International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 6, Issue 2, Feb - Mar 2017; 37-42

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

STATISTICAL ANALYSIS FOR MEDICAL DIAGNOSTICS USING FUZZY TECHNIQUES

V. CHANDAR[1], T. EDWIN PRABAKARAN[2] & N. VISWANATHAN[3]

[1,2]Department of Statistics, Loyola College, Chennai, Tamil Nadu, India

[3]Department of Statistics, Presidency College, Chennai, Tamil Nadu, India

ABSTRACT

Medical diagnosis of diseases is a highly complicated and time consuming process involving skilled resources and high-end diagnostic tools. This study aims at developing working rules based on sample observations from real-time data observed over a period of 6 months in an outpatient facility available at a Hospital in Chennai. The main objective is to identify the best classification procedure/method based on accuracy measures such as maximum correct classification rate and True Positive Rate (TPR). The data has been analyzed using two classification algorithms, namely Fuzzy Composition Rules (Fuzzy Max-Min, Max-Prod & Max-Avg) and Logistic Regression. The results of the analysis showed that Fuzzy Max-Product (or Max-Average) is the better model to predict the TB diagnostics based on the corresponding symptom features.

KEYWORDS: Classification, Fuzzy Logic, Max-Min Composition Rule, Logistic Regression

1. INTRODUCTION

Medical diagnosis of diseases is a highly complicated and time consuming process involving skilled resources and high-end diagnostic tools. The cost and time factors involved are disproportionately increasing with the scarce availability of skilled personals and rising cost of medical equipments. This study suggests ways and means to cut-short the cost and time factors and at the same time, maintaining a reasonably high degree of accuracy in the diagnostic process. This study aims at developing working rules based on sample observations from real-time data observed over a period of 6 months in an outpatient facility available at a Hospital in Chennai. In particular, the focus is mainly on Tuberculosis (TB) disease. Using sample observations, a classification rule has been developed based on the information (symptom + other demographic variables) collected from the hospital samples.

The objectives of this study are (1) Identifying significant factors associated with the disease, (2) Development of classification rules based on significant factors associated with the disease, and (3) Identification of the best classification procedure/method based on accuracy measures such as minimum misclassification rate / maximum correct classification rate, including True Positive Rate (TPR) and False Negative Rate (FNR).

2. METHODOLOGY

In this study, both Fuzzy approaches and Logistic regression analysis have been applied to analyze the data. In particular, three different composition rules of Fuzzy approaches have been applied to classify the patients as “affected by TB” or “not affected by TB”, in comparison with the classification by Binary Logistic Regression. The four predictive modeling techniques applied in this study are listed below.

1. *Fuzzy Max-Min Composition*
2. *Fuzzy Max-prod Composition*
3. *Fuzzy Max-av Composition*
4. *Logistic Regression*

2.1 Fuzzy Logic

2.1.1 Fuzzy Set in Medical Diagnosis

Many mathematical models have been developed by the creators of fuzzy set theory to be applied to different technical domains, including in the medical field to give answers to certain questions concerning a choice of diagnosis. The choice of diagnosis/disease should only be made on the basis of clinical symptoms (i.e., fever, cold, etc.) on the assumption that the symptoms are characteristic of all considered diagnosis/disease (Sanchez, 1978). In order to decide an appropriate diagnosis in one patient, the following three non-fuzzy sets are introduced.

The set of symptoms $S = \{S_1, S_2, \ldots, S_n\}$,

The set of diagnosis $D = \{D_1, D_2, \ldots, D_m\}$, *and*

The set of patients $P = \{P_1, P_2, \ldots, P_t\}$

The symptoms occurring in set S are associated with the diagnoses from set D. In addition, it has to be assumed that information about all symptoms belonging to S is complete in the patient's case. By treating his medical experience as a foundation, a physician then demonstrates connections between the symptoms and the diagnoses.

