FOREST TYPE CLASSIFICATION USING DATA MINING TECHNIQUES

Dr. Harish Kumar
HOD ,IT division ICFRE, India,
harish@icfre.org

Chandradeep Bhatt
Lecturer, ITM Dehradun, India,
bhattchandradeep@gmail.com

Pashupati Bhatt
Sr. Lecturer, ITM, Dehradun, India,
ppbhatt12@gmail.com

Km. Neha
Lecturer, KIT, Dehradun, India
nehagodiyal11@gmail.com

ABSTRACT

India's forest has most diverse in there composition. The composition has unique combination of Indo-Malayan and Austral species with significant geological and paleo-botanical values. The forest in India has Natural Evolution due to the diverse climatic condition. In Indian forest maximum number of endemic species including Dipeterocarpace species which is considered to be earliest species to evolve. India's forest over the different geo-climatic condition with diverse climatic condition evolved in natural way. This leads to evolution of diverse forest type.

The Champion and Seth (1968) classification of forest is being universally accepted and climate variables (rainfall and temperature) along with vegetation characteristics are criteria for the major types and sub-types. These are contemporary research which implied that apart from these parameters many factor effects classification.

In this study the climatology data for the stations mentioned in Champions and Seth (1968) forest classification have been collected, consolidate 1961-90 and 1991 onwards for mean temperature, maximum temperature, total rainfall (mm) and number of rainy days. These five factors are considered for predicting the major forest type (12). Around 896 data sets was available in which 50% are taken for training and another 50% for testing. Sustainability of applying data mining technique for prediction of forest type using these five parameters it was found that two classification algorithm i.e. Artificial Neural Network (ANN) and Support Vector Machine may be applied. The two algorithms were applied and compared.

Key Words: Data Mining, Classification, SVM, Artificial Neural Network, Machine Learning

I. INTRODUCTION

India's forests have a very unique and most diverse vegetation types in the world. The forests of India have Indo-Malayan and Australian species in the forest which signifying the geological and paleo-botanical. The diversity of forest may be due to number of anthropogenic pressures on these forests. Champion and Seth classified India's forest types based on climate, altitude, aspect, physiognomic traits and species composition**.**

1.1 India forest type classification

Carl Linseer made an investigation on the inflame of temperature upon the plant development, in connection with a special study on evolution of plant through geological part (through fossil records). De-Candolle drummed in his contribution scheme of classification which consisted of six vegetation division. Out of which five have been adjusted to different degree of range of annual temperature.

De Candolle's vegetation classes and temperature limits:

Vegetation Classes	**Mean temperature limits**
Magisto – thermal	Above 30^o C
Mega – thermal	20^o to 30^o (Torrid)
Meso – thermal	15^o to 20^o (Warm Temperate)
Micro – thermal	0^o to 15^o C (Cold Temperate)
Hekis – thermal	Below 0^o C (Frigid)

Koppen published an article dividing the earth surface into different temperature zones based on monthly means of temperature and precipitation. This classification identifies the five main groups of climates are designated as below:

A – Type climate: Topical rainy climates (not all seasons)
B - Type climates : Dry climate
C – Type climates : Warm rainy climates (mild winters)
D - Type climates : Boreal climates (severe winters)
E – Type climates : Polar climates.

1.2 Classification of Vegetation

Climate has been a major determinant of vegetation and vegetation is closely correlated with major climatic zones of the world. Climate suitability and productivity of tree species at a particular site is taken as the main basis for classification of vegetation of an area. Most durable of the climatic classification used in vegetation ecology was given by Koeppen and Thornth Waite. FAO/UNEP joint activities in the field of forest resources assessment and monitoring, originating in the recommendation of the 1972 Stockholm UN Conference on Human Environment and implemented within the overall framework of the Global Environment Monitoring System (GEMS).

II. DATA MINING TECHNIQUES

In the 21^{st} century as we are moving towards more and more online system, the databases have grown into terabytes.

Within this huge data, information of importance needs to be identified. Since the evolution of human life, the people discover patterns. As farmer recognizes pattern of growth in the field, bank recognizes the earning and spending pattern of a customer and politicians seeks pattern in voter opinion. This huge amount of data needs to be used either for business growth or scientific discoveries. The process of discovering the patterns and relationships in data using the analysis tools is called Data Mining. The simplest form of data mining is as follows:

1. Describing the data: Making summary using statistical attributes (mean, median, standard deviation) and visually getting the potentially useful relationship amongst variable.
2. Build up Predictive Model: The pattern may be built based on the results. The testing of the model outside the original sample needs to be done.
3. Verify the Model: The model can be verified from the database where the results are already available.

