(x | z_1, z_2, \dots, z_j) \prod_{i=1}^j \pi_X(z_i);$ //the Z_i 's are Compute X 's

parents.

$P(x | a) = \alpha \lambda(x) \pi(x);$ // Compute X 's π values.

}

normalize $P(x | a);$

// Compute $P(x | a)$.

for (each child Y of X)

$send_ \pi_msg(X,Y);$

}

if not ($\lambda(x) = 1$ for all values of x) // Do not send λ messages to X 's other

for (each parent W of X such //parents if X and all of X 's

that $W \neq Z$ and $W \notin A$) //descendents are uninstantiated.

```

    send_λ_msg(X,W);
}

```

Remark: It is important that the comment in module *send_π_message* says “do not send λ messages to X 's other parents if X and all of X 's descendents are uninstantiated.” The reason is that, if X and all X 's descendents are uninstantiated, X d-separates each of its parents from every other parent. Clearly, if X and all X 's descendents are uninstantiated, then all X 's λ values are still equal to 1.

4.3.2.2 An Hypothetical Example solved with Pearl's Algorithm

An hypothetical medical decision example is generated to visualize the updating procedure of Pearl's message passing algorithm. In the network model, there exist tests, which give information regarding the possible diseases on the patient. The results of the tests influence the choice of the treatment. Also the treatments have various dangerous effects on the health of patient's baby. The doctor will give a decision regarding the continuation of the pregnancy.

I. Generation of the Network

An hypothetical model is generated, which has the diseases D1,D2 and D3; tests T1, T2, T3, T4, T5, treatments TR1, TR2 and TR3. The patient is pregnant and has to make a decision about ending or continuing the pregnancy due to the bad effects of the treatments to the child. PR node at the bottom of the network corresponds to this decision. Figure-4.10 below shows the Bayes Network of the problem.

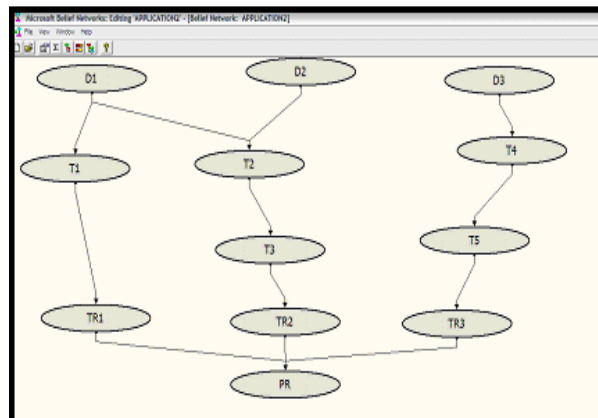


Figure 4.10: BBN of the hypothetical problem generated in MSBN

The conditional and marginal probability tables of all the nodes in the network are given in **Appendix-A**.

II. Entering Evidence

Evidence 1: Suppose T5 is made on the patient and the result is found as negative. Then the probabilities of other nodes change as follows:

Evidence 2: Being the T5 test result negative on hand, another information is obtained regarding node D1. Suppose that a genetic history of the patient's mother about Disease 1 strongly claims that the patient has Disease 1.

The values of the node probabilities after evidence introduction are given in Table-4.1 and the change is shown in Figure-4.11.

Table 4.1: Marginal probabilities of the nodes in the hypothetical problem after evidence 1 and evidence 2 are introduced.

	initial probabilities		after evidence T5=negative		after evidence D1=disease	
D1	0.005	0.995	0.005	0.995	1	0
D2	0.003	0.997	0.003	0.997	0.003	0.997
D3	0.001	0.999	0.0007	0.9993	0.0007	0.9993
PR	0.2733	0.7267	0.2099	0.7901	0.4389	0.5611
T1	0.1441	0.8559	0.1441	0.8559	0.95	0.05
T2	0.1147	0.8853	0.1147	0.8853	0.6709	0.3291
T3	0.3993	0.6007	0.3993	0.6007	0.6385	0.3615
T4	0.2703	0.7297	0.0676	0.9324	0.0676	0.9324
T5	0.28	0.72	0	1	0	1
TR1	0.1549	0.8451	0.1549	0.8451	0.574	0.426
TR2	0.4495	0.5505	0.4495	0.5505	0.6408	0.3592
TR3	0.2852	0.7148	0.12	0.88	0.12	0.88

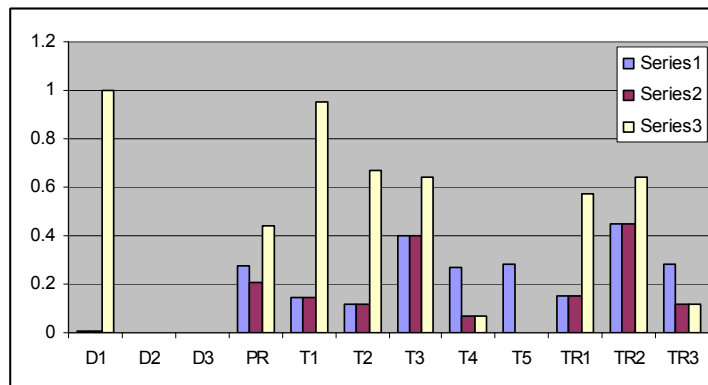


Figure 4.11: Prior and posterior probabilities of the nodes (Series1: prior, Series2: 1st posterior, Series3: 2nd posterior)

III. Evidence Propagation using Pearl's Algorithm

The detailed solution made using this algorithm is given step-wise in **Appendix-B**.

IV. Comment on Results

For this example application, the decision maker should make a decision about continuing or ending pregnancy according to the treatments to be applied on her, since the treatments will have dangerous effects on child.

Before making any test on the woman, according to the prior probabilities of the nodes gathered from population data, the woman should continue the pregnancy being **72.67%** sure.

After the first evidence is obtained and entered into the network, the probability of continuing pregnancy goes up to **79.01%**, since a negative result of Test 5 is a sign for having no disease 3.

However, after a second evidence is obtained which gives information about the presence of disease 1, this results a decrease in the probability of continuing the pregnancy: the probability of continuing the pregnancy in the light of two evidence goes down to **56.19%**. See Figure-4.12 below.

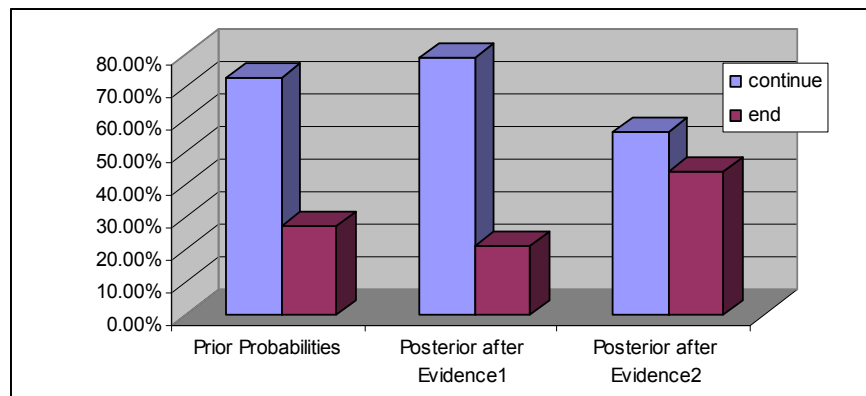


Figure 4.12: Graph of hypothetical example showing prior probabilities and posterior probabilities of the sink node (PR) after evidence propagations

CHAPTER 5

5. MULTI-CRITERIA DECISION ANALYSIS

Bayesian Belief Networks, Multiple criteria decision analysis (MCDA) is a field, which has seen a considerable development during the last ten years. Researchers and practitioners are now more aware of the presence of multiple criteria in real-life problems of management and decision, whatever their nature. The links cannot be ignored which exists between multi-criteria decision analysis and other fields of research such as the theory of social choice, voting procedures, decision in a context of uncertainty, the theory of fuzzy sets, negotiation and expert systems.

5.1 Literature Review

A detailed analysis of the theoretical foundations of different MCDA methods and their comparative strengths and weaknesses is presented in Belton and Stewart (2002). Roy and Vanderpooten (1996) defined four different categories of problems, for which MCDA may be useful:

- **The choice problems** to make a simple choice from a set of alternatives.
- **The sorting problems** to sort actions into classes or categories, such as “definitely acceptable“, “possibly acceptable but needing more information”, and “definitely unacceptable”.
- **The ranking problems** to place actions in some form of preference ordering which might not necessarily be complete.
- **The description problems** to describe actions and their consequences in a formalized and systematic manner, so that decision makers can evaluate these actions.

Belton and Stewart (2002) added the two categories **the design problems** and **the portfolio problems**.

Use of MCDA in Healthcare and Medical Diagnosis can be summarized as follows:

- Lee et al. (2003) mentioned about the difficulty of healthcare decisions. Making medical decisions are difficult because they are complex and have important consequences such as the impact on survival or quality-of-life of individuals and on allocation of limited resources.
- Akhavi and Hayes (2003) compared MCDA methods and concluded that AHP resulted in lower variance in the assessments produced by different decision makers due to the richer information collected by the AHP method.
- Carter et al. (1999) compared three decision-making techniques (Analytical Network Process, AHP and Markov Process) using a common clinical problem: the evaluation of the optimal post-lumpectomy treatment strategy for an elderly woman with a mammographically detected, nonpalpable early-stage breast cancer. The treatment alternatives considered were: observation, radiation, tamoxifen, combination radiation and tamoxifen, and simple mastectomy.
- Dolan (1989) used AHP to determine which of seven recommended antibiotic regimens represented optimal initial therapy for a young woman hospitalized for treatment of acute pyelonephritis. The model included the following criteria: maximize cure, minimize adverse effects (broken down into very serious, serious, and limited), minimize antibiotic resistance, and minimize cost (divided into total cost and patient cost). Alternatives were compared relative to the criteria using published information on the expected frequencies of urinary pathogens and drug toxicity, local antibiotic sensitivities and antibiotic charges, and expert opinion regarding their propensities for inducing anti-microbial resistance.
- Saaty and Vargas (1998) used AHP to find conditional probabilities of symptoms given diseases. In case expert judgment is present, it is possible to combine judgment with statistical data to identify the disease that best describes the observed symptoms. They used priorities obtained from pair-wise comparisons of symptoms given diseases, and used them as conditional probabilities inside Supermatrix. AHP dealt with dependence among the elements or clusters of a decision structure to combine statistical and judgmental information. It is shown that the posterior probabilities derived from Bayes theorem, which is a sufficient condition of a solution in the sense of the AHP.

5.2 An Overview of MCDA Methods

MCDA aims to give the decision-maker some tools in order to enable him to advance in solving a decision problem where several – often contradictory- points of view must be taken into account.

MCDA is divided into four families:

(1) multiple attribute utility theory

The first family, of American inspiration, consists in aggregating the different points of view into a unique function which must subsequently be optimized. The work related this family studies the mathematical conditions of aggregation, the particular forms of the aggregating function and the construction methods.

(2) outranking methods

The second family, of French inspiration, aims first to build a relation, called an outranking relation, which represents the decision-maker's strongly established preferences, given the information at hand. The latter relation is therefore neither complete nor transitive. ELECTRE, MELCHIOR, trichotomic segmentation and PROMETHEE methods are the most representative and most commonly cited outranking methods.

(3) interactive methods

The third and most recent family proposes methods which alternate calculation steps (yielding successive compromise solutions) and dialogue steps (sources of extra information on the decision-maker's preferences). Though they are most recently developed in the frame of multiple objective mathematical programming, some of these methods can be applied to more general cases.

(4) Eigenvalue Methods

These are based on calculation of the eigenvector of the largest modulus of a matrix constructed after pair wise comparisons of the criteria. Though the

method was first published in a previous article by Klee (1971), namely DARE method, the most well-known work is by Saaty (1980), as Analytical Hierarchy Process (AHP).

5.3 Definitions

Some definitions regarding MCDA are given below:

Definition 5.1:(Set of Actions) The *set of actions*, denoted by A , is the set of objectives, decisions or candidates to be explored during the decision procedure. It may be defined by:

- listing its members when it is finite and sufficiently small for an enumeration to be possible
- stating the properties which characterize its elements when it is infinite or finite but too large for an enumeration to be possible.

Examples:

=> Choosing where to build a new factory between 10 possible locations: Set of actions A is defined by listing its elements.

=> Product-mix problem:

A company manufactures plastic boards with properties of flexibility, resistance, weight, color, etc. determined by customers. These properties depend upon amounts x_1, x_2, \dots, x_n of the different components used in manufacturing the plastic. A procedure must be set up in order to satisfy customers as far as possible. In this case, A is the set of vectors (x_1, x_2, \dots, x_n) yielding a plastic which satisfies the requirements determined by the customers: it is infinite and can be described by the mathematical constraints which translate the physical and chemical properties of the mixture resulting from the components involved. A is an evolutive set since the constraints and the components of the mixture vary from one customer to another.

The definition of set A does not only depend upon the problem to be solved and the actors involved in the decision procedure; it also strongly interacts with the steps: defining criteria, modeling preferences, stating the problem and choosing the MCDA method to be applied.

