Mathematical and Computational Applications, Vol. 17, No. 3, pp. 212-222, 2012

A SUGGESTION FOR CONSTRUCTING A BAYESIAN NETWORK MODEL WITH SIMPLE CORRELATION AND AN APPROPRIATE REGRESSION ANALYSIS: A REAL MEDICAL DIAGNOSIS APPLICATION

Semra Erpolat

Department of Statistics, Mimar Sinan Fine Arts University, 34349, Beşiktaş, İstanbul, Turkey.
serpolat@msgsu.edu.tr

Abstract- The major task of medical science is to prevent or diagnose disease. Medical diagnosis is usually made by using some blood metrics and in addition, to be able to reach better results, one can benefit from different scientific methods. In this paper a Bayesian network method is proposed. This method is a hybrid that uses simple correlation and according to dependent variable type either simple linear regression or logistic regression for constructing a Bayesian topology. The Bayesian network is a method for representing probabilistic relationships between variables associated with an outcome of interest. To develop a Bayesian network, a structure must first be constructed. To build the topology of the Bayesian network, some alternative method can be used. One is using domain experts who usually have a good grasp of the conditional dependencies in the domain to develop the structure of the Bayesian network. Another is using structure learning algorithms, such as genetic algorithms, to construct the network topology from training data. In this paper a different construction method is proposed by using correlation analysis and one of the simple linear regression or logistic regression analyses. First, correlations of the examined variables are found. Then according to the significant correlation coefficients, the degree and direction of the interactions between these variables are established by using either simple linear regression or logistic regression. Finally the Bayesian network model is constructed by using this information. For evaluating our model, another model which does not have any relation between the input variables is also constructed. And these two models are compared by using an original thyroid data set. It is concluded that our proposed model provides a high degree of performance and good explanatory power and it may prove useful for clinicians in the medical field.

Key Words- Bayesian Network, Correlation Analysis, Simple Linear Regression, Logistic Regression, Medical Diagnosis.

1. INTRODUCTION

The thyroid gland is one of the most important organs in the body and its primary role is to help regulation of the body's metabolism [13]. Thyroid problems can commonly be treated successfully. Thyroid disorder is a general term representing several different diseases involving thyroid hormones and the thyroid gland. Thyroid disorders are commonly separated into two major categories, hyperthyroidism and hypothyroidism, depending on whether serum thyroid hormone levels (T4 and T3) are increased or decreased [5]. To diagnose thyroid function abnormalities correctly based

on clinical and laboratory tests often proves difficult because many thyroid symptoms are nonspecific. Especially in hypothyroidism, symptoms such as lethargy, confusion, weight gain, and poor memory are easily confused with other psychiatric and medical conditions. Thyroid dysfunction diagnosis presents a challenge to traditional statistical methods because it represents a classification problem with three extremely unbalanced groups. Statistical and other quantitative methods have long been used as decision-making tools in medical diagnosis including thyroid disease detection. These classification methods include both parametric methods such as discriminant analysis and logistic regression and nonparametric models like k-nearest-neighbor and mathematical programming models. The effectiveness of these methods depends to a large extent on the various assumptions or conditions under which models are developed [13]. In this paper a hybrid Bayesian network method that consists of simple correlation analysis and either simple linear regression or logistic regression is proposed. The most decisive facts in the diagnosis of thyroid disease are the patient's blood test results. For this purpose, the blood test results consisting of 13 values, considered as the input data, taken from a total of 76 patients at one of the major hospitals in Turkey, had never been cured for thyroid disease and don't have any other systemic disease are considered together with the output variable to determine whether the patient is a thyroid disease patient or not. To create the model, simple correlation analysis is applied on the input variables first in order to see which of these variables are in relation with each other. Then, considering the relations that are found meaningful, one of the simple linear regression or logistic regression analyses is applied to determine the size of interaction between the variables. Determining which one of these analyses will be applied is made according to the dependent variable being scaled ordinally or nominally. Simple linear regression analysis is applied if the dependent variable is ordinal-scaled and logistic regression is applied if the dependent variable is scaled nominally. The directions of the connections in the Bayesian network are determined according to the size of the obtained interaction coefficients. The verification of the model is performed according to the 2-fold cross validation rule. The simple Bayesian network where there is no connection between the input variables, and all input variables are only in connection with the output variable is used to measure the validity or the success of the model. This model is also verified using the 2-fold cross validation method just like the proposed model. Finally, the results obtained from both models are compared. The analyses are performed with the SPSS 16.0 and Netica 4.16 package programs.