2.1 Model and Pattern

Model is based on the historic data and may be defined as the summary of the data set. Using a model we can make future prediction. Pattern describes the structure and the relationship which exists in the data.

Hand *et al*. defines data mining which lies as a new tool at the intersection of statistics, machine learning, databases, pattern recognition and other areas.Berry *et al.* defines data mining as finding the useful pattern in the data set. However all data base queries finds new patterns. Structured query language (SQL), Online analytical processing (OLAP) and data mining techniques tend to achieve interesting patterns in the database in the varied degree of complexity. Relational algebra is used in constructing the SQL queries and is based on a set of principles given by Codd. On Line Analytical Processing (OLAP) is based on multi-dimensional data model with high level of querying. OLAP is a better technique than SQL and it focuses on multi-dimensional data bases.

Knowledge discovery (KDD) is defined as process of discovering useful pattern from data. In this process data mining is one of the steps in knowledge discovery. KDD is the multi-disciplinary activity and is beyond the scope of any one discipline or KDD can be defined as "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". The process refers to KDD, consisting of multiple steps. Non-trivial means that the data cannot be transmitted straight forwardly based on some quantities such as computation of average value. No one refers about the usefulness of the data i.e. it can be potentially useful. Understandable refers to the fact that it should be understood immediately or at some later stage, after some processing.

2.2 Data mining and Algorithm Components

There are three algorithmic components of data mining,

1. Model representation
2. Model Evolution
3. Search

Discoverable patterns are described using the language which is called as Model Representation. If it a simple representation, an accurate model, then the amount of time taken for training is negligible. In a KDD process the main focus is that patterns satisfy the criteria of goal. The qualitative statement describes the model evolution criteria. Basically in model evolution we cross validate the predicted accuracy and error.

2.3 Challenges in the data mining

Several methods like Machine Learning, Statistics, Rough Sets, Neural network, Genetic Algorithm, have been developed in the field of data mining. In data mining the technique to solve the problem depends on the type of problem. Some techniques are more suitable than the others in terms of expensive search and prediction error. Classification tree is not suited for the problem with true decision boundaries between the classes. Michalski and Kaufman describe the applicability of machine learning and multi strategy methodology to data mining. The multi strategy is used for conceptual data exploration that is finding out high level concept and description from data. The issue of having noise in the data is one of the challenges.

The other challenges are:

1. Learning dataset may or may not represent actual distribution pattern
2. Learning data may be in complete and some of the values of some attributes are unknown or missing
3. Learning set may be in distributed form. It means that learning database is a collection of datasets which are brought together and patterned within them needs to be identified.
4. Learning from the continuous evolving concept. It is seen some of dataset particularly related to the human being such as interest of user in choosing book is a changing over a period of time.
5. Discovery of pattern based on integrated qualitative and quantitative dataset is also tedious

The topic like Fuzzy set and rough set is given for learning data sets which are flexible, which lack precise definition and has contest dependent meaning. One of the challenges is qualitative prediction which means to predict pattern in sequences and processes.

2.4 Data mining Algorithms / Techniques

In this paper, following tools that can be used in the data mining process are presented. The tools that are described here are generally based on statistics/optimization based methodologies. The disciplines of statistics, machine learning, rough set and fuzzy logic in data mining are used to solve large set of problems. Some of the disciplines are described briefly:

2.4.1 Statistics

Statistics solve the problem by summarizing the data and abstracting knowledge. A cluster analysis is process of discovering cluster in large scale, whereas most important variable describing cluster is found by factor analysis. Cluster analysis provides method the discovery of cluster in dataset of objects which are describes by vector values. The important variable describing the cluster is found by Factor analysis. The

technique used commonly in supervised classification as follows:

2.4.2 K-Nearest Neighbors

It is process of creating user specified cluster from a given dataset by simple iterative partitioning. New data is assign a cluster based on its attributes. This algorithm is simple, relative fast and widely used algorithm.

