A Novel Gini Index Decision Tree with Adaptive Neuro Fuzzy Inference System Feature Selection for Detection of Sensitivity of Corticosteroid

Hussein Fawaz , Fatima Nasrallah
Lebanese University, Faculty Of Sciences, Lebanon

ABSTRACT

Health care is the maintenance or improvement of health for humans against disease. Health care is conventionally regarded as an important determinant in promoting the general physical and mental health. Sepsis a disease in healthcare, and it affects on the human life. There exist a treatment for this disease called 'Corticosteroid'. There still a challenge facing this treatment it is the sensitivity caused by it. So, how we can detect the sensitivity/resistance of the corticosteroid? In this paper, we propose a technique using Adaptive Neuro Fuzzy Inference System (ANFIS) for feature selection with decision tree for detection of sensitivity of the treatment. Our simulation is based on real health data with accuracy 80% for the detection.

KEYWORDS

Health Care; ANFIS;Decision tree; Gini index; Classification methods

1 INTRODUCTION

In the present day, healthcare has come to mean every aspect, service and device for taking care of your health. Sepsis is a deregulated immune response to an Infection causing organ dysfunction which is often fatal.It can affect new born, children, pregnant women, imminucompromised, people living in low resource settings.It can cause Diarrheal diseases and Respiratory infections (Bacterial infections). "WHO" estimate that 49 million people are infected with sepsis each year worldwide, causing 11 million death.

The cortecosteroid is one of the treatment of Sepsis,but not every patient is resistant on it. The corticosteroids are a type of anti-inflammatory drug. They are typically used to treat rheumatologic diseases.

We used the ANFIS with decision tree for selection features and detection the sensitivity of the treatment respectively.ANFIS has select features based on the importance of each features.

The Rest of the paper is organized as follow.In Section II,we present related work. Sections III, we define our techniques.In Section IV we will show you the result.Simulations and experiments are presented in Section V.Section VI concludes the paper.Section VII the future work.

2 RELATED WORK

Related Work

3 TECHNIQUE

3.1 Data Collection

In our simulation, we use real patient data which is about 128 features for 612 patients.It's a vital signs of humans.The features are divided into 2 types:categorical such as Gender and numerical like weight.'

3.2 ANFIS

Adaptive Neuro-Fuzzy Inference Systems, developed in 1993 by J.S. Roger Jang, are widely regarded as a universal estimator or Takagi-Sugeno Fuzzy System. The Takagi-Sugeno Fuzzy model is a Type 3 Fuzzy Inference System, where the rule outputs are a linear combination

Supervised by Dr. Hasan Harb.

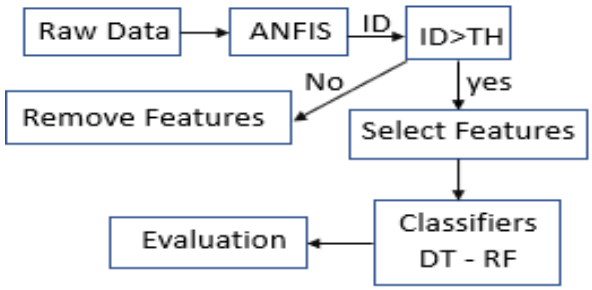

Figure 1: Proposed Technique; ID and TH mean Importance Degree and its Threshold, respectively, DT is decision tree algorithm and RF is random forest algorithm

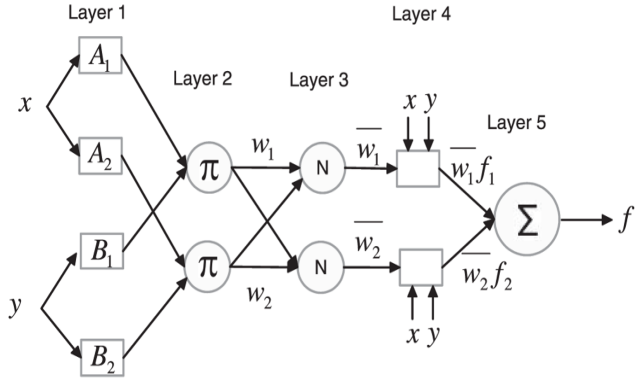

Figure 2: ANFIS architecture

of input variable along with a constant, and the final output is the weighted average of every rule's output.

