Journal of Computer Science 4 (6): 494-502, 2008
ISSN 1549-3636

Self-Organizing Map and Multi-Layer Perceptron Neural Network Based Data Mining To Envisage Agriculture Cultivation

[1]E.T. Venkatesh and [2]Dr. P. Thangaraj
[1]Senior Lecturer, Kongu Engineering College, Perundurai, Erode, Tamilnadu, India
[2]Professor, Kongu Engineering College, Perundurai, Erode, Tamilnadu, India

Abstract: Study on characteristics of soil, to determine the types of crops suitable for cultivation in a particular region can increase the yield to greater extent, which minimizes the expenditures involved in irrigation and application of fertilizers. With the tested techniques available for calibrating the quality of soil and the crops suitable for cultivation in it, it is possible to determine the exact crop, irrigation patterns and even the cycle and quantity of fertilizer application. This paper dealt with the application of SOM based clustering and Artificial Intelligence techniques, to analyze the patterns of soils distributed across huge geographical area and identify the suitable types of crops for the particular soil. Estimation of exact crop(s) suitable for a particular region can help stave off redundant maintenance and the inherent expenditures that would occur due to over irrigation and over usage of fertilizers, to fulfill the natural deficiencies. Our Focus is to improve the optimal utilization of innate characteristics in a soil through cultivation of appropriate crops, which will increase the volume and quality of yield, in particular for a developing country like India, where the huge majority of the population depends primarily on agriculture for livelihood.

Key words: Data mining, agriculture, soil characteristics, clustering, unsupervised learning, self-organizing map (SOM), multi-layer perceptron neural networks (MLPNN)

INTRODUCTION

Information packaged in sizeable databases about all kinds of characteristics and facets are accessible these days. Extracting new and interesting knowledge through interpreting the collected data is cumbersome since more data potentially contains more information. To analyze these databases, immense efforts are being deployed. Understanding the hidden are implicit relationship between attributes in these Knowledge Discovery Databases (KDD) requires data mining techniques, as it has been proven as an effective tool. Data mining in nutshell is, pattern finding. Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets [1]. The modal of the semantic structure of the dataset is obtained and represented from the 'mined' information; tasks such as prediction or classification utilize this model. Finding patterns and elucidating those patterns clearly are the two primary activities of data mining. Facilitating to make profitable predictions through offering insights into the data and providing proper explanations about the data is the mark of efficient data mining tool [2].

Staying on top of the information contained in rapidly growing databases is a crucial problem and data mining techniques offers solutions for it. The expanding surges of data from multiple sources often overwhelm the organizations. The magnitude and intricacy of this data are beyond the limits that it cannot be withstood by conventional protocols, leading more organizations to adapt data mining solutions. Converting data into valuable and profit-making information inevitably requires application of data mining solutions. Data is filtered, selected and interpreted through a sequence of methods provided by data mining tools. Essentially the organizations that properly exploit these skills will surpass and take over their market [2]. Comprehending the furtive constitutions of any kind data and their clandestine correlations thorough explorations is feasible with the available data mining techniques. Conversely, these greater degrees of abstraction pose inherent limitations as the evolved information need not be meaningful or applicable. Evaluations and interpretations of the out come of data mining solutions in accordance to the affirmed objectives or purposes of data analysis become indispensable due to the stated drawbacks [5].

Corresponding Author: E.T. Venkatesh, Kongu Engineering College, Perundurai, Erode, Tamilnadu, India.

Production of goods through domesticating plants, animals and other life forms is termed as agriculture and the study of agriculture is christened as agricultural science. Agriculture has made a paradigm shift in the human civilization and revolutionized the way humans live. This first agricultural revolution called as Neolithic Revolution drastically changed the hunter-gatherer lifestyle of humans and made them settled through domestication of plants and animals [3]. In India, three-fourths of the population that live in rural areas depends primarily on agriculture for the major source of income. Agriculture serves as the predominant source of occupation for two-third of the working population for their livelihood. Apart from providing food for the very sustenance, agriculture also supplies raw-materials for manufacturing industries like textiles, sugar, vegetable oils, jute and tobacco. Soil, capable of sustaining life, is naturally occurring, unconsolidated or loose covering of broken rock particles and decomposing organic matter (humus) on the surface of earth [4]. Food and fiber production crucially depends on soil resources and it's critical to the environment too as it provides mineral and water to the plants. Scientists analyze soil characteristics to describe, classify and interpret soil for various uses. With Global Positioning System scientists or agronomists calculate the regions and the pertaining soil characteristics with a high level of accuracy. Plant growth and yield measurements primarily depend on soil characteristics, quantifying the soil characteristics with their location of presence and other vital data pertaining to that can extract valuable information about the complex process.

