Fuzzy logic to analyze survey data from populations exposed to arsenic-contaminated water

Jose Arturo Molina Mora

Abstract— **Drinking water contaminated with arsenic is a public health problem that currently affects about 11,000 people in Costa Rica, mainly in the north and northwest of the country which source of water is wells-based. Poisoning of this water has, both chronic and acute, major impact on population health and lead to clinical manifestations of neurological, dermal, gastrointestinal and hematological level, among others. Thus, water for subsistence taken from wells of areas with arsenic contamination must be constantly monitored. In this work, a set of data from surveys, including factors such as the degree of contamination of wells that currently use, education, membership of a community association and the distance of a new well, were used for analyzing the intent of change of well. For this, a model of fuzzy logic was applied to handle uncertainty and create rules of association between variables. Also, a comparison was made with classic algorithms including K-means algorithms, decision trees and neural networks. The resolution of the models did not differ from each other and all with about 70% accuracy. To improve the response capacity, a factor analysis was performed using principal component analysis (PCA), obtaining an accuracy above 98% in all cases, including fuzzy logic. Thus, the models using fuzzy logic tools are presented as alternatives to data processing in data mining.**

Index Terms— **Fuzzy-logic, Drinking Water, Data Mining.**

I. INTRODUCTION

Arsenic is common component of the earth's crust and being categorized as metalloid because it shows intermediate properties between metals and non-metals, which makes it an element of great industrial importance [1]. However, his presentation in nature and / or contamination by products industry has led to cases of poisoning. It can be found in source water and this becomes important in certain populations, especially in rural areas of Bangladesh and India, where water is obtained from wells with high contamination with arsenic [2]. In America, some specific areas of Chile and Argentina have had the same findings as in the northern and northwestern part of Costa Rica [1].

Moreover, the concept of fuzzy sets was introduced in 1960 by Zadeh as a generalization of the conventional theory of sets [3]. Fuzzy logic is a type of logic with a series of values specified as a "degree of truth" instead of binary "true or false" values and therefore considers that the most important application of fuzzy systems (fuzzy logic) it is in the management of uncertainty.

Incorporating fuzzy logic within a system, called fuzzy inference system, has a rule-based architecture for reasoning about data [4]. The implementation of these rules defines, for the various inputs of a model or system, degree of belonging to different categories and will be the ones that determine the final output. The fuzzy rules define the connection between the input variables and output, and have the form "if antecedent (input) then consequent (output)" and are usually predefined by the experiential and problem designer features. The decision making process is performed by the inference engine using the rules in the rule base [3].

However, if the rules are not known exactly, it is possible to apply a model ANFIS (Adaptive-Network-based Fuzzy Inference System), which arises from the need to transform human knowledge or experience and data to a system as rules related to neural networks, making the adaptation and adjustment of antecedent or consequent parameters, ie, handles the creation of rules based on the data [5].

The ANFIS uses the theory of neural networks and fuzzy systems in order to determine the properties of data, harnessing the power of the two paradigms. Fuzzy logic simulates human little accurate understanding of the world; the fuzzy inference processes reflects human reasoning. The neural network tries to simulate the neural structure of the human brain and solve the complex problems of learning and training. Thus, the mathematical properties of neural networks are combined with fuzzy rule based systems that approximate how humans process information [6].

The aim of the study was to develop a mathematical model of fuzzy logic to the analysis of survey data related to arsenic-contaminated water of wells and compare it with classical algorithms of data mining.

II. METHODOLOGY

A. Data source

The data set consists of 3020 surveys of households consuming well water. The evaluated variables were the concentration of arsenic in well water, the distance from the well to homes, belonging or not to a community association and the highest level of education of some of the family members. The class variable or category to classify was "yes" or "no", that is if the family is willing to move or not a new well with null or lower arsenic contamination.