2.4.3 Nave Bayes

In this classifies the basic assumption is that attributes are conditionally independent in a class. The aim is to construct rules for the new object to a class from class and vector of variable for a given set of object. Assume X_1, X_2, ..., X_{12} be taken attributes

The algorithm estimates the conditional probability P(Xi/Ci) for each class Ci and each attribute. By Byaes theorem we have

$$P[Ci(Xi\ldots..Xn)] = P(X1/Ci)\ \ P(X2/Ci)\ \ \ldots\ldots\ \ P(Xn/Ci)\ \ P(Ci)$$

While assign the new class Cj, the class with largest value of P(Ci|X1, X2,……., Xn) is chosen [46] [48]. Except the limitation of attribute independence, the classifier is robust to noise. In real world the attributes are not independent

2.4.4 Cart

Breiman gave an algorithm based on decision the algorithm which is known as classification and Regression Tree (CART). The Gini's diversity index determines learning is the concept behind the decision tree.

2.4.5 Machine Learning

The subject judgment and qualitative evaluation was an important factor in classification. Detrich and Balpan, made observation the linear model and statistical tool fail to capture subject judgement and qualitative evolution. The assumption that editing and weightage of attributes is fixed in classical classification methods based on linear model and statistics. The outcome of classification of object will also change with change in the scenario i.e. the environment. It is difficult that these subjectively and non-subject able parameter may be considered by statistical methods. The independence and distribution of attribute are assumed as factor effecting the classification. Breiman*et al.*, and Dietrich and Kaplan carried out experiments comparing machine leaving and discriminate analysis and found that machine learning produce better result in terms of predictive accuracy. The predictive accuracy is better in machine learning and it is free from the parametric and structural assumption.

2.4.6 Neural Network

Artificial Neural Network (ANN) model are on principle of basic biological neuron network. It is computational model based on many nonlinear elements arranged in pattern. ANN is good at pattern reorganization, but bad at mathematical calculations, the neural network architecture is based on input layer, hidden layer and output layer. Training the data is an important component of neural network. Tan *et al.*, using connection weight adjustments for training the ANN. The absence of Semantic and reasoning in neural network is being criticized over the rule based classification.

2.6.7 Support Vector Machine

It is well known pattern classification algorithm. Binary classification and regression estimation are performed by the support vector machine (SVM).

Pujari gave two distinct feature of SVM, first it minimizing the expected error as compared to the classification error and it uses duality theory of mathematical programming which is effective computational method.

2.3 Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations that incorporate periods should not have spaces: write "WARSE". Do not use abbreviations in the title unless they are unavoidable.

A. Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the "Equation" markup style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$\int_0^{r_2} F(r,\varphi)\, dr\, d\varphi = [\sigma r_2/(2\mu_0)] \cdot \int_0^{\infty} \exp(-\lambda\,|\,z_j - z_i\,|)\,\lambda^{-1}\, J_1(\lambda r_2)\, J_0(\lambda r_i)\, d\lambda\,. \quad (1)$$

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit tesla). Refer to "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ...."

2.4 Other Recommendations

Use one space after periods and colons. Hyphenate complex modifiers: "zero-field-cooled magnetization." Avoid dangling participles, such as, "Using (1), the potential was calculated." [It is not clear who or what used (1).] Write instead, "The potential was calculated by using (1)," or "Using (1), we calculated the potential."

Use a zero before decimal points: "0.25," not ".25." Use "cm^3," not "cc." Indicate sample dimensions as "0.1 cm × 0.2 cm," not "0.1 × 0.2 cm^2." The abbreviation for "seconds" is "s," not "sec." Do not mix complete spellings and abbreviations of units: use "Wb/m^2" or "webers per square meter," not "webers/m^2." When expressing a range of values, write "7 to 9" or "7-9," not "7~9."

A parenthetical statement at the end of a sentence is punctuated

outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.) In American English, periods and commas are within quotation marks, like "this period." Other punctuation is "outside"! Avoid contractions; for example, write "do not" instead of "don't." The serial comma is preferred: "A, B, and C" instead of "A, B and C."

III. DATA METHODOLOGY AND TOOLS

3.1 Data Set

The Metrological data was obtained from Indian Metrological Organization (Provided by Climate Change Division of ICFRE, Dehra Dun) for this study. The data sets of various stations according to the Champion and Seth (1968) classification of Forest. The 12 major forest types are taken as input i.e. maximum temperature, minimum temperature, mean temperature, number of rainy days and total rainfall.

3.2 Analysis of climate data in Current Study

Data of 80 years (1930-2010), on climate variables of Rainfall and temperature was collected and was analyzed to record any change in the pattern of rainfall, amount of rainfall and the number of rainy days. The data obtained from India Meteorological Department (IMD) was analyzed by dividing the period into three blocks of period i.e. 1931-1960, 1961-90 and 1991-2010.