2. THE BAYESIAN NETWORK MODEL

Bayesian networks, also known as belief networks or Bayes nets, have emerged as an effective tool for knowledge representation and inference [12]. A Bayesian network is a directed, acyclic graph that can be used to represent the dependency between random variables, represented by nodes. Links between nodes represent conditional probabilities and link directions represent causality between the parent and children nodes. Recently, a great deal of interest has arisen in the artificial intelligence community about the idea of using Bayesian networks for classification [3]. Bayesian networks are a type of expert system that employs probabilistic reasoning, in the form

of Bayesian statistics to provide a classification that is both semantically and statistically justified [11]. According to Bayes' rule, the posterior probability can be expressed in terms of the joint probability, which can be further expressed by conditional probability and a prior probability given in Eq. (1) [9]:

$$P(S/E)=\frac{P(S,E)}{P(E)}=\frac{P(E/S)P(S)}{P(E)} \quad (1)$$

Here *S* denotes semantic task and *E* denotes evidence. In Bayesian networks the strengths of the relationships between the variables are expressed as conditional probability tables (CPT). Thus, a Bayesian network efficiently encodes the joint probability distribution of its variables. For an *n*-dimensional random variable $(X_1,\cdots,X_n)$, the joint probability distribution is determined as in Eq. (2) [9]:

$$P=(x_i,\cdots,x_n)=\prod P\left(x_i/pa(x_i)\right) \quad (2)$$

Here x_i represents the value of the random variable X_i and $pa(x_i)$ represents the value of the parents of X_i. Thus, the structure learning problem of a Bayesian network is equivalent to the problem of searching the optimum in the space of an all directed acyclic graph (DAG) [2]. The question of which nodes considered in the Bayesian network analysis will have connection in between and identifying the directions of these connections is of great importance. One of the most common methods is to consult an expert and another method is using structure learning algorithms [4]. In this study, an alternative way is proposed based on simple correlation analysis and according to the variable type either simple linear regression analysis or logistic regression analysis for constructing a Bayesian network.

3. SIMPLE LINEAR REGRESSION ANALYSIS

The regression methodology models the distribution of a variable, called response, with the help of one or more predictor variables. The equation of the simple linear regression model for investigating a relationship of a response variable Y with a predictor *X* is shown in Eq. (3). In this equation α denotes the constant coefficient that shows the intercept, β denotes the coefficient that shows the slope, and ε denotes the error [6]:

$$Y_i=\alpha+\beta X_i+\varepsilon \quad (3)$$

There are four principal assumptions that justify the use of simple linear regression models: Linearity of the relationship between dependent and independent variables, independence of the errors (no serial correlation), homoscedasticity (constant variance) of the errors versus time, and normality of the error distribution [6].

