Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Enhancing the Interpretability of Soil Classification Data from a Neural Network System by Using Fuzzy Approaches

S.S.Baskar
Research Scholar
Department of Computer
St. Joseph's college (Autonomous)
Trichy, India

Dr L. Arockiam
Associate Professor
Department of Computer
St. Joseph's college (Autonomous)
Trichy, India

S.Charles
Assistant Professor
Department of Computer
St. Joseph's college (Autonomous)
Trichy, India

Abstract: The soil classification results that are available from a Neural Networks system have the probability of being numerical and vague. Analysis of this result to provide a human interpretable output becomes mandatory. The current project deals with constructing a complete system that provides human understandable output. The system is divided into two broad categories. The initial phase deals with organizing the attributes, finding their associations and finding the final set of values that are incorporated with their corresponding weights. These properties are incorporated into the Neural networks for finding the results. The second phase classifies these results, and constructs fuzzy rules based on the inputs and outputs. All the rules are combined to form a fuzzy rule set that provides the user with suitable results.

Keywords Soil classification; Neural networks; fuzzy logic; fuzzy association rule mining; Relief-F

1. INTRODUCTION

1.1. Soil Classification

Soil classification [6] & [7] refers to grouping soils based on their physical and chemical characteristics that distinguish each soil type. For soil resources, grouping soils by their inherent properties, behaviors, or genesis, can provide better results. Despite the differences, classification criteria can group similar concepts so that interpretations do not vary widely. Natural system approaches to soil classification, such as the French Soil Reference System are based on presumed soil genesis. Systems have developed, such as USDA soil taxonomy and the World Reference Base for Soil Resources, which use taxonomic criteria involving soil morphology and laboratory tests to inform and refine hierarchical classes. Another approach is numerical classification [11], also called ordination, where soil individuals are grouped by statistical methods such as cluster analysis. This produces natural groupings without requiring any inference about soil genesis.

1.2. Neural Networks

An Artificial Neural Network (ANN), [1]&[2] often just called a "Neural Network" (NN), is a mathematical model or computational model based on biological neural networks, i.e. it functions similar to that of a biological neural system. It consists of an interconnected group of nodes called artificial neurons and these nodes processes information and provide the output to the user. These nodes communicate with each other and every node has an assigned functionality and it performs the process on the data provided to it.

The ANN changes its structure based on external or internal information that flows through the network during the learning phase. A basic ANN consists of three layers, the input, hidden and the output layers. The hidden layer performs the processing in the network.

1.2.1. Feed Forward Neural Networks

A feed-forward neural network is an artificial neural network where connections between the neurons do not form a directed cycle. This is different from the simple neural networks. The feed-forward neural network is the simplest type of artificial neural network. In this network, the information moves in only in the forward direction. i.e. information moves from the input nodes, to the processing nodes and then to the output nodes. There are no cycles or loops in the network.

1.3. Feature Selection

Feature selection [8], also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features. In general, all data contains both relevant and irrelevant information. The data might also contain redundant features, that may not be of much use. Hence we use the feature selection methods to eliminate the noisy data (i.e. irrelevant and redundant information) from the set. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features.

1.4. Relief-F

Relief algorithms [3] are used for the estimation of the attributes in a system. Estimation of attributes is performed by evaluating the level of dependencies between the attributes. In addition, their quality estimates have a natural interpretation. While they have commonly been viewed as feature subset selection methods that are applied in prepossessing step before a model is learned, they have actually been used successfully in a variety of settings, e.g., to select splits or to guide constructive induction in the building phase of decision or regression tree learning, as the attribute weighting method and also in the inductive logic programming.

1.5. Fuzzy Logic

Fuzzy Logic [10] incorporates a simple, rule-based IF X AND Y THEN Z approach to solving a control problem rather than attempting to model a system mathematically. The Fuzzy Logic model is empirically-based, relying on an operator's experience rather than their technical understanding of the system. Generally, Fuzzy Logic is so forgiving that the system will probably work the first time without any tweaking.

