ORIGINAL RESEARCH

A hybrid knowledge discovery system for oil spillage risks pattern classification

Udoinyang Godwin Inyang [1], Oluwole Charles Akinyokun *[2]

[1] Department of Computer Science, Faculty of Science, University of Uyo, Uyo, Nigeria
[2] Department of Physical Sciences, Landmark University, Omu-aran, Nigeria

Received: July 30, 2014 **Accepted:** September 18, 2014 **Online Published:** November 4, 2014
DOI: 10.5430/air.v3n4p77 **URL:** http://dx.doi.org/10.5430/air.v3n4p77

Abstract

The complexity and the dynamism of oil spillages make it difficult for planners and responders to produce robust plans towards their management. There is need for an understanding of the nature, sources, impact and responses required to prevent or control their occurrence. This paper develops an intelligent hybrid system driven by Sugeno-Type Adaptive Neuro Fuzzy Inference System (ANFIS) for the identification, extraction and classification of oil spillage risk patterns. Dataset consisting of 1008 records was used for training, validation and testing of the system. Result of sensitivity analysis shows that Cause, Location and Type of spilled oil have cumulative significance of 85.1%. Optimal weights of Neural Network (NN) were determined via Genetic Algorithm with hybrid encoding scheme. The Mean Squared Error (MSE) of NN training is 0.2405. NN training, validation and testing results yielded R > 0.839 in all cases indicating a strong linear relationship between each output and target data. Rule pruning was performed with support (15%) and confidence (10%) minimum thresholds and antecedent-size of 3. The performance of the ANFIS was evaluated with eight different types of membership functions (MFs) and two learning algorithms. The model with triangular MF gave the best performance among all other given models while hybrid-learning algorithm performed better than back propagation algorithm. The ANFIS model reported in the paper adopted triangular MF and hybrid learning algorithm for the predication and classification of oil spillage risk patterns. Average training and testing MSE of the model is 0.414315 and 0.221402 respectively. The knowledge mining results show that ANFIS based systems provide satisfactory results in the prediction and classification of oil spillage risk patterns.

Key Words: ANFIS, Triangular membership function, Fuzzy logic, Rule interestingness, Oil spillage patterns

1 Introduction

Developments in Information and Communication Technology have resulted in huge data repositories for analysis and management by public and private sectors of the world economy. A major requirement for a modern knowledge driven society is the effective and efficient management of data held in these repositories and transforming them into information and knowledge.[1] This gives rise to the need for improved techniques, procedures and tools to aid humans in the automatic and intelligent collection and analysis of huge data sets. Knowledge Discovery (KD) effectively uncovers hidden but subtle patterns from large and diverse datasets and out performs traditional statistical techniques.[2,3] Data mining, a major stage in the KD process, is the analysis of datasets that are observational, aiming at finding out hidden relationships among datasets and summarizing the data in such a manner that is both understandable and useful to the users.[4] Some of the intelligent tools for data mining include Neural Networks (NNs), Fuzzy Logic (FL), Ants

*Correspondence: Oluwole Charles Akinyokun; Email: akinwole2003@yahoo.co.uk; Address: Department of Physical Sciences, Landmark University, Omu-aran, Nigeria

Colony Algorithm (ACO), Genetic Algorithm (GA) and so on. NNs are typically used in problems that may be understood in terms of classification or forecasting.[5] Multilayered feed forward Neural Networks have been used in the development of decision support systems.[6,7] The back-propagation algorithm, which is a variant of the gradient search method[8] can find a good set of weights in a reasonable amount of time. The key to back-propagation is the calculation of the gradient of errors with respect to weights of a given input by propagating error backwards through the network. Genetic Algorithms (GA) are proven to provide robust search in complex spaces.[9] The search space of NNs weights is very large and usage of GA will reduce the time needed to optimize the weights of the networks.[10,11] GA is applied on NNs for evolving the weights in a fixed network, the network architecture and the learning rule used by the network.[12] Montana and Davis[13] have used GA instead of back-propagation for finding a good set of the weights for a NN with fixed set of connections.

In recent times, the dependency on oil and gas has increased oil exploitation and exploration activities leading to rampant oil spillages that in turn endanger public health, devastate natural resources, and disrupt the economy. When this occurs, human health and environmental quality are at risk. Ways of minimizing oil spills and their effects need to be explored particularly as the people most affected by the spill are those in the host communities where the exploration and exploitation of crude is being carried out. In addition, oil pollution is a human induced hazard hence as with natural hazards, improved understanding is needed for the sources, extent and responses to contamination in affected areas to be controlled. Like any other type of emergencies, oil spillage is dynamic and changes continuously, thereby making it arduous for planners and responders to produce robust plans towards short term and long term management goals. Hence, the need for an understanding of the nature, sources, impact and responses required to prevent or control their occurrence.[14] Risk modeling must be seen as an understating of the probability of occurrence of events of particular severity and the levels of uncertainty that exist in the data employed and the models themselves.[15]

Risk assessment is the determination of quantitative or qualitative value of risk related to a concrete situation.[16] A quantitative approach generally estimates the cost of risk and its reduction. When reliable data on likelihood and costs are not available, qualitative approach is suitable. In this case, the likelihood of the outcome, or the magnitude of the consequences, is expressed in subjective terms such as ‘high’, ‘medium’ or ‘low’. Risk analysis and assessment based on data mining techniques have been described.[17] Oil spill risk assessment systems are described in Ref.18 and 19. In Ref.20, a study on the means of forecasting ship’s oil spills was undertaken. The analysis revealed that conventional techniques focused on the causal relationship between the regression model and time series analysis, which does not completely reflect the intrinsic characteristics of the structure and the complexity of the dynamic data. The importance of synthetic risk assessment of ship’s oil spill risk and the assessment model of ships’ oil spill risk based on fuzzy neural network model is proposed in Ref.21.

The complexity and the dynamism of oil spillage require sophisticated methods and tools for the construction of knowledge systems that can be used as solutions to such problems. The search for systems that can solve increasingly complex problems has stimulated research in a number of hybrid intelligent systems. Among such systems, Neuro-Fuzzy Genetic Systems, which learn from the environment and reason about its state.[22] Adaptive Neuro fuzzy Inference System (ANFIS) combines the advantages of both neural networks and Fuzzy Inference System. This paper attempts to develop an intelligent system based on the hybridization NN, FL and GA for knowledge discovery and classification of spillage risks patterns. Fuzzified attributes of Oil spillage were the inputs to the system while fuzzified magnitude of oil spillage is the output variable.

