Fuzzy Logic Approach to Predict the Outcome of Tuberculosis Treatment Course Destination

Sharareh R. Niakan Kalhori, Xiao-Jun Zeng

***Abstract*— Tuberculosis (TB) treatment with patient supervision and support as an element of global plan to stop TB designed by World Health Organization in 2006 requires prediction of patient treatment course destination to determine how intensive should be the level of supplying services and supports in DOTS (directly-observed treatment, short course). This study was aimed to develop a model using fuzzy logic technique to forecast TB cases treatment course destination. The five given outcomes included getting cured, completion treatment courses, quite the treatment course, fail in treatment, and dead. 16 parameters verified by former studies were applied as predictors. The data set with 9672 Iranian patients were divided as training to build a model and testing datasets to check the predictive ability of fuzzy model. Using principal component analysis 5 clusters of variables were extracted. 14 inputs (x) categorized in 5 identified components and based on the expert's knowledge 50 fuzzy sets were developed. For each fuzzy set (A) triangular membership function was determined $\mu_A(x) : X \rightarrow [0,1]$ presenting the degree to which x is an element of set A. Predictive model was developed by learning from given historical datasets, based on an integration of simplified fuzzy technique and recursive least square learning algorithm. After applying testing set to developed model by training set, the gained mean absolute percentage error (MAPE) was 1.258. To sum up, fuzzy logic is an acceptable technique with easy to understand output to predict outcomes of tuberculosis treatment course destination.**

***Index Terms*— DOTS, Fuzzy logic, Model, Tuberculosis**

I. Introduction

Tuberculosis (TB) still is a major public-health difficulty in the whole world. It is estimated that it causes about 9 million new infected cases and 2 million deaths annually. DOTS is a recommended package of services by World Health Organization (WHO) pursuing the objective of detecting at least 70% of all infectious Tb cases and cure 85% of detected cases successfully [1]. In fact, ensuring that the TB patient completes therapy to cure in order to prevent drug resistance cases and developing TB in the community is one of the crucial objectives of DOTS [2]. There has been a controversy debate about the result of applying DOTS to control Tb in practice. Because of raising the incidence of TB started from mid-1980s, it seems that the control of TB has been failing practically in many countries .Also, it has been documented that DOTS imposes extra burdens on the patient and health care system because of either lengthened admission or frequent attendance at clinics instead of self-administration with suitable cure rate in some other cases [3].On the other hand, other empirical investigations have confirmed the DOTS` role in treatment success rather than case detection [4]. Overall, it seems that DOTS as one of the most widely-implemented and longest running global health intervention in health history is going to continue as a foundation strategy for TB control. However, because of pointed imperfections in practice, it needs some additional change and support to promote the quality of treatment and gain the defined objectives. Hence, WHO in "Stop TB Strategy" has focused on pursuing high-quality DOTS expansion and enhancement; one of the most crucial components of this worldwide plan is standardized treatment, with supervision and patient support. It has been emphasized that services for TB care should identify and address factors that may make patients interrupt or stop treatment. Moreover, supervision must be carried out in a context-specific and patient–sensitive manner, and is designed to ensure adherence on the part both of providers (in giving proper care and support) and of patients (in taking regular treatment). Also, it has been brought to light that preferred patient groups, for example prisoners, drug users, and affected people by mental health disorders may need intensive support including DOT standing for direct observation of therapy [4].

One of the main reasons that WHO has stressed adherence of supervision and patient support to treatment course is the importance of completing this period entirely. In fact, non-completed treatment course and not entirely cured cases not only do not remove themselves from the prevalent pool but are going to infect more cases increasingly. Noncompliance treatment course has been identified as being associated with recurrence of TB [5]. Recurrent tuberculosis cause significant threats like multi-drug resistance TB (MDR-TB) which is a form of disease resistant to the most essential anti-TB drugs, i.e. isoniazid and rifampicin. Misused or mismanaged of using drugs cause MDR-TB which takes longer to treat with second-line drugs which are more expensive with more side-effects. Furthermore, misused or mismanaged of these drugs can develop XDR-TB known as more resistant to both first-and

Manuscript received August16. This work has been funded by Sharareh R. Niakan Kalhori`s scholarship from Iranian Ministry of Health and Medical Education.

