A Learning System Prediction Method Using Fuzzy Regression

Rosma M. Dom, Sameem A. Kareem, Ishak. A. Razak, Basir Abidin

Abstract - **This paper reports on the development of a learning system for the prediction of dichotomous response variables by combining fuzzy concept with classical regression technique. The algorithm involves linear transformation followed by linear programming. In the algorithm presented it was assumed that the logarithm of the odds (logit) is linearly related to X's, the independent variables after undergoing the logit transformation. In this paper the research backgrounds and methodology are presented**

Index terms - **dichotomous, fuzzy regression, prediction**

I. INTRODUCTION

There is a need for precise and accurate prediction learning system due to questionable accuracy and applicability of existing statistical and artificial intelligent prediction models particularly when dealing with small sample size and / or when there is an ambiguous relationship between the explanatory and response variables.

To fulfill these needs an effort had been taken to develop a new learning system based on the advances of soft computing particularly for cancer screening initiatives. Our proposed learning system is a computer program that makes prediction based on supervised accumulated experiences. The goal is to deal with complex real-world decision-making problems.

This paper is organized as follows: section I gives the brief introduction of the need for a more reliable learning system to be used in prediction exercises, section II describes the background for the undertaken research. The research methodology is shown in great detail in section III. Finally section IV concludes the report.

II. RESEARCH BACKGROUND

Existing predictive models proved to have several limitations when dealing with small sample size and/or when there is an ambiguous relationship between explanatory and response variable [1, 6, 11, 13]. Decision makers are looking for new measures particularly among the artificial intelligent techniques to overcome the above mentioned situation. In decision making exercises more often than not prediction outcomes heavily relied on prior experiences or in computing terms referred as supervised learning.

This study is highly influenced by the works of researchers investigating the feasibility of using machine learning techniques in oral cancer detection and screening [8, 9, 12]

A. Research Aims and Scope

The aim of this research is to develop a prototype of a computer-based prediction model by using fuzzy regression methods that can be used at individual as well as group prediction level. The scope of this research is restricted to crisp input variable and binary response variables.

B. Research Conceptual and Theoretical Frameworks

Regression analysis is a mathematical technique used to model relationship between explanatory and response variables while machine learning is an algorithm that can be developed to estimate the unknown dependency between the given set of input variables to its corresponding response variables [5]. The use of prediction interval in machine learning is referred to as fuzzy linear regression. This research is based on fuzzy linear regression theory introduced by Tanaka in1982. Fuzzy regression techniques are applicable to linear functions only [1, 3].

Logistic regression theory on the other hand is a mathematical modeling approach that can be used to describe the relationship between explanatory variables to a dichotomous dependent variable. In logistic regression the distribution requirement of independent variable: need not be normally distributed neither of equal variance within each group and the relationship between predictor and response variable is non-linear [5].

In the proposed model, fuzzy linear regression and logistic regression theories are combined to produce an adaptive fuzzy regression model.

III. RESEARCH METHODOLOGY

In this study we have taken an exploratory research approach in developing a predictive tool that can be used in multivariate function with binary output. The four components of the methodology are:

Manuscript received September 21, 2007. This work was supported in part by the University Teknology MARA and IRPA grant under the 8th Malaysian Plan.

R. M. Dom is with the University Teknologi MARA, Malaysia, and is currently pursuing a Ph.D degree at the University of Malaya, Malaysia (Phone: (6)03-5543-5341; Fax: (6)03-5543-5501; email: rosma@tmsk.uitm.edu.my).

S. A. Kareem is with the Faculty of Computer Science, University of Malaya, Malaysia (sameen@um.edu.my).

I. A. Razak is with the Faculty of Dentistry, University of Malaya.

B. Abidin is with the Pre Medical Science Department, Cyberjaya University College of Medical Sciences, Malaysia (email: basir@cybermed.edu.my).

- Algorithm development
- Experimentation of the algorithm on an oral cancer data set.
- Model validation by comparing the proposed Fuzzy Regression Predictive Model with Fuzzy Neural Network model
- Analysis of findings

A. Algorithm Development

Reviews of our study have shown that the complexity of a vague and inaccurate functional relationship can be best ameliorated by fuzzy regression analysis in multivariate environment. However existing algorithm is limited to mainly fuzzy linear functions though literatures have suggested that the concept could be extended to non-linear and intrinsically linear function [4]. This paper is the first attempt of such efforts. In developing the algorithm we have followed the suggestions proposed by related literatures but specifically tailored to the nature of the logistic function which we were focusing at. The algorithm is presented in Fig. 1 below, followed by its algebraic form.