2 Methodology

The stages of this work and the major components are outline in Figure 1. 1008 incidences of Oil spillage collected by National Oil Spill Detection and Response Agency (NOSDRA) from the Niger Delta Region of Nigeria, served as the dataset. The attributes of the Oil spillage dataset is given in Table 1.

Table 1: Attributes of Oil Spillage Dataset

Input Indicators	Description	No. of Levels	Values	Codes
Location	Location of Oil spill	2	Onshore	ON
			Offshore	OFF
Cause	Source of oil Spillage	6	Operational/Maintenance Error	OME
			Sabotage	Sab
			Equipment Failure	Eqf
			Corrosion	Cor
			Yet to be determined	Ytd
			Others	Oth
Type	Type of Spilled Oil	6	Refined Product	Re
			Crude	Cr
			Chemical	Ch
			AGO	AGO
			Condensate	Con
			Others	Oth
Date	Date of occurrence of Spillage	3	Day	-
			Month	-
			Year	-
Magnitude	Magnitude (Severity) of Oil Spillage	5	Very Low	VL
			Low	LO
			Medium	ME
			High	HI
			Very High	VH

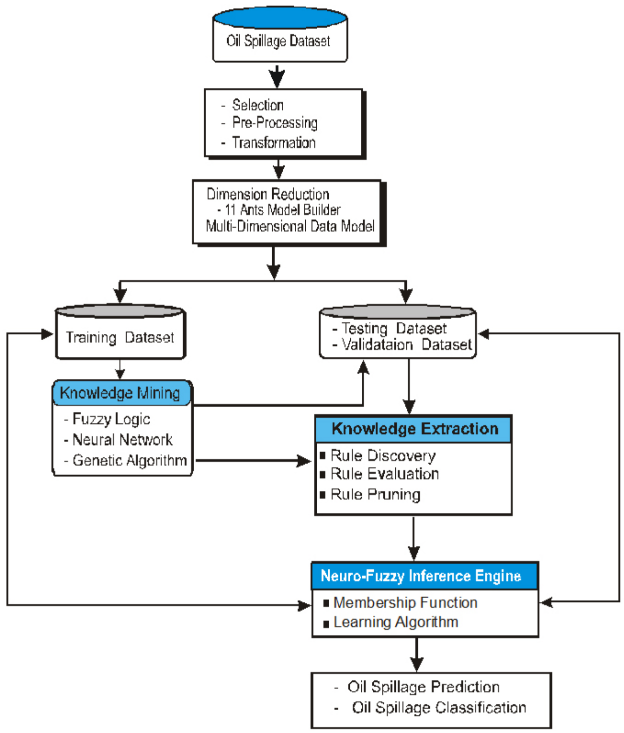

Figure 1: Outline of stages for Oil Spillage Knowledge Discovery System Development

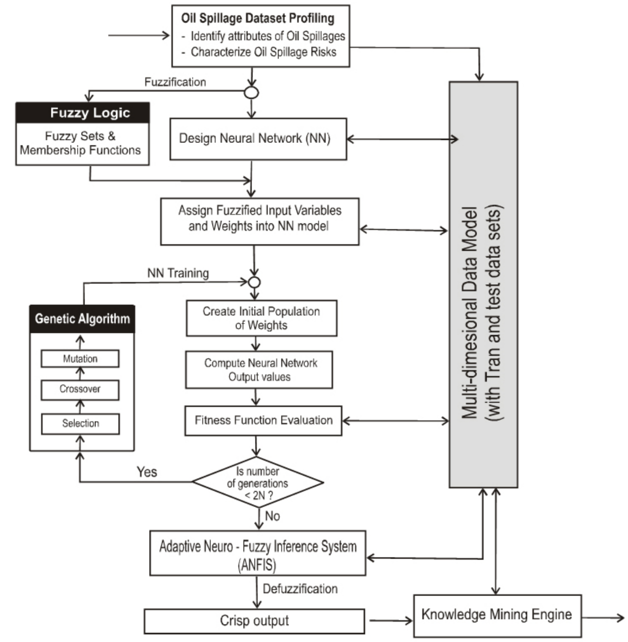

Figure 2: Interaction of KB Components and Procedure of Neuro-Fuzzy-Genetic Hybrid Design

Attribute selection and dataset pre-processing involved the identification of the input and target, which were input and output of the neural network. The target variable is magnitude of spillage, while Day-of-Occurrence, Month-of-Occurrence (M), Year-of-Occurrence(Y), Time-of-Occurrence (I), Location-of-Spill(L), Cause-of-Spill (C) and Type-of-Spill (T) were input variables. 11Ants Model Builder offers a straightforward and effective means for dimension reduction and easy data preparation.[23,24] The pre-processing of the dataset, input rank analysis and dataset splitting were performed with 11 Ants Model Builder. The result of input sensitivity analysis shows that Type has 0.362 as weight while Cause and Location contributed 29.6% and 19.3% respectively to the Magnitude of Spillage. Day has 0.0821 as weight while Time has 0.0235. Year of occurrence showed no contribution to the magnitude of oilspillage. Day, Month, Time are insignificant and noisy in the estimation of oil spillage risks. However, Year was not used for the training of the NN while the other insignificant indicators were basis for rule pruning. The dataset were split into training (70%), testing (15%) and validation (15%) dataset. The major components of the system are Knowledge Base (KB), Knowledge Mining, Inference Engine and Decision Support Engine. The KB has NN, FL and GA as components. The design algorithm of hybrid platforms[25,26] were studied and modified to suit the design of the KB. The interaction of components in the KB and hybrid design procedure is as shown in Figure 2.

The NN is the central component of the system.[27] It receives fuzzified inputs and communicates risks levels associated with oil spillage to the environment. The GA component provides optimal set of weights for training NN while the FL acts as a tool for modeling imprecise and vague knowledge, and for the provision of evaluation and membership functions for the GA and NN.