The analysis of rainfall for the last 80 years across the country covering some important stations representing different forest types has revealed very significant variations in both rainfall and temperature regimes from decade to decade and at the intervals of 30 years block. Total 986 data sets are available for the study.

3.3 Tools for Data Mining Technique

Matlab is used for technical component which integrate computation visualizations and programming. It has a very easy interface to provide solution to the problems use with mathematical notations. The uses of Matlab varied from application to application in mathematical computation. It provides the necessary technical knowledge for algorithm development, modeling, simulation and prototyping. Matlab is based on array and each data element is represented in this form. This feature enables problem solving with matrix and formulation, in minimum time and we can write the programme in scaler non-interactive language like C. Matlab stands for Matrix Laboratory and was developed to provide easy access to Matrix Software by Linpack and Espack Project. These are the high end software in matrix computation. The Matlab consists of four parts.

1. Matlab language
2. Matlab working environment
3. Matlab mathematical working function
4. Matlab application programme interface

3.4 SVM

Lib SVM is simple, efficient and easy to use software. For SVM classification and calibration, it is an integrated software and supports vector machines which in turn, supports multiclass classification. Lib SVM also provides an interface for users to link their own programs easily. The Live SVM has following features:-

1. Diff SVM formulation
2. Efficient multiclass classification
3. Cross validation for model selection
4. Probability estimations
5. Various Kernel(pre-computed Karnak matrix)
6. Waited SVM for unbalanced data

The word "data" is plural, not singular. The subscript for the permeability of vacuum μ_0 is zero, not a lowercase letter "o." The term for residual magnetization is "remanence"; the adjective is "remanent"; do not write "remnance" or "remnant." Use the word "micrometer" instead of "micron." A graph within a graph is an "inset," not an "insert." The word "alternatively" is preferred to the word "alternately" (unless you really mean something that alternates). Use the word "whereas" instead of "while" (unless you are referring to simultaneous events). Do not use the word "essentially" to mean "approximately" or "effectively." Do not use the word "issue" as a euphemism for "problem." When compositions are not specified, separate chemical symbols by en-dashes; for example, "NiMn" indicates the intermetallic compound $Ni_{0.5}Mn_{0.5}$ whereas "Ni–Mn" indicates an alloy of some composition Ni_xMn_{1-x}.

IV. FOREST TYPE CLASSIFICATION

4.1 Classification Using Artificial Neural Networks

Forest type classification is an important problem in order to take future decisions, for the administrators. The conservation of forests is an important activity to be undertaken by forest managers. If they can predict the type of forest based on some initial parameters, probably they can plan their policies efficiently for conserving the forests. In such a situation, the policy makers and foresters can evolve suitable strategies for appropriate management interventions for sustainable forest management, through participatory approach. It has been noted that there exists no significant study on forest type classification in India, using neural network models. There are, however, some studies that were surveyed earlier in similar directions. In this section, we have investigated neural network models to classify the forest types using ANN models. All the simulation experiments are conducted using Neural Network Toolbox provided in Matlab.

4.1.1Construction of ANN Models using Back Propagation Technique

The back-propagation neural network is a universal approximate [61]. The multilayer perception, with BPNN training, is the most widely used ANN model these days. A successful application of the neural networks, for forest type classification requires a good comprehension of the effect of several internal parameters. We have experimented in this thesis with network structure (number of neurons in the hidden layer) and training processes. In this chapter, a three-layer feed forward back-propagation neural network is developed through experimental investigations of various internal parameters to predict forest type.

The experimentation is initiated with certain arbitrarily selected network architecture, on the basis of the knowledge gathered through the literature review. We have experimented with two training functions in order to predict the forest type. We have also conducted the experiments for number of neurons in the hidden layer. After experimenting with a number of neurons in the hidden layer, those results that could converge for the specified number of neurons in the hidden layer have been presented. It is worth mentioning here that we experimented with a number of other combinations, for the hidden neurons in the hidden layer.

The available dataset has been divided into two portions for the purpose of training and testing. The available 986 patterns is partitioned randomly into two disjoint subsets, *viz.*, 'training set', containing 50% records and test set again consisting of the remaining 50% of the patterns. The neural network model based on each of the two training algorithms is trained with one hidden layer containing 80, 90, 100, 110, and 120 neurons. It is worth noting that we have employed 'tansig' transfer function from the input layer to hidden layer and 'purelin' transfer function from hidden layer to output layer in all these experiments. Following graphs (Figure 5.1 page 47) give the details of these experiments.