4. LOGISTIC REGRESSION ANALYSIS

Logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logit function logistic curve. It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical [1, 8]. The dependent variable can be a

categorical variable with two categories (mela/female, live/die, has disease/doesn't have disease) or a continuous variable that has values in the range 0.0 to 1.0, representing probabilities or proportions. An explanation of logistic regression begins with an definition of the logistic function given in Eq. (4):

$$f(z)=\frac{e^{z}}{e^{z}+1}=\frac{1}{1+e^{-z}} \tag{4}$$

5. SIMPLE CORRELATION ANALYSIS

Correlation measures the extent of correspondence between the ordering of two random variables. Pearson's correlation coefficient, also known as the linear product moment correlation coefficient, is generally used when variables are quantitative. It is defined by Eq. (6):

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{\left(n\sum x^{2}-\left(\sum x\right)^{2}\right)\left(n\sum y^{2}-\left(\sum y\right)^{2}\right)}} \tag{6}$$

6. K-FOLD CROSS VALIDATION METHOD

The k-fold cross-validation is a method used for measuring classification accuracy. In the method, the complete dataset is randomly partitioned in k mutually exclusive subsets (called folds) $D_1, D_2, \cdots, D_k$ of approximately equal size [7, 10]. The induction algorithm is trained and tested k times in the following way: in the first iteration, this algorithm is trained on subsets $D_2, \cdots, D_k$ and tested on subset D_1; in the second iteration, the algorithm is trained on subsets $D_1, D_3, \cdots, D_k$ and tested on subset D_2; and so on. The overall number of correct classifications from the k iterations divided by the size of the complete dataset is the accuracy estimate of the k-fold cross-validation. In this study we evaluated the algorithms using 2-fold cross validation. Each model is learned with randomly selected training examples from each of the two training sets. The learned models are then applied to their corresponding testing fold, and classification accuracy is measured for each. Then we average the accuracies over the two training/test splits.

7. COMPARISON OF THE MODELS

7.1. Data encoding

To show that considering simple correlation in forming the Bayesian network structure and one of the simple linear regression or logistic regression according to the types of the variables together gives effective results, the blood test results taken from 76 thyroid disease patients who appealed to one of the major hospitals in Turkey, had never been cured from thyroid disease and don't have any other systemic disease are considered. It is identified that 60 of these patients are healthy and 16 are thyroid patients. To create the model, along with sex and age, the TSH, FT4, glucose, urea, creatinine, cholesterol, triglyceride, HDL, LDL, AST, and ALT values obtained as the result of the blood test are considered as "input" variables; the "conclusion" variable

where the data stating if the patient is a thyroid disease patient is considered as the "output" variable. The "sex" variable that stands between the output variable and the input variables is in categorical form, and the remaining input variables are in continuous form. The simple linear regression and logistic regression used to form the network structure of the proposed model are performed over the unclassified original values of the above-mentioned variables. The Bayesian network analysis that will help with diagnosis is performed over the categorical states obtained using the coding structure given in Table 1. The distribution of the data to the train and test sets is performed according to the 2-fold cross-validation method. Thus, both groups of 60 healthy and 16 thyroid disease patients are divided into two equal parts. Therefore, two data sets each containing data belonging to a total of 38 patients, 30 healthy and 8 with thyroid disease, are formed. In the first repetition of the 2-fold cross validation, one of the formed sets is considered as a train set and the other as a test set, and vice versa in the second repetition. In the study, the "conclusion" variable is used as the output variable. It has "ill" and "healthy" levels, labeled respectively 1 and 2. The description and the order of the input variables are given in Table 1.

Table 1. Description and order of input variables.

Variable	Label	Value	Variable	Label	Value	Variable	Label	Value
G: gender	1: Male 2: Female	- -	P: procreate	1: Low 2: Normal 3: High	13-16 17-43 44-49	LDL	1: Low 2: Normal 3: High	41-59 60-130 131-287
A: age	1: Young 2: Middle Age 3: Elderly	20-34 35-50 51-79	C: creatinine	1: Low 2: Normal 3: High	0.0-0.3 0.4-1.4 1.5-1.7	AST	1: - 2: Normal 3: High	- 0-35 36-213
TSH	1: Low 2: Normal 3: High	0.015-0.33 0.34-5.6 5.7-306.3	CH: cholesterol	1: Low 2: Normal 3: High	95-119 120-200 201-402	ALT	1: - 2: Normal 3: High	- 0-41 42-265
FT4	1: Low 2: Normal 3: High	0.15-0.57 0.58-1.64 1.65-3.78	T: triglyceride	1: Low 2: Normal 3: High	37-49 50-200 201-835			
GS: glukose	1: Low 2: Normal 3: High	65-73 74-106 107-231	HDL	1: Low 2: Normal 3: High	24-29 30-70 70-93			