Sharareh R. Niakan Kalhori is a member of academic team of Ahvaz Jondishapour University of Medical Sciences in Iran. Also, sharareh is a PhD student at Machine learning group and optimization, school of Computer Science, University of Manchester, G33. Kilburn Building, Manchester, UK; Phone: 0044-1613066993 fax: Fax: 0044-1612756204; e-mail: sharareh.niakan@postgrad.manchester.ac.uk.

Xiao-Jun Zeng is with the University of Manchester, Machine learning group and optimization, school of Computer science, Kilburn Building, Manchester, UK, email: x.zeng@manchester.ac.uk.

second-line drugs making treatment options more critically restricted with less cure chance [6].

Although WHO has highlighted the necessity of improving quality of DOTS in terms of supervision and patients support in 'StopTB 'plan, there is no way to measure how intensive health workers` support and supervision should be for patients. To make this supervision more context-specific and patient–sensitive manner, we need a tool to predict patient destination regarding TB treatment course completion. It has been documented that there are different outcome for TB patient who are under DOTS. It may vary from getting cure and completed the treatment course to simply fail the treatment or even quit it. Dead is another inevitable destination which is around 4.8% of all cases [4].

Although several studies [7][8][9] have addressed influential factors like HIV infection, area of residency, history of TB, intravenous drug using, sex, age, original nationality, and affecting with other disease which are able of predicting non-completion of tuberculosis treatment course, there is no workable tool to determine TB cases requiring intensive DOTS.

This study was aimed to introduce a fuzzy model to forecast the outcome of Tb patients after applying DOTS; using this model can help health workers to supervise and support each specific patients based on their situation since they can predict what would be happened at primary stage of applying DOTS. It is very beneficial method to put more support for patients who needs more care based on their prediction results introducing them as high or less risk groups.

II. Material and methods

A. Data

A retrospective analysis was performed in 9672 subjects who were involved in the process of DOTS from registration stage to diagnosis and treatment of TB. In fact, most reported cases in Iran from whole country in 2005 made up the data set. Novel professional software, 'Stop TB', third edition, version 2.1.3.102 developed for data collection of TB treatment process based on DOTS in 2003 was used.

To check the model validity, whole data set was divided into training set including 7254 instances to build the model and testing sets with 2418 instances to check the developed model validity respectively. 16 patients attributions were applied to be explored and after variable selection process via Principle Component Analysis (PCA), the model were developed through 5 identified components containing 14 attributers. Before PCA application, all applied attributes were categorized as follow:

- Patients' demographic data such as *sex*, *age*, *nationality*, *area of residency*.
- Patients` TB history such as *Case type*: the type of TB case like imported, new, returned, and returned after cure; *Treatment category* (Tcat) like treatment by taking group A (first line) or B (second line) drugs; *TB type* which was the type of TB whether it is pulmenory or extrapulmonary TB; *Recent Tb infection* (rtbinf) indicate whether or not the given patient recently had any diagnosed TB affection.
- Patients involvement by other disease like *diabetes*, *HIV*, *Low body weight* (LBW) which can affect TB patient status.
- Patient personal &behavior history like *imprisonment*, *current stay in prison*, *having risky sex* and *IV drug using*.

B. Principle Component Analysis

Principle Component Analysis (PCA) is an exploratory data analysis which can be applied in the process of predictive models development. PCA as a variable selection tool put correlated attributes fairly well together as a component and reveals how each variable might contribute to given components. Applying selected variables and components cause more accuracy for developed fuzzy model in present study. There are different criteria to evaluate the result of PCA. Large eigenvalues is a reason to verify related factors which was used to produce the Scree plot; furthermore, Kaiser`s criterion suggests that all factors with eigenvalues greater than 1, should be retained. Correlation matrix is another tool to check the pattern of relationship. Significant values of variables which are between 0.5 and 0.9 shows acceptable rate. To find the problem of multicollinearity, the given determinants can reveals that there would be value greater than 0.00001. To develop PCA, SPSS 14. which is statistical software was used.