The remaining paper organized as follows, Section 2 discusses the overall system architecture and the overall functioning of the system. Section 3 describes the weight optimization technique in detail, explaining all the involved components. Section 4 provides the simulation results and their explanations. This paper is organized as follows. Section 2 provides an overall system architecture, Section 3 describes the working of the algorithm in detail and section 4 describes the results and the accuracy obtained from our proposed system using the ROC plots and provides the conclusion for our study.

2. SYSTEM ARCHITECTURE

The entire system can be divided into two broad phases. The initial phase deals with evaluating the available attributes and processing them to obtain results in a Neural Networks. The second phase deals with analyzing the Neural Networks to build the fuzzy rules for providing user understandable results.

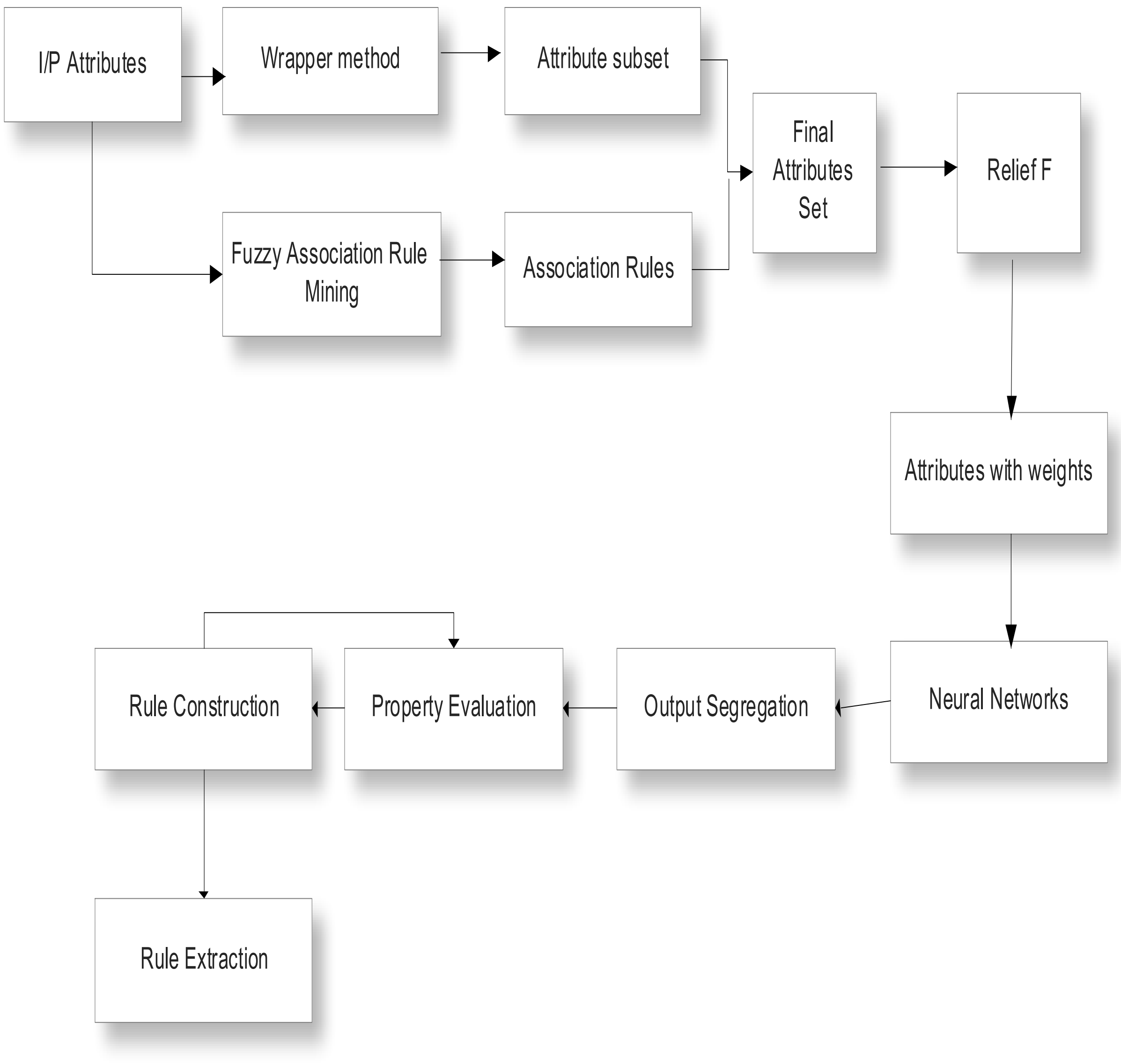

Fig 2: System Architecture

Figure above represents the system architecture that deals with enhancing the interpretability of a Neural Networks system. Initially, all the attributes that make a part of a soils data are identified. All these attributes are then filtered using the wrapper method as discussed in [12]. This results in all the important attributes being shortlisted. Relationship between attributes also play an important role, hence the attribute list is again passed to fuzzy association rule mining algorithm[9] for finding the relationship between the attributes. After this process, all the associated attributes are added to the final attribute list. Hence we obtain a list of all the important properties. These attributes are then passed to the Relief-F[3] algorithm for incorporating the weights. After this process, the attributes along with the weights are passed to the Neural Networks and the Neural Networks is trained. In the second phase, the output values of the Neural Networks are clustered. Every cluster is examined for the properties that influence the output to the maximum extent. These properties are combined to form a single fuzzy rule. This process is repeated for all clusters for obtaining the association rules. All these rules are combined to form the final fuzzy rule set [14]&[15].

3. ENHANCING THE INTERPRETABILITY