When a decision-maker must compare two actions a and b , he/she will react in one of three following ways:

- (1) preference for one of them
- (2) indifference between them
- (3) refusal or inability to compare them

Definition 5.2: (MCDM problem) A multi-criteria decision problem is a situation in which, having defined a set A of actions and a consistent family F of criteria on A , one wishes

- (1) to determine a subset of actions considered to be the best with respect to F (choice problem)
- (2) to divide A into subsets according to some norms (sorting problem)
- (3) to rank the actions of A from best to worst (ranking problem)

It will in fact frequently happen that a real-life problem gives rise to a mixture of choice, sorting and ranking problems. The same real-life problem may imply:

- different definitions of A
- different definitions of F
- different statements of the problem (choice, sorting or ranking)

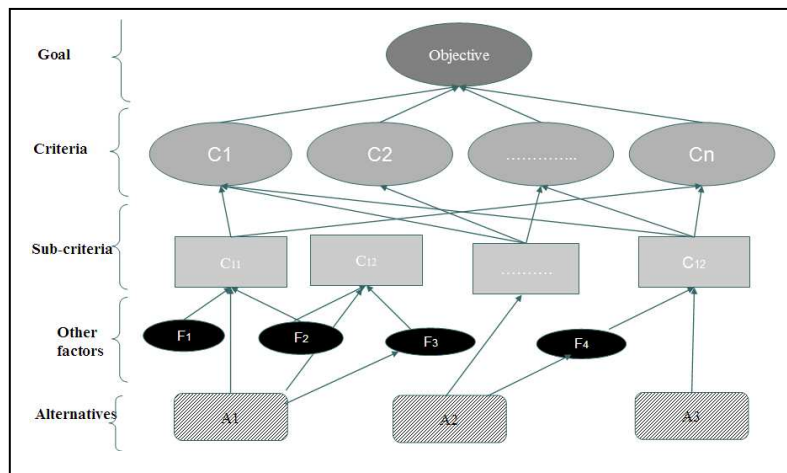


Figure 5.1: MCDM Hierarchy

5.4 Definition of Analytical Hierarchy Method (AHP)

The Analytical Hierarchy Process is one of the most extended multicriteria decision making techniques. It was proposed by Thomas L. Saaty in the mid 1970s and combines tangible and intangible aspects to obtain, in ratio scale, the priorities

associated with the alternatives of the problem. AHP methodology consists of four steps (Saaty,1980):

- (1) **Modelization**, i.e. establishing a hierarchical representation of the problem, which should include all the relevant aspects of the decision problem.
- (2) **Valuation**, in which the decision maker incorporates his judgments through pair-wise comparisons between the elements in the problem taken into consideration. He/she makes her comparisons using Saaty's relative importance scale shown in Table-5.1.

Table 5.1: Saaty's relative importance scale

value of a_{ij}	When criterion i compared with j is:
1	equally important
3	slightly more important
5	strongly more important
7	demonstrably more important
9	absolutely more important

The intermediate values 2,4,6 and 8 can also be used if necessary. If criterion I is neither greater than nor equal to j , a_{ji} is first evaluated as previously and we then write $a_{ij} = 1 / a_{ji}$.

- (3) **Priorization** where the *local priorities* are obtained by using any of the existing prioritization procedures (the eigenvector method – EGVM- and the row geometric mean method- are the most widely used)

W denoting the comparison matrix,

$$W = (w_{ij}) = (w_i / w_j)$$

$$W \cdot w = \begin{bmatrix} w_1 / w_1 & w_1 / w_2 & \dots & w_1 / w_n \\ w_2 / w_1 & w_2 / w_2 & \dots & w_2 / w_n \\ \dots & \dots & \dots & \dots \\ w_n / w_1 & w_n / w_2 & \dots & w_n / w_n \end{bmatrix} x = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{bmatrix} = nw$$

where n is an eigenvalue of W and that w is an associated eigenvector.

The normalized eigenvector of the matrix is also called the priority vector. For a symmetric matrix A , the eigenvalue is the root of the characteristic equation

$$(A - \lambda I) \cdot x = 0, \tag{5.1}$$

where I is the identity matrix and x is the eigenvector of matrix A .

The characteristic equation can be solved if the determinant of the first term is equal to zero;

$$\det(A-\lambda I)=0. \tag{5.2}$$

(4) **Synthesis**, in which the *total priorities* are derived.

One of the main characteristics of AHP is the existence of a measure to evaluate the inconsistency of the decision maker when eliciting his judgments [Aguaron et al., 2003].

The standard method to calculate the values for the weights from an AHP-matrix is to take the eigenvector corresponding to the largest eigenvalue of the matrix, and then to normalize the sum of the components.

5.5 An Example of AHP

Supposing following 3x3 reciprocal matrix constructed as a result of the paired comparisons regarding cost-comfort-duration preference of a Decision Maker (DM) who is going to choose one of the transportation alternatives. According to Saaty's relative importance scale (Pomerol and Romero, 2000), the DM shows the following preference matrix:

$$A = \begin{matrix} & \overbrace{\begin{matrix} \text{cost} & \text{comfort} & \text{duration} \end{matrix}} & \\ \begin{matrix} \text{cost} \\ \text{comfort} \\ \text{duration} \end{matrix} & \begin{bmatrix} 1 & 1/3 & 5 \\ 3 & 1 & 7 \\ 1/5 & 1/7 & 1 \end{bmatrix} & \end{matrix} \xrightarrow{\text{Sum of columns}} A = \begin{bmatrix} 1 & 1/3 & 5 \\ 3 & 1 & 7 \\ 1/5 & 1/7 & 1 \end{bmatrix}$$

sum 21/5 31/21 13

Then each element of the matrix is divided by the sum of its column to normalize the relative weights. The normalized principle eigenvector can be obtained by averaging across the rows:

$$w = \begin{matrix} & \begin{bmatrix} 5/21 & 7/31 & 5/13 \\ 15/21 & 21/31 & 7/13 \\ 1/21 & 3/31 & 1/13 \end{bmatrix} & \end{matrix} \xrightarrow{\text{principle Eigenvector}} w = \frac{1}{3} \begin{bmatrix} 5/21 + 7/31 + 5/13 \\ 15/21 + 21/31 + 7/13 \\ 1/21 + 3/31 + 1/13 \end{bmatrix} = \begin{bmatrix} 0.2828 \\ 0.6434 \\ 0.0738 \end{bmatrix}$$

sum 1 1 1

The normalized principal eigenvector is also called **priority vector**. Since it is normalized, the sum of all elements in priority vector is 1. The priority vector shows relative weights among the things that we compare. In our example above, Cost is 28.28%, Comfort is 64.34% and Duration is 7.38%. In this case, we know more than their ranking. In fact, the relative weight is a ratio scale that we can divide among

them. For example, we can say that for the DM comfort is 2.27 (=64.34/28.28) times more important than cost. Aside from the relative weight, we can also check the consistency of the DM's answer, for which the Principle eigenvalue is used. It is obtained from the summation of products between each element of eigenvector and the sum of the columns of the reciprocal matrix.

$$\lambda_{\max} = \frac{21}{5}(0.2828) + \frac{31}{21}(0.6434) + 13(0.0738) = 3.0967$$

A comparison matrix A is said to be consistent if $a_{ij} a_{jk} = a_{ik}$ for all i, j and k . Saaty proved that for consistent reciprocal matrix, the largest eigenvalue is equal to the number of comparisons, or $\lambda_{\max} = n$. He gave a measure of consistency, called Consistency Index as deviation or degree of consistency using the following Equation 5.3:

$$CI = \frac{\lambda_{\max} - n}{n - 1} \quad (5.3)$$

Then, for the above example,

$$CI = \frac{3.097 - 3}{2} = 0.0484$$

Saaty proposed an appropriate index called Random Consistency Index (RI) to comparing the CI value found. He randomly generated reciprocal matrix using scale $1/9, 1/8, \dots, 1, \dots, 8, 9$ which is similar to the idea of Bootstrap and get the random consistency index to see if it is about 10% or less. The average random consistency index of sample size 100 matrices is shown in the table below:

Table 5.2: Random Consistency Index (RI)

n	1	2	3	4	5	6	7	8	9	10
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

Then, he proposed what is called Consistency Ratio, which is a comparison between Consistency Index and Random Consistency Index and given in Equation 5.4:

$$CR = \frac{CI}{RI} \quad (5.4)$$

If the value of Consistency Ratio is smaller or equal to 10%, the consistency is acceptable. If the Consistency Ratio is greater than 10%, the subjective judgments need to be revised. For the example above, RI for $n=3$ is 0.58, which equates the CR to $8.3\% < 10\%$. Thus, the DM's subjective evaluation about his cost-comfort-duration preference is consistent. Since AHP method includes expert knowledge inside the pair-wise comparison matrix, requires less time for solution and gives more accurate results, in the application of this thesis AHP will be used.

CHAPTER 6

6. INTEGRATING BAYESIAN BELIEF NETWORKS AND ANALYTICAL HIERARCHY PROCESS

Bayesian Belief Networks, which are used in a range of real applications concerned with predicting properties of critical systems, are extremely powerful technique under uncertainty. Although they provide important support for decision-making, in many cases one has to make decisions based on multiple criteria. As it is the application in this thesis, a BBN for predicting the occurrence probability of a disease for an individual cannot be used to make a decision about which one of the alternating diagnostic tests to choose. In such situations, the BBN must be complemented by other decision-making techniques such as multi-criteria decision aid (Fenton and Neil, 2000).

BBN enables us to calculate a probability for each criterion for a given alternative. Then the MCDA methods can be applied to combine the values for a given action and rank them. (Fenton and Neil, 2000)

In literature, Saaty and Vargas (1998) used expert judgments to generate conditional probabilities. They combined Bayesian approach with AHP, where they only converted expert's beliefs into probabilities. In their model, there were diseases and symptoms (diagnostic tests) related to (actually caused by) these diseases. For a selected disease and a pair of symptoms and they asked the following question to the doctor:

- *of the pair, which symptom is more characteristic of that disease and how strongly more when compared with the other?*

Actually, for the cases, where statistical data is present, it is possible to combine judgments with statistical data to identify the disease that best describes the observed symptoms. In this thesis, a generated decision making procedure is

introduced that uses BBNs and MCDA in a complementary way. The procedure consists of defining the variables and their states, and then observing the relations between variables. Marginal probabilities of the variables are calculated from the actual data with a frequentist approach. At the conditional probability generation step, first the pairwise comparisons generated the priorities, then these priorities are utilized to generate beliefs, which is actually the Bayesian approach to probability.

In MCDA, the criteria are evaluated separately as if they were independent of each other. To model the complex and uncertain interactions between criteria as well as between criteria and other factors, BBNs can be used since it takes the interdependencies between variables into account. Interdependencies among these nodes can be qualitatively modeled by the arcs in the diagram and quantitatively by CPTs regarding each chance node.

In this framework, a decision problem can be represented by a graphical model, where each decision node represents the set of alternatives, the utility node represents the set of objectives (the DM's preferences), decision criteria and internal and external factors that may affect the criteria are represented by chance nodes.

To sum up, what makes the probability generation procedure of this thesis different is the fact that the conditional probability generation step utilizes the occurrence frequencies (probabilities), in addition to priority weights. The probabilities are then multiplied with priority weights to compute the **beliefs** of the doctor for the nodes of that variable.

CHAPTER 7

7. AN APPLICATION IN MEDICAL DIAGNOSIS

The problem studied in this application is about breast cancer diagnosis. The analysis includes network model generation, initialization and evidence insertion steps. The accuracy of the suggested model is checked and compared to the actual (observed) results obtained from surgical biopsy.

7.1 Information about the Application Data

Despite the public awareness and scientific research, breast cancer continues to be the most common cancer and the second largest cause of cancer deaths among women (Marshall, 1993). Approximately 12% of women are diagnosed with breast cancer (Muier et al., 1987) and 3.5 % die of it¹.

Most breast cancers are detected by the patient as a lump in the breast. The majority of breast lumps are benign, so it is the physician's responsibility to diagnose breast cancer, that is to distinguish benign lumps from malignant ones (Mangasarian et al., 1994). Masses in the breast are a diagnostic dilemma in clinical medicine. Because of the possibility of cancer, a majority of breast masses are exercised usually with surgical biopsy for accurate diagnosis. However most of these surgical biopsy results (80%) are benign (Aitken, 1990). In other words, surgery has a high false-positive rate, and many women are subject to unnecessary surgery although their mass is benign. So, it would be useful to define clinical, radiologic and cytologic parameters more accurately to determine who should undergo surgery (i.e. surgical biopsy).

There are three available methods for diagnosing breast cancer:

1. Mammography
2. FNA (Fine Needle Aspirate) with visual interpretation

¹ National Center for Health Statistics, GPO. Vital Statistics of US, Mortality, volume 2, 1990.

3. Surgical Biopsy

The following sensitivity values of these diagnostic tests are given in Mangasarian et al. (1994):

1. Mammography: varies from 68% - 79%
2. FNA with visual interpretation: varies from 65% to 98% due to the visual interpretation
3. Surgical Biopsy: close to 100%

It is generally known that;

- mammography lacks sensitivity
- sensitivity of FNA varies widely
- surgical biopsy gives accurate result, however it is invasive, more time consuming, expensive, traumatic than FNA. Also skin incision is necessary for surgical biopsy, where immediate diagnosis is not possible.

Treatment for breast cancer generally consists of surgery (either mastectomy or lumpectomy), followed by radiation therapy, chemotherapy, and/or hormonal therapy. In this thesis, the data used for the application are gathered from Wisconsin Diagnostic Breast Cancer Database. The data respond to the features, which are computed from the **digitized image** of breast masses obtained as the result of **FNA study**. An FNA is taken from the breast mass. This material is then mounted on a microscope slide and stained to highlight the cellular nuclei. A portion of the slide in which the cells are well-differentiated is then scanned using a digital camera and a frame-grabber board, as shown in Figure-7.1.