B. Implementation of the fuzzy model

The survey data set is divided into 70% for the training set and 30% for the test set. The fuzzy logic model was

Manuscript received Jan, 2016.
***Jose Arturo Molina Mora**, School of Mathematic, University of Costa Rica San Jose, Costa Rica. Phone: +506 8885 9445.*

ISSN: 2278 – 1323

implemented using the Fuzzy Logic Toolbox MATLAB (Mathworks) using triangular membership functions. Association rules were obtained with a ANFIS model, using the respective ANFIS Toolbox, also in MATLAB.

C. Implementation of classical algorithms of classification

Using the same sets of training and test, classical models of neural networks and decision trees were run to determine the power of classification and compare with the fuzzy model. All algorithms were executed in the program Weka data mining.

III. RESULTS

The evaluation of power of classification was performed with the survey data for both the original data and the processed data with a principal component analysis. Fuzzy logic, decision trees and neural networks were applied. In all cases, it was used a training data set of 2123 elements and a set of 897 for testing (to add up the total of 3020 data).

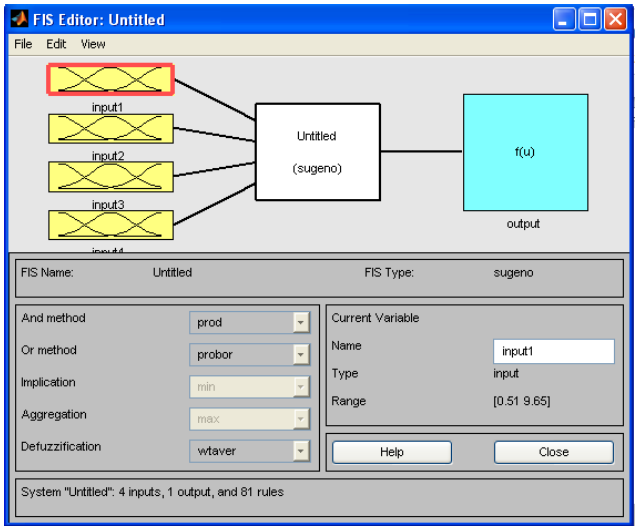

Figure 1. Architecture of the fuzzy model.

In the case of fuzzy model, the architecture is shown in Figure 1. For each independent variable, were considered three categories of membership functions. For example, for variable 2 (distance of the new well) the categories of short, medium and long were considered, and whose representation in fuzzy sets is shown in Figure 2.

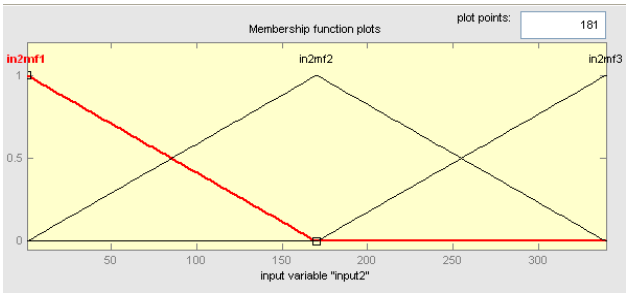

Figure 2. Membership functions for variable 2 (distance of the new well).

The output variable was categorized into 2 classes, "yes" or "no," referring to the decision to switch to a suitable well for drinking water. Value of 0.5 was established as a cutoff point (less than 0.5 is "no" and if greater than 0.5 is "yes"). Creating rules was performed with the ANFIS toolbox, whose relationship of neural network variables in the fuzzy model is depicted in Figure 3.