With more than two-third of the population living in rural areas, where agriculture is the only source of livelihood of the people, India is predominantly agriculture-based country [6]. Sophisticating Indian agriculture with multifaceted efforts ranging from advancements in cultivation practices, utilization of modern gadgets and farm equipments, potent fertilizers and powerful pesticides, refined seeds of improved varieties has helped raise India's food production to considerable levels during the last three decades [7]. Having 5.5 million hectares of net cultivated area with 54 percent of irrigated land, Tamilnadu, the southern most state of India boasts of its agricultural importance in the collective agro-output of the nation. Red laterites, black and alluvial soil constitutes the major portion of agricultural region. The crops grown here range from sorghum, finger millet, maize, pulses, oilseeds, cotton, sugarcane, with rice being the principal crop. Plantation crops like coconut, rubber, tea, coffee, fruits, vegetables, spices and tubers etc., are also grown. Food grains occupy 56 percent of the gross cropped area. The state has an assortment of cropping systems in practice.

The primary intent of this research is to find a solution that could provide a near precise information on the suitability of crop(s) for a given soil type. The concept nurtured designates the clustering utilities of data mining and techniques of artificial intelligence for prediction respectively. The databases we have used are collected from local experts pertaining to agricultural research. The databases are preprocessed and transformed to a format suitable for clustering. Once after the complete datasets are clustered, Artificial Intelligence techniques are applied to find the best match of crop(s) for the given type of soil characteristics. This artificial intelligence technique is trained to find the best match of soil and its suitable crop(s) through thorough analysis of the characteristics of the soil and the nutritional requirements of the crop(s) planned for cultivation. The clustering of dataset is performed by exploiting Self Organizing Map (SOM) and the prediction is done by deploying Multilayered Perceptron Neural Networks (MLPNN). Neural networks, despite the presence of considerable noise in the training set, have the ability to trace the hidden and strongly non-linear dependencies and learn from examples once the training is completed, is an appropriate tool when it comes to prediction.

Related work: Piatetsky-Shapiro *et al.*[1] presents an overview of the state of the art in research on knowledge discovery in databases and analyze Knowledge Discovery and define it as the nontrivial extraction of implicit, previously unknown and potentially useful information from data and also they discuss application issues, including the variety of existing applications and propriety of discovery in social databases.

Gerhard Munz *et al.*[5] gives an introduction to Network Data Mining, i.e. the application of data mining methods to packet and flow data captured in a network, including a comparative overview of existing approaches and they present a novel flow-based anomaly detection scheme based on the K-mean clustering algorithm.

Mai *et al.*[6] have demonstrated for the scope for application of spatial mining tools for a utility study and analysis and that specific application of Polyanalyst gave a clear scope for evaluation and comparison of predicted and real values and also their project helps the Government to draw attention to the agriculture sectors.

Reddy and Ankaiah [7] proposes the framework for a cost-effective agricultural information dissemination system (AgrIDS), to disseminate expert agricultural knowledge to the farming community in order to improve crop productivity and they have made an effort to present a solution to bridge the information gap by exploiting advances in information technology.

Seifert [8] discussed the overview of Data Mining and limitations of data mining. And also he discusses the uses of data mining and provides Terrorism Information Awareness (TIA) Program, Computer-Assisted Passenger Prescreening System.

Yu Guan *et al.*[9] provides a clustering heuristic for intrusion detection, called Y-means and their proposed heuristic is based on the K-means algorithm and other related clustering algorithms and also their experimental results show that Y-means is a promising clustering method for intrusion detection without supervision.

Yeung and Ruzzo[10] study the effectiveness of principal components (PC's) in capturing cluster structure and they compared the quality of clusters obtained from the original data to the quality of clusters obtained after projecting onto subsets of the principal component axes using both real and synthetic gene expression data sets.

Hansen and Mladenovic[11] have proposed a new local search heuristic, called J-MEANS for solving the minimum-sum-of-squares clustering algorithm and they compared the new heuristic with two other well known local search heuristics, K-means and H-means as well as with H-means, an improved version of the later in which degeneracy is removed.

