PREDICTION OF MALIGNANCY IN SUSPECTED THYROID TUMOUR PATIENTS BY THREE DIFFERENT METHODS OF CLASSIFICATION IN DATA MINING

Saeedeh Pourahmad[1,3], Mohsen Azad[1] ,Shahram Paydar[2] and Hamid Reza Abbasi[2]

[1]Department of Biostatistics, School of Medicine,
Shiraz University of Medical Sciences, Shiraz, Iran
pourahmad@sums.ac.ir
azadmo@sums.ac.ir
[2]Department of Surgery and Trauma Research Center, School of Medicine,
Shiraz University of Medical Sciences, Shiraz, Iran
paydarsh@sums.ac.ir
abbasihr@sums.ac.ir
[3]Colorectal Research Center,
Shiraz University of Medical Sciences, Shiraz, Iran

ABSTRACT

In the present study, the abilities of three classification methods of data mining namely artificial neural networks with feed-forward back propagation algorithm, J48 decision tree method and logistic regression analysis are compared in a medical real dataset. The prediction of malignancy in suspected thyroid tumour patients is the objective of the study. The accuracy of the correct predictions (the minimum error rate), the amount of time consuming in the modelling process and the interpretability and simplicity of the results for clinical experts are the factors considered to choose the best method.

Keywords

Data Mining, Artificial Neural Networks, Decision Trees, Logistic Regression Analysis, Malignancy, Thyroid Tumour Patients

1. INTRODUCTION

Thyroid nodule is a common problem in population and decision making for type of management is subject of controversies. Management varies from observation to total thyroidectomy. Sequence of diagnostic procedures is based on management protocols of the center. FNA (Fine Needle Aspiration) of the nodule is one of the most useful tools in determining type of management of thyroid nodules but has some limitations in accurate report and some significant mistakes while decision making, are made [1].

Therefore, it is necessary to help the physicians by data mining techniques and search deeply in patients' attributes to find the meaningful relations and develop the diagnosis process.

Sundarapandian et al. (Eds): ICAITA, SAI, SEAS, CDKP, CMCA, CS & IT 08,
pp. 01–08, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2501

There are different classification methods in data mining techniques [2]. Some of them are deeply dependent to underlying theoretical assumptions such as linear and logistic regression models. They are called parametric methods. And some the others are assumptions free like artificial neural networks, decision trees, K-nearest neighbourhood, etc.

Recently, non parametric methods of data mining techniques which are not depending on theoretical distributional assumptions receive more attention in practice. The real dataset seldom follows the underlying theoretical assumptions of parametric methods. Clinical dataset is such an example. The avoidance of ideal theoretical assumptions in one hand and the variability in the nature of clinical dataset and their vague relations on the other hand cause this favourability. The attribute nature of biological data and their vague relation does not consist with theoretical distributional assumptions of parametric methods. For instances, dichotomous attribute in logistic regression analysis which is modelled based on the other attributes should follow Bernoulli theoretical distribution. The independency of error terms of models, the independency of other attributes in the model and enough sample size are the other ideal theoretical assumptions of these modelling methods. The application of such methods and the accuracy of their results depend on fulfilment of their ideal assumptions. In comparison, other methods such as artificial neural networks and decision trees use the learning process from a set of existent prototypes without any specific underlying assumptions. Therefore, the relation among the attributes is discovered from a part of the dataset (training set) and the parameters are estimated in such a way that the error prediction is minimized. Then, the power of model's prediction is evaluated by the other part of dataset (testing set).

These two mentioned methods have some advantages and disadvantages in their own [3]. In general, decision trees will require much less training time than neural networks. However, despite of neural network methods, decision tree techniques are less sensitive to the noise. In other words, the hidden layers in neural network discover complicated relations of the data and assign more weights to the important attributes. Nevertheless, their performance depends on the dataset. For some data, much more time and complicated computing is needed to train a neural network or/and the results are hardly interpretable by the expert of that field. For some other data, decision tree techniques are not able to discover reasonable relations. There are some studies which compare the abilities of these two methods and also, with logistic regression analysis in different research fields. For instances see [4, 5, 6].