C. Fuzzy logic

Fuzzy logic known as a basic concept for embedding structured human knowledge into workable algorithms constitutes fuzzy models which is one of the soft computing tools. This powerful tool to tackle imprecision and uncertainty was initially introduced by Zadeh in 1965 to improve tractability, robustness and low-cost solutions for real world problems. The theory of fuzzy sets is a theory of graded concepts and membership elasticity [10].

Defined fuzzy sets or classes for each variable allows intermediate grades of membership in them which means each set could have elements that belong partially to it; the degree of belonging is called membership functions ranging from 0 to 1. If X is the Universe of discourse and its elements be denoted as x, in contrast with crisp set, the fuzzy set A of the X has a characteristics function associated to it:

$$\mu_A : \mathrm{X} \rightarrow [0,1] \tag{1}$$

$$\mu_A(\mathrm{X}) = \text{if x is totally in A}$$

$$\mu_A(\mathrm{X}) = 0 \text{ if x is not in A}$$

$$0 < \mu_A < 1 \text{ if x is partly in A}$$

Therefore, a fuzzy membership function μ_A(x) indicate the degree of belonging of some element x to the universe of discourse X. It maps each element of X to a membership grade between 0 and 1 in various shapes such as triangular, trapezoidal, singleton and Gaussian. Triangular membership function which is widely in use can be calculated as following:

$$\mu(x) = \begin{cases} 0 \quad ifx < a \\ \dfrac{x-a}{c-a} ifx \in [a, c] \\ \dfrac{b-x}{b-c} ifx \in [c, b] \\ 0 \quad ifx > b \end{cases} \qquad (2)$$

When a, b, and c have been defined by experts. Fig 1 shows fuzzy sets defined by triangular membership function for one of this study`s attribution which was the age of TB patients. For example for fuzzy set young, the value of a, b, c is 1, 25, 50 respectively. Ability to present the linguistic variables is a prominent feature of fuzzy logic model since they can convert numeric values to linguistic variables which are highly understandable to final system users. By using those linguistic variables, fuzzy if-then rules which are the main output of the fuzzy system would be set up; generally presented in the form of: If X is A then y is B when A and B are linguistic terms defined by fuzzy sets.

D. *Recursive Least Square learning Algorithm*

Least square learning algorithm aims to minimize the summation of the errors between the modeled values for all input-output pairs in the training dataset; thus, it is capable of producing the goal optimal result. Let's have n historical instances in the training dataset, and then least square learning algorithm can be used to minimize the overall error. The Least-squares normal equation is as follows:

$$X'X\hat{\beta} = X'y \qquad (3)$$

After multiplying the both sides of (3) by the inverse of; hence, the least square estimation of β is:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (4)$$

provided that the inverse matrix $(X'X)^{-1}$ exists. Generally, y is a $n \times 1$ vector of the observations; X is a $n \times p$ matrix of the levels of the attributions, β is a $p \times 1$ vector of the regression coefficients and ε is a $n \times 1$ vector of random errors. In present study mean absolute standard error was considered for testing the accuracy of developed model by using the following formula in (5).

$$MAPE = \frac{1}{n}\sum_{r=1}^{n}\left|\frac{y_r - f(x_r)}{y_r}\right| \qquad (5)$$

When y_r is the target output of the rth instances and $f(x_r)$ is the obtained output by modeling. Assume that β_r is the parameter of the r_{th} training instance, with the use of recursive computation, it can be represented as a function of β_{r-1}. This algorithm can be used in standard fuzzy system and simplified fuzzy system to produce the global optimum parameters.