Pornphan and Rangsanseri[12] gives a priori spatial information with the FCM clustering for improving the segmentation result and their segmented images show more homogeneous regions when they compare with the standard FCM, which do not use the spatial information.

Wen *et al.*[13] provides a high-resolution temporal map of fluctuations in mRNA expression of 112 genes during rat central nervous system development, focusing on the cervical spinal cord using the reverse transcription-coupled PCR and that data provide a temporal gene expression ''fingerprint'' of spinal cord development based on major families of inter- and intracellular signaling genes ans also they found that genes belonging to distinct functional classes and gene families clearly map to particular expression profiles.

Demiriz *et al.*[14] introduced a novel method for semi-supervised learning that combines supervised and unsupervised learning techniques and their experimental results show that using class information improves the generalization ability compared to unsupervised methods based only on the input attributes.

Saracoglu [17] describes Artificial Neural Network (ANN) based prediction of the response of a fiber optic sensor using Evanescent Field Absorption (EFA) and their performance comparisons show that all of the neural models used in this work can predict the sensor responses with considerable errors and also their artificial neural network approaches can play an important role in the design and development of intelligent sensors.

Pramanik [18] investigate the use of neural network techniques for the prediction of cell mass and ethanol concentration under varying fermentation conditions and to compare the experimental results with those obtained by Neural Network (NN) simulation and they found a simple propagation network using the Levenberg-Marquardt for training the network to be very effective to generalize and predict the cell mass and ethanol concentration during batch fermentation.

MATERIALS AND METHODS

This research delineates the general idea of the proposed methodology for prediction of appropriate crops to soil types. The dataset analyzed is eclectically collected from agricultural experts pertaining to the area under research. With the object of increasing the accuracy of mining process sufficient preprocessing is performed over the collected dataset which transforms the raw data, attuned for clustering process. The concept proposed categorizes the soil primarily by their geology (Clay, Granite, Soil and etc.,) using SOM based clustering. This process of stratification goes up till a level enough to predict near accurate crop(s) suitable for that particular kind of soil. The attributes analyzed include the Elevation, Slope, Erosion, Drainage, Permeability, Geology, Land use, Vegetation and Taxonomic Classification. Where, the parameter 'Land use' exhibits the crops tested for that particular variety of soil and the parameter Vegetation signifies the natural growth of crops in that region. The parameter Horizon too is taken in to consideration and the attributes like color, texture, stickiness, type, amount of pores and types of blocks present and etc., are measured while stratification. Each of these strata falls in to dataset based on its characteristics.

The primary intent of this approach is to find a solution that could provide a near precise information on the suitability of crop(s) for a given soil type. Once after the complete datasets are clustered, Artificial Intelligence techniques are applied to find the best match of crop(s) for the given type of soil characteristics. This artificial intelligence technique is trained to find the best match of soil and its suitable crop(s) through thorough analysis of the characteristics of the soil and the nutritional requirements of the crop(s) planned for cultivation. This can be done through a database built through observation of crop(s) cultivated previously in the soil types and its growth patterns, or through laboratory experiments. The output of this process will be the list of cultivable crops and even list of crops that could grow naturally. Such intelligence system can lead to minimal application of fertilizers and appropriate irrigation patterns. This practice of exploiting the natural abundance of the soil will not only result in better quality but will also lead to increased production in a very cost effective and sustainable method. Out of the several techniques of artificial intelligence like Expert Systems, Fuzzy Logic, Genetic Algorithm, Intelligent Agent, Artificial Neural Networks, explored we narrowed down on Artificial Neural Networks (ANN). This technique in particular exhibits tolerance to a substantial degree over noises if present in the training set. Furthermore, its astonishing capability to map out the veiled and tenaciously non-linear dependencies and learn from examples once the training is completed makes it an apt methodology for prediction. Back-propagation Neural Network, Feed-Forward Neural Network, Radial Basis Function Neural Network and etc. are some of the normal techniques used in ANN. Feed Forward Neural Networks (FFNNs), currently recognized as state-of-the-art approach for the prediction Process. For this reason we have chosen Multi-Layer Perceptron Neural Network also known as Multilayer feed-forward neural network among the above discussed ANN techniques. Fig. 1 shows the overall architecture of the proposed system.