7.2. Simple Model

In normal conditions, diagnosis of the patients who sought medical advice as thyroid patients or healthy is done considering the results obtained from the measurements done on the patients. In the study, the Bayesian network model named "simple model" is created for this situation. In this model, no connection is established between the considered input variables, and all of the input variables are associated directly with the result variable. The state of the simple model which is trained according to the first repetition of the 2-fold validation method is as in Fig. 1.

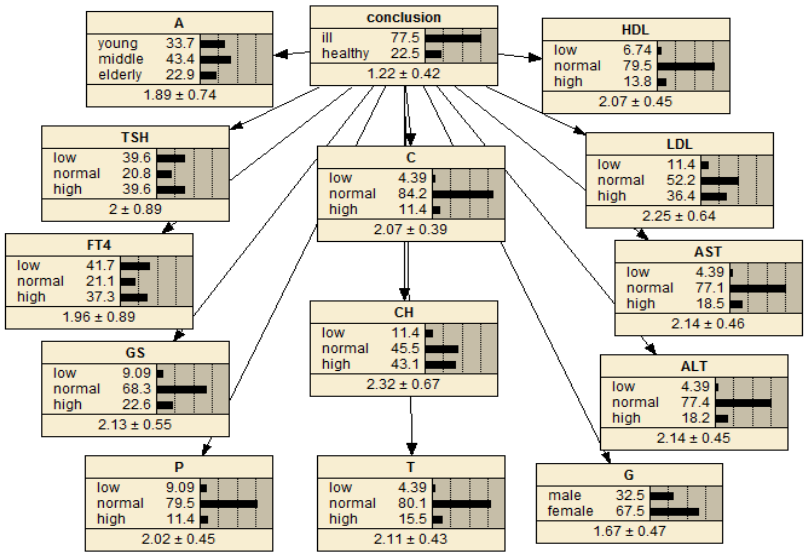


Fig. 1. Simple model network structure.

The results obtained by performing the test process by choosing the "conclusion" node in the simple model according to the 2-fold cross validation method are as in Table 2.

Table 2. Classification results for simple model.

First Repetition			Second Repetition			Average		
Actual	ill	healthy	Actual	ill	healthy	Actual	ill	healthy
ill	27	3	ill	25	5	ill	26	4
healthy	1	7	healthy	4	4	healthy	2.5	5.5
Error rate = 10.53%			Error rate = 23.68%			Error rate = 17.11%		

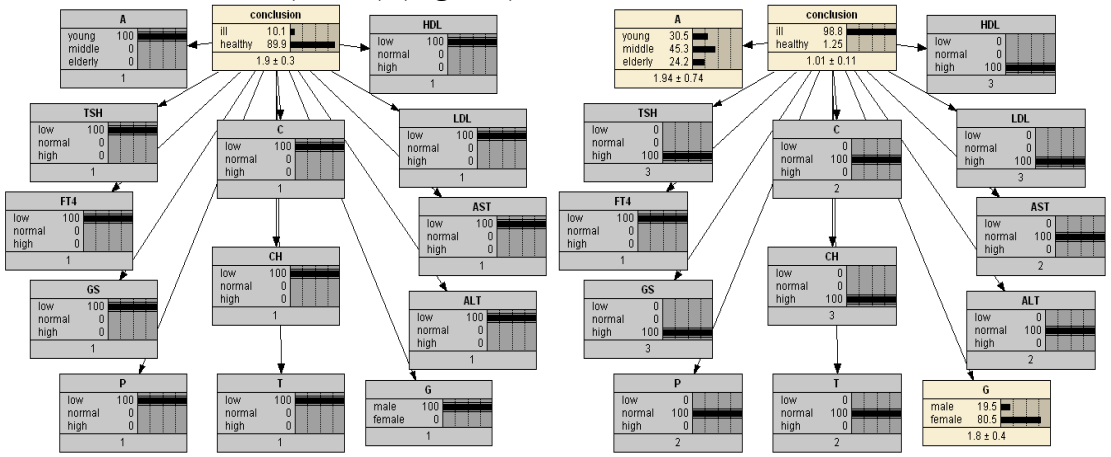
According to the table, it is seen that the error rate obtained from the classification made for the simple model as the result of the 2-fold cross validation is 17.11%. For the first and second repetitions, quality of test is given in Table 3.