In this paper, we compare feed forward back propagation neural networks and a well-known algorithms of decision tree namely J48 on our clinical dataset. A feed forward back propagation neural network repeatedly examines all the training data in the process of updating its weights [7]. The mentioned decision tree learning algorithms recursively partitions the training data into ever smaller subsets on which a test is made [8].

2. METHODS AND MATERIALS

In this section, three applied classification methods are described briefly. Then, the clinical diagnosis problem and the patients' population used in practical part of the study are explained at the end.

2.1. Artificial Neural Networks

Neural networks are the branch of artificial intelligence. Their models are inspired by the neural systems of human brain. And have been applied in many research fields such as biology, psychology, statistics, mathematics, medical science, computer sciences, and also, a variety of business areas like finance, management and decision making, marketing and production [9]. Recently, artificial neural networks (ANNs) become a very popular model to diagnose disease. However, ANNs have some disadvantages and some advantages for medical analysis. Their

discrimination power, discovery the complex and nonlinear relationship among the attributes, and prediction of the cases are the most important advantages of ANNs. Nevertheless, they can be over-fitted for training data, and time consuming because of computational requirements [9]. The selection of an appropriate training algorithm, transfer functions, initial values of network weights and also, the number of parameters and hidden layers to define the network size determine the performance of an ANN. But compared to logistic regression analysis, neural network models are more flexible [10].

In this study, we use a type of neural networks namely feed-forwards network with back propagation algorithm to model the relations among attributes in our clinical dataset. A feed-forward back propagation neural network is a supervised network. That is, it uses training and testing data to build a model. The data involves a set of input attributes with their corresponding output. The network uses the training data to "learn" how to predict the known output, and the testing data is used for validation. The aim is to predict the output for any given inputs in such a way that the distance between the observed and predicted outputs becomes minimized. This algorithm repeatedly examines all the training data to update its weights. These weights are adjusted during training and the process is only in the forward direction through the network without any feedback loops [9].

The simplified process for training a feed-forward back propagation network is as follows [9]:

1. Input data is entered to the network and propagated through the network until it reaches the output layer. This forward process produces a predicted output.
2. The predicted output is subtracted from the actual output and an error value for the networks is calculated.
3. The neural network then uses supervised learning, which in most cases is back propagation, to train the network. Back propagation is a learning algorithm for adjusting the weights.
4. Once back propagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimized.

Equations 1 to 3 summarize the formulations as follows:

$$a^{m+1} = f^{m+1}(w^{m+1}a^m + b^{m+1}) \qquad for\ m = 0, \dots, M-1 \quad (1)$$

Where,

$\{p_1, t_1\}, \{p_2, t_2\}, \dots, \{p_Q, t_Q\}$ is the data set (p_i and t_i are the i^{th} input and output respectively), and M is the number of layers with f^m as the transfer function, w^m as the weight and b^m as the bias in the m^{th} layer. The input of the first layer is the network input and the output of the last layer is the network output.

$$a = a^m, a^0 = p \qquad (2)$$

The parameters (w^m and b^m) are estimated in such a way that the mean square errors (the mean distances between the observed and estimated outputs, t_i and a_i respectively) is minimized.

$$F(x) = E(e^T e) = E((t-a)^T(t-a)) \qquad (3)$$

2.2. Decision Trees

Decision tree is a typical method for the classification of objects in to decision classes [12]. A decision tree classifier is a function as follows:

$$dt : dom(X_1) \times dom(X_2) \times \dots \times dom(X_n) \rightarrow dom(Y) \qquad (4)$$

In which,

$X_1, X_2, ..., X_n$ are input attributes and Y is the output, where X_i has domain dom (X_i) and Y has domain dom (Y). We assume without loss of generality that dom (Y)={Y_1, Y_2} (a dichotomous discrete attributes).

A decision tree is a directed, acyclic graph T in a form of a tree. Each node in a tree has either zero or more outgoing edges. If a node has no outgoing edges, then it is called a decision node (a leaf node); otherwise, a node is called a test node (or an attribute node). Each decision node N is labeled with one of the possible decision classes $Y \in \{Y_1, Y_2\}$. Each test node is labelled with one input attribute $X_i \in \{X_1, X_2, ..., X_n\}$ i.e. called the splitting attribute. Each splitting attribute X_i has a splitting function f_i associated with it. The splitting function f_i determine the outgoing edge from the test node, based on the attribute value X_i of an object O in question. It is in form of $X_i \in Y_i$ where $Y_i \subset dom(X_i)$; if the value of the attribute X_i of object O is within Y_i , then the corresponding outgoing edge from the test node is chosen.