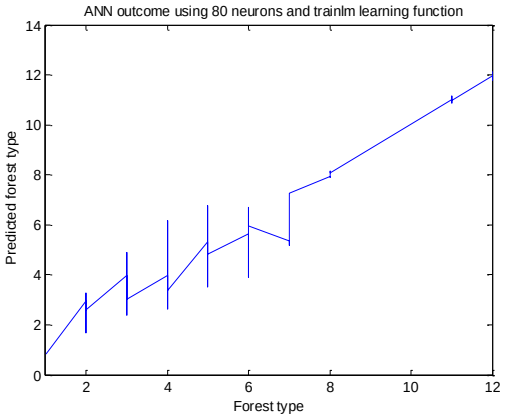

Figure 1: Predicted forest type as a function of forest type using ANN model with 80 neurons in hidden layer and trainlm function

Figure 1 contains the results on testing patterns, for the case when trainlm function is used for training the network and we have taken 80 neurons in the hidden layer. The model could achieve an MSE of 0.40 after 381 iterations.

Figure 2, contains the results of testing when ANN model with 90 neurons in the hidden layer is considered with trainlm as the training function. In this case, we could achieve an MSE of 0.40 with 418 iterations.

Figure 3, contains the similar results for the case when 100 neurons are taken in the hidden layer. The experiment is repeated for two more cases when 110 and 120 neurons are considered in the hidden layer. These results are presented in Fig 5.4 and Fig. 5.5, respectively.

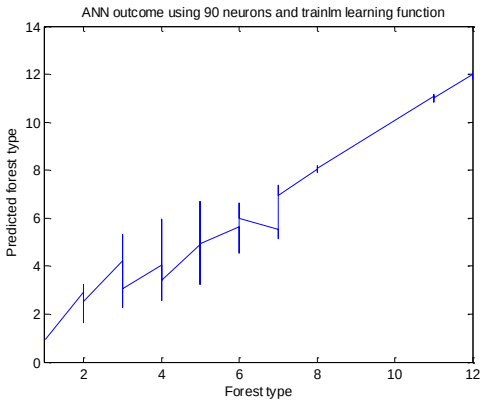

Figure 2: Predicted forest type as a function of forest type using ANN model with 90 neurons in hidden layer and trainlm function

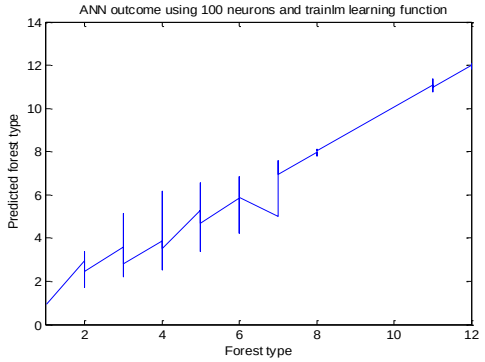

Figure 3: Predicted forest type as a function of forest type using ANN model with 100 neurons in hidden layer and trainlm function

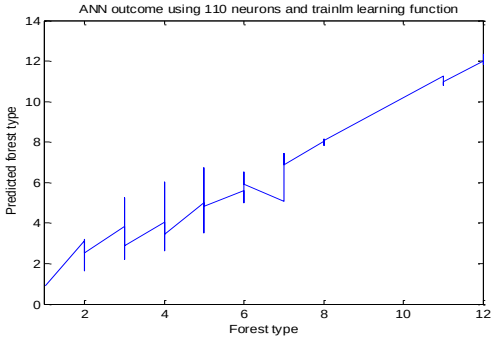

Figure 4: Predicted forest type as a function of forest type using ANN model with 110 neurons in hidden layer and trainlm function

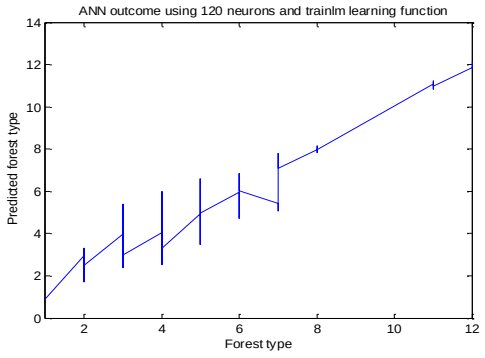

Figure 5: Predicted forest type as a function of forest type using ANN model with 120 neurons in hidden layer and trainlm function

It was also observed that with 100 neurons in the hidden layer, the MSE of 0.4 was achieved with 201 iterations, with 110 neurons the result was achieved with 202 iterations and with 120 neurons it was achieved with 178 iterations.