Table 3. Quality of test for simple model.

First Repetition					Second Repetition				
Cutoff	Sensitivity	Specificity	Predictive	Predict-Neg	Cutoff	Sensitivity	Specificity	Predictive	Predict-Neg
0	100	0	78.95	100	0	100	0	78.95	100
80	63.33	100	100	42.11	70	70	62.5	87.5	35.71
95	10	100	100	22.86	95	43.33	75	86.67	26.09
100	0	100	100	21.05	99.9	6.67	100	100	22.22
					100	0	100	100	21.05

According to the table, four cutoffs are performed for the first repetition, whereas five cutoffs are performed for the second repetition. Various results can be obtained according to the network structure in Figure 1 and as the result of the first repetition of the 2-fold cross validation. For example, suppose that we've selected the "young" level of the input variable A, the "male" level of the input variable G, and the "low" level of all the remaining input variables. In this case, the simple model will give the ratio of the patients appearing in the "ill" and "healthy" levels of the "conclusion"

output variable as 10.1% and 89.9% (Fig. 2-a). Another result that can be obtained is in the simple model where TSH is selected as high, FT4 is selected as low, GS is selected as high, P is selected as normal, C is selected as normal, CH is selected as high, T is selected as normal, HDL is selected as high, LDL is selected as high, AST is selected as normal, ALT is selected as normal; then it is determined that the "conclusion" output variable is "ill" (98.8%), the input variable G is "female" (80.5%), and the input variable A is "middle" (45.3%) (Fig. 2-b).



(a) (b)

Figure 2. Inferences for simple model.

7.3. Proposed Model

In this study, a model where there are connections created between the input variables in addition to the connections created in the simple model is proposed. In the creation of these connections, first correlation analysis, then either one of the simple linear regression or logistic regression analyses is made use of. For this purpose, we examine if the considered variables verify the assumptions mentioned in Section 3. Logarithmic transformation is applied on the variables because it is determined that the variables are not normally distributed. As a result, it is seen that the assumptions are verified and all analyses are performed over these created logarithmic variables. It is started off by performing correlation analysis where the variables are considered as nodes. This way, it is determined which nodes are related to each other, in other words, between which nodes there should be connections created. Then, either simple linear regression or logistic regression is applied according to the data structure of the variables in order to decide the direction of the connections for the nodes. For this purpose, each of the variables that are designated as input variables are considered as dependent variables, one by one. For any input variable designated as a dependent variable, every remaining input variable is separately taught as an independent variable and we examine if this variable has an important effect on the dependent variable. Simple linear regression or logistic regression is performed bidirectionally for any two input variables. That is to say, any one of the variables is considered as the dependent variable and the other as the independent variable and the analysis is performed. Then vice versa, the one considered as the dependent variable is considered as the independent variable, the one considered as the independent variable is considered as

the dependent variable and the analysis is performed. This way, the β coefficients that will be used when deciding the direction of connection, in other words the direction of the arrow, in the connections that will be created between the nodes when the Bayesian network structure is being formed are calculated. According to this, the direction of the arrow is formed from the variable considered as the dependent variable towards the variable considered as the independent variable regarding in which equation the β coefficient is larger in terms of absolute value. Since the input variable G is categorical, a dummy variable is used in all the simple linear regression analyses where the variable in question is considered as the dependent variable. Logistic regression is performed in the case G is the dependent variable. As a result of the performed correlation, simple linear regression and logistic regression analyses, the β coefficients regarding the cases where it is determined that there is an important connection between the dependent and the independent variable (sig. $P < 0.05=\alpha$) are as given in Table 4.