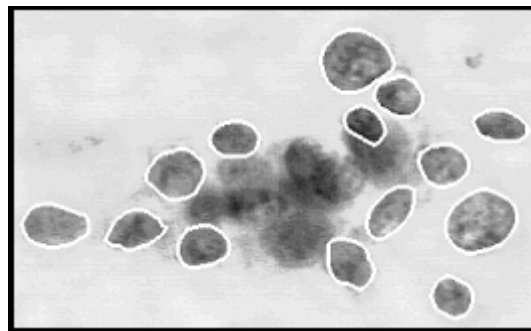


Figure-7.1: A magnified image of a breast FNA

Two different groups of data are recorded after FNA study. First group is the prognostic data set, and second group is the diagnostic data set. Prognostic data gives information about progress of the disease in long-term period. Therefore it will be used to calculate evidence information. Diagnostic data on the other hand will

be used to construct and compute initial network. Both data set have the same attributes (features found after FNA study). The data sets also include the actual state of the patient's tumor (Benign or Malignant), which is found after surgical biopsy study. Prognostic data includes 2 additional attributes (tumor size and lymph node status). The details of the data groups are given below.

7.1.1. Wisconsin Prognostic Breast Cancer (WPBC) Dataset (Set 1)

The source of data (completed on December, 1995) belong to Dr. William H. Wolberg from Clinical Sciences Center-General Surgery Dept. of University of Wisconsin, W. Nick Street from Computer Sciences Dept. University of Wisconsin and Olvi L. Mangasarian from Computer Sciences Dept. University of Wisconsin. Various versions of this data have been used in Street et al. (1995), Mangasarian et al. (1995), Wolberg et al. (1995).

There are 198 records (instances) and 15 attributes in the data set. Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer. In the data set, the outcome is divided into two classes: a number of 151 patients not recurred and 47 occurred. There exist missing valued attributes (Lymph node status is missing in 4 cases).

Attribute (features) information

The features are computed from a digitized image of a **fine needle aspirate (FNA) of a breast mass**. They describe **the characteristics of the cell nuclei present in the image**. The attributes are:

- 1) **ID number** (of the patient)
- 2) **Outcome** (R = recurrence, N = no recurrence)
- 3) **Time** (recurrence time if field 2 = R, disease-free time if field 2= N)

Features computed for each cell nucleus are given from (4) to (15). Features (4) to (12) are nuclear features, whereas (14) and (15) are surgical biopsy parameters. In the network model, only the outcome will be used from surgical biopsy parameters, since (14) and (15) are not comparable to FNA parameters.

- 4) **radius** (mean of distances from center to points on the perimeter): is computed by averaging the length of radial line segments from the center of the nuclear mass to each of the points of the nuclear border.

- 5) **texture** (standard deviation of gray-scale values): is measure by finding the variance of the gray scale intensities in the component pixels.
- 6) **perimeter**: is measured as the distance around the nuclear border.
- 7) **area**: is measured by counting the number of pixels in the interior of the nuclear border and adding one-half of the pixels on the perimeter.
- 8) **smoothness** (local variation in radius lengths): is quantified by measuring the difference between the length of each radius and the mean length of adjacent radii.
- 9) **compactness**: perimeter and area are combined to give the measure of the compactness ($\text{perimeter}^2/\text{area}$)
- 10) **concavity** (severity of concave portions of the contour): is determined by measuring the size of any indentations in the nuclear border.
- 11) **concave points** (number of concave portions of the contour): count the number of points on the nuclear border that lie on an indentation.
- 12) **symmetry**: is measured by finding the relative difference in length between line segments perpendicular to and on either side of the major axis.
- 13) **fractal dimension**: is approximated using the “coast-line approximation” which measures nuclear border irregularity.
- 14) **Tumor size** (diameter of the excised tumor in centimeters)
- 15) **Lymph node status** (number of positive axillaries lymph nodes observed at time of surgery)

7.1.2 Wisconsin Diagnostic Breast Cancer (WDBC) Dataset (Set 2)

The source of Information is same as Set 1 and it is completed on November 1995. It consists of 569 cases; 357 of them are benign, 212 cases are found as malignant after the surgical biopsy. There are no missing values in this data set.

Various versions of this data have been used in Wolberg et al. (1995), Street et al. (1993) and Mangasarian et al. (1994).

Diagnostic features of the digital images for samples can be classified in three groups (Wolberg, et al., 1999):

1. nuclear **size features**: radius, area perimeter
2. nuclear **shape features**: smoothness, compactness, concavity, concave points, symmetry, fractal dimension.
3. other features: (**not nuclear**) lymph node status, tumor size.

Area is expressed as square micro meters (μm^2), and for Radius and Perimeter as micro meters (μm). Values for remaining features are dimensionless.

In the data set, there types of parameter are defined for each feature:

1. mean value
2. standard error
3. worst value (mean of the three largest values)

Attribute information

- 1) ID number (of the patient)
- 2) Diagnosis (**M** = malignant, **B** = benign)

Features computed for each cell nucleus are given from (3) to (12):

- 3) radius
- 4) texture
- 5) perimeter
- 6) area
- 7) smoothness
- 8) compactness
- 9) concavity
- 10) concave points
- 11) symmetry
- 12) fractal dimension

7.2 Analysis of the Data

In order to avoid misinterpretation of the dataset, we have taken information from pathologist specialized on cytology. Prof. Şevket Ruacan and his assistant Dr. Sevgen Önder from Pathology Department of Hacettepe University Medicine Faculty. They made following comments on FNA study and on the data set to be used in this thesis:

- FNA study is performed prior to the surgical biopsy. According to the parameters found in FNA study, a decision is made about applying biopsy. If FNA parameters show a tendency to malignant tumor, then surgical biopsy will be performed to get exact solution. However, it is invasive for the woman, costly, time consuming and dangerous for some cases. Therefore it is important which decision is given in the light of the FNA study.
- Usually pathologist do not prefer morphometric study (means measuring the values of quantitative variables) due to the time, cost and effort constraints. Instead they prefer to use their experience to interpret the image in a visual way. According to Dr. Onder's comments, some of the important features of the

tumor image, which make the doctor absolutely suspicious about the malignancy are

- the contrast (texture value) of the tumor compared to the adjacent cells
- the uniformity of cell shape
- the distance between this type of suspicious cells
- He also added that a peak in one variable being others normal can be a sign to malignancy.
- Shape of the tumor is more important than its size. Also contrast and smoothness are more important than size.
- Worst values (mean of the three largest values) of the nuclear characteristics are more important and effective than mean and standard error parameters.
- Area, radius, perimeter, texture, smoothness, compactness, concavity, concave points, symmetry and fractal dimensions are nuclear-characteristics of the image.
- Tumor size and lymph node status are non-nuclear characteristics which can only be observed after surgery.

Moreover, he added that the final diagnosis comes after the surgical biopsy. So, the outcome (malignant or benign) given in the data set is not the outcome of FNA study, they are actual disease status gathered after surgery.

Since worst values (average of the largest three values) are more meaningful for the diagnosis than mean values and standard error of each attribute, only **worst values** data will be used in this thesis to decrease the variables of interest.

Both the statistical and the RSA machine learning analyses performed Wittekind, et al. (1987) demonstrate that computer-derived nuclear features are prognostically more important than are the classical prognostic features; tumor size and lymph node status. Also classical prognostic features are not comparable to the diagnostic ones. Due to these reasons, tumor size and lymph node status are not used in the network model generated.

7.2.1 Outline of the Analysis:

The aim of the data analysis is to find the causal relations between the variables in order to construct the Bayesian Belief Network. After learning the networks structure, we will be able to enter evidences regarding additional information. After the relations between variables are found, their priorities are found with Saaty's

AHP method. The priorities will be used to find the probabilities of hidden nodes representing each variable group. Finally, the probabilities of the FNA study (malignant or benign) will be compared to the surgical biopsy results found according to the same attribute values.

7.2.2 Analysis Procedure:

Diagnostic Data set (**Set 1**) consists of 10 attributes {three different value for each attribute: mean, standard error, worst value (mean of three largest)} of FNA and the diagnosis found by surgical biopsy.

Prognostic Data set (**Set 2**) includes the same 10 attributes as diagnostic, however it also includes the tumor size and lymph node status, which are after surgery parameters (Tumor size and lymph node status parameters will not be used in the study). Prognostic means the behavior of the disease in long term period. Therefore, prognostic data enables the doctor to predict the direction or long term behavior of the tumors. We will use the prognostic data to state the evidence information.

The problem is to compute conditional probabilities of the nodes. In order to find the conditional probabilities, the values of the nodes should be converted into binary variables, which represent whether the parameter has a serious value for the malignancy or not. After the parameters are converted into binary variables, we will be able to count their frequency, then multiply with priorities of parameters to find the conditional probabilities. The marginal probability of the sink node is calculated using conditional probabilities of that node and the marginal probabilities of the parent nodes.

To determine the critical value, which separate the serious value from non-serious one, we took the opinion of Dr. Onder about 3rd Quantile. According to his comments, a value bigger than 75% of the values on the sample can be used as threshold value. So, for the parameters in Set 1 and Set 2, 3rd quantile will be used as threshold values. In literature, threshold values of the parameters included in a scaled diagnostic breast cancer data set of Wisconsin University (Set 3) have been determined with *data clustering and discretisation using Chi² technique* (Hoffman et al., 2001).

Set 2 will be used to construct the network model. Set 1 will be used both to check the accuracy of the model and compute the evidences (prognostic information), which will be entered into the model.

The outcome variables included in both Set 1 and Set 2 are *not the outcome of the FNA study*. They show the **exact outcome of the surgical biopsy**. Therefore, the marginal probability of the sink node of the network model (**FNA study result** with states “suspect for biopsy”, “no suspect for biopsy”) will be compared with exact biopsy result to check the accuracy of the model. This accuracy comparison will be performed two times: 1.for diagnostic data, 2.for prognostic data.

7.2.3 Analysis 1: Finding causal dependencies between attributes

The purpose is to state the dependency relations between features, which will be utilized to construct Bayes Network. Dependencies between the variables are measured in terms of scatter plots and correlation coefficients and regression.

Correlation, also called correlation coefficient, indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. There are several coefficients, measuring the degree of correlation, adapted to the nature of data. A number of different coefficients are used for different situations. The best known is the *Pearson product-moment correlation coefficient*, which is obtained by dividing the covariance of the two variables by the product of their standard deviations.

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y}$$

Several authors have offered guidelines for the interpretation of a correlation coefficient. Cohen (1988) for example, has suggested the following interpretations for correlations in psychological research, in the table below.

Table-7.1: Classification of the correlation coefficient values

Correlation	Negative	Positive
Small	-0.29 to -0.10	0.10 to 0.29
Medium	-0.49 to -0.30	0.30 to 0.49
Large	-1.00 to -0.50	0.50 to 1.00

Correlation matrix given in Table-7.2, is summarized in Table-7.3 and converted into verbal values in Table-7.4. Two versions of **Scatter Plots** are drawn for the attributes in Set 2; Figure-7.2 states the general relation between variables, while Figure-7.3 shows the effect of outcome (benign, malignant) on this relation.

A general comment can be made looking at graph in Figure-7.3 for this effect. As the attribute values increase, the tendency to malignant tumor increases also. This is because the malignant (green points) cases are generally cumulated on the right upper part and benign (red points) cases are located on the left lower part of the graphs.

7.2.4 Analysis 2: Construction of the BBN model

Perimeter and area are calculated using radius parameter. However, they are not calculated with an exact formula, since the image of the lump includes infinitely many radiuses due to its non-circular (irregular) shape. We can say that the area and perimeter variables are d-separated. Compactness is independent of radius given its parents perimeter and area. In the light of Figures 7.2 to 7.5, correlation tables and according to the comments of the Cyto-pathologist, the network shown in Figure-7.6 is constructed.

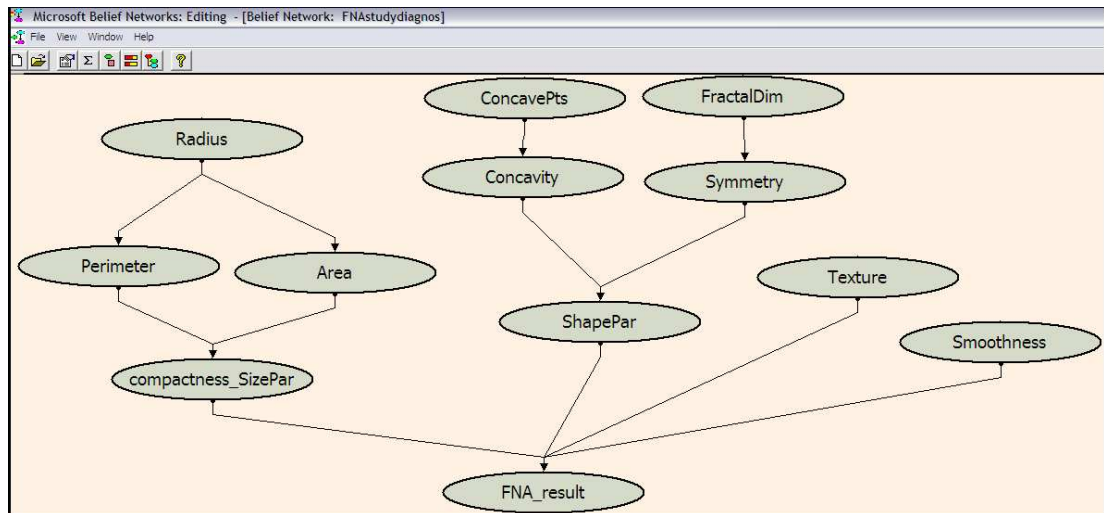


Figure-7.6: BBN of FNA study

The objective of this study is to find an approximation of FNA study prediction, which is close to the exact biopsy outcome.