Fuzzy sets of oil spillage indicators are expressed as functions while the elements of the set are mapped to their degree of membership. A fuzzy set A in a universe of discourse X is given in Equation 1 and can be expressed in the form given in Equation 2.[28]

$$A = \{\mu_A(x) : x\varepsilon X\} \tag{1}$$

$$A = \{\frac{\mu_A(x)}{x} : x\varepsilon X\} \tag{2}$$

Where $A = \{\mu_A(x) : x\varepsilon X\}$ is a mapping known as membership functions (MF) of the fuzzy set A and $\mu_A(x)$ is the degree of membership of x in X in the fuzzy set A. In this work, $\mu_A(x)$ further mapped to the fuzzy linguistic values of *"very low", "low", "medium", "high" or "very high"* specified in the rules. Equation 3 is an example of MF for a linguistic term 'high'.

$$high(x) = \begin{cases} 0 & if\ x < 0.6 \\ \frac{x-0.6}{0.2} & if\ 0.6 \le x < 0.8 \\ 1 & if\ x \ge 0.8 \end{cases} \tag{3}$$

The NN is a 3-layered feed-forward architecture with sigmoid function for neuron activation.[21] Fuzzified likelihood of oil spillage attributes are inputs to the NN while severity level of oil spillage is output. The hidden layer consists of 12 neurons. Optimal weights of NN were generated via GA in four stages; initial population generation, selection, crossover and mutation.[25,29,30] A gene is represented as a connection weight between the *ith* input node and *jth* hidden node (ω_{ij}) or between the *jth* hidden node to *kth* output node (ω_{jk}). A chromosome is encoded as a string of genes $\{\omega_{11}, \omega_{12}, \omega_{13}, \cdots, \omega_{1m}, \omega_{21}, \omega_{22}, \cdots, \omega_{mq}, \cdots, \omega_{qv}\}$ where m represents the number of input nodes, q represents the number of nodes in the hidden layer and v represent the number of output nodes.[29,30] The set of weights is 12×6 matrix as $\{\omega_{11}, \omega_{21}, \cdots, \omega_{12,1}, \omega_{1,2}, \cdots, \omega_{12,2}, \cdots, \omega_{12,6}\}$. The GA encoding scheme is a combination of binary string and real value encoding. Binary encoding and transformed to real value encoding using Equations 4 and 5.

$$g_i = \begin{cases} 1 & if\ b_1 = 1 \\ -1 & if\ b_1 = 0 \end{cases} \tag{4}$$

$$R_i = \frac{g_i}{10} \sum_{t=2}^{m} (b_t \times 2^{m-t}) \tag{5}$$

where R_i is the real value encoding of the *ith* gene, $i = 2, 3, \cdots m$. g_i is the sign bit of genei. The selection operator evaluated each individual providing fitness values, which are then normalized. The normalized fitness value is given as:

$$T_i = \frac{y_i}{\frac{1}{N}\sum_{j=1}^{p} p y_i} \tag{6}$$

Where $j = 1, 2, 3 \cdots p$ and y_i is the probability of the ith chromosome to be selected for crossover and mutation. The algorithm terminates when 2N iterations are completed with individuals with the largest fitness value being selected. This set of chromosomes represents the optimal weights of the NN.

The Neuro-fuzzy inference engine is a five layered, first-order Sugeno ANFIS system for the evaluation and extraction of rules and the production of fuzzy output. The rule base consists of rules of the form:

$$\text{IF } (C_i\ is\ A_1^r) \text{ and } (T_j\ is\ A_2^r) \text{ and } (L_k\ is\ A_3^r) \text{ THEN } f = (p_0^r + P_1^r C_i + P_2^r T_j + P_3^r L_k) \tag{7}$$

where r is the rule number, C_i is the *ith* Cause of Spillage, T_j is the *jth* Type of spilled oil, L_k is the kth Spillage Location, f is the linear output within the fuzzy region specified by the fuzzy rule. The variables $p_0^r, p_1^r, p_2^r, p_3^r$ are the linear parameters in the consequent part of the sugeno-fuzzy model that is determined during the training process. A_1^r, A_2^r, A_3^r are linguistic values *very low, low, medium, high, very high* characterized by appropriate membership function μ_{A_n}. Each layer consists of the nodes described by the node function.

Layer 1 is the input layer. It has Cause, Location and Type as inputs. Each node in this layer generates fuzzy membership grades for the inputs. This is given by:

$$\begin{cases} O_i^1 = \mu_{A_i}(C_i) & i = 1, 2, \cdots 6 \\ O_i^1 = \mu_{A_j}(T_j) & j = 1, 2, \cdots 6 \\ O_i^1 = \mu_{A_k}(L_k) & k = 1, 2 \end{cases}$$

The general form of the triangular MF is presented in Equation 8 and 9 while generailized bell-shaped and trapezoidal MFs are given in Equation 10 and Equation 11 respectively.[31]

$$\begin{cases} 1 & if\ x = b \\ \frac{x-a}{b-a} & if\ a \le x < b \\ \frac{c-x}{c-b} & if\ b \le x < c \\ 0 & if\ c = x \end{cases} \tag{8}$$

$$\mu_A(x) = \max\left(\min(\frac{x-a}{b-a}, \frac{c-x}{c-b}), 0\right) \tag{9}$$

where a and c are the parameters governing triangular MF; b represents the value for which $\mu(x)$= 1 and is defined as $b = \frac{a+c}{2}$.

$$\mu_A(x) = \frac{1}{1 + \left\{\left(\frac{x-c}{a}\right)^2\right\}^b} \tag{10}$$

where a, b and c are the parameters governing generalized bell-shaped MF.

$$\mu(x) = \begin{cases} \frac{x-a}{b-a} & if\ a \le x \le b \\ 1 & if\ b \le x \le c \\ \frac{d-x}{d-c} & if\ c \le x \le d \\ 0 & otherwise \end{cases} \tag{11}$$

where a, b, c and d are the parameters governing trapezoidal-shaped MF.

Layer 2 is the rule node. It computes the firing strengths, O_i^2 of each rule as given in Equation 12. These are the products of the corresponding membership degrees obtained from layer 1. The normalization layer (layer 3) computes the ratio of the each rule firing strength to the sum of all rules' firing strength. The normalized output, $\bar{w}_i$ is given in Equation 13. Layer 4, the defuzzification layer, consists of consequent nodes for calculating the contribution of each rule to the overall output as in Equation 14. The overall output of the ANFIS model is determined by summing all incoming signals by layer 5. This is done by transforming each rule's fuzzy results into crisp value. This paper adopts the centroid method depicted in Equation 15.