Enhancing the interpretability of a Neural Network system can be performed by developing rules based on fuzzy logic. A completely trained Neural Networks is mandatory for developing the fuzzy rules. Hence our process can be performed in two broad phases. The initial phase deals with training the Neural Networks to provide the best output[12] and the second phase deals with categorizing the outputs to provide fuzzy rules.

3.1. PHASE I

All the available attributes that are required for the soil classification are collected initially. Collection of all the attributes are mandatory, because, missing out certain attributes will result in the absence of some attributes in the fuzzy rules, which might result in partially biased output. All these attributes are passed to the wrapper method for feature selection [8]&[9]. Results obtained from this process are added to the final attribute list. The following rules are considered for finding if an attribute is relevant.

Almuallim and Dietterich [4] define relevance under the assumptions that all features and the label are Boolean and that there is no noise.

A feature Xi is said to be *relevant* to a concept C if Xi appears in every Boolean formula that represents C and *irrelevant* otherwise.

Gennari et al. [5] allow noise and multi-valued features and define relevant features as those whose "values vary systematically with category membership".

We formalize this definition as follows.

X_i is *relevant* iff there exists some x_i and y for which p(X_i = x_i) > 0 such that

$$p(Y = y \mid X_i = x_i) \neq p(Y = y).$$

Under this definition, Xi is relevant if knowing its value can change the estimates for the class label Y, or in other words, if Y is conditionally dependent on X;. Note that this definition fails to capture the relevance of features in the parity concept where all unlabeled instances are equiprobable, and it may therefore be changed as follows.

Let $S_i = \{ X_1,\ldots,X_{i-1}, X_{i+1},\ldots,X_m \}$, the set of all features except X_i. Denote by s_i a value-assignment to all features in S_i.

Xi is *relevant* iff there exists some Xi, y, and si for which p(Xi = xi) > 0 such that

$$p(Y = y,\ S_i = s_i \mid X_i = x_i) \neq\ p(Y = y,\ S_i = s_i\ ).$$

Under the following definition, Xi is relevant if the probability of the label (given all features) can change when we eliminate knowledge about the value of Xi.

