International Journal of Applied Operational Research
Vol. 4, No. 1, pp. 51-60, Winter 2014

Journal homepage: www.ijorlu.ir

Evaluation of correlation between risk factors of breast tumors by Multi Nominal Logit Model

M. Nasiri, S. Zojaji*, F. Boroumand, B. Minaei-Bidgoli, F. Minaei-Bidgoli

Received: 20 August 2013 ; **Accepted:** 16 December 2013

Abstract Breast cancer is a malignant tumor and the most common cancer among American women that starts from cells of the breast. The growth of the size of data exceeds the ability of humans to analyze the data. So we need some intelligent machines that can analyze datasets intelligently for us. We propose a Multi Nominal Logit Model to evaluate correlation between risk factors of breast tumor. Our model has 95% confidence. We use the Aranobidgol's breast tumor dataset.

Keywords Breast Cancer, Multi Nominal Logit Model, Data Mining, Regression, Prediction, Classification.

1 Introduction

Breast cancer is a malignant tumor that starts from cells of the breast. A malignant tumor is a group of cancer cells that may grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. The disease occurs almost entirely in women, but men can get it, too [1].

Breast cancer is the most common cancer among American women, except for skin cancers. The chance of developing invasive breast cancer at some time in a woman's life is a little less than 1 in 8 (12%) [1].

The American Cancer Society's most recent estimates for breast cancer in the United States are for 2010 [1]:

- About 207,090 new cases of invasive breast cancer will be diagnosed in women.
- About 54,010 new cases of carcinoma in situ (CIS) will be diagnosed (CIS is noninvasive and is the earliest form of breast cancer).

* **Corresponding Author. (✉)**
E-mail: sahba_zojaji@comp.iust.ac.ir (S. Zojaji)

M. Nasiri
PhD Student , Computer Engineering Department Iran University of Science and Technology Tehran, Iran.

S. Zojaji
MS, Computer Engineering Department Iran University of Science and Technology Tehran, Iran.

F. Boroumand
MS, Statistics and applications from Ferdowsi University of Mashhad, Iran, Mashhad.

B. Minaei-Bidgoli
Assistant Prof., Computer Engineering Department Iran University of Science and Technology Tehran, Iran.

F. Minaei-Bidgoli
Assistant Prof., MD. Gynecologist, head of Shahid Rajaei Hospital in Aranobidgol, Iran, Aranobidgol.

- About 39,840 women will die from breast cancer.

After increasing for more than 2 decades, female breast cancer incidence rates decreased by about 2% per year from 1998 to 2007. This decrease was seen only in women aged 50 or older, and may be due at least in part to the decline in use of hormone therapy after menopause that occurred after the results of the Women's Health Initiative were published in 2002 [1].

Breast cancer is the second leading cause of cancer death in women, exceeded only by lung cancer. The chance that breast cancer will be responsible for a woman's death is about 1 in 35 (about 3%). Death rates from breast cancer have been declining since about 1990, with larger decreases in women younger than 50. These decreases are believed to be the result of earlier detection through screening and increased awareness, as well as improved treatment [1].

The growth of the size of data and number of existing databases far exceeds the ability of humans to analyze this data, which creates both a need and an opportunity to extract knowledge from databases. Medical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data is often concerned with treatment of incomplete knowledge, with management of inconsistent pieces of information and with manipulation of various levels of representation of data [2].

Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Data mining tasks are included: Classification, Clustering, Association Rule Discovery, Sequential Pattern Discovery, Regression and Anomaly Detection [3, 4]. Figure 1 shows the process of knowledge discovery in databases (KDD).

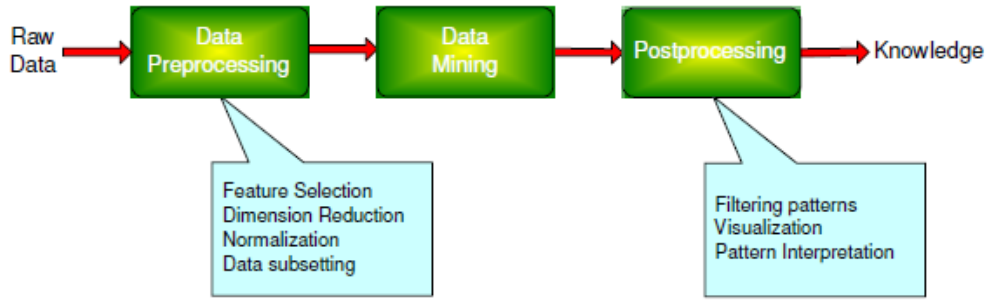

Fig. 1 The process of knowledge discovery in databases (KDD) [4]

Existing intelligent techniques of data analysis are mainly based on quite strong assumptions (some knowledge about dependencies, probability distributions, large number of experiments), are unable to derive conclusions from incomplete knowledge, or can not manage inconsistent pieces of information. The most commonly intelligent techniques used in medical data analysis are neural network, Bayesian classifier, genetic algorithms, decision trees, fuzzy theory [2].