$$O_i^2 = w_i = \mu_{A_n}\mu_{B_n}(T_j)\mu_{D_n}(L_k) \tag{12}$$

$$O_i^3 = \bar{w}_i = \frac{w_i}{\sum_i w_i} \tag{13}$$

$$O_i^4 = w_i f_i = w_i(p_0^i + p_1^i C_i + p_2^i T_j + p_3^i L_k) \tag{14}$$

$$O_i^5 = M = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{15}$$

ANFIS applies either a hybrid learning algorithm or the back-propagation method to identify and update the membership function parameters of the output. The hybrid method involves the combination of least-squares and back propagation gradient descent methods for the fuzzy inference system training.[32,33] In Hybrid learning algorithm, when the premise parameters are fixed, the overall output of the ANFIS is expressed as a linear combination of consequent parameters $p_0^r, p_1^r, p_2^r, p_3^r$ and the output can be expressed as follows:

$$\begin{aligned} M =& \sum_i \bar{w}_i f_i = \bar{w_1} f_1 + \bar{w_2} f_2 + \bar{w_3} f_3 \\ =& (\bar{w_1} C_i) p_1^i + (\bar{w_1} T_j) p_2^i + (\bar{w_1} L_k) p_3^i + (\bar{w_1} p_0^i) + (\bar{w_2} C_i) p_1^i \\ &+ (\bar{w_2} T_j) p_2^i + (\bar{w_2} L_k) p_3^i + (\bar{w_2} p_0^i) + (\bar{w_3} C_i) p_1^i + (\bar{w_3} T_j) p_2^i + (\bar{w_3} L_k) p_3^i + (\bar{w_3} p_0^i) \end{aligned} \tag{16}$$

It consist of the forward and backward pass, in the forward pass, each node's output goes forward until it reaches the fourth layer and the consequent parameters are identified by the least squares method. During the backward pass, the premise parameters are updated by gradient descent as the error signal propagates backwards.

Suppose the oil spillage training dataset has m entries, let B be the output matrix, (Oil spillage risks), X represents the matrix of consequent parameters and A is the premise parameters as follows:

$$B = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_m \end{bmatrix}, X = \begin{bmatrix} p_0^1 \\ p_1^1 \\ p_2^1 \\ p_3^1 \\ \vdots \\ p_0^3 \\ p_1^3 \\ p_2^3 \\ p_3^3 \end{bmatrix}$$

and

$$\begin{bmatrix} \bar{w_1}C_1 & \bar{w_1}T_1 & \bar{w_1}L_1 & \bar{w_1} & \bar{w_2}C_1 & \bar{w_2}T_1 & \bar{w_2}L_1 & \bar{w_2} & \bar{w_3}C_1 & \bar{w_3}T_1 & \bar{w_3}L_1 & \bar{w_3} \\ \bar{w_1}C_2 & \bar{w_1}T_2 & \bar{w_1}L_2 & \bar{w_1} & \bar{w_2}C_2 & \bar{w_2}T_2 & \bar{w_2}L_2 & \bar{w_2} & \bar{w_3}C_2 & \bar{w_3}T_2 & \bar{w_3}L_2 & \bar{w_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{w_1}C_m & \bar{w_1}T_m & \bar{w_1}L_m & \bar{w_1} & \bar{w_2}C_m & \bar{w_2}T_m & \bar{w_2}L_m & \bar{w_2} & \bar{w_3}C_m & \bar{w_3}T_m & \bar{w_3}L_m & \bar{w_3} \end{bmatrix}$$

Then $AX = B$, X is unknown with element from the consequent parameters set. This is a standard linear least squares problem, thereby the least squares estimator (LSE), X^* is given by Equation (16).[28,34]

$$X^* = (A^T A)^{-1} A^T B \tag{17}$$

The result gives the consequent parameters from which the fuzzy output of the system was derived. The output of the system is in the form as shown in Equation 18

$$T_i = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1w} \\ t_{21} & t_{22} & \cdots & t_{2w} \\ t_{31} & t_{32} & \cdots & t_{3w} \\ \cdots & & & \\ t_{v1} & t_{v2} & \cdots & t_{vw} \end{bmatrix} \tag{18}$$

$i = 1, 2, 3 \cdots u; j = 1, 2, 3, \cdots v; k = 1, 2, 3, \cdots w.$

3 Development of knowledge mining system

Neural Network training is presented in Section 3.1. In Section 3.2, rules discovery, pruning and clustering are presented. ANFIS model and results obtained from the experimental study are presented in Section 3.3.

3.1 Neural network training

The system was implemented with Matlab 7.7.0 (R2008b) as front-end tool, Microsoft Excel and Microsoft Access were the database management tools. The NN, FL and ANFIS toolboxes of Matlab were deployed in this system. The system used seventy percent of the data (706 samples) for training. Testing and validation were carried out with 151 records (15%) each. In every training session, GA selects training samples randomly from the entire dataset thereby generating different values of mean square error (MSE) depending upon which 70 percent of the input data was selected for training. The graphical representation of the NN performance during training, validation and testing on the dataset is presented in Figure 3 while the optimal training weights are presented in Table 2. As shown in Figure 3, the best performance is noticed at the 1000 epoch with MSE 0.2405, which is good.[35] The weight is a 12 $\times$ 6 matrix and within the range [-1,1] as specified in Equation 5. The regression plot, presented in Figure 4, depicts the relationship between the output and the target. Figure 4 consists of three axes representing the training, validation and testing data. The dashed line represents the perfect result (R=1). The solid line represents the best-fit linear regression line between outputs and targets. The R-value gives an indication of the relationship between the output and the corresponding target. In all cases, (training, validation and testing) R-value is > 0.839, which indicates a good fit showing a strong linear relationship between output and target data.

3.2 Rules discovery, pruning and clustering

Pattern discovery from the trained NN was performed in three stages using the modified Apriori Association rule-mining algorithm. It involved the identification of frequent k-antecedent set, formulation of multidimensional rules, pruning less interesting rules and clustering rules based on categorical levels of each spillage indicator. 329 rules were extracted based on uniform minimum support and confidence thresholds of 15% and 10% respectively. The visualization of the rule confidence and support is presented in Figure 5.

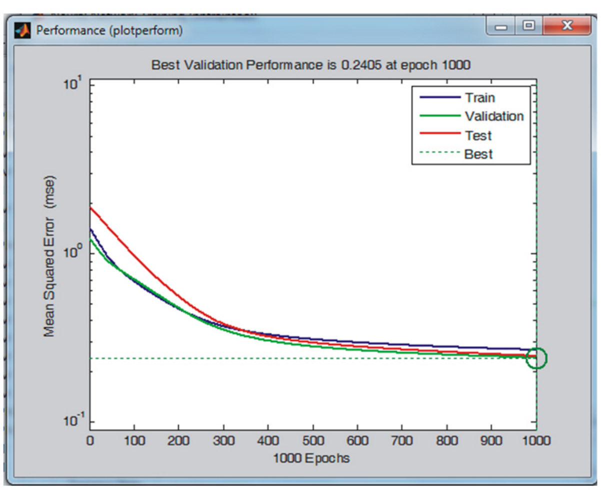

Figure 3: Performance of NN during Training, Validation and Testing

Table 2: Matrix of Optimal Input Layer Weights

	Inputs Neurons						
Hidden layer Neurons		1	2	3	4	5	6
	1	-0.6	-0.1	0.6	1.3	-0.7	-1.2
	2	0.7	-1.0	-0.1	-1.2	-0.9	-0.7
	3	0.4	-0.3	-1.0	0.8	-0.7	-1.0
	4	-0.3	-0.9	-0.8	0.9	1.0	-0.9
	5	0.2	0.9	1.0	1.0	0.3	-0.8
	6	-1.0	-0.6	-1.0	-1.0	0.2	-0.3
	7	-1.0	0.1	0.8	-0.2	-1	-1.0
	8	0.1	-1.0	0	-0.7	1	1.0
	9	0.8	0.3	0.7	0.9	0.4	1.0
	10	1.0	-0.8	-1.0	-0.3	-0.5	0.1
	11	-1	0.7	-0.2	1.0	0.1	0.1
	12	0.3	0.7	-1.0	-1.0	-0.4	-0.6