Xi is *relevant* iff there exists some xi, y, and si for which p (Xi = xi, Si =si) > 0 such that

$$p(Y = y, \mid X_i = x_i\ , S_i = s_i\ ) \neq p(Y = y,\ S_i = s_i\ ).$$

After this process, attribute relationships are investigated. Properties of certain attributes are in such a manner that they do not provide any significance when taken alone, but when used with another attribute, it provides a significant change in the output. These can be found using the association rule mining technique[9]. Initially, the support values are found for considering the attribute for the next level. Calculate *support* for every (candidate) q-itemset using the following equations:

$$Support(I^q)^k = \sum_{T \in M} v_{ij}^q\ .\mu_{I'}^k$$

M is the set of qualified transactions; it can be proved that the previous equation satisfied the following property:

$$\sum_{i \in D} Support\ (i) = \mid M \mid$$

For q =1, I^q can be considered as a single item. if q>1 then generate candidate set C_2^k. All items above the provided threshold support value are taken to the next level. After this phase the confidence value is calculated for finding the relationship between the items. Confidence is calculated using the formula

$$conf\,(A \Rightarrow B) = P(B \mid A) = \frac{Support(A \cup B)}{Support(A)}$$

[13] describes in detail about the calculation of the support and the confidence values using fuzzy association rule mining.

All the attributes present in the final attribute list are checked for their related attributes and these attributes are added to the list to obtain the usable list, which is then passed to the Relief-F method for providing the weights. The algorithm below describes the working of the Relief-F method.

Algorithm:

1. Initialize weights for all attributes *W*[*A*] := 0.0;
2. for i from 1 to *m*
2.1 randomly select an instance *Ri*;
2.2 find *k* nearest hits *Hj*;
2.3 for each class *C* <> *class*(*Ri*) do
2.3.1 from class *C* find *k* nearest misses *Mj*(*C*);
2.4 for *A* from 1 to a
2.4.1 $W[A] := W[A] - (x+a)^n = \sum_{j=1}^{nk} \frac{\mathrm{diff}(A,R_i,H_j)}{(m.k)} + \sum_{C \neq \mathrm{class}(R_i)} [\frac{P(C)}{1-P(\mathrm{class}(R_i))} \sum_{j=1}^{k} diff(A,R_i,M_j(C))]/(m.k)$
3. end;

Function diff($A; I_1; I_2$) calculates the difference between the values of the attribute A for two instances $I1$ and $I2$. For nominal attributes it was originally defined as:

$$diff(A, I_1, I_2) = \begin{cases} 0, & value(A, I_1) = value(A, I_2) \\ x, & otherwise \end{cases}$$

and for numerical attributes as:

$$diff(A, I_1, I_2) = |value(A, I_1) - value(A, I_2)| / (\max(A) - \min(A))$$

The function diff is used also for calculating the distance between instances to find the nearest neighbors. The total distance is simply the sum of distances over all attributes (Manhattan distance).

The returned results can be directly used by the Neural Networks for further processing of the data.

3.2. PHASE II

The second phase deals with generating the fuzzy rules for soil classification. Let us consider an ANN with input, hidden, and output layers. Let us suppose that the Neural Network has n input neurons $(x_1,\dots,x_n)$, h hidden neurons $(z_1,\dots,z_h)$ and m output neurons $(y_1,\dots,y_m)$. Let τ_j the bias for neuron and for neuron Z_j and φ_k. Let w_{ij} be the weight of the connection from neuron x_i to neuron z_j and β_{jk} the weight of the connection from neuron z_j to neuron y_k. The function the net calculates is

$$F : \Re^n \to \Re^m ; F\ (x_1,\dots,x_n) = (y_1,\dots,y_m)$$

$$y_k = g_A\left(\sum_{j=1}^{h}(z_j \beta_{jk}) + \varphi_k\right) \text{with } z_j = f_A\left(\sum_{i=1}^{n}(x_i w_{ij}) + \tau_j\right)$$

where g_A and f_A are activation functions.