As an initial step, the values of variables are converted into binary values using 3rd quartile as threshold value (i.e. “1” is assigned to the values bigger than threshold

and “0” to the values less than the threshold). As it is expected, all the nodes without parents have an initial marginal probability of 0.25.

- $P(\text{radius}=\text{serious})=0.25$ $P(\text{radius}=\text{not serious})=0.75$
- $P(\text{fractal_dimension}=\text{serious})=0.25$ $P(\text{fractal_dimension}=\text{not serious})=0.75$
- $P(\text{concave pts}=\text{serious})=0.25$ $P(\text{concave pts}=\text{not serious})=0.75$
- $P(\text{smoothness}=\text{serious})=0.25$ $P(\text{smoothness}=\text{not serious})=0.75$
- $P(\text{texture}=\text{serious})=0.25$ $P(\text{texture}=\text{not serious})=0.75$

We have already learned the conditional probabilities of real nodes from the data. That is, the conditional probabilities for area, perimeter, compactness, concavity and symmetry are found from frequencies from data-set which provides joint probabilities:

$$P(\text{Symmetry} = \text{serious} | \text{fractal_d} = \text{serious}) = \frac{P(\text{Symmetry} = \text{serious}, \text{fractal_d} = \text{serious})}{P(\text{fractal_d} = \text{serious})}$$

$$P(\text{Symmetry} = \text{serious} | \text{fractal_d} = \text{not serious}) = \frac{P(\text{Symmetry} = \text{serious}, \text{fractal_d} = \text{not serious})}{P(\text{fractal_d} = \text{not serious})}$$

$$P(\text{Symmetry} = \text{serious}) = P(\text{Symmetry} = \text{serious} | \text{fractal_d} = \text{serious})P(\text{fractal_d} = \text{serious}) + P(\text{Symmetry} = \text{serious} | \text{fractal_d} = \text{not serious})P(\text{fractal_d} = \text{not serious})$$

Likewise, the conditional probabilities of other nodes are calculated as:

- $P(\text{Concavity} = \text{serious} | \text{concave_points} = \text{serious}) = 0.86014$
- $P(\text{Concavity} = \text{serious} | \text{concave_points} = \text{not serious}) = 0.046948$
- $P(\text{Perimeter} = \text{serious} | \text{radius} = \text{serious}) = 0.979167$
- $P(\text{Perimeter} = \text{serious} | \text{radius} = \text{not serious}) = 0.004706$
- $P(\text{Area} = \text{serious} | \text{radius} = \text{serious}) = 0.986111$
- $P(\text{Area} = \text{serious} | \text{radius} = \text{not serious}) = 0.002353$

However, the node Shape_parameters and FNA_study_result are generated (hidden) variables. Shape_parameters node is generated to summarize and represent the concavity, concave points, symmetry and fractal dimension features in a single variable. Similarly, FNA result node is a hidden node representing the outcome of the FNA study, which is effected from shape, size, texture and smoothness

parameters. The conditional probabilities of these hidden nodes are calculated using AHP weights of their parent nodes.

Calculation of AHP weights

FNA node: For the FNA study result being suspicious, following question is asked to pathologist to construct AHP matrix:

- of the pair, which feature of the tumor is more characteristic of the FNA study outcome and how strongly more when compared with the other?

The comments of the pathologist are then converted into the AHP matrix as shown in Table-7.5:

Table-7.5: AHP priority matrix for parents of FNA node

pairwise comparisons	size	Shape	texture	smoothness
size	1	0.5	0.11	0.11
shape	2	1	0.20	0.33
texture	9	5	1	2
smoothness	9	3	0.5	1

Then the pair-wise comparison matrix is normalized, and the priority weights are found for the attributes as shown in Table-7.6.

Table-7.6: Priorities of size, shape, texture and smoothness nodes

pairwise comparisons	size	shape	texture	smoothness		pairwise comparisons	size	shape	texture	smoothness		principle eigen vector	
size	1	0.5	0.11	0.11	➔	size	0.05	0.05	0.06	0.03	➔	size	0.05
shape	2	1	0.20	0.33		shape	0.10	0.11	0.11	0.10		shape	0.10
texture	9	5	1	2		texture	0.43	0.53	0.55	0.58		texture	0.52
smoothness	9	3	0.5	1		smoothness	0.43	0.32	0.28	0.29		smoothness	0.33

Shape_parameters node: According to the pair-wise comparison matrix given in Table-7.7, the weights of concavity and symmetry are found 67% and 33%, respectively.

Table-7.7: Priorities of concavity and symmetry nodes

	concavity	symmetry		principle eigenvector
concavity	1	2	➔	0.67
symmetry	0.5	1		0.33

The probabilities of their parents are multiplied with priorities of their parent nodes to find the conditional probability of the hidden child node. Since a Bayesian

probability is a probability of the person who assigns the probability, the weighted probability represents the **belief** of the decision maker.

We need marginal probabilities of real nodes to calculate the conditional probabilities of the hidden nodes. The dataset provides the joint probabilities for real nodes. After the marginal probabilities of the parent nodes of *Shape_parameter* are calculated, AHP weights are multiplied to generate conditional probability of *Shape_parameter*.

FNA study belief is found by summing the multiplications of the probabilities of size, shape, texture and smoothness parameters with their priorities, respectively as shown in Figure-7.7.

7.3 Entering evidence into the network

Prognostic data is used to find evidences, since they show the long term behavior of the tumor.

Evidence 1: For the case $P(\text{FNA_result}=\text{suspicious for biopsy})=1$, this evidence effects other nodes as follows

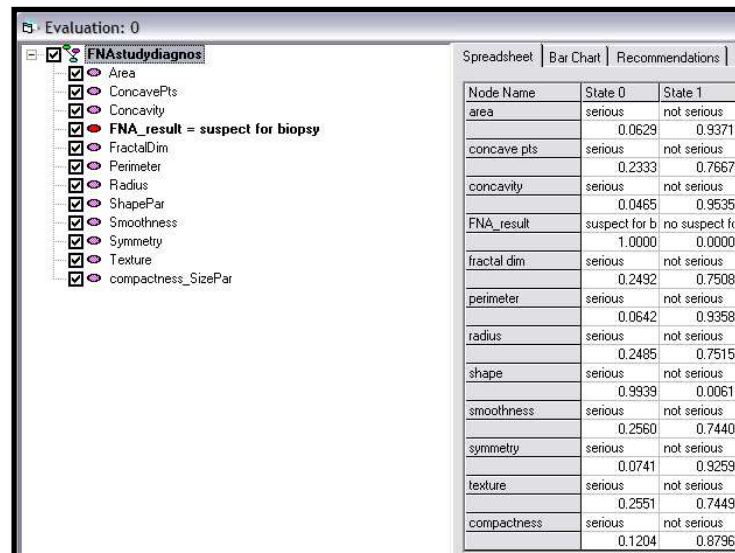


Figure-7.8: Marginal probabilities of the nodes after the introduction of evidence 1

Evidence 2: Being evidence 1 on hand, another evidence is obtained saying concave points attribute is serious, then the probabilities of the nodes change as follows

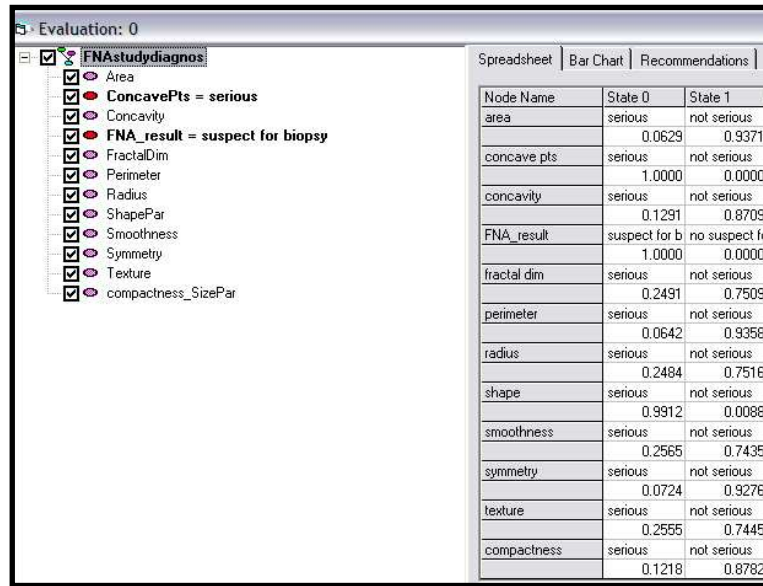


Figure-7.9: Marginal probabilities of the nodes after the introduction of evidence 2

7.4 Checking The Accuracy Of The AHP Part Of The Model

The accuracy of the model generated can be assessed in terms of the consistency of the decision maker and in terms of the validity of the results found. The pair-wise comparisons which formed the AHP comparison matrix should be checked in terms of consistency. As mentioned in the definitions part of the thesis, consistency index (CI) and the consistency ratio (CR) help us to check this.

First, the λ_{max} values for all four tumor characteristics (Table-7.6) are computed as follows:

For $n = 4$, $RI=0.9$

$$\left. \begin{array}{l} \text{Consistency Index: } CI = \frac{\lambda_{max} - n}{n - 1} \\ \text{Consistency Ratio } CR = \frac{CI}{RI} \end{array} \right\} \begin{array}{|c|c|c|} \hline \lambda_{max} & CI & CR \\ \hline 4.060018 & 0.020006 & 0.022229 \\ \hline \end{array}$$

Since the CR value is 2% and less than 10%, the DM's subjective evaluation about his symptoms priorities on the diseases can be said to have a high consistency. The AHP pair-wise comparison matrix generated for shape parameters node (Table-7.7) is a 2x2 matrix and completely consistent.

7.5 Checking the Accuracy of the BBN model

The accuracy of the model and the priorities can be assessed by comparing the estimated outcome of FNA study result with the exact outcome of biopsy. Since we have the biopsy results in the dataset for each individual (1 or 0), we are able to calculate the probability of the tumor being malignant (or benign) from real frequencies.

- ✓ **For diagnostic data**, using the frequencies, the probability of outcome is found as $P(\text{surgical biopsy result is malignant}) = \mathbf{0.372583}$

Network model computed the probability of being suspicious about to FNA test result as:

$P(\text{FNA_study result is suspicious about malignancy}) = \mathbf{0.387916}$

In case the priority weights assigned for parent nodes of FNA are:

- size: 4.85%
- shape: 10.19%
- texture: 52.19%
- smoothness: 32.77%

For diagnostic data: the percentage error between observed and estimated outcomes is **4.12%**, which is small enough to consider the model accurate.

- ✓ Using same priority values for the parents of FNA, the FNA outcome estimated from the same network model for prognostic data is: $P(\text{FNA_study result is suspicious about malignancy}) = \mathbf{0.228967}$

- ✓ Using prognostic frequencies, the probability of outcome is found as $P(\text{surgical biopsy result is malignant}) = \mathbf{0.237374}$

For prognostic data: The percentage error between observed and estimated outcomes is **3.54%**, which is also small enough to consider the model accurate. This result is parallel to the diagnostic data set.

The **contingency table** of the proposed model shown below (Table 7.8) enables us to calculate the *sensitivity*, *specificity*, *positive and negative predictive values* of the FNA model as 82%, 94%, 91% and 88%, respectively. These values show the accuracy of the test outcomes for individuals.

Table-7.8: Contingency table of the test

	Surgical biopsy = +	Surgical biopsy = -	Total
Model result = +	201	20	221
Model result = -	43	305	348
Total	357	212	569

7.6 Comment on Results

In this application we have tried to modeled the FNA test result and observe the changes in the outcome probabilities of this test in the light of new information. Initially, the estimated prior probabilities of the FNA outcome (about malignancy) was **0.388**. The effects of evidence 1 and evidence 2 on other nodes is shown in Table-7.1 and Figure-7.14

Table-7.8: Effect of evidences on prior probabilities of the nodes

	Prior Probability	Posterior after evidence 1	Posterior after evidence 2
area	0.0639	0.0629	0.0629
concave points	0.2500	0.2333	1.0000
concavity	0.0850	0.0465	0.1291
FNA_result	0.3879	1.0000	1.0000
fractal dimensions	0.2500	0.2492	0.2491
perimeter	0.0651	0.0642	0.0642
radius	0.2500	0.2485	0.2484
shape	0.8344	0.9939	0.9912
smoothness	0.2500	0.2560	0.2565
symmetry	0.1225	0.0741	0.0724
texture	0.2500	0.2551	0.2555
compactness	0.1007	0.1204	0.1218

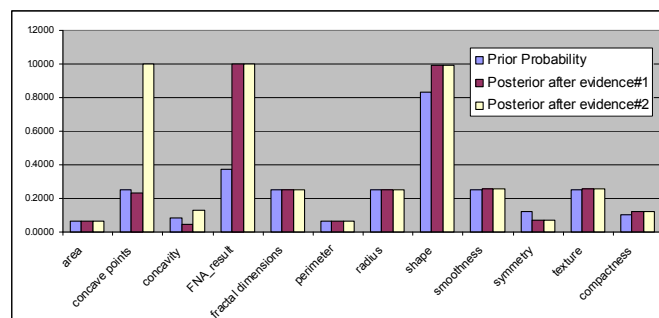


Figure-7.10: Effect of evidences on prior probabilities of the nodes

The accuracy parameters of the FNA test calculated according to the proposed method are high and good enough to consider the method as an applicable alternative to the visual interpretation. Our study showed a very high and comparable sensitivity, specificity, positive and negative predictive values. The most important accuracy parameter is the rate of the people for whom the FNA outcome gave negative result although their surgical biopsy were positive. Here, our approximated test provided a ratio of %7. Also, the error between estimated and actual result is small for both the prognostic and the diagnostic cases. only results given in the contingency table In summary,

CHAPTER 8

8. CONCLUSIONS

Judgments are needed in medical diagnosis to determine which treatments to perform given certain symptoms. These judgments belong to the decision-makers, here the pathologists, physicians and cytologists, who decide the procedure to be followed, which is usually critical for the life of the patient. At the decision making step, using a single criterion to justify a decision offers an advantage by simplifying the problem. However, this does not guarantee that the problem is well formulated with respect to the reality. In order to state the problem nearest to the real situation, multiple criteria should be taken into account.