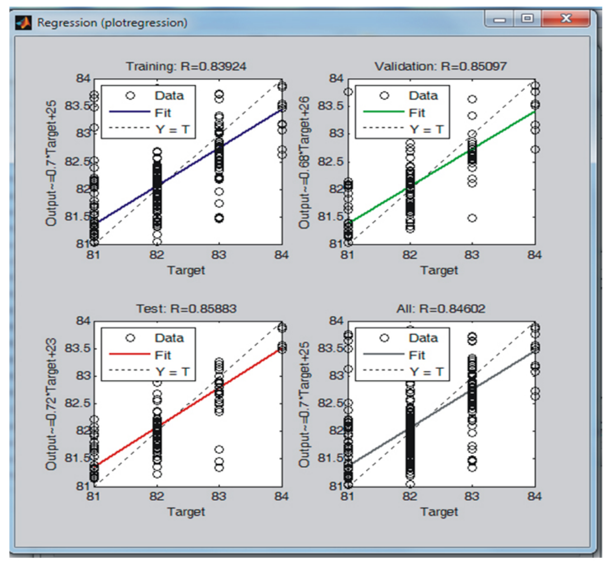

Figure 4: Relationship of NN Output and Target for Training, Validation and Testing Datasets.

Table 3: Distribution of Extracted Rules

Size of Rule Antecedent	Number of Rules Count	Percentage (%)
1	77	23.40
2	110	33.43
3	95	28.88
4	40	12.16
5	7	2.13
Total	**329**	**100**

As shown in Figure 5, the support of rule decreases as the number of rules increases while the confidence of rules also decreases as number of rules increases. The result shows that confidence of a rule is higher than its support while an increase in the number of rules causes a decrease in the rule support and confidence. This result shows that the rules are efficient for classification.[36] The size of rules' antecedent ranges from 1 to 5. The distribution of indicators in the rule antecedent part is presented in Table 3.

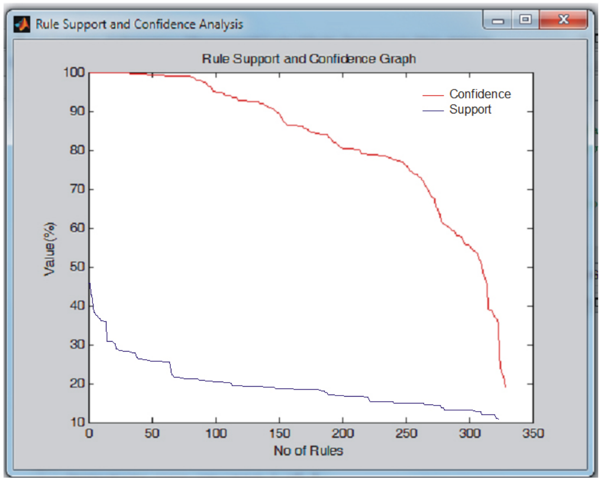

Figure 5: Graphical Analysis of Rule Support and Confidence

As depicted in Table 4 rules with antecedent size less than 3 are 56.83% and those greater than 2 are 43.27%. Though Rules with antecedent size 1 and 2 have high support and confidence, they may occur by chance and may be misleading.[37,38] Hence, these rules were eliminated from the rule-set. The importance of rule pruning and rule interestingness measures are reported in Ref.36 and 39. The rule pruning method used[39] was adopted in this work, with three interestingness measures of confidence, support and rule antecedent size. Rule Support is often used to represent the relevance of an association pattern and very efficient for pruning exponential search space of candidate patterns due to its downward closure property.[40,41] Confidence is an accuracy measure of a given rule. Support and confidence based pruning is a viable technique for examining the quality of association rules.[36] In this paper, pruning of weak and uninteresting rules was performed, in the following stages; firstly, specifying a user-defined minimum support and minimum confidence thresholds of 10% and 15% respectively. Secondly, rules with antecedent size less than 3 were discarded. Thirdly, rules with antecedent part consisting of any insignificant oil spillage indicators were also discarded. The resultant ruleset contains 73 rules and form the rules of the ANFIS model for prediction and classification of oil spillage patterns.

Table 4: Performance of ANFIS Model on Membership Functions and Learning Algorithms

MF	**MF Description**	**Back-propagation Algorithm**		**Hybrid Algorithm**		***Average MSE***
		Training Error	***Testing Error***	***Training Error***	***Testing Error***	
Trapmf	Trapezoidal-Shaped MF	0.45099	0.31698	0.414315	0.221403	**0.35092**
Dsigmf	Difference Sigmoidal MF	0.48397	0.36879	0.414315	0.221404	**0.37212**
Trimf	Triangular MF	0.44881	0.29646	0.414315	0.221402	0.34525
psigmf	Product Sigmoidal MF	0.46108	0.31870	0.414315	0.221404	0.35388
pimf	Pi-shaped MF	0.49958	0.39238	0.414315	0.221403	**0.38192**
gauss2mf	Gaussian Combinational	0.47951	0.30252	0.414315	0.221403	**0.35444**
gbellmf	Generalized Bell MF	0.82719	0.66664	0.414315	0.221404	**0.53239**
gaussmf	Symmetric Gaussian MF	1.03783	1.0291	0.414315	0.221404	**0.67566**
Average MSE		**0.58612**	**0.461446**	**0.414315**	**0.221403**	

3.3 ANFIS model and results

The ANFIS model is a 5-layered structure consisting of a total 166 nodes. The structure of the ANFIS model for oil spillage predication and classification is presented in Figure 6. The inputs to the system are *Cause, Type and Location*. There are 15 nodes in the fuzzification layer, which represents linguistic values set *Very Low, Low, Medium, High, Very High* for each input node. The rule layer has 73 nodes; each node represents a rule antecedent part. The normalization layer also have 73 nodes, each node is the rule consequent part corresponding to the rule antecedent node of the rule layer. The defuzzification and output layer has one node each. The output of the system is the severity of oil spillage risks. The final surface views of ANFIS rules are presented in Figure 7, 8 and Figure 9 respectively.