2.1.2 The Patient-Symptom Relation

Each symptom belonging to the set S will be represented as a fuzzy set. The three basic types of biological parameters are assumed as follows (Gerstenkorn & Rakus, 1992): 1. Simple qualitative Features, 2. Compound qualitative features, and 3. Quantitative (measurable) features.

2.1.3 The Symptom-Diagnosis Relation

The relation between symptom S_j and diagnosis D_k in each pair (S_j, D_k) is depicted as a value of the membership degree accompanied by this pair. The expression $R \subset S \times D$ is defined as the fuzzy relation between symptoms S and diagnoses D, where this relation is termed as medical knowledge expressing association between symptoms and diagnosis (Sanchez, 1979).

2.1.4 Numerical Representations of Linguistic Variables

In order to come-up with suitable numerical values for a certain parameters to be used to replace a linguistic variable with a reasonable membership degree, different values of the parameters were tested and obtained the representatives of the variables "never",…, "always" as shown in the following table (Rakus-Anderson, 2007).

Table 1: Numerical Description of Fuzzy Variables in "*presence*"

Fuzzy Variables	x	μ_common(x)
never	7.5	0
almost never	15	0.016
very seldom	22.5	0.062
seldom	30	0.14
rather seldom	37.5	0.25
moderately	50	0.5
rather often	62.5	0.75
often	70	0.86
very often	77.5	0.938
almost always	85	0.984
always	92.5	1

Table 1 provides the information on how to tie the words given in the list constructed for the presence of disease, to real numbers that replace them in the fuzzy relations "Symptom to Diagnosis", which we need to be generated.

2.1.5 The Patient – Diagnosis Relation

Fuzzy relations PS and SD are the two components of the equation of fuzzy relation that provides the solution termed as fuzzy relations of the type PD = "patient-diagnosis". This new relations contain pairs (P_j, D_k), j=1, 2, …,t; k = 1, 2, …, m. In order to obtain equations with the operation of max-min type, the fuzzy relations were developed as suggested in the definition of Max-Min Composition that is given below.

Definition of Max-Min Composition (Zadeh, 1965; Zimmermann, 1996)

Let X = {x_1, x_2, … , x_m}, Y = { y_1, y_2, … , y_m } and Z = { z_1, z_2, … , z_m }. We introduce $\tilde{R}$ with $\mu_{\tilde{R}}(x_i, y_j)$, $(x_i, y_j) \in X \times Y$, and $\tilde{Q}$ with $\mu_{\tilde{Q}}(y_j, z_k)$, $(y_j, z_k) \in Y \times Z$, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, $k = 1, 2, ..., p$, as two fuzzy relations. The max-min composition of $\tilde{R}$ with $\tilde{Q}$, denoted by $\tilde{R} \circ \tilde{Q}$, will be a fuzzy relation.

$$\tilde{S} = \tilde{R} \circ \tilde{Q} = \left\{ \left((x_i, z_k), \mu_{\tilde{R} \circ \tilde{Q}}(x_i, z_k) = \max_{y_i \in Y} \left\{ \min \left\{ \mu_{\tilde{R}}(x_i, y_j), \mu_{\tilde{Q}}(y_j, z_k) \right\} \right\} \right) \right\}$$

In addition to the above **Max-Min composition**, two special cases of the operation **max-*** are taken into consideration and are defined as follows.

Definition of Max-* Composition (Zadeh, 1965; Zimmermann, 1996)

Let $\tilde{R}$ with $\mu_{\tilde{R}}(x_i, y_j)$, $(x_i, y_j) \in X \times Y$, and $\tilde{Q}$ with $\mu_{\tilde{Q}}(y_j, z_k)$, $(y_j, z_k) \in Y \times Z$, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, $k = 1, 2, ..., p$, as two fuzzy relations. The max-prod composition $\tilde{R} \circ \tilde{Q}$ and the max-av composition $\tilde{R} \underset{av}{\circ} \tilde{Q}$ are proposed as fuzzy relations as follows:

$$\tilde{R} \circ \tilde{Q} = \left\{ \left((x_i, z_k), \mu_{\tilde{R} \circ \tilde{Q}}(x_i, z_k) = \max_{y_i \in Y} \left\{ \mu_{\tilde{R}}(x_i, y_j) . \mu_{\tilde{Q}}(y_j, z_k) \right\} \right) \right\} \text{ and}$$

$$\tilde{R} \underset{av}{\circ} \tilde{Q} = \left\{ \left((x_i, z_k), \mu_{\tilde{R} \underset{av}{\circ} \tilde{Q}}(x_i, z_k) = \frac{1}{2} . \max_{y_i \in Y} \left\{ \mu_{\tilde{R}}(x_i, y_j) + \mu_{\tilde{Q}}(y_j, z_k) \right\} \right) \right\} \text{ for } x_i \in X,\ y_j \in Y,\ z_k \in Z$$

2.2 Logistic Regression

Logistic regression was developed by a Statistician David Cox in 1958 (Walker and Duncan, 1967). Logistic regression, also called as logit regression or logit model, is a regression model where the response variable is categorical. In this study, the binary logistic regression has been applied since the outcome of the dependent variable has only two levels (Yes / No). Logistic regression is used to predict the odds of being a case (i.e., a diseased person) based on the values of the predictors, where the odds are defined as the ratio of the probability of a particular outcome to be a case over the probability of a particular outcome to be a non-case.

The logistic equation in terms of the probability that Y=1, which is referred to as $\hat{p}$. The probability that Y=0 is $1-\hat{p}$. The logistic regression equation is given as follows (Hosmer & Lemeshow, 1989).

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)=\hat{\beta}_0+\hat{\beta}_1 X$$

In the above equation, 'ln' refers to a natural logarithm and $\hat{\beta}_0+\hat{\beta}_1 X$ is a well known equation for the regression line. The probability P can be computed from the above equation as follows.

$$\hat{p}=\frac{\exp(\hat{\beta}_0+\hat{\beta}_1 X)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 X)} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

This study is focusing only on Pulmonary TB to analyze the general symptoms of this type of TB including the medical experts' knowledge and experience in measuring the level of symptoms.

3. EMPIRICAL DATA ANALYSIS

3.1 Sample Size

Based on the results of pilot study conducted in a private hospital, the sample size for the main study has been planned to be around 200. To accommodate sample attrition, the main study included 230 samples.

3.2 Identifying Significant Features

The collected data contains totally 26 independent variables and a dependent variable (TB or No-TB). Using Chi-square test for independence of attributes, the relationship between 'tb' and each of the 26 variables have been examined and the corresponding results showed that only 10 symptom features are significant variables to be used for model building.

In order to apply Fuzzy Composition Rule, the value of membership degree has been calculated for each symptom features using the following formula.

$$y=\mu_{S_j}(x)=s(x,\alpha,\beta,\gamma)=\begin{cases}0 & \text{for } x\le\alpha\\ 2\left(\frac{x-\alpha}{\gamma-\alpha}\right)^2 & \text{for } \alpha<x\le\beta \ \ \beta=\frac{\alpha+\gamma}{2}\\ 1-2\left(\frac{x-\gamma}{\gamma-\alpha}\right)^2 & \text{for } \beta<x\le\gamma\\ 1 & \text{for } x>\gamma\end{cases}$$

The weights attached to each level and the corresponding membership degrees are given as follows.

Table 2: Membership Degrees

X	1	2	3	4	5
Weight (x)	-1	-0.5	0	0.5	1
S(x)	0	0.125	0.5	0.875	1

The **Symptom-to-Diagnosis (SD)** information was obtained from a medical expert and the same has been replaced with the suitable membership degrees to get the SD matrix as shown below.