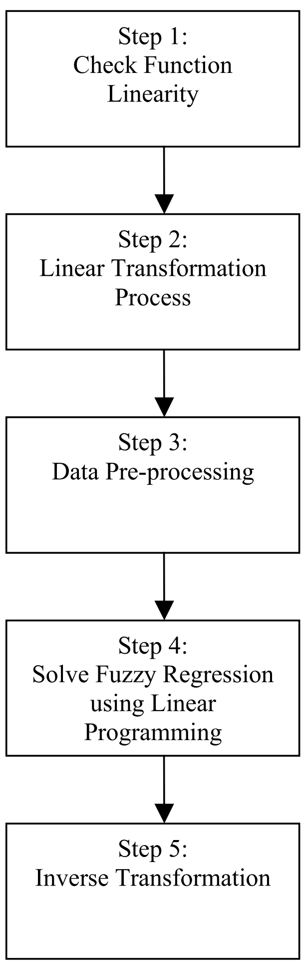

Fig. 1: Flow Chart Diagram for the proposed Algorithm

Algebraically the fuzzy logistic regression algorithm can be written as follows:

1. Let X= [$x_1, x_2, x_3, \ldots\ldots, x_n]^T$ be the vectors of the explanatory input variables. Assume all X's are discrete crisp values.
2. Let Y be the corresponding response variable such that Y=0 or Y=1 for all X's.
3. Plot is not a straight line implying that the functional relationship is not linear.
4. Carry out linear transformation

$$\ln\left(\frac{P}{1-P}\right) = a + b_1x_1 + b_2x_2 + \ldots + b_jx_j$$

$$\frac{P}{1-P} = e^{a+b_1x_1+b_2x_2+\ldots+b_jx_j}$$

$$P = \frac{1}{1+e^{-(a+b_1x_1+b_2x_2+\ldots+b_jx_j)}}$$

Let Y= $\left(\frac{P}{1-P}\right)$.

5. Data pre-processing
Convert raw data into the corresponding probabilities, odds and logits (logarithm of odds).
6. Solve Fuzzy Linear Regression Analysis based on Tanaka's possibilistic regression:
$Y = A_0x_0 + A_1x_1 + A_2x_2 + \ldots.. + A_jx_j + \ldots + A_kx_k$
where Y is the fuzzy output, $x = [x1, x2, \ldots, xk]^T$ is the real-valued input vector of independent variables and each regression coefficient Aj, j=0, . . ., k, was assumed to be a symmetric triangular fuzzy number with center αj and half-width cj, $Cj \geq 0$.
The fuzzy linear regression model can now be rewritten as:
$y = (\alpha_0, c_0) + (\alpha_1, c_1)x_1 + (\alpha_2, c_2)x_2 + \ldots. + (\alpha_k, c_k)x_k$.
The following linear programming (LP) formulation was employed to estimate
$A_j = (\alpha_j, c_j)$:

Minimize $J = \sum_{j=0}^{k} (c_j \sum_{i=1}^{n} x_{ij})$

Subject to $\sum_{j=0}^{k} \alpha_j x_{ij} + (1-h)\sum_{j=0}^{k} c_{jx} x_{ij} \geq y_i$

and

$$\sum_{j=0}^{k} \alpha_j x_{ij} - (1-h)\sum_{j=0}^{k} c_{jx} x_{ij} \leq y_i$$

$a_j \ \varepsilon\ R$, $c_j \geq 0$ j= 0,1,2,…..,k
x_{i0}= 1 , i=1,2,…..,n
$0 < h < 1$
where J is the total fuzziness of the fuzzy regression model. The h value is the threshold level that determines the degree of fitness of the fuzzy linear model to its data.
7. Transform the output using an inverse transformation process.

This completes the algorithm thus giving prediction outputs in an interval form since the regression coefficient in used is fuzzy [7]. However single prediction output can be quoted by taking averages or midpoints of the interval if required.

Table 1: Input variables description

Predictor	Description
Age	< 40 , > 40
Gender	Male , Female
Ethnicity:	Aborigines, Non-Aborigines
Cigarette Smoking habits	Smoker, Non-Smoker
Alcohol Drinking habits	Drinker, Non-Drinker
Betel-quid Chewing habits	Chew, Do not Chew
Molecular marker GSTM1	Positive, Negative
Molecular marker GSTT1	Positive, Negative

B. Experimentation of the algorithm on an oral cancer data set.

A retrospective cohort study was conducted using data obtained from the Oral Cancer Database resourced at the Faculty of Dentistry, University of Malaya. The target population includes all patients who were seeking care at the nine selected centers in Malaysia. The compilation of this oral cancer database was funded by a grant provided under the 8th Malaysian Plan. The sample comprises all identified new patients with oral cancer and potentially malignant lesions. The control group consists of patients without oral cancer or potentially malignant lesions matched on referral pattern, sex and age.

Data on patients' demographic profile and risk habits were collected during a survey interview. DNA extracted from subjects' blood samples was analyzed using PCR (Polymerase chain reaction technique) to detect GSTM1 and GSTT1 gene polymorphisms. The final count of the sample consisted of 84 oral cancer patients and 87 controls.

Risk factors in demographic and disease variables of patients that were reported to be associated with oral cancer in previous studies were considered in developing the predictive models. The input variables are tabulated as in Table 1.

The full dataset was divided randomly into a modeling dataset (65% of the total) and testing dataset (the remaining 35%).The dichotomous output refers to the health state of either "cancer (1)" or "healthy (0)".