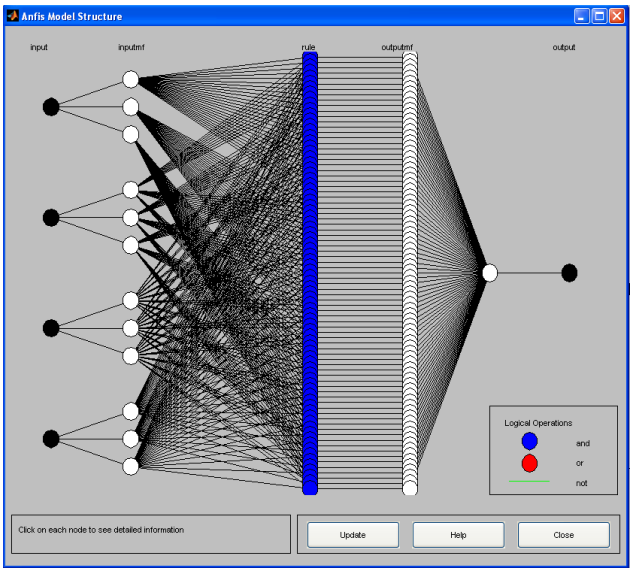

Figure 3. Representation of the neural network in the model of ANFIS.

By running the model, which was performed with 10 cycles, a training error of less than 0.4726 was obtained. When performing the confusion matrix it was obtained that 74% of the data were properly classified. The toolbox lets you view the fuzzy rules and also extract information regarding relationships between variables. The general form can be viewed in Figure 4.

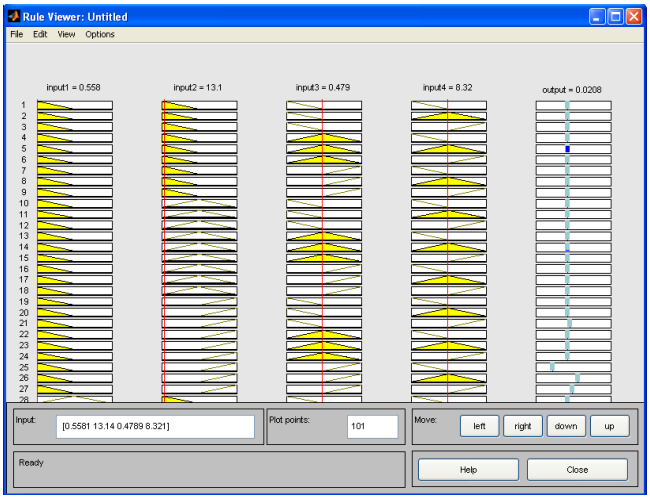

Figure 4. Graphic representation of the rules obtained with ANFIS model.

Furthermore, it is possible to display the model surface for different variables. For example, Figure 5 shows that the arsenic concentration (input 1) is critical to the decision to change or not of well, when contrasted with belonging or not to an association of community (input3).

ISSN: 2278 – 1323

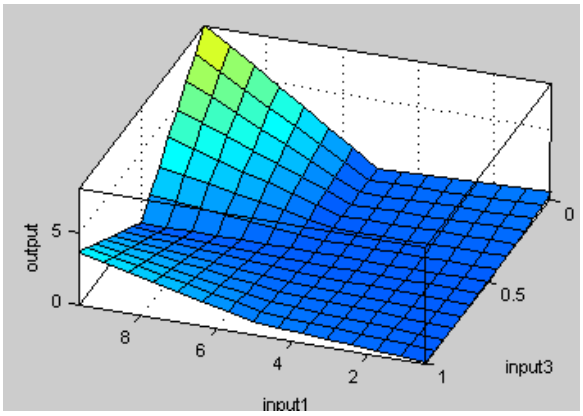

Figure 5. Surface of fuzzy model.

In order to compare confidence of the fuzzy model implementation, the decision tree algorithms and neural networks were evaluated. In the case of decision trees (algorithm J48), confidence obtained was 67.5% and whose tree is shown in Figure 6. When implementing the algorithms of neural networks, a 72.98% of the data were correctly classified.