This thesis study is aimed to introduce the usage of Graphical Models especially Bayesian Belief Networks in Medical Diagnosis problems. To make the idea behind the belief network structure clearer, first a brief introduction of Bayesian approach to probability and its usage in diagnostic test evaluation are given. The applications appeared in the literature are mentioned which use Bayesian probability and multi-criteria decision methods together. In the application part of the thesis study, the interdependencies between the features of a visual interpretation test Fine Needle Aspirate and its outcome are found in terms of probabilities as a result of AHP study, and then they are used as the conditional probabilities to construct the Bayesian Belief Network model for the Fine Needle Aspiration test, which gives the cyto-pathologist an idea about the breast tumors behavior. This integrated approach enabled us to assess the interdependencies between uncertain variables. Consistency index is found for the result to evaluate the inconsistency of the Decision Maker for his answers. Evidence information originated from prognostic data of the same test, is added to the model to calculate posterior probabilities. The approach we used in this thesis to generate conditional probabilities seems visible and applicable due to its small error compared to real observation. However, the model constructed in the application had limitations at modeling reality and

evidence calculation steps. The reason for preferring to use AHP as the MCDA method is that AHP provides pair-wise comparisons of the factors, which is easier for a doctor to comment. Also the usage of AHP in medical decision making and the comparisons with other MCDA methods in the literature show the superiority of AHP method.

In a BBN model, we observe the casualty effect of the nodes on each other; however the important thing is that they only affect each other, this means: there may be other factors having a causal effect on a node but not included in the model. In the model, there are 13 cases (patients), which had “0” in all diagnostic variables but have a positive biopsy result. This is an evidence for the fact that there are factors other than the observed morphometric characteristics explaining the tumor’s malignancy. Actually, according to the decision makers, they are the morphologic characteristics of the tumors, which are not measurable and determined by the subjective assessments of the doctor.

As it is already mentioned in Chapter-7, the prognostic data provides evidence information. From the prognostic data set, we observed dramatic increases in the probability of FNA and concave points node, which are 57% and 68% respectively. For simplicity and due to the MSBN software limitations, the evidence are converted into “1”. That is, MSBN software has the option to enter evidences in 1 or 0 formats. Using a more flexible BBN software, the actual prognostic evidences can be entered into the model.

As a further application of this thesis, a study can be performed using all the attributes affecting the outcome. In this application we only preferred to use worst values of the attributes, since we decided them as more important than standard error and mean values. In such a study, continuous probability distributions of the variables can be found and assigned to the nodes. That would model the system more accurately with a smaller error.

Additionally, first data set includes missing data. The EM algorithm can be used to approximate these missing values, and increase the accuracy of the model.

We believe that multi-criteria decision aid will still see some important development, on the theoretical side (the theory is still at its very beginning) as well as on the practical side with the help of the increasingly user-friendly software, which is currently being developed. Also the use of Bayesian Belief Network phenomena to model real world systems is at its beginning.

REFERENCES

- [1] **AGUARON, J., ESCOBAR, M.T., MORENO, J.** (2003), Consistency Stability Intervals for a Judgment in AHP Decision Support Systems, *European Journal of Operational Research*, Vol.145, pp.382 – 393.

- [2] **AKHAVI, F, HAYES, C.** (2003), A Comparison of Two Multi-criteria Decision-making Techniques, *IEEE International Conference on Systems, Management and Cybernetics*, Vol.1, pp.956-961.

- [3] **AITKEN, R.J.** (1990), Outcome of Surgery for Non-Palpable Mammographic Abnormalities, *Breast Surgery Journal*, Vol. 77, pp.673-676.

- [4] **BLANK, L.** (1980), *Statistical Procedures for Engineering, Management and Science*, McGraw-Hill, New York.

- [5] **BELTON, V., STEWART, T.J.** (2001), *Multiple Criteria Decision Analysis: An Integrated Approach*, Kluwer Academic Publishers, The Netherlands: Dordrecht,

- [6] **BERNARDO, J.M., SMITH, A.F.M.** (1994), *Bayesian Theory*, Wiley England: Chichester.

- [7] **BERRY, D.A., STANGL, D.K.** (2000), *Meta-analysis in Medicine and Health Policy*, Marcel Dekker, New York.

- [8] **CARLIN, J.B.** (1992), Meta-analysis for 2x2 Tables: A Bayesian Approach, *Statistical Medicine*, Vol.11, pp.141-158.

- [9] **CARTER, K. J., RITCHEY, N. P., CASTRO, F., CACCAMO, L. P., ERICKSON, E. B. A.** (1999), Analysis of Three Decision-making Methods: A Breast Cancer Patient as a Model, *Medical Decision Making*, Vol.19, pp.49 - 57.

- [10] **COHEN, J.** (1998), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), NJ: Lawrence Erlbaum Associates, Hillsdale.

- [11] **COLMAN, R.W., HIRSH, J., MARDER,V.J.** (1994), *Homeostasis and Thrombosis*, J.B Lippincott Company, Philadelphia.
- [12] **CASTILLO, E., GUTIERREZ, J.M, HADI, A.S.** (1997), *Expert Systems and Probabilistic Network Models: Monographs in Computer Science*, Springer.
- [13] **CHARNIAK, E., GOLDMAN, R.** (1989), Plan Recognition in Stories and in Life, *In Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pp.54-60. California: Association for Uncertainty in Artificial Intelligence.
- [14] **COOPER, G.F.** (1990), The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks, *Artificial Intelligence*,Vol.42, pp.393-405.
- [15] **DEAN, T.** (1990), Coping with Uncertainty in a Control System for Navigation and Exploration, *In Proceedings of the Ninth National Conference on Artificial Intelligence*, pp.1010-1015, Menlo Park, California: American Association for Artificial Intelligence.
- [16] **DECHTER,R.** (1996), Bucket Elimination: A Unifying Framework for Probabilistic Inference Algorithms, *In Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pp. 211–219.
- [17] **DECHTER, R.** (1997), Mini-buckets: A General Scheme for Generating Approximations in Automated Reasoning, *In Proceedings of the International Joint Conference on Artificial Intelligence*, pp.1297–1303.
- [18] **DECHTER, R., RISH, I.** (2003), Mini Buckets: A General Scheme for Bounded Inference, *Journal of the ACM*, Vol.50, pp.107–153.
- [19] **DIAMOND, G.A., FORRESTER, J.S.** (1983), Clinical Trials and Statistical Verdicts: Probable grounds for appeal, *Annals of Internal Medicine*, Vol.98, pp.385-394.
- [20] **DOLAN, J.G.** (1989), Medical Decision Making Using the Analytic Hierarchy Process, *Medical Decision Making*, Vol. 9, pp.51-56.
- [21] **DRAPER, D.** (2006), Statistical Methods for the Biological and Environmental Sciences, Department of Statistics, University of California, Santa Cruz.
- [22] **DUMOUCHEL,W., HARRIS,J.E.** (1983), Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species, *Journal of The American Statistical Association*, Vol.78, pp.293-315.

- [23] **EDDY, D.M, HASSELBLAD, V., SCHACHTER, R.** (1992), *Meta-analysis by the Confidence Profile Method: The statistical synthesis of evidence*, Boston: Academic Press
- [24] **ESPALLARDO, N.L.** (2003), Decision on Diagnosis in Family Practice: Use of Sensitivity, Specificity, Predictive Values and Likelihood Ratios, *Asia Pacific Family Medicine*, Vol.2, pp.229-232.
- [25] **FENTON, N.** (2006), Technical Report on BBNs, Viewed on 16 August 2006, <http://www.dcs.qmw.ac.uk/%7Enorman/BBNs/BBNs.htm>
- [26] **FENTON, N., NEIL,M.** (2000), *Making Decisions: Using Bayesian Nets and MCDA*, Queen Mary and Westfield College: London
- [27] **FINN V. JENSEN** (1995), Cautious Propagation in Bayesian Networks, In *Proceedings of the Eleventh Conference on Uncertainty in AI (UAI-95)*, pp.323-328.
- [28] **FRANKLIN, D.J., SPIEGELHALTER, F.J., MACARTNEY, F.J., BULL, C.** (1989), Combining Clinical Judgment and Statistical Data in Expert Systems: Over the Telephone Management Decisions for Clinical Heart Disease in the First Month of Life, *International Journal of Clinical Monitoring and Computing*, Vol. 6, pp.157-166.
- [29] **FISCHOFF, B., SLOVIC, B.,LICHTENSTEIN, S.** (1980), *Knowing What You Want: Measuring Labile Values: Cognitive Processes in Choice and Decision Behavior*, Lawrence Erlbaum Associates, Hillsdale.
- [30] **SHAFER G.** (1996), *Probabilistic Expert Systems, CBMS-NFS Regional Conference Series in Applied Mathematics*; National Science Foundation.
- [31] **GENEST, C.,RIVEST, L.P.** (1994), A Statistical Look at Saaty's Method of Estimating Pair-wise Preferences Expressed on a Ratio Scale, *Journal of Mathematical Psychology*, Vol.38, pp. 477-496.
- [32] **GOLDMAN, R.** (1990), A Probabilistic Approach to Language Understanding, Technical Report, CS-90-34, Dept. of Computer Science, Brown University.
- [33] **GOODMAN, S.** (1999), Toward Evidence-based Medical Statistics, I: The P-value Fallacy, *Annals of Internal Medicine*, Vol.130, pp.995-1004.
- [34] **HOFFMAN F., HAND D.J., ADAMS N., FISHER D.** (2001), Lecture Notes in Computer Science, Advances in Intelligent Data Analysis: Fourth International Conference, Berlin.

- [35] **HANSSON, O., MAYER, A.** (1989), Heuristic Search as Evidential Reasoning, *In Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pp.152-161. California: Association for Uncertainty in Artificial Intelligence.
- [36] **HARREL, F. E., SHIH, Y.C.T.** (2001), Using Full Probability Models to Compute Probabilities of Actual Interest to Decision Makers, *International Journal of Technology Assessment in Health Care*, Vol.17:1, pp.17-26.
- [37] **HECKERMAN, D.** (1990), Probabilistic Similarity Networks, Ph.D. Thesis, Program in Medical Information Sciences, Stanford University, California.
- [38] **HECKERMAN, D.** (2004), A Tutorial on Learning with Bayesian Networks, Technical Report No:MSR-TR-95-06, Microsoft Research.
- [39] **HORNBERGER, J.** (2001), Introduction to Bayesian Reasoning, *International Journal of Technology Assessment in Health Care*, Vol.17:1, pp. 9-16.
- [40] **JANSSENS, D., WETS, G., BRIJS, T., VANHOOF, K., ARENTZE, T., TIMMERMANS, H.** (2005), Integrating Bayesian Networks and Decision Trees in a Sequential Rule-based Transportation Model, *European Journal of Operational Research*, a.i.p.
- [41] **JENSEN, F.** (1996), *Introduction to Bayesian Networks*, Springer-Verlag: New York.
- [42] **JENSEN, F., LAURITZEN, S., OLESON, L.** (1990), Bayesian Updating in Causal Probabilistic Networks by Local Computations, *Computational Statistics Quarterly*, Vol.4, pp.269-282.
- [43] **JORDAN, M.I.** (1999), *Learning in Graphical Models*, The MIT Press, Massachusetts.
- [44] **KJAERULFF, U.** (1990), Triangulation of Graphs: Algorithms Giving Small Total State Space, Dept. of Mathematics and Computer Science, Strandvejan, Denmark
- [45] **KLEE, A.J.** (1971), The Role of Decision Models in the Evaluation of Competing Environmental Health Alternatives, *Management Science*, Vol.18:2, pp.52-67.
- [46] **LAININEN, P., HAMALAINEN, R. P.** (2003), Analyzing AHP-matrices by Regression, *European Journal of Operational Research*, Vol.148, pp.514-524.