We have repeated these experiments with one more learning algorithm, namely, 'trainscg'. Following graphs give the results of this study. While training the ANN models with this algorithm. It was not possible to get the models trained for an MSE of 0.4, however, these models were trained for an MSE of 0.5. As such, all these models contain the results that are obtained after training the models with a tolerance value of 0.5.

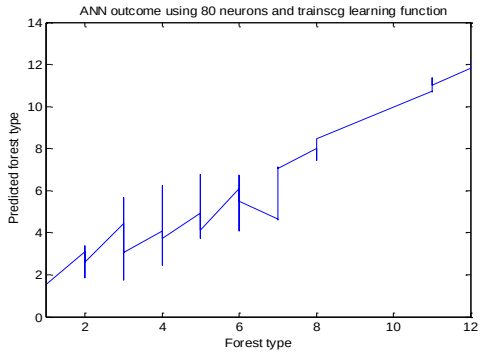

Figure 6: Predicted forest type as a function of forest type using ANN model with 80 neurons in hidden layer and trainscg function

The figure 6, represents the predicted forest type as a function of forest type. In this model, the ANN model was trained with the help of 'trainscg' learning algorithm. In this model, we have also taken 80 neurons in the hidden layer. The network was trained in 7754 iterations, in order to achieve an MSE of 0.5. Figure 5.7 below contains the results of experiments that have been carried out for ANN training with 'trainscg' function and with 90 neurons in the hidden layer. This network could be trained in 5996 iterations in order to achieve an MSE of 0.5.

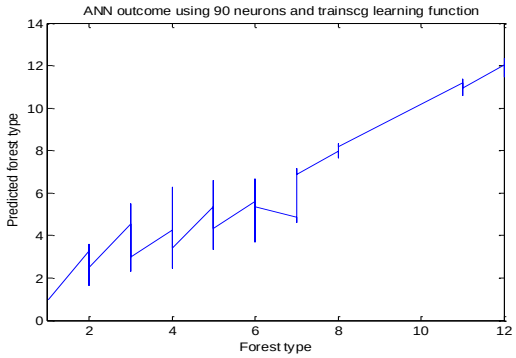

Figure 7: Predicted forest type a forest type s a function of using ANN model with 90 neurons in hidden layer and trainscg function

Figure 8 below, contains similar results when 100 neurons were taken in the hidden layer and the network was trained with 'trainscg' function. This model was trained in 5762 iterations.

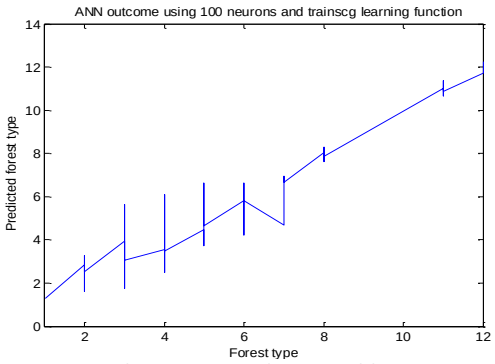

Figure 8: Predicted forest type as a function of forest type using ANN model with 100 neurons in hidden layer and trainscg function

This study has further been extended to include 110 and 120 neurons in the hidden layer. The training function 'trainscg' has been repeated for the training of these two ANN models. The results of these two experiments are prseneted in Figure 9, and Figure 10 respectively. It was noted that the ANN model with 110 neurons in hidden layer could be trained in 6366 iterations and the ANN model with 120 neurons in the hidden layer could get trained in 3940 iterations, in order to achieve an MSE of 0.5.

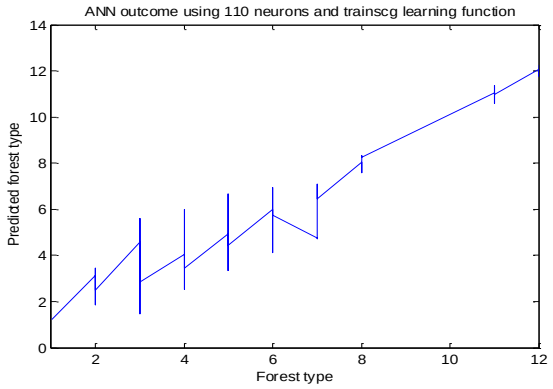

Figure 9: Predicted forest type as a function of fores type using ANN model with 110 neurons in hidden layer and trainscg function *t*