C. Models Evaluation

The essential part of this research is the advocating of fuzzy regression learning system as a predictive tool for binary output in multivariate environment. We have validated our model by comparing the predictive performance of the proposed Adaptive Fuzzy Regression Model with the predictive performance of Fuzzy Neural Network model using the difference between membership function values of the observed and estimated outcomes.

D. Analysis of findings

Table 2: Membership error for 1input variable

One-input variable	**Member-ship error (Fuzzy Neural Network model)**	**Member-ship error (Fuzzy Regression Model)**
Gstm	0.8248	0.510722
Gstt	0.5429	0.51203
Smoking	0.5348	0.507585
Drinking	0.4416	0.402832
Chewing	0.4253	0.419937
Gender	0.5332	0.506448
Agegroup	0.4526	0.404563
Ethnicgrp	0.4731	0.465483

Table 3: Membership error for 2 input variables

Two-input variable model	**Member-ship error (Fuzzy Neural Network model)l**	**Member-ship error (Fuzzy Regression Model)**
Chew & gender	0.4251	0.422307
Chew & ethnic	0.4091	0.39109
Chew & age	0.3975	0.372855
Chew & drink	0.3727	0.350955
Chew & smoke	0.4336	0.42235
Chew & gstt	0.4398	0.422927
Chew & gstm	0.7171	0.402903

The average membership error for fuzzy regression and fuzzy neural network models were measured and compared in order to determine whether there is any statistical significance. The average membership error values for one input variable set are tabulated in Table 2 while Table 3 gives the average membership error values for two input variable sets for both the adaptive fuzzy regression and fuzzy neural network models. Spearman's correlation values for one input variable set is 0.905 for the two models while Spearman's correlation value for two input variable set is 0.786. High Spearman's correlation values imply that the proposed adaptive fuzzy regression model has significantly similar ability in predicting dichotomous outcome as fuzzy neural network model.

IV RESEARCH CONCLUSION AND FUTURE WORKS

In this paper we are advocating an artificial intelligent algorithmic learning system approach for the prediction of dichotomous response variables in multivariate functions. Our investigation has been directed towards developing a suitable algorithm that can be used to produce an accurate, reliable and transparent predictive model.

The main contribution of this research is the development of an adaptive prediction model for binary output based on Fuzzy Regression. This technique is suitable for small sample size and for variables governed by ambiguous relation. Though the proposed prediction learning system was experimented on an oral cancer data set, its' use can be extended to other intrinsic linear functions such as exponential, reciprocal, growth and decay functions etc**.** Future works involve empirical experiments carried out on other datasets and other linearly intrinsic functions.

V.REFERENCES

[1]. A. F. Shapiro, "Fuzzy Regression Models", *ARC*, 2005.

[2]. F.M. Tseng and L. Lin, "A Quadratic Interval Logit Model for Forecasting Bankruptcy", *Omega*, Vol. 33, Issue 1, 2005, pp 85-91.

[3]. H. Tanaka, S. Uejima, and K. Asai, "Linear Regression Analysis with Fuzzy Model", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 12, No 6, 1982, pp 903 – 907.

[4]. H.F Wang and R.C. Tsaur, "Insight of a Fuzzy Regression Model", *Fuzzy Set and Systems*, 2000, pp. 355-369.

[5]. J. Miles and M. Shevlin,. "*Applying Regression and Correlation. A guide for Students and Researchers*". SAGE Publication Ltd., 2001.

[6]. J.V. Tu, "Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes", *J Clin Epidemol*, Vol 49, No 11, 1996, pp.1225-1231,.

[7]. L. Durga. and P. Dimitri, "Machine Learning Approaches for Estimation of Prediction Interval for the Model Output", *Neural Networks Special Issue*, 2006, pp. 1-11.

[8] L.Muzio et al., Expression of cell cycle markers and human papilloma virus infection in oral squamous cell carcinoma: Use of fuzzy neural networks, *Int. J. Cancer*, Vol. 115 (2005), pp 717- 723

[9] M. Kambara et al., Evaluation of fuzzy regression analysis and its application to oral age model, *Biomedical Soft Computing and Human Sciences*, Vol. 5, No.1, (2000), pp 41 – 49

[10]. M. Modarres, E. Nasrabadi, and M.M. Nasrabadi, "Fuzzy Linear Regression Models with Least Square Errors", *Applied Mathematics and Computation*, Vol 163, 2005, pp 977-989.

[11]. M. Nasiri et al., "Comparison of Statistical Regression, Fuzzy Regression and Artificial Neural Network Modeling Methodologies in Polyester Dyeing", *Proceedings of 2005 International Conference for modeling, control and automation*.

[12]P. Speight and P. Hammond, The use of machine learning in screening for oral cancer, *Artificial Neural Networks in cancer diagnosis, prognosis and patient management*, (2005)

[13]. S. Dreiseitl and O. Machado, "Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review", *Journal of Biomedical Informatics*, 35, 2003, pp352-359.

[14]. Y. Xue et al., "Fuzzy Regression Method for Prediction and Control the Bead Width in the Robotic Arc Welding Process, *Journal of Material Processing Technology*, Vol. 164-165 2005, pp. 1134 -1139.