- [47] **LAURITZEN, S., SPIEGELHALTER, D.** (1988), Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems, *Journal of the Royal Statistical Society*, Vol.50, pp.157–224.
- [48] **LEE, R.C., DONALDSON, C., COOK, L.S.** (2003), The Need for Evolution in Healthcare Decision Modeling, *Medical Care*, Vol.41, No. 9, pp.1024-1033.
- [49] **LEVITT, T., MULLIN, J., BINFORD, T.** (1989), Model-based Influence Diagrams for Medicine Vision, In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pp.233-244. California: Association for Uncertainty in Artificial Intelligence.
- [50] **LILFORD, R.J., BRAUNHOLTZ, D.** (1996), The Statistical Basis of Public Policy: A paradigm Shift Is Overdue, *British Medical Journal*, Vol.313, pp.603-607.
- [51] **MANGASARIAN, O.L., STREET, N., WOLBERG, W.H.** (1994), Breast Cancer Diagnosis and Prognosis via Linear Programming, Mathematical Programming Technical Report, pp.94-10.
- [52] **MANGASARIAN, O.L., STREET, W.N., WOLBERG, W.H.** (1994), Breast Cancer Diagnosis and Prognosis via Linear Programming, *Operations Research*, Vol.43, pp.570-577.
- [53] **MARSHALL, E.** (1993) Search for a Killer: Focus Shifts From Fat to Hormones in Special Report on Breast Cancer, *Science Journal*, Vol.259, pp.618-621.
- [54] **MUIER, C., WATERHOUSE, J., MACK, T., POWELL, J., WHELAN, S.** (1987) *Cancer Incidence in Five Continents*, Vol.5, International Agency for Research on Cancer, Lion, France.
- [55] **MURPHY, K., WEISS, Y., JORDAN, M.** (1999), Loopy Belief Propagation for Approximate Inference: An Empirical Study, In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp.467–475.
- [56] **NEATHN, A.A., SAMANIEGO, F.J.** (1997), On the Efficacy of Bayesian Inference for Non-identifiable Models, *The American Statistician*, Vol.51, pp. 225-232.
- [57] **PARMIGIANI, G.** (2002), *Modeling in Medical Decision Making: A Bayesian Approach*, England: John Wiley & Sons, Ltd.: Statistics in Practice.
- [58] **PEARL, J.** (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers.

- [59] **POMEROL, J.C., ROMERO, S.B.** (2000), *Multi-criterion Decision in Management: Principles and Practice*, Kluwer's International Series.
- [60] **NEAPOLITAN, R.E.** (2004), *Learning Bayesian Networks*, Pearson Prentice Hall.
- [61] **RAIFFA, H., SCHLAIFER, R.** (1961), *Applied Statistical Decision Theory*, Boston: Harvard Business School
- [62] **RAJABALLY, E., SEN, P., WHITTLE, S., DALTON, J.** (2004), Aids to Bayesian Belief Network Construction, *Second IEEE International Conference on Intelligent Systems*, pp.457-461.
- [63] **ROSE, D.** (1974), Triangulated Graphs and The Elimination Process, *Journal of Mathematical Analysis and Applications*, Vol.32, pp.597-609.
- [64] **ROY, B., VANDERPOOTEN, D.** (1996), The European School of MCDA: Emergence, Basic Features and Current Works, *Journal of Multi Criteria Decision Analysis*, Vol. 5, pp.22-38.
- [65] **SAATY, T.L.** (1980), *The Analytical Hierarchy Process*, McGraw-Hill, New York.
- [66] **SAATY, T.L., VARGAS, L.G.** (1998), Diagnosis with Dependent Symptoms: Bayes Theorem and the Analytical Hierarchy Process, *Operations Research*, Vol.46:4, pp.491-502.
- [67] **SAVAGE, L.J.** (1954), *The Foundations of Statistics*, New York: Wiley.
- [68] **SCHLAIFER, R.** (1959), *Probability and Statistics for Business Decisions*, New York: McGraw-Hill.
- [69] **STINNETT, A.A., MULLAHY, J.** (1998), Net Health Benefits: A New Framework for The Analysis of Uncertainty in Cost-effectiveness Analysis, *Medical Decision Making*, Vol.18(suppl 2), pp.68-80.
- [70] **STREET, W. N., MANGASARIAN, O. L., WOLBERG, W.H.** (1995), An Inductive Learning Approach to Prognostic Prediction, *In Proceedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann, pp.522-530, San Francisco.
- [71] **STREET, W. N., MANGASARIAN, O. L., WOLBERG, W.H.** (1995), Breast Cancer Diagnosis and Prognosis via Linear Programming. *Operations Research*, Vol.43:4, pp.570-577.

- [72] **STREET, W.N., WOLBERG, W.H., MANGASARIAN, O.L.** (1993), Nuclear Feature Extraction for Breast Tumor Diagnosis, *International Symposium on Electronic Imaging: Science and Technology*, Vol.1905, pp.861-870, San Jose.
- [73] **WATTHAYU, W., PENG, Y.** (2004), A Bayesian Network based Framework for Multi-criteria Decision-making, *MCDM 2004 Conference*:: Canada.
- [74] **WINKLER, R.L.** (1972), *An Introduction to Bayesian Inference and Decision*, Holt, Rinehart and Winston, Inc., USA.
- [75] **WINKLER, R.L.** (2001), Why Bayesian Analysis Hasn't Caught on in Healthcare Decision Making, *International Journal of Technology Assessment in Health Care*, Vol.17:1, pp.56-66.
- [76] **LICHODZIJEWSKI, P., KHARRAZI, H., RICHARD, M.** (2005), Evolutionary Computing for Knowledge Discovery in Breast Cancer, Dalhousie University.
- [77] **WITTEKIND, C., SCHULTE, E.** (1987), Computerized Morphometric Image Analysis of Cytologic Nuclear Parameters in Breast Cancer, *Analytical and Quantitative Cytology and Histology*, Vol.9:6, pp.480-484.
- [78] **WOLBERG, W.H., MANGASARIAN, O.L., STREET, N.** (1995), Computer-Derived Nuclear "Grade" and Breast Cancer Prognosis, *Anal. Quant. Cytol. Histol.*, Vol.17:4, pp.257-64.
- [79] **WOLBERG, W.H., MANGASARIAN, O.L., STREET, N.** (1999), Importance of Nuclear Morphology in Breast Cancer Prognosis, *Clinical Cancer Research*, Vol. 5, pp.3542-3548.
- [80] **WOLBERG, W.H., STREET, W.N., HEISEY, D.M., MANGASARIAN, O.L.** (1995) Computerized Breast Cancer Diagnosis and Prognosis from Fine Needle Aspirates, *Archives of Surgery*, Vol.130, pp.511-516.
- [81] **WOLBERG, W.H., STREET, W.N., HEISEY, D.M., MANGASARIAN, O.L.** (1995), Image Analysis and Machine Learning Applied to Breast Cancer Diagnosis and Prognosis, *Analytical and Quantitative Cytology and Histology*, Vol.17: 2, pp.77-87.

APPENDIX A

Conditional and Marginal Probability Tables of Hypothetical Problem in Chapter 4

The conditional probabilities regarding the network are assigned as follows:

	T1	
D1	Positive	Negative
Exists	0.95	0.05
No Disease	0.14	0.86

		T2	
D1	D2	Positive	Negative
Exists	Exists	0.98	0.02
Exists	No Disease	0.67	0.33
No Disease	Exists	0.75	0.25
No Disease	No Disease	0.11	0.89

	T3	
T2	Positive	Negative
Positive	0.78	0.22
Negative	0.35	0.65

	T4	
D3	Positive	Negative
Exists	0.55	0.45
No Disease	0.27	0.73

	T5	
T4	Positive	Negative
Positive	0.82	0.18
Negative	0.08	0.92

	TR1	
T1	Apply	Do not apply
Positive	0.6	0.4
Negative	0.08	0.92

	TR2	
T3	Apply	Do not apply
Positive	0.93	0.07
Negative	0.13	0.87

The marginal probabilities regarding the network are calculated as follows:

D1	
Exists	No Disease
0.005	0.995

D3	
Exists	No Disease
0.001	0.999

D2	
Exists	No Disease
0.003	0.997

T2	
Positive	Negative
0.11471505	0.88528495

T1	
Positive	Negative
0.14405	0.85595

T4	
Positive	Negative
0.27028	0.72972

T3	
Positive	Negative
0.399327472	0.600673

TR1	
Apply	Do not apply
0.15491	0.845094

T5	
Positive	Negative
0.28001	0.719993

TR2	
Apply	Do not apply
0.44946198	0.55053802

TR3	
Apply	Do not apply
0.285204248	0.714795752

	TR3	
T5	Apply	Do not apply
Positive	0.71	0.29
Negative	0.12	0.88

The conditional probabilities of the sink node “PR”

Assessment (Model: APPLICATION2, Node: PR)					
Parent Node(s)			PR		
TR1	TR3	TR2	end	continue	bar charts
APPLY	APPLY	APPLY	1.0	0.0	
		DO NOT APPLY	0.75	0.25	
	DO NOT APPLY	APPLY	0.65	0.35	
		DO NOT APPLY	0.5	0.5	
DO NOT APPLY	APPLY	APPLY	0.6	0.4	
		DO NOT APPLY	0.4	0.6	
	DO NOT APPLY	APPLY	0.2	0.8	
		DO NOT APPLY	0.0	1.0	

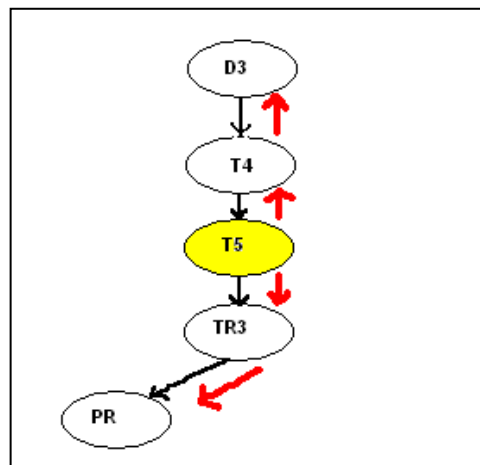
APPENDIX B

Evidence Propagation Using Pearl's Message Passing Algorithm

Evidence 1: Suppose Test 5 (T5) is performed on the patient and the result seems to be negative (n). In that case, the changes on other nodes are shown below:

Update_tree (APPLICATION2,T5,negative(n))

$\lambda(T5_n)=1; \pi(T5_n)=1; P(T5 | \{T5_n\})=1;$ //Instantiate T5 for negative (T5_n)
 $\lambda(T5_p)=0; \pi(T5_p)=0; P(T5 | \{T5_n\})=0;$



The flow of evidence 1 on the network

The call

send_lambda_message(T5,T4)

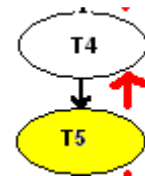
$$\lambda_{T5}(T4_p) = P(T5_p | T4_p) \lambda(T5_p) + P(T5_n | T4_p) \lambda(T5_n) = (0.82)(0) + (0.18)(1) = 0.18$$

$$\lambda_{T5}(T4_n) = P(T5_p | T4_n) \lambda(T5_p) + P(T5_n | T4_n) \lambda(T5_n) = (0.08)(0) + (0.92)(1) = 0.92$$

$$\lambda(T4_p) = \lambda_{T5}(T4_p) = 0.18 \quad // \text{Compute T4's } \lambda \text{ values}$$

$$\lambda(T4_n) = \lambda_{T5}(T4_n) = 0.92$$

$$P(T4_p | \{T5_n\}) = \alpha \lambda(T4_p) \pi(T4_p) = \alpha(0.18)(.27028) = 0.0486504\alpha$$



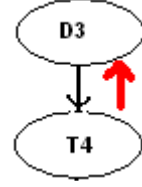
$$P(T4_n | \{T5_n\}) = \alpha \lambda(T4_n) \pi(T4_n) = \alpha(0.92)(.72972) = 0.6713424\alpha$$

$$P(T4_p | \{T5_n\}) = (0.0486504\alpha) / (0.0486504\alpha + 0.6713424\alpha) = 0.067571;$$

//Compute $P(T4 | \{T5_n\})$

$$P(T4_n | \{T5_n\}) = (0.6713424\alpha) / (0.0486504\alpha + 0.6713424\alpha) = 0.932429;$$

send_λ_message(T4,D3)



$$\lambda_{T4}(D3_y) = P(T4_p | D3_y) \lambda(T4_p) + P(T4_n | D3_y) \lambda(T4_n) = (0.55)(0.18) + (0.45)(0.92) = 0.513$$

$$\lambda_{T4}(D3_n) = P(T4_p | D3_n) \lambda(T4_p) + P(T4_n | D3_n) \lambda(T4_n) = (0.27)(0.18) + (0.73)(0.92) = 0.7202$$

$$\lambda(D3_y) = \lambda_{T4}(D3_y) = 0.513;$$

$$\lambda(D3_n) = \lambda_{T4}(D3_n) = 0.7202;$$

$$P(D3_y | \{T5_n\}) = \alpha \lambda(D3_y) \pi(D3_y) = \alpha(0.513)(0.001) = 0.000513\alpha$$

$$P(D3_n | \{T5_n\}) = \alpha \lambda(D3_n) \pi(D3_n) = \alpha(0.7202)(.999) = 0.7194798\alpha$$

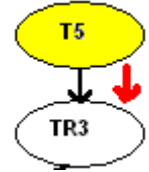
$$P(D3_y | \{T5_n\}) = (0.000513\alpha) / (0.000513\alpha + 0.7194798\alpha) = 0.0007125071251;$$

$$P(D3_n | \{T5_n\}) = 1 - 0.0007125071251 = 0.9992874929 ; //Compute $P(D3 | \{T5_n\})$$$

send_π_message(T5,TR3)

$$\pi_{TR3}(T5_p) = \pi(T5_p) = 0;$$

$$\pi_{TR3}(T5_n) = \pi(T5_n) = 1;$$



$$\pi(TR3_y) = P(TR3_y | T5_p) \pi_{TR3}(T5_p) + P(TR3_y | T5_n) \pi_{TR3}(T5_n) = (0.71)(0) + (0.12)(1) = 0.12$$