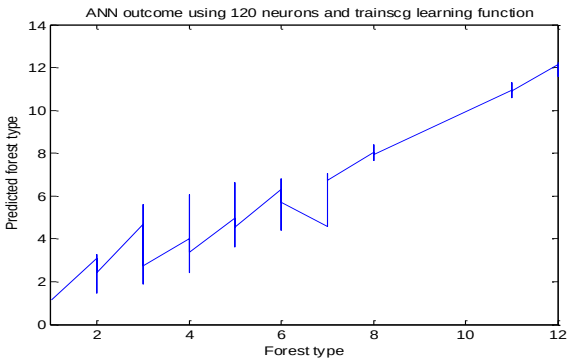

Figure 10: Predicted forest type as a function of forest type using ANN model with 120 neurons in hidden layer and trainscg function

4.2 Classification Using Support Vector Machine

SVM is regarded as a powerful technique in order to deal with the classification problems. In this work, we have explored different parameters of SVM in order to find the best possible recognition accuracy. The LibSVM tool has been used,

for conducting experiments in this work. There are different parameters of SVM that can be considered for the purpose of optimizing the accuracy of classification by the SVM. The first parameter that has been considered is the type of kernel that can be used with SVM. Three different types of kernels have been used in this work. These are, linear kernel, polynomial kernel with degree 3, and the RBF kernel. The second parameter that has been considered is the number of folds that can be varied. We have considered three values of *n* in the *n*-fold cross validation. These values are *n* = 3, 4 and 5. In a*n*-fold cross validation, the data is partitioned into *n* sets. The SVM takes patterns from *n* – 1 set for training and testing accuracy is calculated for remaining one set. This procedure is repeated *n* times giving *n* testing accuracies. We obtain a final testing accuracy as the average of these *n* accuracies obtained in *n* different experiments.

There are two more parameters that have been experimented in this work. These are (i) scaling the data set to a specific interval and (ii) tolerance value (ε). The data values have been scaled to four different intervals, namely, [-1, 1], [-3, 3], [1, 3] and [1, 7]. As such, there are 4 levels of scales for the purpose of scaling the data that have been used in this work. The four values of tolerance level that have been considered in this work are 0.0001, 0.001, 0.01 and 0.1.

All these values of the parameters, namely:- 3 different kernels, 3 different values of folds in *n*-fold cross validation, 4 different values of the intervals to which the data has been scaled and 4 values of the tolerance level has given us 144 different SVM models. These 144 models have been used for experimentation in this work. One more parameter, γ can also be considered for experimentation. We have considered the default value of this parameter in our experiments. The default value of this parameter in LibSVM is considered as 1 / (number of features in the data set). There are five features that have been considered in the forest type classification in this work. As such, we are taking the value of γ as 0.2.

4.2.1 SVM with linear kernel

This section presents the results of the experiments that have been carried out for SVM with linear kernel. Table 1, contains the results of these experiments when we have employed 3-fold cross validation. One can browse from this table that the highest accuracy achieved in this case, when we employ linear kernel with 3-fold cross validation, is 59.5% (when the scaled interval is either [-3, 3] or [1, 7] and the tolerance level is 0.1).

Table – 1: Accuracy obtained using linear kernel with 3-fold cross validation

Sl. No.	Value of Tolerance (ε)	Scaled Interval	Recognition Accuracy (in %)
1	0.0001	[-1,1]	56.3
2	0.0001	[-3,3]	59.2
3	0.0001	[1,3]	56.4
4	0.0001	[1,7]	59.2
5	0.001	[-1,1]	56.5
6	0.001	[-3,3]	59.2
7	0.001	[1,3]	56.5
8	0.001	[1,7]	59.2
9	0.01	[-1,1]	56.4
10	0.01	[-3,3]	59.3
11	0.01	[1,3]	56.5
12	0.01	[1,7]	59.4
13	0.1	[-1,1]	56.6
14	0.1	[-3,3]	**59.5**
15	0.1	[1,3]	56.6
16	0.1	[1,7]	**59.5**

Table 2 below contains the results for linear kernel with 4-fold cross validation.