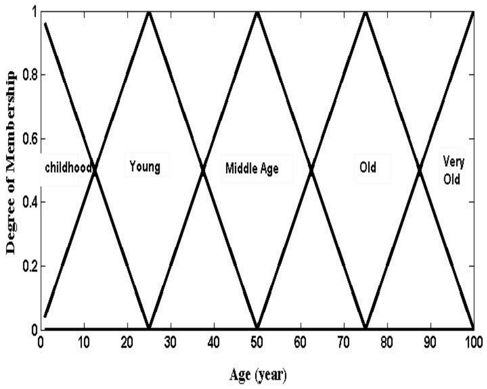

Fig1. Fuzzy sets of patients' age with their linguistic variables

E. *Simplified Fuzzy model*

When the number of inputs is too large, the simplified fuzzy system is a proper modeling method to prevent the curse of dimensionality. It constructs a linear fuzzy inference scheme to overcome the parameters and data knowledge dimensionality problem both efficiently and effectively. In simplified fuzzy system, there is a sub-fuzzy system for each input and the overall output would be the summation of the outputs for all sub-fuzzy systems which can be written as follows:

$$f_{ij}(x_j) = \sum_{n=1}^{Nj} \mu_{i_j}^{j}(x_j)\, y_{i_j} \qquad (6)$$

Where $\mu_{i_j}^{j}\,(x_j)$ is the output of the i_j^{th} fuzzy set in j^{th} input variable to the input x_j and y_{i_j} is the centre of the obtained output fuzzy set which can be set by experts. The final output of the model ($\hat{y}$) would be resulted from the following mathematical formula which is result of applying learned parameters from data to the above formula.

$$\hat{y} = \sum_{j=1}^{n} \beta_j\, f_{i_j}(x_j) = \sum_{j=1}^{n} \beta_j \left(\sum_{s=1}^{Nj} \mu_{i_j}^{j}(x_j)\, y_{i_j} \right) \qquad (7)$$

Where β_j are the parameters which were defined by the recursive least square learning algorithm. To develop the simplified fuzzy model and calculating recursive least square MATLAB 7.2 was used.

III. RESULTS

A. *Variable Selection*

Having looked at the result of principal component analysis (PCA), 2 of independent variables including *IV drug* *using* and *Nationality* were deleted when the value of suppress absolute values were considered less than 0.4. As shown in table1, five components were produced containing 14 variables which applied in pre model-driven development with given arrangement. Final model were result of composition of whole five blocks. To find the variables with large enough eigenvalue, known criterion to identify important factors, scree plot was developed. This plot helps us to decide whether or not an eigenvalue of each factor adequately address a meaningful factor. As shown in Fig.2, the plot presents each eigenvalue in Y-axis $(0.10 \leq \text{eigenvalue} \leq 1.98)$ against the associated factor in X-axis. It appears the relative importance of each factor when the factors are composed of points in the plot. Points of inflexion are the start point of each separated component and in Fig. 2 there are five points of inflexion. Several criteria showed the accuracy of PCA result:

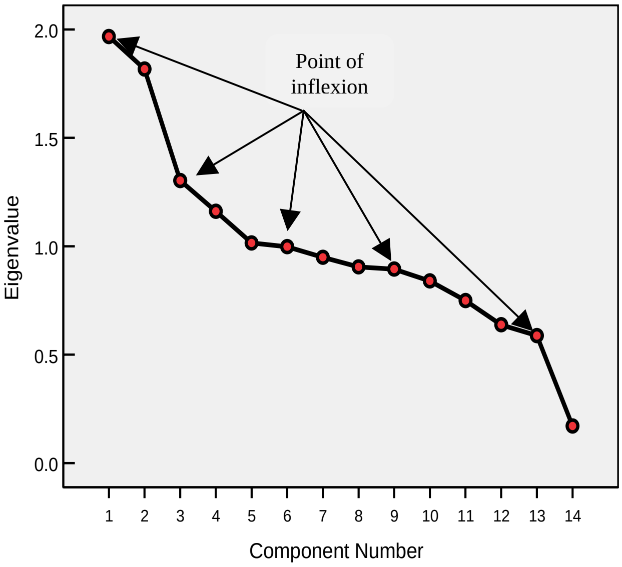

Fig2. The Scree plot of the given data with 5 components

- Correlation matrix revealed the pattern of factors relationship which correlated each other fairly well but not perfect. This correlation was verified by significance values in which they were mainly less than 0.05.
- The determinant of the correlation matrix should be greater than the necessary value of 0.00001, indicating that there was not any problem of multicollinearity.
- The KMO, Bartletts test of sphericity examined whether the population correlation matrix look like an identity matrix. Since the value of Kaiser-Mayer-Olkin Measur of sampling adequacy was 0.7 which could be assessed as a good result.