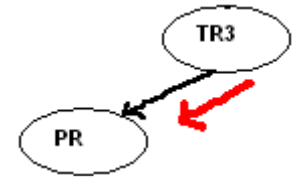
$$\pi(TR3_n) = P(TR3_n | T5_p) \pi_{TR3}(T5_p) + P(TR3_n | T5_n) \pi_{TR3}(T5_n) = (0.29)(0) + (0.88)(1) = 0.88$$

$$P(TR3_y | \{T5_n\}) = 0.12;$$

//Compute $P(TR3 | \{T5_n\})$

$$P(TR3_n | \{T5_n\}) = 0.88;$$

send_π_message(TR3,PR)



$$\pi_{PR}(TR3_y) = \pi(TR3_y) = 0.12;$$

$$\pi_{PR}(TR3_n) = \pi(TR3_n) = 0.88;$$

$$\begin{aligned} \pi(PR_y) = & P(PR_y | TR3_y, TR2_y, TR1_y) \pi_{PR}(TR3_y) \pi_{PR}(TR2_y) \pi_{PR}(TR1_y) \\ & + P(PR_y | TR3_n, TR2_y, TR1_y) \pi_{PR}(TR3_n) \pi_{PR}(TR2_y) \pi_{PR}(TR1_y) \\ & + P(PR_y | TR3_y, TR2_y, TR1_n) \pi_{PR}(TR3_y) \pi_{PR}(TR2_y) \pi_{PR}(TR1_n) \\ & + P(PR_y | TR3_n, TR2_y, TR1_n) \pi_{PR}(TR3_n) \pi_{PR}(TR2_y) \pi_{PR}(TR1_n) \\ & + P(PR_y | TR3_y, TR2_n, TR1_y) \pi_{PR}(TR3_y) \pi_{PR}(TR2_n) \pi_{PR}(TR1_y) \\ & + P(PR_y | TR3_n, TR2_n, TR1_y) \pi_{PR}(TR3_n) \pi_{PR}(TR2_n) \pi_{PR}(TR1_y) \\ & + P(PR_y | TR3_y, TR2_n, TR1_n) \pi_{PR}(TR3_y) \pi_{PR}(TR2_n) \pi_{PR}(TR1_n) \\ & + P(PR_y | TR3_n, TR2_n, TR1_n) \pi_{PR}(TR3_n) \pi_{PR}(TR2_n) \pi_{PR}(TR1_n) \end{aligned}$$

```

=      (1)      (0.12)  (0.4495) (0.15491)
+      (0.65)   (0.88)  (0.4495) (0.15491)
+      (0.6)    (0.12)  (0.4495) (0.84509)
+      (0.2)    (0.88)  (0.4495) (0.84509)
+      (0.75)   (0.12)  (0.5505) (0.15491)
+      (0.5)    (0.88)  (0.5505) (0.15491)
+      (0.4)    (0.12)  (0.5505) (0.84509)
+      (0)      (0.88)  (0.5505) (0.84509)

```

```

 $\pi(\text{PR}_y) = 0.2099206;$ 
 $\pi(\text{PR}_n) = 1 - 0.2099206 = 0.7900794;$ 

```

```

P(PR_y | {T5_n}) = 0.2099206;           //Compute P(PR | {T5_n})
P(PR_n | {T5_n}) = 0.7900794;

```

Evidence 2: Being the T5 test result on hand (as negative), another information is obtained regarding node D1. Suppose that the genetic history of the patient's mother about Disease 1 strongly claims that the patient has Disease 1.

This evidence affects the probabilities of the other nodes as follows:

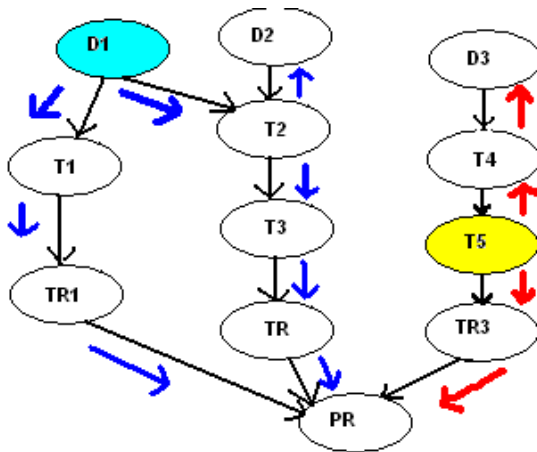


Figure 4.12: The flow of evidence 2 on the network, being evidence 1 present

Update_tree (APPLICATION2, T5:negative(n), D1:exists(y))

```

 $\lambda(D1_y) = 1; \lambda(D1_n) = 0;$            //Instantiate D1 for disease exists (D1_y)
 $\pi_{T1}(D1_y) = \pi(D1_y) = 1;$ 
 $\pi_{T1}(D1_n) = \pi(D1_n) = 0;$ 

```

The Call

send_pi_msg(D1, T1)

```

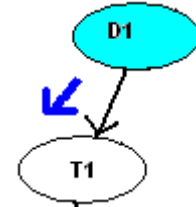
 $\pi_{T1}(D1_y) = \pi(D1_y) \lambda_{T2}(D1_y) = (1)(1) = 1;$  //D1 sends T1 a pi message.
 $\pi_{T1}(D1_n) = \pi(D1_n) \lambda_{T2}(D1_n) = (0)(1) = 0;$ 

```

$$\begin{aligned}\pi(T1_y) &= P(T1_y | D1_y) \pi_{T1}(D1_y) + P(T1_y | D1_n) \pi_{T1}(D1_n) \\ &= (0.95)(1) + (0.14)(0) = 0.95; \\ \pi(T1_n) &= (0.05)(1) + (0.86)(0) = 0.05; \quad // \text{Compute } T1\text{'s } \pi \text{ values.}\end{aligned}$$

$$\begin{aligned}P(T1_y | \{T5_n\}, \{D1_y\}) &= \alpha \lambda(T1_y) \pi(T1_y) = \alpha(1)(0.95) = 0.95\alpha \\ P(T1_n | \{T5_n\}, \{D1_y\}) &= \alpha \lambda(T1_n) \pi(T1_n) = 0.05\alpha\end{aligned}$$

$$\begin{aligned}P(T1_y | \{T5_n\}, \{D1_y\}) &= 0.95; \\ P(T1_n | \{T5_n\}, \{D1_y\}) &= 0.05;\end{aligned}$$



send_pi_msg(T1, TR1)

$$\begin{aligned}\pi_{TR1}(T1_y) &= \pi(T1_y) = 0.95; \\ \pi_{TR1}(T1_n) &= \pi(T1_n) = 0.05;\end{aligned}$$

$$\begin{aligned}\pi(TR1_y) &= P(TR1_y | T1_y) \pi_{TR1}(T1_y) + P(TR1_y | T1_n) \pi_{TR1}(T1_n) \\ &= (0.6)(0.95) + (0.08)(0.05) = 0.574; \\ \pi(TR1_n) &= (0.4)(0.95) + (0.92)(0.05) = 0.426; \quad // \text{Compute } TR1\text{'s } \pi \text{ values.}\end{aligned}$$

$$\begin{aligned}P(TR1_y | \{T5_n\}, \{D1_y\}) &= 0.574; \\ P(TR1_n | \{T5_n\}, \{D1_y\}) &= 0.426;\end{aligned}$$

send_pi_msg(D1, T2)

$$\begin{aligned}\lambda(D1_y) &= 1; \pi(D1_y) = 1; P(D1_y | \{D1_y\}) = 1; \\ \lambda(D1_n) &= 0; \pi(D1_n) = 0; P(D1_n | \{D1_y\}) = 0;\end{aligned}$$

send_lambda_msg(T2, D2)

$$\begin{aligned}\lambda_{T2}(D2_y) &= [P(T2_y | D2_y, D1_y) \pi_{T2}(D1_y) + P(T2_y | D2_y, D1_n) \pi_{T2}(D1_n)] \lambda(T2_y) \\ &+ [P(T2_n | D2_y, D1_y) \pi_{T2}(D1_y) + P(T2_n | D2_y, D1_n) \pi_{T2}(D1_n)] \lambda(T2_n) \\ &= [(0.98)(1) + (0.75)(0)](1) + [(0.02)(1) + (0.25)(0)](1) \\ &= 1 \quad // T2 \text{ sends } D2 \text{ a } \lambda\end{aligned}$$

message.

$$\lambda_{T2}(D2_n) = 0;$$

$$\begin{aligned}\lambda(D2_y) &= \lambda_{T2}(D2_y) = 1; \quad // \text{Compute } D2\text{'s } \lambda \text{ values.} \\ \lambda(D2_n) &= \lambda_{T2}(D2_n) = 0;\end{aligned}$$

$$\begin{aligned}P(D2_y | \{T5_n\}, \{D1_y\}) &= \alpha \lambda(D2_y) \pi(D2_y) = \alpha(1)(0.003) = 0.003\alpha \\ P(D2_n | \{T5_n\}, \{D1_y\}) &= \alpha \lambda(D2_n) \pi(D2_n) = \alpha(1)(0.997) = 0.997\alpha\end{aligned}$$

$$\begin{aligned}P(D2_y | \{T5_n\}, \{D1_y\}) &= 0.003; \\ P(D2_n | \{T5_n\}, \{D1_y\}) &= 0.997;\end{aligned}$$

send_pi_msg(D2,T2) $\pi_{T2}(D2_y) = \pi(D2_y) = 0.003;$
 $\pi_{T2}(D2_n) = \pi(D2_n) = 0.997;$

$\pi_{T2}(D1_y) = \pi(D1_y) = 1;$
 $\pi_{T2}(D1_n) = \pi(D1_n) = 0;$

$\pi(T2_y) = P(T2_y | D1_y, D2_y) \pi_{T2}(D1_y) \pi_{T2}(D2_y) + P(T2_y | D1_y, D2_n) \pi_{T2}(D1_y) \pi_{T2}(D2_n)$
 $+ P(T2_y | D1_n, D2_y) \pi_{T2}(D1_n) \pi_{T2}(D2_y) + P(T2_y | D1_n, D2_n) \pi_{T2}(D1_n) \pi_{T2}(D2_n)$
 $= (0.98)(1)(0.003) + (0.67)(1)(0.99) + (0.75)(0)(0.003) + (0.11)(0)(0.99)$
 $= 0.67093$

$\pi(T2_y) = 0.32907$
 $P(T2_y | \{T5_n\}, \{D1_y\}) = 0.67093;$
 $P(T2_n | \{T5_n\}, \{D1_y\}) = 0.32907;$

send_pi_msg(T2,T3)

$\pi_{T3}(T2_y) = \pi(T2_y) = 0.67093;$
 $\pi_{T3}(T2_n) = \pi(T2_n) = 0.32907;$

$\pi(T3_y) = P(T3_y | T2_y) \pi_{T3}(T2_y) + P(T3_y | T2_n) \pi_{T3}(T2_n)$
 $= (0.78)(0.67093) + (0.35)(0.32907) = 0.63384999;$
 $\pi(T3_n) = 1 - 0.63384999 = 0.3615001; \quad // \text{Compute } T3\text{'s } \pi \text{ values.}$

send_pi_msg(T3,TR2)

$\pi_{TR2}(T3_y) = \pi(T3_y) = 0.63384999;$
 $\pi_{TR2}(T3_n) = \pi(T3_n) = 0.3615001;$

$\pi(TR2_y) = P(TR2_y | T3_y) \pi_{TR2}(T3_y) + P(TR2_y | T3_n) \pi_{TR2}(T3_n)$
 $= (0.93)(0.63384999) + (0.13)(0.3615001) = 0.636476;$
 $\pi(T3_n) = 1 - 0.636476 = 0.3635245; \quad // \text{Compute } T3\text{'s } \pi \text{ values.}$

For the sink node

send_pi_msg(PR; TR1, TR2, TR3)

$\pi_{PR}(TR1_y) = 0.574$
 $\pi_{PR}(TR2_y) = 0.636476$
 $\pi_{PR}(TR3_y) = 0.12$

$\pi_{PR}(TR1_n) = 0.426$
 $\pi_{PR}(TR2_n) = 0.3635245$
 $\pi_{PR}(TR3_n) = 0.88$

$\pi(PR_y) = (1)(0.12)(0.636476)(0.574) + (0.65)(0.88)(0.636476)(0.574)$
 $+ (0.6)(0.12)(0.636476)(0.426) + (0.2)(0.88)(0.636476)(0.426)$
 $+ (0.75)(0.12)(0.3635245)(0.574) + (0.5)(0.88)(0.3635245)(0.574)$
 $+ (0.4)(0.12)(0.3635245)(0.426) + (0)(0.88)(0.3635245)(0.426)$
 $= 0.4380805478$

$\pi(PR_y) = 0.5619194522$

$P(PR_y | \{T5_n\}, \{D1_y\}) = 0.438081;$
 $P(PR_n | \{T5_n\}, \{D1_y\}) = 0.561919;$