We use the breast tumor data set that consists of breast cancer risk factors information. This data set has 450 records and 40 fields that belong to Aranobidgol's women information from April to June 2010.

In this paper, we propose Multi Nominal Logit Model (MNLM). In statistics, a multinomial logit model, also known as multinomial logistic regression, is a regression model which generalizes logistic regression by allowing more than two discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables [5-7]. Our model can predict new data with 95% confidence.

Rest of this paper is organized as follows. The next section reviews related work. Section 3 gives information about the data set that we used and the pre-processing of that. Our proposed method is discussed in section 4. Section 5 presents experimental results. Conclusion and future works are given in the last section.

2 Related work

Aboul Ella Hassanien and et al,[2] proposed a rough set method for generating classification rules from a set of observed 360 samples of the breast cancer data.

In another work, [8] used is the SEER Public-Use Data that the preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. It investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms and it showed that that C4.5 algorithm has a much better performance than the other two techniques [8]. After 4 years, in a similar work, [9] presented a new data pre-classification method after the preprocessing of the SEER Public-Use Data 2005, it investigates several algorithms. As a result, this paper [9], found c5 algorithm has the best performance of accuracy.

Orlando Anunciacao and et al, [10], used a Decision Tree to find a group with high-susceptibility of suffering from breast cancer. It explores the applicability of decision trees to this problem. A case-control study was performed, composed of 164 controls and 94 cases. Its goal was to find one or more leaves with a high percentage of cases and small percentage of controls [10].

In another work, [11] presented a comparison of different learning models used in Data Mining and a practical guideline how to select the most suited algorithm for a specific medical application and some empirical criteria for describing and evaluating learning methods were given. Three case studies, Breast Cancer, Diabetes Pima, IRIS, for medical data sets were used [11].

3 Dataset

We use the breast tumor data set that consists of breast cancer risk factors information. This data set has 450 records and 40 fields that belong to Aranobidgol's women information from April to June 2010. Most of variables of this data set is categorical,

except fields like name, family and age. And also these categorical variables are nominal and ordinal.

These data includes medical information about women whom had breast tumor that have been examined by a physician first and then the ultrasound, mammography, or pathology tests is performed based on physician's opinion. Among this collection contains some benign cancers, some malignant and some also have not had any cancer. This data set contains information about doctor's physical examination, the final result of ultrasound, mammography and pathology tests. This dataset is still being completed.

3.1 Data pre-processes

The data set contains empty values, out of range values, redundant fields and etc, which was pre-processed with appropriate treatment with each of them. Whereas we can communicate with the medical team that creating this data set, the medical team is also engaged in some stages of preprocessing task. We discuss all preprocessing tasks that we do to prepare data set for further post processing tasks in the following subsections.

3.1.1 Empty and out of range values

After consulting with medical team, we discover that all empty and out of range value in this data set is due to human errors. Thus these errors are fixed by them.

Categorizing field value

Some fields of this data set are categorical by themselves, like job. But there were some fields, such as age, which contains the continuous value. Since the Multi Nominal Logit Models require binary data, in the first step fields like that is changed to categorical fields.

3.1.2 Increase readability

In some cases, it was observed that there are some inconsistencies between field value and its meaning. For example, '0' is used for 'true' and '1' is used for 'false' or etc. this problem is caused low readability for human users. Hence in order to increase the readability such problems were fixed. For example, we do so for 'nipple dimpling' field.

3.1.3 Remove redundant fields

There were some redundant fields that caused collinearty in data set. By removing these fields before modeling, we avoid the occurrence of collinearty in our model. For example, removing the 'family history of cancer' field that has a strong correlation with fields like 'sister history of cancer', and etc.

3.1.4 Creating new fields

Some fields of this data set have multiple values, that leads to higher run-time or regression was impossible to do. Thus these fields split and create several new binary fields that are suitable for MNLM. For example, splitting the 'family history of cancer' field and create six new binary field from its values.

3.1.5. Data set binarization

To generating MNLM every independent variable, both qualitative and quantitative, changes in to two-level Bernoulli variable, yes or no, because the MNL model is the generalization of logistic model which is fitted to data with the nominal dependent variable with two categories [5-7]. For example, independent variable, age of menopause, first change into a categorical variable with five categories: 'lack of menopause', 'less than 40 years', 'from 40 to 45 years', 'from 46 to 50 years', 'above 50 years'. Then it is assumed that the first category, 'lack of Menopause', has two answers: Yes or No. The rest of independent variables also change into the Bernoulli variable.