Table 4. β coefficients for important independent variables (α=0.05).

Dependent Variable	Important Independent Variables and Their β Coefficients
A	GS (0.431), P (0.587), CH (0.459), T (0.216), LDL (0.304)
TSH	FT4 (-3.72), C (2.742), CH (6.269), T (1.785), LDL (4.375), AST (1.927),
FT4	TSH (-0,228), C (-1.834), CH (-0.004), T (1), LDL (-0.005), AST (-0.565)
GS	A (0.134), T (0.074)
P	A (0.474), C (0.477), CH (0.303), LDL (0.212), G (0.089)
C	TSH (0.103), FT4 (-0,224), P (0.373), CH (0.440), LDL (0.336), AST (0.141), G (0.128)
CH	A (0.321), TSH (0.058), FT4(-111.741), P (0.262), C (0.488), T (0.261), HDL (0.487), LDL (0.692), AST (0.201)
T	A (0.684), TSH (0.074), FT4(-90.419), GS (0.759), CH (1.184), LDL (0.673), AST (0.261)
HDL	CH (0.360), LDL (0.172)
LDL	A (0.401), TSH (0.760), FT4(-88.424), P (0.347), C (0.702), CH (1.305), T (0.280), HDL (0.440), AST (0.287)
AST	TSH (0.620), FT4(-0,267), C (0.544), CH (0.702), T (0.201), LDL (0.530), ALT (0.614)
ALT	AST (1.091)
G	P (-6.151), C (-11.247)

The direction of the arrows in the connections to be created in the proposed model referring to the information in Table 4 are as stated in Table 5.

Table 5. The direction of the arrows in the connections.

Variable	Connection Direction	Variable	Connection Direction	Variable	Connection Direction
A	A→GS	FT4	FT4→C	C	C←CH
	A→P		FT4←CH		C←LDL
	A→CH		FT4←T		C←AST
	A←T		FT4←LDL		C←G
	A←LDL		FT4←AST	CH	CH←T
TSH	TSH→FT4	GS	GS←T		CH→HDL
	TSH→C	P	P→C		CH←LDL
	TSH→CH		P→CH		CH←AST
	TSH→T		P←LDL	HDL	HDL←LDL
	TSH→LDL		P←G	LDL	LDL←AST
	TSH→AST	T	T→LDL	AST	AST←ALT
			T→AST		

The trained version with respect to the first repetition of the 2-fold cross validation method of the Bayesian network model that is created according to the results is as given in Fig. 3.

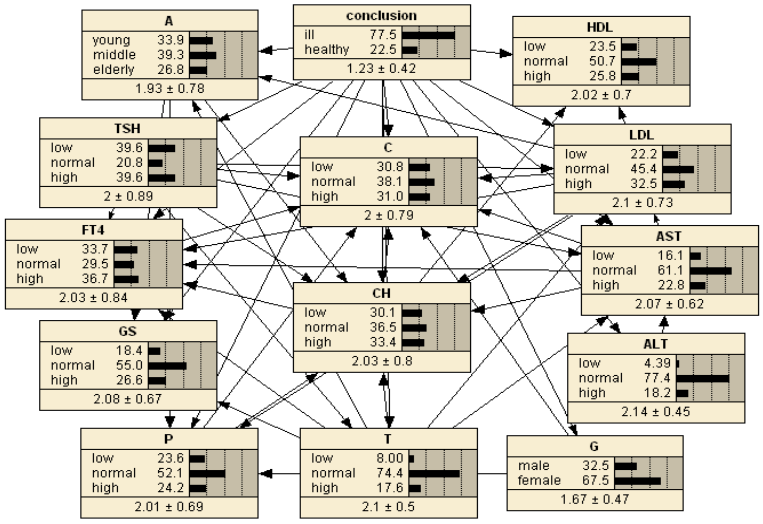


Fig. 3. Proposed model network structure.