The output generated from the Neural Networks are organized and classified [11] according to their final interpretations. This data serves as the basis for generating Fuzzy rules [14]&[15] using Fuzzy Logic [10].

Each cluster is initially examined and all the input neurons x_i that contributes to the final output are identified. The weights of these neurons are not considered during the selection process.

For x hidden layer neurons

$$E_{io} = \sum w_{ix} . w_{xo}$$

Where I represents an input neuron and o an output neuron

After the processing of all the clusters, the obtained rules are combined and the final rule set is extracted. After this phase, processing can be directly performed using the extracted rules.

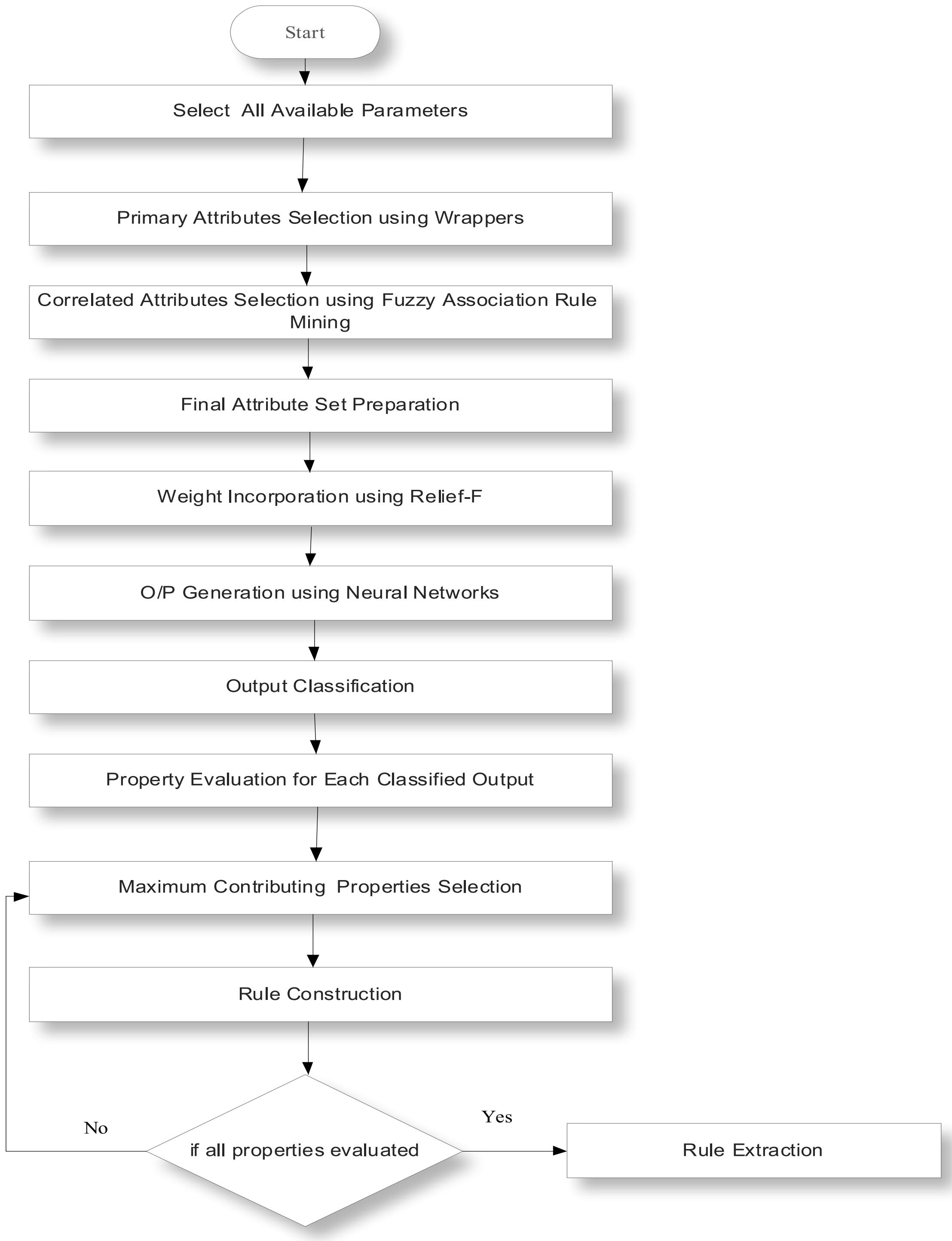

Fig 3: Interpretability Enhancement

4. DISCUSSIONS AND CONCLUSION

Analysis of the current system with the soil data shows higher accuracy and better understandable results when compared to [12]. Neural networks have the property of providing numerical results. Due to this property, the results produced by the Neural Networks cannot be directly used by the end user, instead, these results have to be interpreted to user understandable form. We eliminate this downside by the usage of fuzzy rules. Numerical inputs from the Neural Networks are passed on to the fuzzy interpreter, which forms fuzzy rules that directly points towards easily comprehensible results. Further advantage of this system is that once this entire process is performed and fuzzy rules are generated, the fuzzy rules can be directly used for obtaining the results.