Bartlett`s measure tested the null hypothesis that the original correlation matrix was an identity matrix. Significant value of this test (P<0.0001) addressed that the R-matrix was not an identity matrix; thus, there were some relationship between variables and PCA could found them properly.

By factor rotation, orthogonal rotation (varimax), it was calculated that to what degree 14 variables were load onto 5 components shown in table 1 and it is crucial to note that factor loading less than 0.4 was not displayed since it was set up for the value bigger than given boundary .

Table 1. The rotated component matrix analysis for 5 components and their related attributes

	Components				
	1	2	3	4	5
CaseType	.950				
Tcat	.946				
prison		-.747			
RiskySex		.673			
sex		.526			
Rtbinfection			-.679		
DiabetM			-.651		
LBW			.607		
imprisonment				.654	
TbType				.716	
age				.623	
length				-.495	
area					.761
HIV					.613

Table 2. List of components and related attributes with leaned parameters in training model.

Component	Attributes(j)	No. of fuzzy sets(i)	Learned Parameters by training set
1	Case Type	4	0.467
	Tcat*	2	-0.079
2	Risky sex	3	-0.945
	Prison	2	0.992
	sex	2	0.400
3	Rtbinf**	3	0.793
	Diabetes	3	2.445
	LBW	2	1.313
4	TB type	2	0.950
	Length	5	-0.235
	Imprisonment	3	2.398
	age	5	0.237
5	area	4	0.8895
	HIV	3	0.731

*Tcat: treatment category, **Ttbinf: recent TB infection

B. Model Development

This study was a part of a MISO (multi-input single-output) simplified fuzzy system development. To fuzzify the 14 input variables by triangular membership function, 50 fuzzy set patients compared with their outcome in reality. By using recursive learning algorithm $f_{ij}(xj)$ (equation 6) were defined when i was the number of fuzzy sets defined by triangular membership function $\mu_{i_j}^{j}(x_j)$ shown in table 2 and j was the input variables (j = 1, 2,…14). Based on the PCA result shown in table 1, $F_{ij}(xj)$ calculated as follows:

$$F_{ij}(xj) = \sum_{j=1}^{n} \beta_j f_{i_j}(x_j)$$

When j was the number of variables in each component; final $\hat{y}$ were calculated by applying equation 7.

$$\hat{y} = \sum_{j=1}^{n} \beta_j f_{i_j}(x_j) = \sum_{j=1}^{n} \beta_j \left(\sum_{s=1}^{Nj} \mu_{i_j}^{j}(x_j) y_{i_j} \right)$$

Where n was 5 based on the result of PCA. The produced learned parameters were presented in table 2.

C. Model Accuracy

First, the training set was applied to develop the model; afterwards to check the model`s accuracy the testing set was used to predict the given outcome which was the TB patients destiny after applying DOTS. The real outcome for each patient in testing set was already available and by using model predicted outcome for each case were defined. Then, predicted outcomes were compared with real outcome in testing data set. By using mean absolute percentage error (MAPE), the value was 1.24. Fig3 presents the real values of outcome for the first 100 patients in the testing set and the predicted value by developed model for the same 100 patients. As can be seen, the values of outcomes might be any values from 1to 5 which means different outcomes. Value 1 means patient completed the treatment course in frame of DOTS, 2 means the patient has been cured, 3 means patients has quitted the course, 4 means patients has failed and finally 5 is a sign of dead as outcome of TB treatment course .