Clustering based on Self-Organizing Map (SOM): Preprocessing and Transformation: One of the biggest challenges of data mining is the comprehensive nature of the quality of data. Accuracy and completeness are the principle aspects of data quality. The structure and consistency facets of the data being scrutinized can impinge on the quality of data. All these characteristics induce to perform a preprocessing to enhance the exactitude of mining process, as 80% of mining efforts are dissipated toward grooming data quality. The efficacy of intricate data mining techniques is susceptible even to the delicate differences that may exist in the data. The inconsistency of data can be due reasons that include presence of duplicate records, deficient data standards, data values gone astray and human errors. Sprucing up the data for mining comprise several activities that include scrapping duplicate records, regularizing values used characterize information in the database, bookkeeping of data points gone astray and eliminating redundant data fields. To prepare the data suitable for mining process data needs to be transformed accordingly. This preprocessed data is refined further to convert it to a format compatible for clustering.

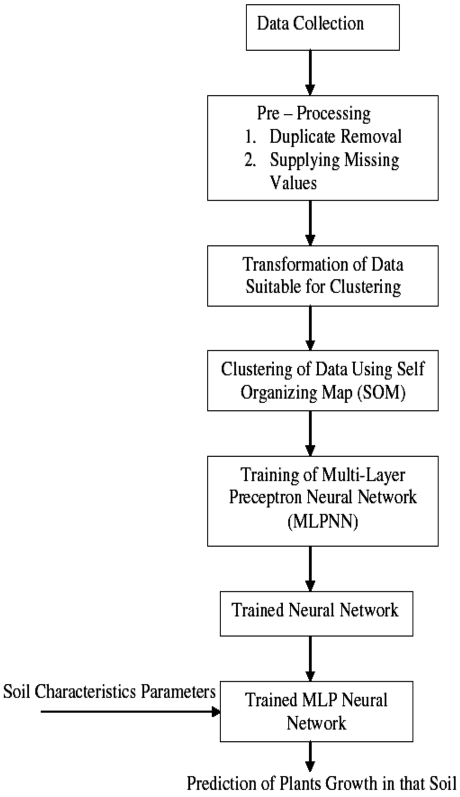
Data Collection
Pre – Processing
1. Duplicate Removal
2. Supplying Missing Values
Transformation of Data Suitable for Clustering
Clustering of Data Using Self Organizing Map (SOM)
Training of Multi-Layer Preceptron Neural Network (MLPNN)
Trained Neural Network
Soil Characteristics Parameters
Trained MLP Neural Network
Prediction of Plants Growth in that Soil

Fig. 1: Training system architecture

Clustering: Data elements with high similitude are bunched up within same cluster while data elements with less similarity are crowded up in different clusters. Clustering is the unsupervised classification of patterns into groups. Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar and the members from different clusters are quite different from each other [9]. Clustering gradually discovers the data and institutes maximum possible number of groups, albeit the process commence with spur-of-the-moment regarding the number of groups present. Clustering has been performed with several algorithms like K-Means[11], Y-Means[9], H-Means+[11], J-Means[11], Fuzzy C-Means[12], Hierarchical clustering[13], Principal Component Analysis[10] (PCA) based and genetic algorithms[14]. The algorithm fostered by us in this study exploits Self Organizing Map (SOM) for clustering.

Self-Organizing Maps (SOM) mimics the structure of systematic organization of information in the cerebral cortex [14]. Organizing archetypes of data to an n-dimensional grid of neurons or units is the underlying principle of SOM. As opposed to the familiar connotations of input space, the grid materializes it to an output space, where it stores the data patterns. This configuration tries to systematize and retain the proximity of mapped units in input space to the corresponding output-space and vice-versa. SOM adapts itself through a self paced learning.

With SOM as the basis we have proposed an approach for clustering data items. The methodology projected takes advantage of k-means measurement of identicalness (distance) amongst the sampled data, as that's the core constituent of cluster analysis. Void of any need for instructional data i.e. a trainer signal that provides the required transfer knowledge regarding correct answer, the methodology performs clustering and capable of self-governance. This line of attack seeks out to construe the exact number of clusters, devoid of any information about it. The steps in the proposed approach are given as follows.

Assumptions:

N = Total number of items in the shown dataset
NOC = Number of clusters
J = Number of clusters with zero standard deviation
C_a = Actual clusters
C_f = Intermediately Formed Clusters
C_r = Cluster to be removed

The dataset is shown by $S = \{S_1, S_2, S_3,\ldots,S_N\}$. The initial value of NOC = 1 and J = 0.