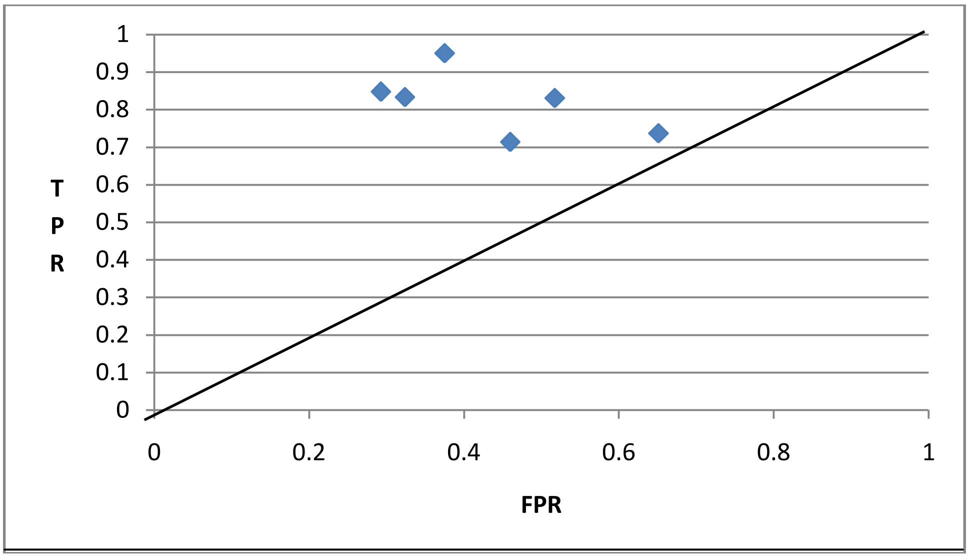

Fig 4: ROC using fuzzy rules

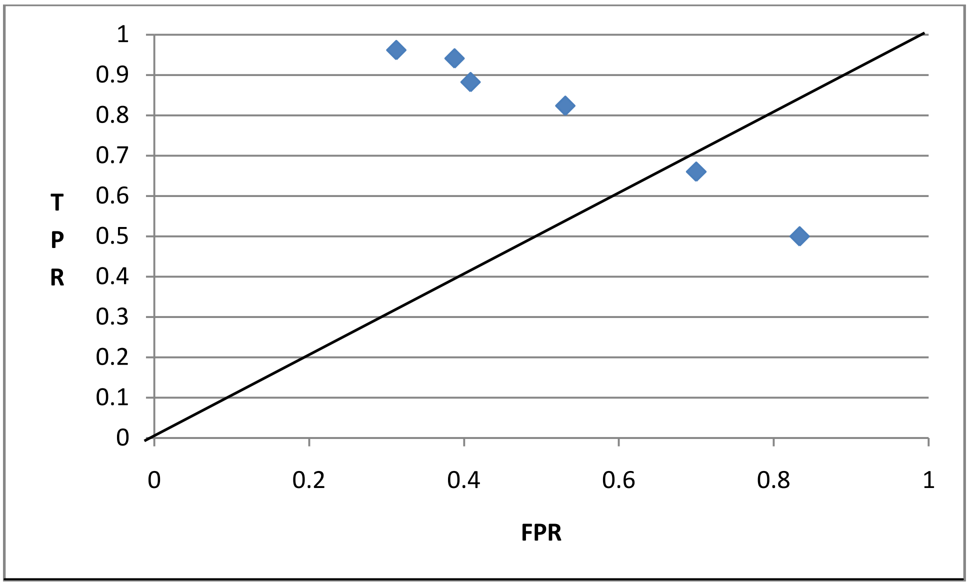

Fig 5: ROC without fuzzy rules

The ROC curves represent the accuracy levels of the current system and [12]. It shows that the fuzzy rules provide better accuracy than performing the process using a Neural Networks alone. All the points can be seen above the diagonal, which proves that the system moves towards near accuracy, i.e. (1,1). Soil Classification data contains numerous attributes. Not all attributes will be present at an instant. The current processing is performed using all the available attributes. The developed fuzzy rules have the probability of containing missing attributes. The fuzzy rules can be further optimized by using the association rule mining technique for developing multiple set of fuzzy rules, which performs the processing with the available set of attributes.

REFERENCES

[1] Ms.Smita. Nirkhi, “Potential Use of Artificial Neural Network in Data Mining”, 978-1-4244-5586-7/10, 2010 IEEE.

[2] Zhijun Chen, “Research of Data Mining Based On Neural Network”, 978-1-4244-7161-4/10, 2010 IEEE.

[3] M. Robnik-Sikonja and I. Kononenko, “Theoretical and Empirical Analysis of ReliefF and RReliefF”, Mach. Learn., 53(1- 2):23.69, 2003.

[4] H. Almuallim and T.G. Dietterich, Learning with many irrelevant features, in: Proceedings AAAI-91, Anaheim, CA (MIT Press, Cambridge, MA, 1991) 547-552, 1991.