The problem of decision tree construction is as follows: Given a data set $X = (x_1, x_2, ..., x_n)$, where x_i are input random samples from an unknown probability distribution P, find a decision tree classifier T such that the misclassification rate $R_T(P)$ is minimal [12].

2.3. Logistic Regression Analysis

Logistic regression is also called logistic model or logit model, is a type of predictive model in which the target variable (output attribute) is a dichotomous variable for instances healthy or unhealthy, dead or alive, win or loss, etc. Logistic regression is used for the prediction of the probability of occurrence the desired event from two existent ones in output variable by fitting the data into a logistic curve. Like many forms of regression analysis, the input attributes may be either numerical or categorical. For example, the probability that a person has a heart attack in a specified time that might be predicted from the knowledge of person's age, sex and body mass index [9]. Logistic regression is widely applied in the medical sciences.

The output attribute Y, of a subject can take one of two possible values, denoted by 1 and 0 (for example, Y=1 if a disease is present; otherwise Y=0). Let X=(x_1, x_2,…, x_n) be the vector of input attributes. The logistic regression model is used to explain the effects of the input attributes on the probability of occurrence the value 1 for output Y.

$$Logit\{P(Y=1)\} = \log\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\} = b_0 + b_1x_1 + \cdots + b_nx_n \qquad (5)$$

Where, P stands for probability, b_0 is called the "intercept" and b_1, b_2,… are called the "regression coefficients" of x_1, x_2, … respectively. Each of the regression coefficients describes the importance of corresponding input attribute on output. A positive regression coefficient means that this input increases the probability of outcome, where as a negative regression coefficient means that the considered input decreases the probability of outcome. In addition, the absolute value of the coefficient detects its effect on the probability of outcome. A large values means strongly influences and a non-zero regression coefficient means little influence on the probability of outcome [9].

Analysis of our clinical data was done by Matlab 2008a for artificial neural network method, WEKA software, version 3.7.1 for decision tree technique and SPSS software, version 16.0 for logistic regression analysis.

The accuracy rate in prediction for these three methods is calculated to determine the best predictive methods.

2.4. Patients' Population and Clinical Problem

The malignancy or benignity of thyroid nodule is sometimes in ambiguity for the physicians. Only based on the pathology result after the surgery and removal of the thyroid tumour, the type of tumour is certainly determined, whereas it is better to avoid the unnecessary surgery. Although there are various factors which help the physician in diagnosis before the surgery but, the ultimate decision is still in ambiguity. For instance, FNA (Fine Needle Aspiration) of the nodule is one of the most useful tools in determining type of tumour thyroid but it has some limitations in accurate report and some significant mistakes while decision making, are made [1].

In the present study, all the patients in two recent years (2011 & 2012) which are referred to Shahid Rajaee hospital in Shiraz, Iran for the surgery of thyroid tumour are participated to our study. During this time, 259 cases with positive FNA result are entered the study. Most of them are female (211 versus 48 cases) with overall mean age of 42.3± 13.6 (41±13 for female and 47.9±15 for male). Some patients' characteristics are considered as the inputs attributes such as gender, age, thyroid nodules size, tumour size, type of operation, the duration of the disease, patient family history,The dichotomous output is the malignancy or benignity of thyroid nodule derived from the pathology results after the surgery.

3. RESULTS

Three explained methods applied to the clinical dataset and the results are summarized separately as follows:

3.1 Artificial Neural Network Results

Table 1 explains the information of the network trained by the thyroid tumour data. From the 259 cases, 75 percents (196 cases) which were chosen randomly are used in training set and 25 percents (63 cases) are used for validation. The absolute values of the final updated weights on the 14^{th} step determine the important attributes on the malignancy of the tumour. According to the trained network results, the first five important attributes are Multiple Nodule, Cancer Family History, Size of Left Lobe, Lobectomy and Type of Operation, respectively.