Table 3: Symptom-to-Diagnosis Matrix

	Experts' Verbal Scale		Membership Equivalence	
Symptom	**TB presence**	**TB absence**	**TB presence**	**TB absence**
Cough	almost always	rather often	0.984	0.75
Cough With Blood	often	seldom	0.86	0.14
Fever	very often	rather often	0.938	0.75
Night Sweat	often	often	0.86	0.86
Loss of Appetite	very often	rate	0.938	0.75
Shortness of Breath	often	very often	0.86	0.938
Chest Pain	moderately	rather often	0.5	0.75
Fatigue	ofen	very often	0.86	0.938
Weight of	often	rather often	0.86	0.75
Loss				
Chills	very often	rather often	0.938	0.75

Now, based on this PS and SD matrix, the PD matrix has been computed by applying three Fuzzy Compositional Rules, namely Fuzzy Max-Min, Fuzzy Max-prod and Fuzzy Max-av composition.

4. ANALYSIS

The algorithm has been developed in **R software** to get the predicted scores for each the above three compositional rules. In order to apply these approaches and validate the same, the actual data of size 230 subjects has been split into train and test data in 70:30 ratio, so that Train data contains 161 observations and Test data contains 69 observations.

5. RESULTS AND DISCUSSION

The purpose of this study is to identify the best predictive model that maximizes the correct classification rate and also the True Positive Rate (TPR). The consolidated results of these two approaches (Fuzzy logic and Logistic Regression) are tabulated below.

Table 4: Model Comparison

	Fuzzy Composition Rule			Logistic Regression
Accuracy Measures	**Max-Min**	**Max-Prod**	**Fuzzy-Avg**	
Correct Classification%	48%	**54%**	**54%**	74%
True Positive Rate	85%	**85%**	**85%**	48%
False Negative Rate	15%	**15%**	**15%**	52%

The above table clearly shows that Fuzzy Max-Product and Fuzzy Max-Average rules give a higher TPR rate

(85%) with reasonable correct classification rate (54%). According to these results, one can conclude that Fuzzy Max-Product (or Fuzzy Max-Average) is the better model to predict the TB diagnostics based on the corresponding symptom features.

6. CONCLUSIONS AND RECOMMENDATION

This study, in the overall context, prefers the Fuzzy Max-Product approach for the classification of TB and Non-TB cases. However, in its future recommendations, it suggests that instead of using the experts' weights under "Symptom-to-Disease" matrix, such parameters can be derived by optimizing the misclassification rate in the trial data. This multivariate optimization can be achieved using Optimization Algorithms, such as Genetic Algorithm (GA) and Ant-Colony Search. These optimization algorithms helps us avoiding locked with local optima and they converge rapidly towards the corresponding global optimum. Cross verification between two or more search algorithms will confirm the predictive power of the classification procedure.

7. REFERENCES

1. Hosmer, D.W. & Lemeshow, S. (1989). Applied Logistic Regression. New York: Wiley.
2. Gerstenkorn, T., Rakus, E.: The Method of Calculating the Membership Degrees for Symptoms in Diagnostic Decisions. Cybernetics and Systems Research '92-Proceedings of the XIth European Meeting on Cybernetics and System Research, vol. 1, Vienna (1992) 479-486
3. J SANCHEZ, E.: Inverses of Fuzzy Relations. Application to Possibility Distributions and Medical Diagnosis. Fuzzy Sets Syst. 2 (1979) 75-86.
4. Rakus-Anderson, E. (2007). Fuzzy and Rough Techniques in Medical Diagnosis and Medication (1st ed.): Springer.
5. SANCHEZ, E.: Medical Diagnosis and Composite Fuzzy Relations. In M. M. Gupta, R. K. Ragade, R. R. Yager (Eds.): Advances in Fuzzy Set Theory and Applications, pp. 437-444. (Amsterdam: North-Holland 1979).
6. Sanchez, E.: Resolution of Eigen Fuzzy Set Equations. Fuzzy Sets and Systems 1 (1978) 69-74
7. Walker, S. H and Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". *Biometrika*. **54**: 167–178. doi:10.2307/2333860.
8. Zadeh, L. A.: Fuzzy sets. Inf. Control 8 (1965) 338-353
9. Zimmermann, H. J.: Fuzzy Set Theory and Its Applications. 3rd edn, Kluwer Academic Publishers, Boston (1996)