[5] S.J. Russell and P Norvig, “Artificial Intelligence: A Modern Approach”, (Prentice Hall, Englewood Cliffs, NJ), 1995.

[6] Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade, Vahida Attar, “Soil Moisture Estimation Using Classification And Regression Trees And Neural Networks”.

[7] Ramesh Vamanan, K.Rama, Ramesh Vamanan et al, "Classification of Agricultural Land Soils: A Datamining Approach", International Journal on Computer Science and Engineering (IJCSE).

[8] Cheng-Lung Huang, Chieh-Jen Wang, "A GA-Based Feature Selection and Parameters Optimization for Support Vector Machines", Expert Systems with Applications 31, 231–240, 2006.

[9] Ashish Mangalampalli, Vikram Pudi, "Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IIIT/TR/2009/173, 2009, IEEE.

[10] P. Ponce-Cruz, F.D. Ramirez-Figueroa, "Fuzzy Logic", Intelligent Control Systems with LabVIEW™, Springer 2010.

[11] P.Bhargavi1, Dr.S.Jyothi, "Fuzzy C-Means Classifier for Soil Data", International Journal of Computer Applications (0975 – 8887), Volume 6– No.4, September 2010.

[12] Improved Techniques For Enhancing The Accuracy Level in a Feed-Forward Neural Networks.

[13] Quick Convergence Of Feed-Forward Neural Networks By Weight Optimization Technique.

[14] Chris Matthews, Ilona Jagielska , "Fuzzy Rule Extraction from a Trained Multilayered Neural Network".

[15] Juan L. Castro, Carlos J. Mantas, and José M. Benítez, Interpretation of Artificial Neural Networks by Means of Fuzzy Rules, IEEE Transactions on Neural Networks, Vol. 13, No. 1, January 2002, IEEE.