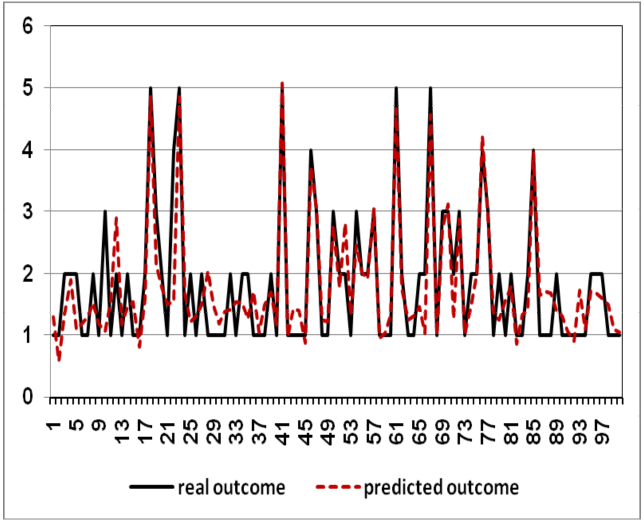

Fig3. The predicted values for 100 cases in testing set using developed model compared with their real outcome.

IV. CONCLUSION

Pursuing the idea of providing the DOTS in different levels to TB patients based on their status is a necessary purpose requiring a tool to determine the patient destination after getting DOTS. This study was aimed to develop this tool as a valid fuzzy model. This prediction would be carried out at commence of patient treatment in frame of DOTS. This valid fuzzy model can determine the level of patient support and supervision assisting the health workers to understand how intensive should be their care for each specific patient.

ACKNOWLEDGMENT

We wish to thank department of Leprosy and Tuberculosis Office, Ministry of Health and Medical Education, Tehran, Iran for data access.

REFERENCES

[1] A. D. Harries,C.Dye,‘‘Tuberculosis (Centennial review)’’. *Annals of Tropical Medicine & Parasitology*, vol. 100(5, 6), 2006, pp. 415-431.

[2] P. D.O. Davies, ‘‘the role of DOTS in tuberculosis treatment and control’’. *American journal of respiratory and critical care medicine*, vol. 2(3), 2003, pp. 203-209.

[3] Z. Obermeyer , J. Abbott-Klafter , C.J.L. Murray. ‘‘Has the DOTS Strategy Improved Case Finding or Treatment Success?’’ An Empirical Assessment. *PloS ONE*, 2008; 3 (3): e1721.

[4] World Health Organization. The Stop TB Strategy, Document WHO/HTM/TB/2006.35. Geneva: WHO.

[5] P.D. Picon, S. L. Bassanesi, M. L. A. Caramori, R.L.T Ferreira, C. A. Jarczewski, P. R. B. Vieira. ‘’Risk factors for recurrence of tuberculosis’’. *Jornal brasileiro de pneumologia*.2007; 33(5): 572-578.

[6] C. D. Wells, J. P. Cegielski, L. J. Nelson, K. F. Laserson, T.H. Holtz, A. Finlay, K.. G. Castro, K. Weyer. HIV infection and Multidrug-Resistant Tuberculosis –The perfect storm’’. The Journal of Infectious disease s, 2007; 196; S86-107.

[7] I. Baussano, E. Pivetta, L.Vizzini, F. Abbona, M. Bugiani. ‘’Predicting tuberculosis treatment outcome in a low-incidence area’’. *Int J Tuberc Lung Dis*, 2008; 12(12):1441-8.

[8] D. Antoine,C. E. French, J. Jones, J.M. Watson. ‘’Tuberculosis treatment outcome monitoring in England, Wales and Northern Ireland for cases reported in 2001’’ *J Epidemiol Community Health* 2007; 61: 302–307.

[9] H.G. Tanguis, J.A. Cayla, P. Garcia de Olalla, J.M. Jansa , M.T. Brugal. ‘’Factors predicting non-complition of tuberculosis treatment among HIV-infected patients in Barcelona (1987-1997).’’*INT J TUBERC LUNG DIS* 2000; 4(1): 55-60.

[10] L.A. Zadeh, ‘’Fuzzy Sets’’. *Information and Control*, 1965, volume 8, 338-353.