The IF-THEN rules for a 2-input Takagi-Sugeno system are described as follows.

(i) Rule 1: IF x_1 *is* A_1 , y_1 *is* B_1 *THEN*
$f_1 = p_1x_1 + q_1y_1 + r_1$
$(ii) Rule2 : IFx_2$ *is* A_2, y *is* B_2 *THEN*
$f_2 = p_2x_2 + q_2y_2 + r_2$

Where x and y are the inputs in the crisp set,A and B are the linguistics labels,P and q are the consequent parameters,f1 and f2 are the output fuzzy membership functions.

The standard ANFIS architecture, as given in Figure 2, consists of five layers of interconnected neurons, evident of artificial neural networks having alike functionalities. The architecture is briefly explained as follows.

- **Layer 1:** It is the Fuzzification Layer where each neuron is an adaptive node and holds the fuzzy value of the crisp inputs.
  The node output is calculated as follows:

$$o_1 = \mu(x) = \frac{1}{1+}|\frac{x-c}{a}|^{2b}$$

where is a membership function for the fuzzy sets A, B. Numerous membership functions exist, i.e., Gaussian, Trapezoidal, Triangular, etc. We prefer a bell-shaped function in ANFIS. Hence, the Gaussian function is the optimum choice. The formula for Gaussian function is where a and b are the premise parameters for the membership functions of ANFIS.

$$f(x) = a.e^{-\frac{(x-b)^2}{2c^2}}$$

where is the weight of the neuron.

- **Layer 2** This is an Implication Layer where the neurons contain the product of inputs, i.e., the weight of premise parameters. The node output is calculated as follows:

$$w_i = \mu x_i(x) * \mu y_i(y)$$

where i=1,2

- **Layer 3:** It is Normalizing Layer where the neurons are fixed and are normalized by the sum of weights of all neurons in this layer. The node output is calculated as follows:

$$\bar{W_i} = \frac{W_i}{\sum_{j=1}^{2} W_j}$$

where is the normalized weight of the neuron.

- **Layer 4:** This is the Defuzzification Layer where each neuron is also an adaptive node and holds the consequent parameters of the architecture. The node output is calculated as follows:

$$O_4 = \bar{W_i}f_i = \bar{W_i}(p_ix_i + q_iy_i + r_i)$$

where i=1,2

- **Layer 5:** It is an Output Layer where a single neuron is present for output, which is the sum of all the inputs. The node output is calculated as follows:

f(x,y)=$\sum_i \bar{W}_i f_i = \frac{\sum_i W_i f_i}{\sum_i W_i}$

$f_{out} = \sum_i^n \frac{\bar{W}_i f_i}{\bar{W}_i}$

Classical ANFIS favors hybrid learning process, where parameters are updated through two passes and use two different optimization algorithms.

During the forward pass, the consequent parameters are updated, when the inputs are provided to ANFIS, and the premise parameters are kept fixed, using LSE, the consequent parameters are updated in Layer 4, and the final output is calculated accordingly.

As the final output is calculated, the backward pass starts, during which the error is propagated back to Layer 1, and the premise parameters are updated. In this pass, the consequent parameters are kept fixed.

3.3 Features Selection

We have proposed two techniques, the first one is about two features respectively like f1,f2 then f3,f4 and we put it in the ANFIS function with the label, the number returned by this function should be compared with the thresh-hold (TH), if it is bigger than the TH then we select f1 f2 and we add it to the new data frame and we continue to the last features in the same manner. Then we apply machine learning algorithms on the selected features.
In the second technique, we take each feature with all other features one by one and applying ANFIS on them with the label, and then we calculate the average of the output values from ANFIS and compare it with the TH and it's bigger the TH, we select this feature, and we continue to the last features in the same manner. Then we apply machine learning algorithms on the selected features.