Table 1. The information of trsined artificial neural network on thyroid tumour data

Network type	Feed-forward
Algorithm	Back Propagation
No. of inputs	29
No. of output	1
No. of hidden layer	1
No. of neuron in hidden layer	10
Transfer function	Log-sigmoid
Iteration	50
Epochs	1000
Convergence precision	0.00001
No. of step to convergence	14

The accuracy of the prediction is 98 percents for training set and 92 percents for validation set. These results confirm the power of the trained network in prediction. Table 2 shows the prediction results. The target output is the observed result from pathology which detects the real status of the patient after surgery and the estimated output is derived from the trained network.

Table 2. The estimated output versus the observed one by the trained neural network

Estimated output / **Target**	**In training set**		**In validation set**	
	Malignant	Benign	Malignant	Benign
Malignant	81	2	32	3
Benign	1	112	2	26

3.2 Decision Tree Results

150 cases were randomly chosen from 259 cases to train the decision tree. J48 algorithm is used and the results are summarized in Tables 3 and 4. The first five important attributes near the root of the tree are Size of Left Lobe, Type of Operation, Multiple Nodule, Encapsulation and Cancer Family History, respectively. Almost near to neural network results but with lower accuracy rate (80 percents in training and 75 percents in validation set).

Table 3. The estimated output versus the observed one by the derived decision tree

Estimated output / **Target**	**In training set**		**In validation set**	
	Malignant	Benign	Malignant	Benign
Malignant	43	12	52	11
Benign	18	77	16	30

Table 4. The information of derived decision tree on thyroid tumour data

Root mean squared error	0.4826
Relative absolute error	99.1188 %
Root relative squared error	101.0544 %
Coverage of cases (0.95 level)	98.7013 %
Mean rel. region size (0.95 level)	98.7013 %
Total Number of Instances	259

3.3 Logistic Regression Analysis Results

In our clinical dataset for logistic regression analysis, the real type of tumour for each case from pathology result is considered as the binary output of the model. Therefore, Y=1 for malignant and Y=0 for benign tumour. And 29 patients' characteristics are considered as the input vector of attributes, i.e. $X=(x_1, x_2,\ldots, x_{29})$. Conditional forwards method is used as the variables' entrance method to the model. No significant relation found among the data. That is, none of the input attribute enters the model and logistic regression analysis was not able to find the significant relation between the inputs and the output. Several methods of variables' entrance were also used without any significant results. Table 5 and Eq. 6 describe the estimated model containing the only coefficient as the intercept. Coefficient of determination which detects the model ability to explain the output's variability is 0.69. It means that without any information of these 29

attributes of the patient, 69 percents of tumor malignancy is estimable. This result is not confirmed by the physician.

Table 5. Estimated coefficient of logistic regression fitted on thyroid tumor dataset

Variables in the equation	B	Standard Error	Wald statistics	Degree of freedom	Significance	EXP(B)
Constant	-0.974	0.139	48.868	1	<0.001	0.378

$$\log\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\} = -0.974 \qquad (6)$$

4. CONCLUSIONS

As mentioned earlier, the ideal theoretical assumptions' of parametric classification methods bring them some limitations in practice. In other words, the real dataset seldom fulfil their assumptions. For logistic regression classification method for instance, in addition of underlying distributional assumptions such as Bernoulli probability distribution for observed dichotomous outputs, independency of observed inputs, independency and identically distribution of error terms and enough sample size in both categories of output, the number of inputs is important too. Usually, parametric methods are seldom able to manage more than 30 attributes. In the clinical example of our study, the theoretical assumptions may not be held. Furthermore, 29 input attributes with 12 categorical variables which are entered to the model by 27 indicator variables may seem out of the model's ability. And this fact may cause to the strange results far from the physician's expectance.

In comparison, two nonparametric classification methods discussed in the present study are more flexible and nearer to the real data circumstances. In our clinical dataset, neural network method consumed more time for computation with complicated results not simply interpretable by the clinical experts. Decision tree method in contrast, represents the simple decision rules in shorter time. However, it is less sensitive to the noise (irrelevant attribute) with lower accuracy rate than neural network.

To choose the best method, authors suggest further attempt to test more networks with different algorithms and transfer functions and also various algorithms of decision tree on the discussed dataset. However, it depends to the physician's preference too.

ACKNOWLEDGEMENTS

The authors would like to thank trauma research center of Shahid Rajaee hospital in Shiraz University of Medical Sciences and its staff, for their valuable cooperation in data gathering process.