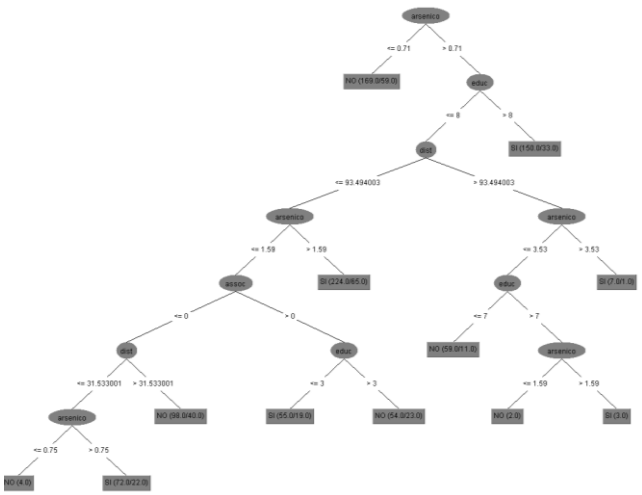

Figure 6. Decision tree (J48) for the survey data.

In order to evaluate the improvement in classification accuracy of the models, prior to the classification analysis, it was carried out a principal component analysis. With this new data set, in the case of fuzzy model, the error was less than 0.0001 obtained using 3 cycles with a 100% confidence. For decision trees, in which the variables generated by the principal component analysis cannot be interpreted directly, it was found that 98.39% of the data were properly classified. Meanwhile, neural networks achieved a confidence rating of 100%. A comparison between the different models is presented in Table 1.

Table 1. Comparison of the confidence gained with different classification techniques, with or without principal component analysis.

Model	*Original data*	*Data with PCA*
Fuzzy model	74%	100 %
Decision tree J48	68%	98%
Neural networks	73%	100%

IV. DISCUSSION

Fuzzy logic is based on human reasoning because it involves the use of variables that can be considered part of fuzzy set, ie linguistic variables or defined by words instead of numbers used categories [3]. Thus, it allows the association of variables with semi-quantitative rules. In this way, fuzzy logic is a powerful and suitable tool for handling complex problems in a position where there is incomplete information or not precise [7]. According to Zadeh, the essence of fuzzy logic is that the exact reasoning is seen as a limiting case of approximate reasoning, everything is a matter of degree and any logical system can undergo the theory of fuzzy logic [3].

Specifically the problem analyzed, the intake of small amounts of arsenic can cause chronic effects for its accumulation in the body, or serious acute poisoning that can occur when the amount taken is at least 100 mg [8]. A high level, the main health problems related to arsenic poisoning include cardiovascular, cutaneous and mucosal, neurological, urinary and hematological, gastrointestinal symptoms among others. A chronic level, Arsenicosis (prolonged exposure to arsenic) occurs and with direct involvement in development of various cancers (skin, lung, kidney or bladder), as well as blindness, lameness and even gangrene [8], [2].

Thus, the use of different data mining models in the context of this public health problem provides an opportunity to improve and implement strategies that encourage families to change well, and the management of more radical alternatives that include an improvement access to potable water and different of using wells. The set of variables analyzed allow the identification of factors that lead to the decision to change of well or not, depending on the degree of contamination by arsenic wellbore currently used and distance of a new potential well. For example, as shown in Figure 4, if the arsenic concentration is low (input 1) and the new well is near (input 2), the family is not changed, having an output of 0.0208 (less than 0.5) which it is understandable because they are not so contaminated. Another case in the simulation (not shown) is that if contamination of wells is high and the distance of the new is short, then the family can change of well (where the output is 0.752).

In the case of the decision tree and neural networks, analysis of results it shows that the main factor that determines whether or not transfer well is the amount of arsenic in the current well. In general, it is considered that if the amount of arsenic in excess 1 ppm is considered toxic, and tree (Figure 6) shows that less than 0.71 ppm families do not consider the option of changing of well. The second variable of weight is education because if the educational level is greater than 8 (university) and arsenic is greater than 0.71 ppm families tend to change, regardless of the distance.

In analyzing the confidence values obtained, the data set worked is noisy due to the inability of algorithms traditional classification and fuzzy logic to have precision values acceptable (over 80%), so a multivariate analysis was performed to extract as much information by principal component analysis.