ANFIS systems produces different results depending on the type of MF and learning algorithm.[33] The mean squared eror (MSE) and root mean square error (RMSE) are standard statistical metrics to measure model performance.[42] To find the most fitted model, the ANFIS model was tested with eight (8) types of MFs and two learning algorithms (back-propagation and hybrid algorithms). MSE was the performance measure used to evaluate the ANFIS model. MSE resulting from the training and testing of the ANFIS model using each type of MF and a learning algorithm is presented in Table 4.

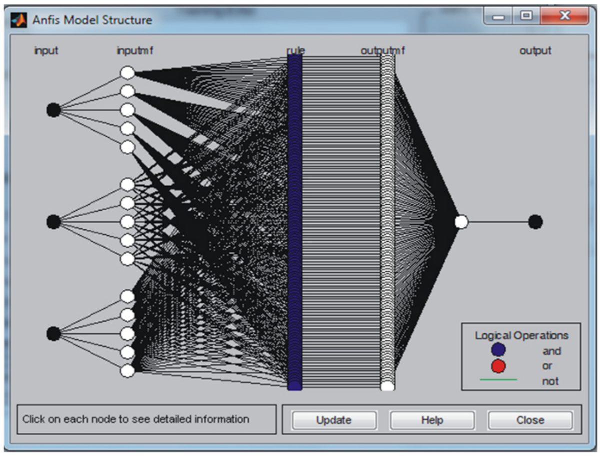

Figure 6: Structure of ANFIS Model for Oil Spillage Risk Analysis

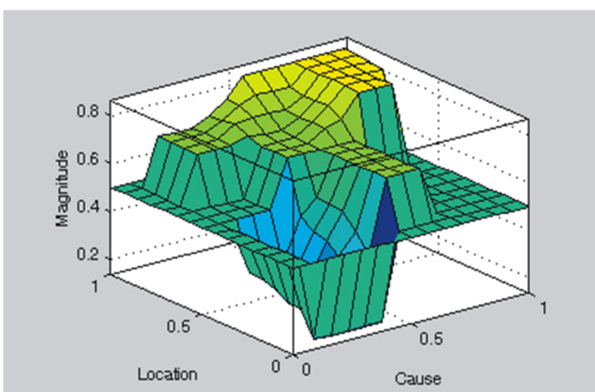

Figure 7: Final Surface View of ANFIS Rules for Cause and Location

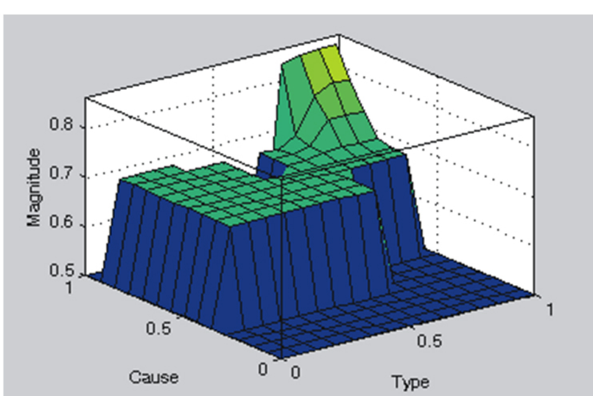

Figure 8: Final Surface View of ANFIS Rules for Cause and Type

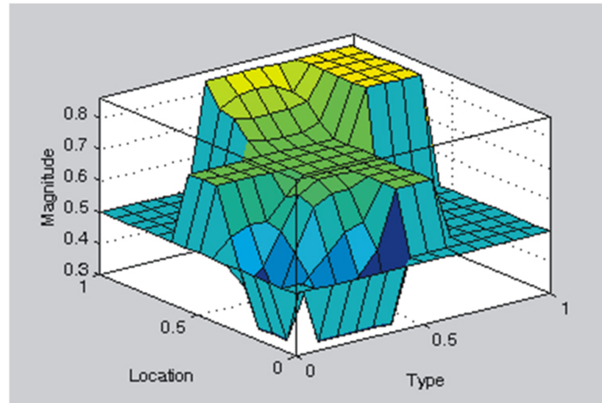

Figure 9: Final Surface View of ANFIS Rules for Location and Type

As shown in Table 4, the performance of the ANFIS model is much better with the hybrid algorithm than back-propagation algorithm. There is no difference in performance resulting from a change in the type of MF with the hybrid-learning algorithm. However, the performances of the ANFIS model vary by type of MF with back-propagation learning algorithm. The Triangular MF yielded the least MSE of 0.44881 and 0.2965 for training and testing respectively while the worst performance was observed when the Symmetric Gaussian MF was used in conjunction using the back-propagation learning algorithm yielding training MSE of 1.0378 and testing MSE of 1.0291. The overall best performing MF in both back-propagation and hybrid learning algorithms is the Triangular MF with an average MSE of 0.3453. This suggests why triangular MF is widely and most commonly used in the construction of fuzzy inference systems. In this paper, the resultant ANFIS model is the one with triangular MF and hybrid algorithm, and was used for the prediction and classification of oil spillage patterns.

Two Fuzzy Inference System (FISs) structures were generated and used in ascertaining the performance of the resultant ANFIS model. First, a Sugeno-type FIS (genfis3) was built by extracting rules that model the oil spillage's dataset behaviour using membership functions for rules' antecedent and consequent parts. The second FIS (genfis2), a Sugeno-type was generated using subtractive clustering in determining the number of rules and antecedent membership functions; and linear least squares estimation method for determining each rule's consequent. The summary of the performances of the resultant ANFIS model based on these FISs is presented in Table 5.

The result depicted in Table 5 shows that fismat3 (genfis2) yielded the lowest with the training (0.5056) and checking (0.2660) errors. However, the resultant ANFIS model performed better than the generated FISs. This confirms the suitability of the resultant ANFIS model for classification and prediction of oil spillage risks.