APPENDIX C

Tables and Figures that are not included in Chapter 7

Table-7.2: Correlation matrix of attributes in Set 2

		Correlations									
		RADIUS	TEXTURE	PERIMETE	AREA	SMOOTH	COMPACT	CONCAVIT	CONCAPTS	SYMMETRY	FRACTDIM
RADIUS	Pearson Correlation	1.000	.344**	.998**	.988**	.157**	.492**	.667**	.817**	.151**	-.306**
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
TEXTURE	Pearson Correlation	.344**	1.000	.351**	.338**	.005	.257**	.323**	.316**	.093*	-.065
	Sig. (2-tailed)	.000	.	.000	.000	.907	.000	.000	.000	.028	.125
	N	556	556	556	556	556	556	556	556	556	556
PERIMETE	Pearson Correlation	.998**	.351**	1.000	.987**	.195**	.545**	.708**	.846**	.187**	-.254**
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000	.000	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
AREA	Pearson Correlation	.988**	.338**	.987**	1.000	.166**	.486**	.677**	.818**	.153**	-.278**
	Sig. (2-tailed)	.000	.000	.000	.	.000	.000	.000	.000	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
SMOOTH	Pearson Correlation	.157**	.005	.195**	.166**	1.000	.659**	.521**	.554**	.557**	.591**
	Sig. (2-tailed)	.000	.907	.000	.000	.	.000	.000	.000	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
COMPACT	Pearson Correlation	.492**	.257**	.545**	.486**	.659**	1.000	.880**	.827**	.610**	.584**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.	.000	.000	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
CONCAVIT	Pearson Correlation	.667**	.323**	.708**	.677**	.521**	.880**	1.000	.919**	.509**	.354**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.	.000	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
CONCAPTS	Pearson Correlation	.817**	.316**	.846**	.818**	.554**	.827**	.919**	1.000	.471**	.182**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.	.000	.000
	N	556	556	556	556	556	556	556	556	556	556
SYMMETRY	Pearson Correlation	.151**	.093*	.187**	.153**	.557**	.610**	.509**	.471**	1.000	.477**
	Sig. (2-tailed)	.000	.028	.000	.000	.000	.000	.000	.000	.	.000
	N	556	556	556	556	556	556	556	556	556	556
FRACTDIM	Pearson Correlation	-.306**	-.065	-.254**	-.278**	.591**	.584**	.354**	.182**	.477**	1.000
	Sig. (2-tailed)	.000	.125	.000	.000	.000	.000	.000	.000	.000	.
	N	556	556	556	556	556	556	556	556	556	556

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table-7.3: Correlation matrix of attributes in Set 2 (summarized)

Correlation coeff.	Radius	Textur e	Perime ter	Area	Smoot hness	Comp actnes s	Conca vity	Conca ve Pts	Symett ery	Fract. Dim
Radius	1.00									
Texture	0.34	1.00								
Perimeter	1.00	0.35	1.00							
Area	0.99	0.34	0.99	1.00						
Smoothness	0.16	0.01	0.20	0.17	1.00					
Compactness	0.49	0.26	0.55	0.49	0.66	1.00				
Concavity	0.67	0.32	0.71	0.68	0.52	0.88	1.00			
Concave Pts	0.82	0.32	0.85	0.82	0.55	0.83	0.92	1.00		
Symmetry	0.15	0.09	0.19	0.15	0.56	0.61	0.51	0.47	1.00	
Fract. Dim	0.31	0.07	0.25	0.28	0.59	0.58	0.35	0.18	0.48	1.00

Table-7.4: Correlation matrix of attributes in Set 2 (converted into verbal values)

Correlation coeff.	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave Pts	Symmetry
Radius									
Texture	medium								
Perimeter	high	medium							
Area	high	medium	high						
Smoothness									
Compactness			high	medium	high				
Concavity	high	medium	high	high	high	high			
Concave Pts	high	medium	high	high	high	high	high		
Symmetry					high	high	high	medium	
Fract. Dim	medium				high	high	medium		medium

Radius Texture Perimeter Area Smoothness Compactness Concavity Concave Pts Fractal Dim

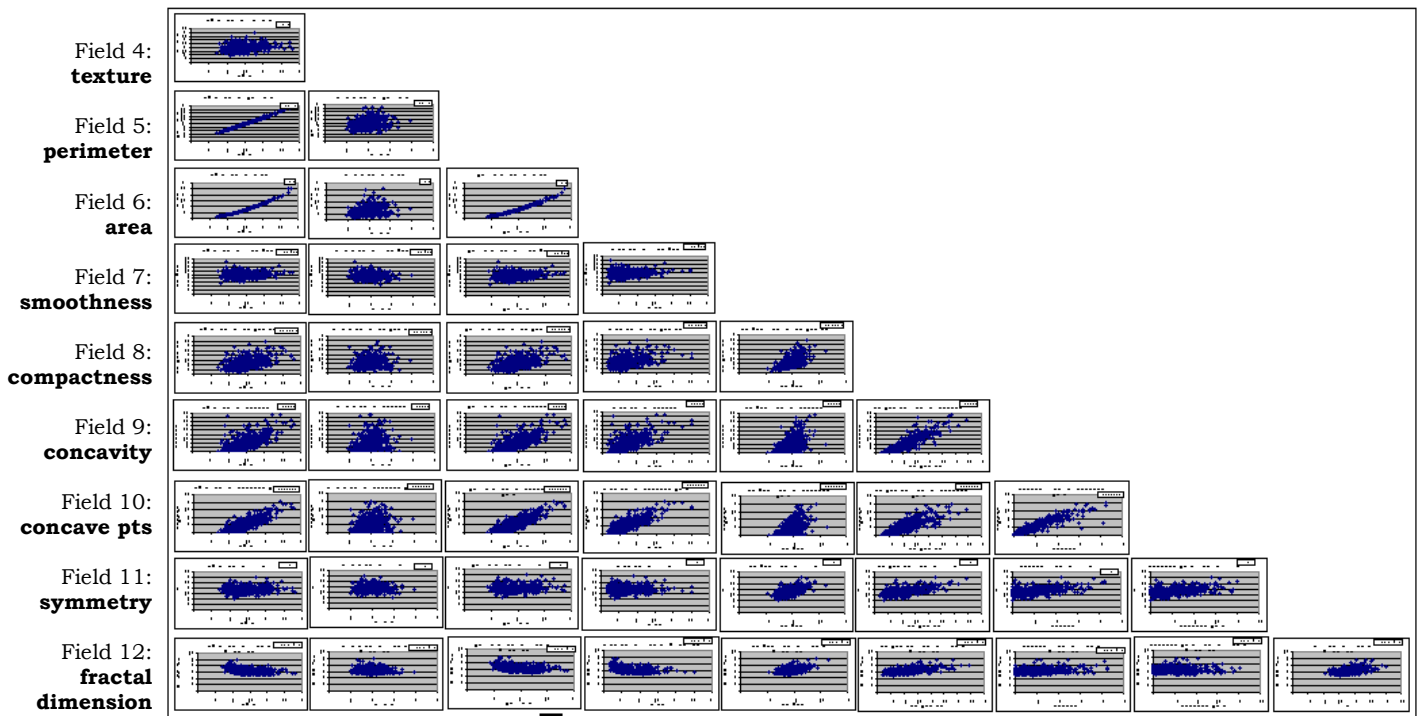


Figure-7.2: Scatter plot identifying the relations between attributes (Set 2)

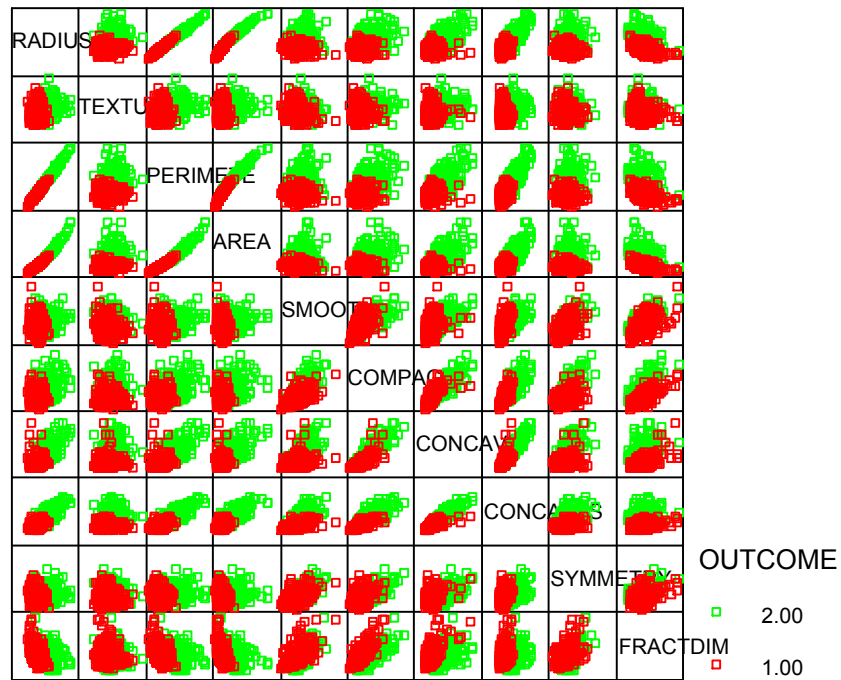


Figure-7.3: Scatter plot matrix identifying the difference of the attribute values (Set 2) in terms of their type of diagnosis. Red points: benign tumors, Green points: malignant tumors

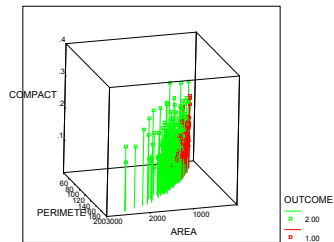


Figure-7.4: Three dimensional Scatter plot showing the dependency of compactness on area and perimeter

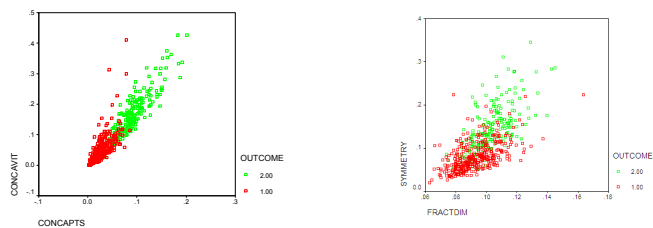


Figure-7.5: Dependency between concave points & concavity; fractal dimension & symmetry

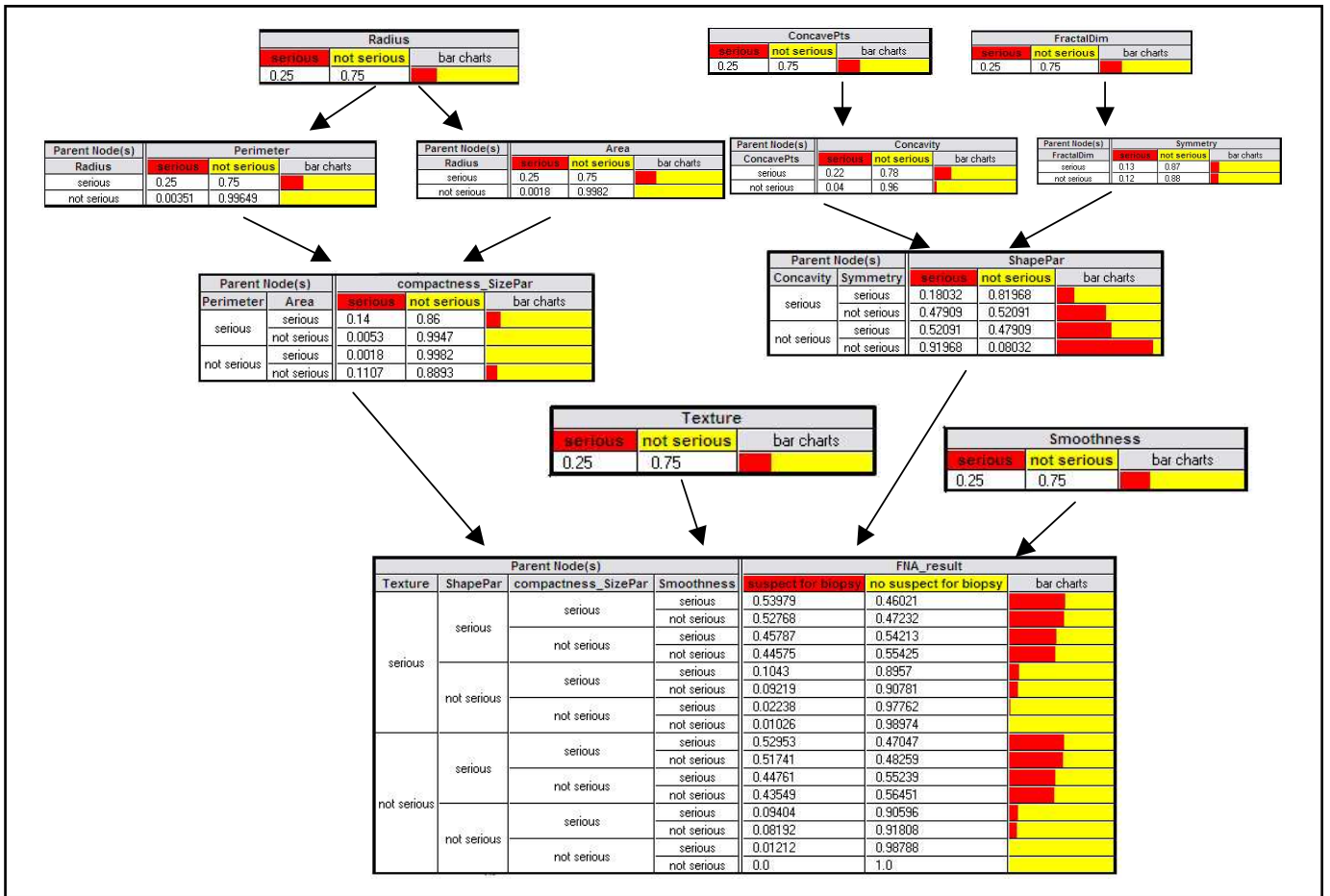


Figure-7.7: Conditional Probability Tables (CPTs) of the BBN model