3.2. Clean data set fields

After doing pre-processes above, we obtain a clean dataset for further processes. Table 1 shows these dataset fields.

Table 1 Clean data set fields, dependent and independent variables

Age	Family history of Mother	Menarche age
Occupation	Family history of Sister	Age of first pregnancy
Tumor location	Family history of sister of mother	No. Childbirth
Tumor formidability	Family history of sister of father	Breastfeeding
Pain in tumor	Family history of others	History of taking birth control pills
Adhesion around	Cancer type: Ovary	Menopause age
Tumor margin	Cancer type: Endometrial	History of radiation to the ovaries and uterus
Nipple Dimpling	Cancer type: Breast	
Orange skin view	Cancer type: Cervical	History of alcohol
Nipple Bleeding	Cancer type: others	Smoking history
BMI	Cancer history: Ovary	Mammogram findings ***(Dependent)***
	Cancer history: Endometrial	Ultrasound findings ***(Dependent)***
	Cancer history: Breast	Pathology findings ***(Dependent)***
	Cancer history: Cervical	
	Cancer history: others	

4 Proposed method

In this paper, we proposed a regression model to recognize the relationship between variables which could have an estimate of future observations. Our proposed model, can predict future observations with 95% confidence. In the following sections we introduce calculating the correlation between variables and also our regression model.

4.1 Calculating relation between variables

There are different ways to calculate the correlation between categorical variables that some of them expressed briefly below. In this paper, 'contingency coefficient' and the 'phi and Cramer's V coefficient' to determine the correlation between variables is used.

Spearman correlation coefficient

This correlation coefficient calculates the correlation between the ordinal data. This correlation coefficient that used for ordinal variables is the equivalent of Pearson correlation coefficient that used for quantitative variables [5].

Correlation measure for nominal and ordinal variables

Correlation between the nominal variables is calculated by the measures such as: contingency coefficient, phi and Cramer V coefficient, uncertainty coefficient, lambda and the correlation between ranking variables is measured by gamma, Sommer d, Kendall tau-b, and Kendall tau-c. The Contingency coefficient and phi and Cramer V coefficient are based on Chi-Square that is checking the independency criterion. The phi and Cramer V coefficient measure the correlation between two level categorical variables and the Contingency coefficient measures the correlation between more than two level categorical variables [5].

$$c = \sqrt{\frac{x^2}{N + x^2}} \quad (1)$$

$$Q = \sqrt{\frac{x^2}{N*(x-1)}}, 0 < Q < \sqrt{\frac{x-1}{x}} \quad (2)$$

$$\chi = \text{Min (i,j)}$$

Where both C and Q statistics are the Contingency coefficient and phi and Cramer V coefficient, respectively. N is number of samples and i and j are indexes of i-th and j-th sample.

The result of these measures for nominal variables are between 0 and 1 while for ranking variables are between -1 and +1 the same as quantitative variables [5].

4.3 The multi nominal logit model

In statistics, a multinomial logit model, also known as multinomial logistic regression, is a regression model which generalizes logistic regression by allowing more than two discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.). So, choosing the suitable regression model to fit to the data is based on dependent variable type. Based on this, when the dependent variable is a nominal variable with more than two levels, the suitable regression model is MULTI NOMINAL LOGIT MODEL, briefly MNLM. In fact this model is generalized logistic regression model that fits to data with the nominal dependent variables with two levels [5-7].

$$\Pr(y_i = j) = \frac{\exp(X_i \beta_j)}{1 + \sum_{j=1}^{J} \exp(X_i \beta_j)} \quad (4)$$

and

$$\Pr(y_i = 0) = \frac{1}{1 + \sum_{j=1}^{J} \exp(X_i \beta_j)} \quad (5)$$

where for the ith individual, yi is the observed outcome and Xi is a vector of explanatory variables. The unknown parameters β j are typically estimated by maximum a posteriori (MAP) estimation [5-7].

The most important challenge in using this MNLM is that this model involves many parameters and this engages a researcher with different complex conclusions, nonlinearity of the model is due to the difficult estimation of MNLM [5-7].

Estimation of this model has two levels. It means that dependent variable is estimated based on all independent variables separately and simultaneously. Thus, if the model has j independent variables, there exists j-1 level [5].

After doing all the pre-processing tasks as said before and preparing the data set to explore the important risk factors in breast tumor, the MNLM with 95% confidence threshold is implemented that you can see the results in the next section.

5 Implementation and experimental result

In this section we introduce some examples of the results. Data on breast tumor patients based on 'mammography findings', including 5 categories: negative, cyst, solid mass, benign, malignant. In fact, these five categories represent different categories of our dependent variable. Independent variables mentioned before, see table 1.

By first performing of regression model, we discover that existence of all Xi is not necessary, because the significance of them are greater than 0.05, so none of independent variables should be presented in the model.