Table 5: Performance of FISs on the Resultant ANFIS Model

S/N	FIS	Function	Type	Training Error (trnMSE)	Checking Error (chkMSE)	Average Error
1	Fismat1	genfis3	Sugeno	0.82956	0.6608	0.74518
2	Fismat3	genfis2	Sugeno	0.5056	0.2660	0.3858

The performance metrics are MSE[42,43] and Mean Absolute Percentage Error.[44] The performance of the resultant ANFIS model is shown in the plot of the ANFIS predicted output and the target dataset depicted in Figure 10. As shown in Figure 10, the number of epochs used by the system was 40 while the average testing error is 0.22140. The plot also shows that there is no significant deviation between the predicted values and the actual target of the testing dataset. The Mean Absolute Percentage Error (MAPE) of the ANFIS model is 0.00813 which proves that the ANFIS model is satisfactory[44] and suitable for the prediction and classification of oil spillage patterns.

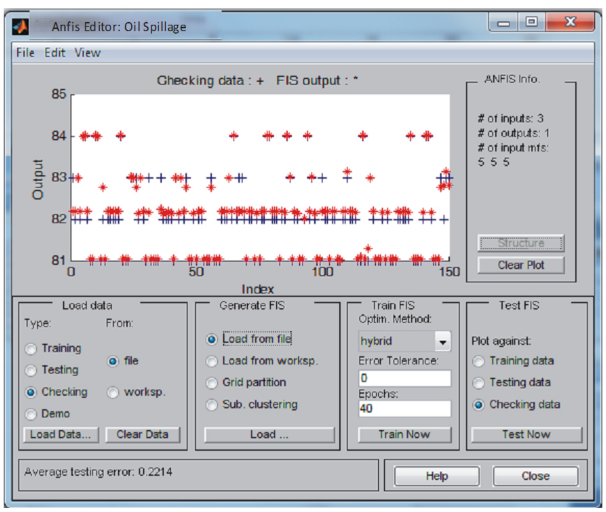

Figure 10: Plot of ANFIS Output and Target Dataset

4 Conclusion

This paper proposes an intelligent hybrid system driven by Sugeno-Type ANFIS for the identification, extraction and classification of oil spillage risk patterns. The methodology is based on NN, GA and FL hybridization for the discovery of patterns in terms of relationships, rules and interdependencies from oil spillage dataset. The model deploys and integrates the advantages of NN, FL and GA thereby compensating for the drawbacks of each tool. Moreso, the efficiency and capabilities of sugeno-type ANFIS in the predication and classification is oil spillage severity was demonstrated. Dataset consisting of 1,008 records was used for training, validation and testing of the system. Result of sensitivity analysis shows that Cause, Location and Type of spilled oil have cumulative significance of 85.1%. MATLAB R2008b was the system tool. Optimal weights of Neural Network (NN) were determined via Genetic Algorithm with hybrid encoding scheme. The Mean Squared Error (MSE) of NN training is 0.2405. NN training, validation and testing results yielded R > 0.839 in all cases indicating a strong linear relationship their corresponding target data. 329 rules were extracted from NN. Pruning of non interesting rules was performed with support (15%) and confidence (10%) minimum thresholds and antecedent-size of three (3) resulted in seventy three (73) interesting rules used in the ANFIS model. The performance of the ANFIS was evaluated with eight different types of membership functions (MFs) and two learning algorithms. Triangular MF gave the best performance, followed by Trapezoidal MF. This work confirms the adaptability of triangular and trapezoidal MFs to complex problems and explains why both MFs are the most commonly used MFs. In term of learning algorithm, the hybrid-learning algorithm yielded a better performance. The ANFIS model reported in the paper adopted triangular MF and hybrid learning algorithm for the predication and classification of oil spillage risk patterns. Average training and testing MSE of the model is 0.414315 and 0.221402 respectively with MAPE = 0.8128%. The results show that ANFIS based systems provide satisfactory results and are suitable in the prediction and classification of oil spillage risk patterns. As a further research, comparative analysis of ANFIS performance with GA as the learning algorithm is necessary.

References

[1] Kao, W,. Hung, J. C. and Hsu, V. Using Data Mining in MURA Graphic Problems. Journal of Software. 2008; 3(8): 73-99.

[2] Han, J., Cai, Y. and Cercone, N. Knowledge Discovery in Databases: An Attribute-Oriented Approach. In Proceedings of the 18th Conference of Very Large Databases, Vancouver, British Columbia, Canada. 1992 August 24-27: 547-559.

[3] Miller, H. J. and Han, J. Geographic Data Mining and Knowledge Discovery-An Overview. Geographic Data Mining and Knowledge Discovery. In: Miller, H.J., Han, J.(eds). Taylor and Francis. 2001: 3-32.

[4] Mosley, R. The Use of Predictive Modeling in the Insurance Industry", Pinnacle Actuarial Resources. 2005.

[5] Gurney, K. and Gurney, K. H. An Introduction to Neural Networks, CRC Press. 1997.

[6] Zhang, J., Jiang, Y. and Yan, H. Committee Machines with Ensembles of Multilayer Perception for the Support of Diagnosis of Heart Diseases. In Proceedings of the International Conference on Communications, Circuits and Systems. 2006; 1(3): 2046–2050.