3.4 Decision Tree

Data mining is used to extract useful information from large datasets and to display it in easy-to-interpret visualizations.First introduced in 1960's, decision trees are one of the most effective methods for data mining; they have been widely used in several disciplines because they are easy to be used, free of ambiguity, and robust even in the presence of missing values. Both discrete and continuous variables can be used either as target variables or independent variables. More recently, decision tree methodology has become popular in medical research. We apply here:

- **Decision Tree based on Voting:**
  Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.
- **Decision Tree based on Gini Index:**
  Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure. Let's perceive the criterion of the Gini Index, like the properties of entropy, the Gini index varies between values 0 and 1, where 0 expresses the purity of classification, i.e. All the elements belong to a specified class or only one class exists there. And 1 indicates the random distribution of elements across various classes. The value of 0.5 of the Gini Index shows an equal distribution of elements over some classes. The Gini Index is determined by deducting the sum of squared of probabilities of each class from one, mathematically, Gini Index can be expressed as:

  Gini Index= 1-$\sum_{i=1}^{n} (P_i)^2$

Where Pi denotes the probability of an element being classified for a distinct class.

4 SIMULATION

We use real health data of 612 patients with 128 features which are about the vital signs of these patients. These features are divided into 2 types, categorical like gender for example but it's one hot encoded and numerical features like weight. Also the collected data is labeled 0 or 1; 0 means that the patient is resistant to corticosteroids and 1 means the patient is sensitive to corticosteroids.

We use python programming language for the simulation using scikit-learn library for the classifiers and the ANFIS library from GitHub for feature selection of data that's is presented in an excel file.
Our simulation is running on a machine with 8 GB of ram, Core i5 CPU with 2.4 GHz base clock speed using jupyter notebook. We start applying the technique by calling ANFIS for each feature then the selected features are inserted into the classifiers presented in the previous section.

5 RESULTS AND DISCUSSION

This section presents the experimental results obtained from applying proposed method on patients data set. Proposed neuro-fuzzy feature selection method as well as the classifier algorithms that have been explained in previous sections have been applied to detect the sensitivity of patient on corticosteroid. Performance of the proposed method has been evaluated using following indexes:

- **True Positive (TP):** number of sensitive states that correctly classified as sensitive;
- **True Negative (TN):** number of resistant states that correctly identified as resistant;
- **False Negative (FN):** number of sensitive states that incorrectly identified as resistant;
- **False Positive (FP):** number of resistant states that incorrectly identified as sensitive;

Confusion Matrix:

Actual/Predicted	No	Yes
No	TN	FP
Yes	FN	TP

Using the above indexes, accuracy percentage of each method is calculated as follow:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

6 CONCLUSION

This study presented Adaptive neuro-fuzzy inference systems for feature selection with decision tree for detect the sensitivity on corticosteroid of patients that have sepsis disease. Final results showed that the proposed method can produce higher accuracy by using less features than other feature selection methods. The data mining is useful for investigating the information, when the physical examination of the information isn't possible in the data mining strategies. The information mining procedures are PC based calculations which recognize the relationship among the information and extraction of the comparable example information on which they are prepared. This article exhibited a decision tree based arrangement system alongside neural system classifiers for exact detection of sensitivity/resistance of corticosteroid.

7 FUTURE WORK

Many different adaptations, tests, and experiments have been left for the future due to the lack of time. One of the techniques that can be implemented as a future work is that instead of use the entire data of a feature as an input of ANFIS which need a huge processing and computational power, we calculate for each feature four filter indexes like Fisher index, T-test index, Correlation index and Chi-Squared Test in case of categorical feature and these four filter indexes now are used as input of the ANFIS, and we continue in the same way like the technique presented in this paper by comparing the output with a fixed Threshold. Also we can test later other classifier algorithms such as SVM, Naive Bayes, Neural Network, etc. to see if we can get more accurate results.

REFERENCES