Steps:

- Calculate NOC = NOC–J +1
- Form NOC Clusters and calculate the centroid of the clusters. The centroid is calculated using the following formula:

  $$COC_j = \frac{1}{|C_j|}\sum_{x \in C_j} x \text{ where, } 0 < j \le k$$

  where, COC_j is the calculated centroid of cluster j. K is the number of items in the cluster j
- Calculate the Euclidean distance of each data item with the centroids of the available clusters. The Euclidean distance is calculated using the following formula:

  $$d(S_i,S_j) = \left[\sum_{k=1}^{d} |S_{i,k} - S_{j,k}|^2\right]^{\frac{1}{2}}$$

  where, d is the dimensionality of the data
- Assign the data item to the cluster with the minimum distance
- Repeat steps 3 and 4 until there is no change in the clusters.
- Calculate the standard deviation of all the clusters formed. Neglect j clusters with standard deviation less thanϕ. The standard deviation is calculated as follows:

  $$\bar{S} = \frac{1}{|C_j|}\sum_{x \in C_j} x SD = \sqrt{\frac{1}{|C_j|}\sum_{x \in C_j}(S_i - \bar{S})^2}$$

- Repeat steps 1 to 6 until the standard deviation of all clusters reaches a value less than ϕ

The pseudo code of the proposed clustering algorithm is as follows:

```
Initialize NOC = 1 and J = 0
Repeat
    Calculate NOC = NOC-J + 1
    Form C_f clusters of size NOC
    Calculate centroid $COC_j = \frac{1}{|C_j|}\sum_{x \in C_j} x$
    where $0 < j \le k$
    Repeat
        for k = 1 to N
            for c = 1 to NOC
                $dis|C| = \left[\sum_{k=1}^{d} |S_a - S_b|^2\right]^{\frac{1}{2}}$
            end
            $dis_{min} = min(dis)$
            Assign data item to cluster with $dis_{min}$
        end
    Until there is no change in formed clusters
    for C =1 to NOC
        $\bar{S} = \frac{1}{|C_j|}\sum_{x \in C_j} x$
        $SD = \sqrt{\frac{1}{|C_j|}\sum_{x \in C_j}(S_i - \bar{S})^2}$
        if (SD < $\phi$)
            Remove $C_r$ from $C_f$ where $C_r \in C_f$ and add to $C_a$
            J + +,
        end
    end
    Compute $R_c$ remaining clusters after neglection
Until $R_c$ reaches zero
```

The flow diagram of the proposed clustering process is shown in Fig 2.

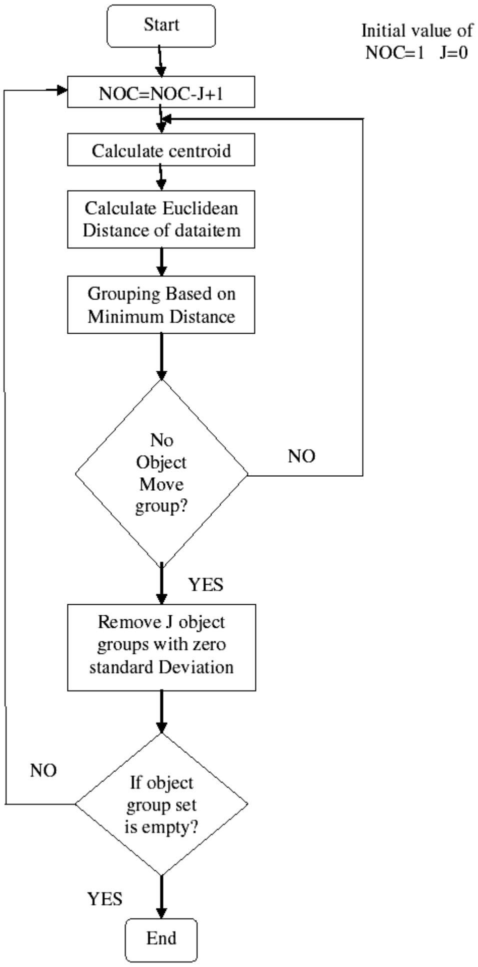
Start
Initial value of
NOC=1 J=0
NOC=NOC-J+1
Calculate centroid
Calculate Euclidean Distance of dataitem
Grouping Based on Minimum Distance
No Object Move group?
NO
YES
Remove J object groups with zero standard Deviation
NO
If object group set is empty?
YES
End