[7] Aeinfar, V., Mazdarani, H., Deregeh, F., Hayati, M. and Payandeh, M. Multilayer Perception Neural Network with Supervised Training Method for Diagnosis and Predicting Blood Disorder and Cancer', IEEE International Symposium on Industrial Electronics, 5–8 July, Seoul, Korea, 2009; 2075–2080.
[8] Rumelhart, D. E. Schemata: The Building Blocks of Cognition', In R.J. Spiro, B. Bruce and W.F. Brewer (Eds.): Theoretical Issues in Reading and Comprehension, Erlbaum, Hillsdale, NJ. 1980.
[9] Goldberg, D. E. Genetic Algorithms in Search, Optimization & Machine Learning, Pearson Education Inc. 2006.
[10] Jang, J. S., Sun, C. T. and Mitzutani, E. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall Inc., Upper Saddle River, USA. 2007; 8-33.
[11] Vijaya, K. Nehemiah, H. K., Kannan, A. and Bhuvaneswari, N. G. Fuzzy Neuro Genetic Approach for Predicting the Risk of Cardiovascular Diseases. International Journal of Data Mining, Modeling and Management. 2010; 2(4): 388-402. http://dx.doi.org/10.1504/IJDMMM.2010.035565
[12] Mitchell M. Introduction to Genetic Algorithms. MIT Press, Cambridge, MA USA. 1998.
[13] Montana, D. J. and Davis, L. Training Feedforward Neural Networks Using Genetic Algorithms. In Proceedings of International Conference on Artificial Intelligence, Detroit, Michigan. 1992; 1: 762–767.
[14] Akinyokun, O. C. and Inyang, U. G. Experimental Study of Neuro-Fuzzy-Genetic Framework For Oil Spillage Risk Management. Artificial Intelligence Research. 2013; 2(4): 13-35.
[15] Granger K, Jones T., Leiba, M. and Scott G. Community Risk in Cairns: A Multi-Hazard Risk Assessment. Australian Journal of Emergency Management. 1999; 14(2): 25-26.
[16] Apostolakis, G.E. How Useful is Quantitative Risk Assessment? Risk Analysis. 2004; 24(3): 515-520. PMid:15209926 http://dx.doi.org/10.1111/j.0272-4332.2004.00455.x
[17] Cummings, M. C., McGarvey, D. and Vinch, Peter M. Homeland Security Risks Assessment Volume II. Methods, Techniques and Tools. Homeland Security Institute. 2006.
[18] Castanedo, S., Abascal, A. J., Medina, R., Fernandez, F., Liste, M. and Olabarrieta, M. Development of a GIS-based Oil Spill Risk Assessment System. In Proceedings of the 2nd International Conference on Computational Intelligence and Software Engineering. (CISE 2010) Wuhan, China. 2010 December 10-12: 1-6.
[19] Ji, Z., Johnson, W., Marshall, C. and Lear, E. Oil Spill Risk Analysis: Gulf Mexico Outer Continental Shel(OCS) Lease Sales, Central Planning Area and Western Planning Area, 2007-2012, and Gilfwide OCS Program, 2007-2046. OCS Report U.S. Interior Mineral Management Service Environmental Division. 2007.
[20] Wenxue, C., Yanwu, Z., Yongqiang, Shi and Huiling, Z. Threat Level Forecast for Ship's Oil Spill Based on BP Neural Network Model. In Proceedings of International Conference on Computational Intelligence and Software Engineering. Wuhan, China. 2009 December: 2374-2377.
[21] Cao, X and Fan, S. The Synthetic Assessment Modeling of Ships' Oil Spill Risk Based on Fuzzy Neural Network. In Proceedings of International Conference on Intelligent Systems and Applications (ISA). Wuhan, China. 2013 May 22-23: 368-371.
[22] Zanchettin, C, Minku, L and Ludermir, T. Design of Experiments in Neuro-Fuzzy Systems. International Journal of Computational Intelligence and Applications. 2010; 9(2): 137-151. http://dx.doi.org/10.1142/S1469026810002823
[23] Inyang, U. G. Generating Rules for Students' Performance Prediction in Tertiary Institutions. ICASTOR Journal of Mathematical Sciences. 2011; 5(2): 205-216.
[24] Inyang, U. G. and Joshua, E. E. Fuzzy Clustering of Students' Data Repository for At-Risks Students Identification and Monitoring. Computer and Information Science. 2013; 6(4): 37-50.
[25] Shanthi, D; Sahoo, G. and Saravanan, N. Evolving Connection Weights of Artificial Neural Networks Using Genetic Algorithm with Application to the Prediction of Stroke Disease. International Journal of Soft Computing. 2009; 4(2): 95-102.
[26] Rivero, D., Dorado, E., Efernandez-Banco, E. and Pazos, A. A Genetic Algorithm for ANN Design, Training and Simplification. In Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence. 2009; 391-398.
[27] Jain, L. and Martin, N. M. Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications. CRC Press, LLC. 1998.
[28] Yadav, Ramjeet Singh, Soni, A. K., and Pal, Saurabh. Modeling Academic Performance Evaluation Using Hybrid Fuzzy Clustering Techniques. 2014; 8(3): 98-111
[29] Cabestany, J., Rojas,I. and Gonzalo, J. Advances in Computational Intelligence. In Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence. Salamanca, Spain. 2009 June 10-12th: 433-448.
[30] Shahriar, A. and Kianfar, J. A Hybrid Neuro-Genetic Approach to Short Term Traffic Volume Prediction. International Journal of Civil Engineering. 2009; 7(1): 41-48.
[31] Dubois, D. and Prade, H. What are Fuzzy Rules and How to Use Them? Fuzzy Sets Systems. 1996; 84(2): 169–185. http://dx.doi.org/10.1016/0165-0114(96)00066-8
[32] Jang, J. and Sun C. Neuro-Fuzzy Modeling and Control. In Proceedings of the IEEE. 1995. 83(3): 378–406. http://dx.doi.org/10.1109/5.364486
[33] Mayilvaganan, M. K. and Naidu, K. B. Comparison of Membership Functions In Adaptive Network-Based Fuzzy Inference System (ANFIS) For The Prediction of Groundwater Level of A Watershed. Journal of Computer Applications Research and Development (JCARD). 2011; 1(1): 35-42.
[34] Yan, H, Zou, Z. and Wang, H. Adaptive Neuro Fuzzy Inference System for Classification of Water Quality Status. 2010; 22(12): 1891–1896.
[35] Abhishek, K., Kumar, A., Ranjan, R., Kumar, S. A Rainfall Prediction Model using Artificial Neural Network. IEEE Control and System Graduate Research Colloquium. 2012: 82-87.
[36] Tan, Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava. Selecting the Right Objective Measure for Association Analysis. Information Systems. 2009: 293–313.
[37] Chandra, E. and Nandhini, K. Knowledge Mining from Student Data. European Journal of Scientific Research. 2010; 47(1): 156-163.
[38] Inyang, U. G. and Umoren, M. U. Students' Dataset Mining for Academic Performance Risk Patterns Identification. World Journal of Applied Science and Technology. 2011; 3(2): 57-64.
[39] Kannan, S. and Bhaskaran, R. Role of Interestingness Measures in CA Rule Ordering for Associative Classifier: An Empirical Approach. Journal of Computing. 2010; 2(1): 8-15.
[40] Agrawal, R. and Srikant R. Fast Algorithms for Mining Association Rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), September 12-15, 1994, Santiago de Chile, Chile. 1994: 87–499.
[41] Bayardo, R. and Agrawal, R. Mining the Most Interesting Rules, in Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, CA. 1999 August: 145–154.
[42] Hossain, S. J., & Ahmad, N. Adaptive Neuro-Fuzzy Inference System (ANFIS) Based Surface Roughness Prediction Model for Ball End Milling Operation. Journal of Mechanical Engineering Research. 2012; 4(3): 112-129.
[43] Chai, T and Draxler, R. Root Mean Square Error (RMSE) or Mean Absolute Error(MAE)? –Arguments Against Avoiding RMSE in the Literature. Geoscientific Model Development. 2014; 7(1): 1247-1250. http://dx.doi.org/10.5194/gmd-7-1247-2014
[44] Hyndman, R. J. and Koehler, A. Another Look At Measures of Forecast Accuracy. International Journal of Forecasting. 2006; 22(4); 679-688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001