The results of the test process performed where the "conclusion" node is chosen in the proposed model with respect to the 2-fold cross validation method is as given in Table 6.

Table 6. Classification results for proposed model.

First Repetition			Second Repetition			Average		
Actual	ill	healthy	Actual	ill	healthy	Actual	ill	healthy
ill	30	0	ill	30	0	ill	30	0
healthy	2	6	healthy	0	8	healthy	1	7
Error rate = 5.263%			Error rate = 0%			Error rate = 2.6315%		

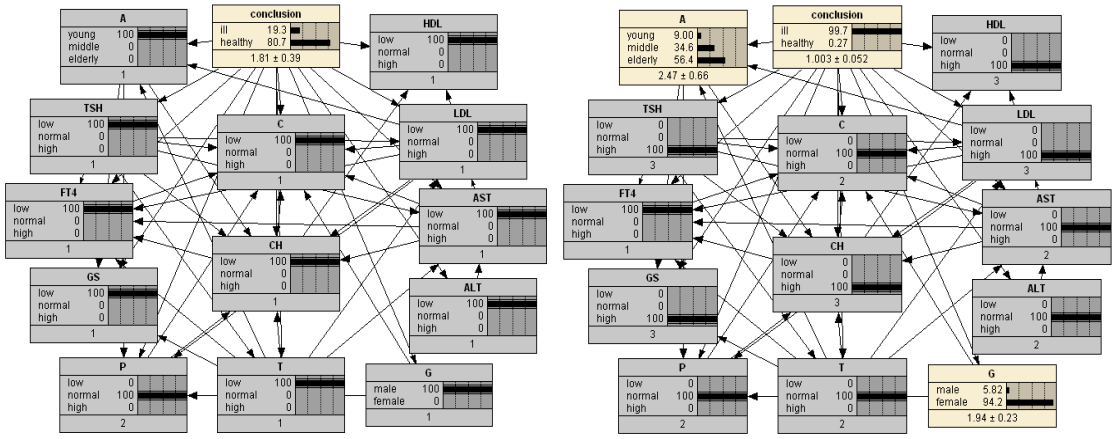
According to Table 6, the error rate of the classification done for the proposed model as the result of the 2-fold cross validation is 2.6%. For the first and second repetitions, the quality of the test is given in Table 7.

Table 7. Quality of test for proposed model.

First Repetition					Second Repetition				
Cutoff	Sensitivity	Specificity	Predictive	Predict-Neg	Cutoff	Sensitivity	Specificity	Predictive	Predict-Neg
0	100	0	78.95	100	0	100	0	78.95	100
99.5	53.33	100	100	36.36	98	63.33	100	100	42.11
100	0	100	100	21.05	99.5	26.67	100	100	26.67
					100	0	100	100	21.05

According to Table 7, three cutoffs are performed for the first repetition, whereas four cutoffs are performed for the second repetition. When the deduction state given in Figure 2 for the simple model is performed for the result obtained for the first repetition of the 2-fold cross validation method (Fig. 3), the network structures given in Figure 4-a and b are obtained. According to this, the "young" level of the input variable A, "male" level of the input variable G, and the "low" level of all the remaining input

variables are selected; according to the "conclusion" output variable, the ratio of the patients appearing in the "ill" level is 19.3%, the ratio of the patients appearing in the "healthy" level is 80.7% (Fig. 4-a). As a result of organizing the proposed model for the second condition performed for the simple, the obtained result is "ill" (99.7%) for the "conclusion" output, "female" (94.2%) for the input variable G, and "elderly" (56.4%) for the input variable A (Fig. 4-b).