ISSN: 2278 – 1323

The data mining models used classical and fuzzy logic had no significant variation among themselves, they provide precision values very similar, both raw data (68-74%) as the processed data with principal component analysis (98 % -100%). However, the lack of interpretation is one of the main drawbacks of the analysis of principal components.

Models using fuzzy logic are alternative data processing in data mining tools, and although confidence values were slightly higher than the other methods, it is not considered a value significantly higher. Thus, the general conclusion is that they are not necessarily better than more traditional methods but depending on context and available data, if they could be suggested more clearly. In this regard, in 2001, Aceves presented a paper setting out the uses and abuses of fuzzy logic. Particularly it criticized the sometimes fuzzy logic has been proposed as the solution to all the problems that other methods cannot solve, but this is false because it only works with uncertainty and inaccurate data. He explains how other authors mention the fuzzy logic as an intelligent system, which also is not true because the model itself does not learn unless it is combined with other methods otherwise. Finally it criticizes the membership functions are made almost arbitrarily and not necessarily those variables have such regular behavior, though so used to not complicate the computations [9].

V. CONCLUSION

A fuzzy logic model was presented for the study of population survey data who extract drinking water from wells contaminated with arsenic. Comparison with other algorithms based on neural networks and decision trees show similar levels of confidence (all close to 70%) to fall into the established categories, so that fuzzy logic is proposed as another tool that manages data mining uncertainty and allows for semi-quantitative relationships between variables. In addition, the prior application of the technique of principal component analysis was able to increase confidence levels over 98% of all cases, but with the loss of the interpretation of association between the variables.

Variables of distance of the new well and the level of the contamination proved to be key to make the decision to change of well, which may suggest strategies in local government on how to handle the awareness of families, the decision to change water sources and level of comprehensive health management.

REFERENCES

[1] R. N. Ratnaike, "Acute and chronic arsenic toxicity," *Postgrad. Med. J.*, vol. 79, no. 933, pp. 391–396, Jul. 2003.

[2] V. D. Martinez, E. a Vucic, D. D. Becker-Santos, L. Gil, and W. L. Lam, "Arsenic exposure and the induction of human cancers.," *J. Toxicol.*, vol. 2011, p. 431287, Jan. 2011.

[3] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing," *Commun. ACM*, vol. 37, no. 3, pp. 77–84, Mar. 1994.

[4] M. Ghomshei, J. Meech, and R. Naderi, "Fuzzy Logic in a Postmodern Era," *Forg. New Front. Fuzzy …*, pp. 363–376, 2008.

[5] M. Neshat and A. Adeli, "A Comparative Study on ANFIS and Fuzzy Expert System Models for Concrete Mix Design," *Int. J. Comput. Sci.*, vol. 8, no. 3, pp. 196–210, 2011.

[6] I. Güler and E. D. Ubeyli, "Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients.," *J. Neurosci. Methods*, vol. 148, no. 2, pp. 113–21, Oct. 2005.

[7] G. B. Fogel, "Computational intelligence approaches for pattern discovery in biological systems.," *Brief. Bioinform.*, vol. 9, no. 4, pp. 307–16, Jul. 2008.

[8] M. F. Hughes, B. D. Beck, Y. Chen, A. S. Lewis, and D. J. Thomas, "Arsenic exposure and toxicology: a historical perspective.," *Toxicol. Sci.*, vol. 123, no. 2, pp. 305–32, Oct. 2011.

[9] A. Aceves, "Uses and abuses of Fuzzy Logic control processes: An alternative model of incomplete and inaccurate information from an observation," *Con Manten. Product.*, vol. 2, no. 8, pp. 12–17, 2001.

Jose Arturo Molina Mora is microbiologist and mathematician of the University of Costa Rica. He has a Master degree in Bioinformatics and currently is working as professor and researcher in the same university. Main research work focus on mathematical models of medical/biological issues, including study of metabolic routes with dynamical modeling and soft computing.

ISSN: 2278 – 1323