This led to this hypothesis that all independent variables are independent of outcome variable. So independency test between each of independent and outcome variable is preformed two by two. By performing Chi square independency test, based on significance=0.05, those independent variables that were independent of the outcome variable is omitted from model and therefore the number of independent variables reduced to 13. For instance table 2 shows the independency test results between 'mammography findings' and 'Family history of Sister'.

Table 2 the independency test results between

Asymp. Sig. (2-sided)	df	Value	
.000	6	41.182(a)	Pearson Chi-Square
.071	6	11.621	Likelihood Ratio
.005	1	8.068	Linear-by-Linear Association
		413	N of Valid Cases

After that, we perform the MNLM again. The new obtained model was significant. The next subsection discusses about the results of our proposed model.

5.1 Discussion

Now we want to interpret the results of our model. We discuss some example below, because complete interpretation of our MNLM is so time consuming.

Based on the obtained model for the 'mammography findings', it is predicted that 'age' and 'menopausal age' are determinant variables. It means that if somebody has value 0 for menopause, non-menopausal, and after evaluating the age, it can be predicted that the findings of mammography shows the cyst.

Also for the 'solid mass' we can say that 'sister family history of cancer', 'history of breast cancer' and 'history of ovarian cancer' are determinant variables. As before said, if someone has these risk factors, we can say she is susceptible for a 'solid mass'.

Similarly, interpretation is the same for other independent and outcome variables values. They must interpret one by one. It means that each value of outcome variable with all of independent variable.

5.2 Residual analysis

One of the most common statistical errors is forgetting the residual analysis. Performing fundamental hypotheses on residual is meant to confirm the work [5-7]. In this article residual analysis was performed as follows:

First of all the residuals, ei (the difference of actual value and predicted value), calculates. Then, the four fundamental hypotheses are evaluated [5-7].

1. ei is independent;
2. ei is normal distributed;
3. Variance stability;
4. Average of residuals is zero.

Our model is more confident, after doing this, because we do residual analysis and all hypotheses are true.

6 Conclusion

We discover the relationship between risk factors and incidence of breast tumors and we propose a statistical model, MNLM with 95% confidence, to explore this correlation. To do this we use the Aranobidgol's breast tumor dataset. MNL model could offer acceptable results in terms of medical professionals. After this, we want to do these tasks to complete our proposed model or to propose a new model.

- Use factor analysis to reduce dimension of the data set;
- Use discriminant analysis to reduce dimension of the data set;
- Use PCA to reduce dimension of the data set;
- Compare these three methods in reducing the dimensions of data set;
- Discover association rules in order to understand the relationships between risk factors and incidence of benign and malignant breast tumor;
- Learn decision trees and use them for tumor screening and prevention of malignant tumors.

References

1. 1-800-ACS-2345, (2010). Breast Cancer, American Cancer Society, Atlanta.
2. Hassanien, A. E., Jafar M., (2004). Rough Set Approach for Generation of Classification Rules of Breast Cancer Data, Informatica, 15, 23-38.
3. Han, J., Kamber, M., (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, British Columbia.
4. Tan, P., Steinbach, M., Kumar, V., (2006). Introduction to Data Mining. Addison-Wesley, Massachusetts.
5. Agresti, A., (2002). Categorical Data Analysis. John Wiley & Sons, New Jersey.
6. Long, J., Freese, J., (2001). Regression Models for Categorical Dependent Variables Using STATA, STATA Press, Texas.
7. Daniel, A., Xie, Y., (1999). Statistical methods for categorical data analysis, Academic Press. Inc., New York.
8. Bellaachia, A., Guven, E., (2006). Predicting Breast Cancer Survivability Using Data Mining Techniques, Proceedings of the 2006 SIAM International Conference on Data Mining, April 20-22, SIAM, Philadelphia, 125-134.
9. Fan, Q., Zhu, C., Yin, L., (2010). Predicting breast cancer recurrence using data mining techniques. Proceedings of ICBBT 2010, Chengdu, 16-18 April 2010, IEEE, China, 310 – 311.
10. Anunciação, O., Gomes, B., Vinga, S., Gaspar, J., Oliveira, A., Rueff, J., (2010). A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups, Advances in Soft Computing, Volume 74, 43-51.

11. Andreeva, P., Dimitrova, M., Radeva, P., (2004). Data Mining Learning Models and Algorithms for Medical Applications, Proceedings of the 18-th Conference on Systems for Automation of Engineering and Research SAER 2004, Sofia, 24-26 September, Bulgaria, 11-18.
12. Nasiri, M., , Taghavi,L., ,Minaee.B(2011) , Numeric Multi-Objective Rule Mining Using Simulated Annealing Algorithm, International journal of operational research,Volume 1, Number 2, 16-23