(a) (b)

Fig. 4. Inferences for proposed model.

8. CONCLUDING REMARKS

It is of great importance, especially in the diagnosis of vital diseases, to make use of scientific methods in order to make more accurate diagnoses. For this purpose, researchers took the course of developing different models to be used in the diagnosis of any disease. In this study, a different hybrid model is proposed with the purpose of diagnosing thyroid disease. The proposed model is based upon Bayesian networks, and the relations created in the model and determining the directions of these relations depend on simple correlation analysis and, according to the data type, one of the simple linear regression or logistic regression analyses. In order to prove the success of the proposed model, the model in question is compared with a model called the simple model where no connections between the input variables are created and where all input variables are directly associated to the result variable. For this purpose, the blood test results taken from a total of 76 patients who appealed to one of the major hospitals in Turkey, had never been cured for thyroid disease and don't have any other systemic disease are used as the data set. Since the gathered data are low in number, the 2-fold cross validation model is applied and the comparison of the classification successes of the models is performed over the average of the results from every repetition. It is seen, as the result of performing the test process by choosing the "conclusion" node in the simple model according to the 2-fold cross validation method, that the average error rates are 17.1% for the simple model and 2.6% for the proposed model. This shows that the proposed model gives considerably better results in classification compared to the simple model. Also in this study, the ability of a Bayesian network to examine what result the examined variable or variables for the selected levels of the considered variable or variables will give is performed both for the simple and proposed models.

For this purpose, two cases are determined and the success of the models in question is compared. The values of the determined cases are found over real data. Then, the cases in question are carried through using the simple model and the proposed model. It is determined if these models give correct results, as a result of the comparison between the real values and the values obtained from these models. It is observed that the proposed model gives better results compared to the simple model, even 100.0% correct results for some cases. Eventually, it is observed that performing the connections that can be created for the Bayesian network model by simple correlation and establishing the directions of these connections by simple linear regression or logistic regression gives considerably good results.

9. REFERENCES

1. A. Agresti, *Categorical Data Analysis* (Wiley-Interscience, New York, 2002).
2. M. Ankush and K. Ashraf, *Bayesian Network Technologies: Applications and Graphical Models*, IGI Publishing, 2007.
3. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer Verlag, Berlin, 1985.
4. G. Cooper and E.A. Herskovits, A bayesian method for the induction of probabilistic networks from data, *Lach Learn* 9(4), 109-347, 1992.
5. J. DeRuiter, *Thyroid Hormone Tutorial: Thyroid Pathology, Endocrine Module* (PYPP 5260), Thyroid Section, 2002.
6. F.A. Graybill and H.K. Iyer, *Regression Analysis: Concepts and Applications*, Publisher: Duxbury Pr, 1994.
7. J. Han and M. Kamber, *Data Mining. Concepts and Techniques*, Morgan Kaufmann, 2001.
8. J.M Hilbe, *Logistic Regression Models*, (Chapman & Hall/CRC Press, 2009).
9. F.V. Jensen, *An Introduction to Bayesian Networks*, Springer Verlag, New York, 1997.
10. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, *14th International Joint Conference on Artificial Intelligence IJCAI'95*, Montreal, Canada, Morgan Kaufmann, 1995.
11. L.M. Kristen and S.D. Brown, Novel 'hybrid' classification method employing bayesian networks, *J. Chemometrics* 13, 579-590, 1999.
12. J. Pearl, Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference, San Mateo: Morgan Kaufman, 1988.
13. G.P. Zhang and V.L. Berardi, An investigation of neural networks in thyroid function diagnosis, *Health Care Management Science*, 1, 29-37, 1998.