REFERENCES

[1] Papageorgiou E, Kotsioni I & Linos A,(2005) "Data mining: a new technique in medical research", Hormones, Vol. 4, No. 4, pp 189-191.
[2] Seifert J.W. (2010) "Data Mining and Homeland Security: An Overview", BiblioGov.
[3] Lawrence O. Hall, Xiaomei Liu, Kevin W. Bowyer2 & Robert Ban_eld, (2003) "Why are neural networks sometimes much more accurate than decision trees: an analysis on a bio-informatics problem", IEEE International Conference on Systems, Man & Cybernetics, Washington, D.C., pp. 2851-2856, October 5–8.

[4] Kazemnejad A, Batvandi Z & Faradmal J, (2010) "Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes", Eastern Mediterranean Health Journal, Vol. 16, No. 6, pp 615-620.
[5] Kue W.J, Chang R.F, Chen D.R, et al. (2001) "Data Mining with decision trees for diagnosis of breast tumor in medical ultrasonic images", Breast Cancer Research and Treatment, Vol. 66, pp 51-57.
[6] Sakai S, Kobayashi K, et al. (2007) "Comparison of the levels of accuracy of an artificial neural network model and a logistic regression model for the diagnosis of acute appendicitis" J Med Syst, Vol. 31, pp 357–364.
[7] Anthony M. & Bartlett P, (1999) "Neural Network Learning: Theoretical Foundations", Cambridge University press.
[8] Quinlan J.R, (1996) "C4.5: Programs for Machine Learning", Morgan Kaufmann, San Mateo, CA.
[9] Raghavendra B.K & Srivatsa S.K, (2011) "Evaluation of logistic regression and eural network model with sensitivity analysis on medical datasets", International jcomputer science and security, Vol 5, no, 5. pp 503-511.
[10] Tasdelen B, Helvaci S, Kaleagasi H & Ozge A, (2009) "Artificial neural network analysis for prediction of headache prognosis in elderly patients", Turk J Med Sci , Vol. 39, No. 1, pp 5-12.
[11] Zhang J, Lok T, R.Lyu M, (2007) "A hybrid particle swarm optimization-back-Propagation algorithm for feedforward neural network training", Applied Mathematics and Computation, Vol. 185, pp1026-1037.
[12] Kokol P, Pohorec S, Stiglic G & Podgorelec V, (2012) "Evolutionary design of decision trees for medical application", WIRE Data Mining Knowl Discov, Vol. 2, pp 237-254.
[13] Shahbaz Khan F, Anwer RM, Torgersson O, Falkman G, (2007) "Data Mining in Oral Medicine Using Decision tree", World Academy of Science,Engineering and Technology, Vol. 37, pp 12-16.

Authors

Saeedeh Pourahmad is an assistant professor at the Biostatistics Department of Shiraz University of Medical Sciences in Iran. She obtained a B.Sc. in Statistics from Shiraz University in 2002. She also obtained an M.S. and a Ph.D. degree in Biostatistics from Shiraz University of Medical Sciences in 2004 and 2011 in Iran, respectively. Her research is modelling in Fuzzy environments, neural networks, nonlinear and linear relations in crisp environments and their clinical applications.

Mohsen Azad is a M.S. student at the Biostatistics Department of Shiraz University of Medical Sciences in Iran. He obtained a B.Sc. in Statistics from Shahid Bahonar University in Kerman (Iran) in 2010. He is currently involved with the present research.

Shahram Paydar is an assistant professor at the Department of Surgery in Medical School of Shiraz University of Medical Sciences in Iran. He obtained a MD d egree and Na tional board of surgery in 2000 and 2007 from Shiraz University of Medical Sciences in Iran, respectively. He is a full time member of Trauma Research Center in Rajaee Hospital of Shiraz, Iran, oct. 2007 and his research is in Trauma and emergency management and Thoracic surgery.

Hamid Reza Abbasi is an associated professor at the Department of Surgery in Medical School of Shiraz University of Medical Sciences in Iran. He obtained a MD degree and General Surgery, MD in 1994 and 1998 from Shiraz University of Medical Sciences in Iran, respectively. His research is in determining prognosis of acutely ischemic limb, correlation of sonographic measurement of IVC diameter, etc.