Fig. 2: Flow diagram of the clustering process

Prediction of crops growth using multi-layer perceptron neural networks (MLPNN): Predicting is a process of conceiving something as it might happen in future, based fundamentally, on knowledge gathered from past experiences and from present scenario. The degree of success varies day to day, in problem solving based on prediction[15]. The debility of the dependency of the predicted variable when extrapolated to the other events of the future can be calibrated with the use of various approximations. For prediction and simulation in engineering applications, tools of artificial intelligence have been found to be strong. Void of need to add additional information (such as type of dependency like with the regression) and to learn the dependencies automatically from measured data is an advantage. Trained using historical data, neural networks exhibits ability to discern clandestine dependencies for future use. Not in a state of being confined to an explicit model, neural network is a kind of black box with an ability to learn autonomously.

Based on neural constitution of the brain, artificial neural networks are rudimentary electronic networks, gaining attractiveness in the field of artificial intelligence [16]. Neural networks assimilate knowledge by collating the presumptions of records one by one against the actual known records (which, at the inception is largely subjective). The network reorganizes itself based on corrigendum it receives from the preliminary predictions and the process is iterated for optimum refinements

Multi layered perceptron neural network: Literature analysis reveals a pervasive application of feed forward neural networks, from among the diverse categories of connections for artificial neurons [17]. In feed-forward neural networks the neurons of the first layer drive their output to the neurons of the second layer, in a unidirectional manner, means the neurons are not received from the reverse direction. Multilayer Perceptron Neural Networks (MLPNN) or Multilayer Feed-forward Neural Network (MFNN) is one such feed-forward neural network mechanism.

Incorporating three layers, input, output and intermediate, the MLPNN designates distinct roles for each. Input layer maintains equal number of neurons corresponding to that of the variables in the problem. The output layer comprises a number of neurons equal to the preferred number of quantities, computed from the input and makes accessible the Perceptron responses. The intermediate/hidden layer takes care of approximating non-linear problems. Processing linear problems necessitates the presence of only the input and output layer of the MLPNN. Data having discontinuities like saw tooth wave pattern necessitate the presence of two hidden layers for prototyping. The risk of congregating to local minima is greater while using two hidden layers and it seldom refines the model. Hypothetical rationale behind implementing more than two hidden layers is also void [20]. Separate weights are applied to the sums forwarded to each layer while the output from the first hidden layer is fed to the input of the next hidden layer, in scenarios where more than one hidden layers are deployed. The approach fostered by us employs a network with two hidden layers. A general structure of MLPNN comprising three layers is shown in Fig. 3.

The only task of the neurons in the input layer is to distribute the input signal x_i to neurons in the hidden layer. Each neuron j in the hidden layer sums up its input signals x_i after weighting them with the strengths of the respective connections w_{ji} from the input layer and computes its output y_j as a function f of the sum, given by:

$$y_j = f\left(\sum W_{ji} X_i\right)$$

where, f can be a simple threshold function such as a sigmoid, or a hyperbolic tangent function. The output of neurons in the output layer is computed in the same manner. Following this calculation, a learning algorithm is used to adjust the strengths of the connections in order to allow a network to achieve a desired overall behavior. There are many types of learning algorithms in the literature.

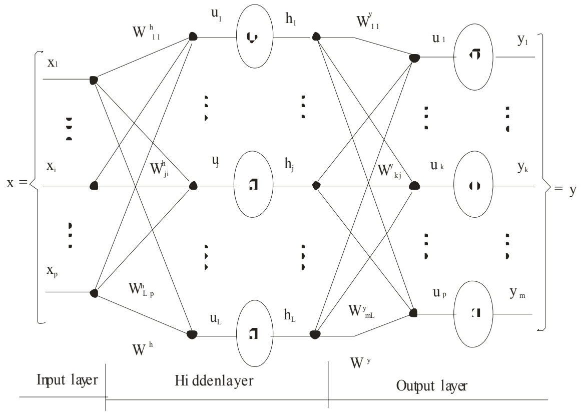

Fig. 3: Structure of MLPNN

Back propagation training: We have used Back Propagation Training algorithm [18]. The back propagation of error algorithm, based on multi-layered feed-forward net and considered to be the most versatile algorithm[19], was used to train the network for predicting correct outputs those obtained from experiments and generated one. The BP algorithm adjusts the network weights and bias values to minimize the square sum of the difference between the given output (X) and output values calculated by the net (X') using gradient decent method as follow

$$SSE = 1/2\ N\sum\left(X\text{-}X'\right)^2$$

where, N is the number of experimental data points used for the training.

RESULTS AND DISCUSSION

An MLPNN consist of one input layer, two hidden layers and one output layer. The output layer predicts the crops that growth in a soil given by its characteristics. Table: 1 shows some of the soil characteristics and weights assigned to them for MLPNN training and Table: 2 shows the output weights of the MLPNN training. Table: 3 shows the sample input soil characteristics for MLPNN prediction and Table: 4 shows the predicted land use and crops for the given soil characteristics.

Table: 1 Weights of the parameters in the MLPNN training

Elevation	Training weight	Slope	Training weight	Erosion	Training weight	Drainage	Training weight	Permeability	Training weight	Geology	Training	Tax anomic classification	Training weight
1 m above MSL	0.1	Flat lands	0.1	Moderate erosion (e2)	0.1	Excessively drained	0.1	Medium	0.1	Clay	0.1	Clayey, mixed, isohyperthermic, Calcareous, lithic	0.1
100 m above MSL	0.100	Foot and side slopes (5-15%)	0.2	Moderate to severe erosion	0.2	Imperfectly drained	0.2	Moderate permeability	0.2	Eastern Ghats	0.2	Clayey, mixed, isohyperthermic, calcareous, Vertic	0.2
1004 m above MSL	0.1004	Foot slopes (3-8%)	0.3	None to slight erosion (e1)	0.3 drained	Moderately well	0.3	Moderate to slow permeability	0.3	Granite	0.3	Clayey, mixed, isohyperthermic, noncalcareous, Lit	0.3
101 m above MSL	0.101	Gentle slope with depression	0.4	Severe erosion (e3)	0.4	Poor	0.4	Moderately rapid permeability	0.4	Laterite	0.4	Clayey-over-sandy, mixed, isohyperther- mic, noncalc	0.4
102 m above MSL	0.102	Gently sloping	0.5	Sheet and gully erosion	0.5	Poorly darined	0.5	Moderately slow permeability	0.5	Quartzite	0.5	Clayey-skeletal, kaolinitic, isohyperthermic, nonc	0.5
103 m above MSL	0.103	Gently sloping (0-1%)	0.6	Very severe erosion (e4)	0.6	Poorly drained	0.6	Rapid	0.6	Sand	0.6	Clayey-skeletal, mixed, isohyperthermic, calcareou	0.6

Table: 2 Output weights of the MLPNN Training

Land use	Training weight	Natural vegetation	Training weight
Agriculture	1	Acacia arabica, croton sparciflorus, Cyprus rotant	1
Banana, betelvine, paddy	2	Acacia arabica, grasses, wetland weeds	2
Banana, guava, mango	3	Acacia, calotropis	3
Banana, sugarcane, paddy	4	Acacia, cassia auriculata, tamarind	4
Banana	5	Acacia, croton sparciflorus, neem	5
Brinjal, maize	6	Acacia, croton sparciflorus, prosopis juliflora	6

Table: 3 Sample input soil characteristics for MLPNN prediction

Elevation	Slope	Erosion	Drainage	Permeability	Geology	Tax anomic classification
101 m above MSL	Foot and side slopes (5-15%)	Sheet and gully erosion	Imperfectly drained	Moderately rapid permeability	Eastern Ghats	Clayey-over-sandy, mixed, isohyperthermic, noncalc
100 m above MSL	Gently sloping (0-1%)	Very severe erosion (e4)	Poor	Moderate permeability	Quartzite	Clayey-skeletal, kaolinitic, isohyperthermic, nonc
1 m above MSL	Flat lands	Moderate erosion (e2)	Excessively drained	Medium	Clay	Clayey, mixed, isohyperthermic, calcareous, Lithic

Table: 4 Predicted land use and crops for the given soil characteristics

Land Use	Natural vegetation
Agriculture	Acacia arabica, croton sparciflorus, cyprus rotant
Banana	Cyanodon, neem
Groundnut	Neem, prosophis juliflora, accacia

CONCLUSION

The application of Self Organizing Maps for clustering of data and Multi-Layered Perceptron Neural Network or Multilayer Feed-Forward Neural Network has helped achieving near precise estimation of suitability of soil characteristics and the choice of crops planned for cultivation based on its nutrient requirements. Self Organizing Maps for clustering has been chosen for the solution after a pervasive scrutiny of all neural network algorithms. In particular the traits analyzed like the ability to withstand noise levels to certain extents, the capacity to learn in a self-paced manner and the degree to which it can self-govern, made SOM an irresistible choice for the solution. For the purpose of predicting the clustered datasets we incorporated MLPNN, especially for dealing with discontinuities that resembles a saw tooth wave pattern and problem being non-linear, the approximations demanded two intermediate layers apart from the input and output layers. MLPNN, having the adaptability to accommodate multiple intermediate layers became the preferred option of neural network algorithms available. Experimental results show fascinating results on the application of the algorithm and even surpass the anticipated performance.

REFERENCES

1. W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus, "Knowledge Discovery in Databases: An Overview," AI Magazine, Vol. 13, No. 3, 1992, pp. 57-70.
2. "Data Mining: Discovering Opportunities in Your Company Data", White Paper from International Resource Management (IRM), Inc. Centerbrook Ct. – 06409.
3. Bolens, L., 'Agriculture' in Encyclopedia of the history of Science, technology, and Medicine in Non Western Cultures, Editor: Helaine Selin; Kluwer Academic Publishers. Dordrecht/Boston /London, 1997, http://en.wikipedia.org/wiki/Agriculture.
4. Voroney, R.P., “The Soil Habitat in Soil Microbiology, Ecology and Biochemistry”. Eldor A. Paul (Ed.), ISBN: 0125468075, 2006
5. Gerhard Münz, Sa Li and Georg Carle. "Traffic Anomaly Detection Using k-means clustering," In GI/ITG Workshop MMBnet, 2007.
6. Mai, C.K., I.V. Murali Krishna and A.V. Reddy, 2006. Data Mining Of Geospatial Database For Agriculture Related Application. Map India. http://www.gisdevelopment.net/proceedings/mapindia/2006/agriculture/mi06agri_124.htm.
7. Reddy, P.K. and R. Ankaiah, “A framework of information technology-based agriculture information dissemination system to improve crop productivity”. Curr. Sci., 88: 1905-1913, 2005.
8. Jeffrey W. Seifert, "Data Mining: An Overview, " Resources, Science, and Industry Division, Congressional Research Service, Library of Congress, 2004. http://digital.library.unt.edu/govdocs/crs/permalink/meta-crs-6059

9. Guan, Y. Ghorbani, A.A. Belacel, N., "Y-means: a clustering method for intrusion detection," Canadian Conference on Electrical and Computer Engineering, IEEE CCECE 2003. Vol. 2, pp: 1083- 1086, May 2003.
10. Yeung, K.Y. and W.L. Ruzzo, “Principal component analysis for clustering gene expression data. Bioinformatics”, 2000, 17: 763-774.
11. Hansen, P. and N. Mladenovic, “J-means: A new local search heuristic for minimum sum-of-squares clustering”. Patt. Recog. 2002, 34: 405-413.
12. Dulyakarn P, Rangsaseri Y., "Fuzzy c-means Clustering Using Spatial Information with Application to Remote Sensing," In 22nd Asian Conference on Remote Sensing. Singapore: Center for Remote Imaging, Sensing & Processing, 2001: 5-9.
13. Wen, X., S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker and R. Somogyi, “Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. USA., 95: 334- 339.
14. A. Demiriz, K. P. Bennett, and M. J. Embrechts., "Semi-supervised clustering using genetic algorithms," In Artificial Neural Networks in Engineering (ANNIE-99), pages 809–814, 1999.
15. Marek Obitko, “Prediction Using Neural Networks”, 2007, http://www.obitko.com/tutorials/neural-network-prediction/introduction.html.
16. XLMiner Online Help, “Neural Networks Prediction”, 2007, http://www.resample.com /xlminer/he p/NNCPredict/NNCPredict_intro.htm.
17. Saracoglu Ö.G., “Artificial neural network approach for prediction of absorption measurements of an evanescent field fiber sensor”. Sensors, 2008, 8: 1585-1594.
18. Pramanik, K., “Use of artificial neural networks for prediction of cell mass and ethanol concentration in batch fermentation using saccharomyces cerevisiae yeast”. J. Inst. Eng. India Chem. Eng., 2004, 85: 31-35.
19. Savkovic-Stevanovic, “Neural networks for process analysis and optimization: modeling and applications”. Comput. Chem. Eng., 1994, 18: 1149-1155.
20. Phillip H. Sherrod, "Multilayer Perceptron Neural Networks",DTREG-Predictive Modelling Software , 2008, http://www.dtreg.com/mlfn.htm.