Table – 2: Accuracy obtained using linear kernel with 4-fold cross validation

Sr. No.	Value of Tolerance (ε)	Scaled Interval	Recognition Accuracy (in %)
1	0.0001	[-1,1]	57.1
2	0.0001	[-3,3]	58.9
3	0.0001	[1,3]	57.1
4	0.0001	[1,7]	58.9
5	0.001	[-1,1]	57.1
6	0.001	[-3,3]	58.9
7	0.001	[1,3]	57.1
8	0.001	[1,7]	58.9
9	0.01	[-1,1]	57.2
10	0.01	[-3,3]	58.9
11	0.01	[1,3]	57.1
12	0.01	[1,7]	58.9
13	0.1	[-1,1]	56.9
14	0.1	[-3,3]	58.9
15	0.1	[1,3]	56.9
16	0.1	[1,7]	58.8

This is evident from Table 2 that for the linear kernel with 4-fold cross validation, an accuracy of 58.9% is achieved in 7 cases. As such, here in this case the value of tolerance level and that of scaled interval is not making significant impact.
Table 6.3 below consists of the results that have been obtained when we consider linear kernel with 5-fold cross validation.

Table – 3: Accuracy obtained using polynomial kernel with 4-fold cross validation

Sr. No.	Value of Tolerance (ε)	Scaled Interval	Recognition Accuracy (in %)
1	0.0001	[-1,1]	38.7
2	0.0001	[-3,3]	59.7
3	0.0001	[1,3]	58.8
4	0.0001	[1,7]	**62.8**
5	0.001	[-1,1]	38.7
6	0.001	[-3,3]	59.7
7	0.001	[1,3]	58.8
8	0.001	[1,7]	**62.8**
9	0.01	[-1,1]	38.7
10	0.01	[-3,3]	59.8
11	0.01	[1,3]	58.9
12	0.01	[1,7]	**62.8**
13	0.1	[-1,1]	39.0
14	0.1	[-3,3]	59.6
15	0.1	[1,3]	59.3
16	0.1	[1,7]	**62.8**

This table reveals that a maximum accuracy of 62.8% has been achieved for 4 different combinations. There is a commonality in these cases i.e. cases are for the same value of the scaled interval [1, 7]. Table 6.6 below contains the results of the experiments that have been performed using 5-fold cross validation and polynomial kernel with degree 3.

V. RESULTS AND DISCUSSION

We have developed the ANN models that have a five input one output structure. We have considered one hidden layer in these models. The numbers of neurons in the hidden layer and also with the training strategies for training the ANN models have been experimented with. As mentioned above, five values of the number of neurons in the hidden layer have been considered as 80, 90, 100, 110 and 120. The training strategies that have been adopted in this work are 'trainlm' and trainscg. It is worth mentioning that these are not the only experiments that have been carried out in this work. We have also experimented with twelve different training methods and with different number of neurons in the hidden layer. The results reported in this work are for those cases where the ANN was trained for the specified tolerance level.

In these experiments, it is worth mentioning that ANN model could be trained with 'trainlm' training function, at a faster pace when 100 neurons were taken in the hidden layer. Also, the ANN models could not be trained for an MSE of 0.4 and we took the MSE as 0.5 here. The ANN model with 120 neurons in the hidden layer could be trained most rapidly when 'trainscg' function for training was taken up. we have considered 144 different SVMs. These SVMs were considered for different values of the parameters considered for the SVMs. The highest recognition accuracy achieved in these experiments is 62.8%. This was achieved when we considered the polynomial kernel with degree 3 and 4-fold cross validation. A significant observation was made when we employ polynomial kernel with degree 3 and 5-fold cross validation, an accuracy of 62.7% was achieved. As such, polynomial kernel is a promising kernel for classifying the forest type. This can further be explored by experimenting with other combinations of the parameters.

In most of the accuracies that have fallen into the maximum category, it has come to the forefront that a good interval of scaling is [1, 7]. We can further investigate other parameters for the scaling interval.

VI. CONCLUSION

This paper deals with the problem of forest type classification. Classifying the type of forest is an important problem that has ecological effects and has an impact on economy of our country. A good prediction of type of forest will help the forest managers in planning their activities for the benefit of the masses of our country. This will also help them in making the plans for future growth of forest, which is essential for a better ecosystem.

Data mining techniques are widely being used for different domain such as medicine, supply chain manage, pattern recognition etc. We, however, have used two promising techniques, namely Artificial Neural Network (ANN) modeling and SVM for conducting the studies. These techniques have successfully being applied in above mentioned classification of study.

we have considered two training algorithms, namely, 'train lm' and 'train scg' for training the ANN model and five example values of number of neurons in the hidden layer, namely, 80, 90, 100, 110 and 120 have been considered. It has been ANN model with 'train lm' training methodology converged with faster rate when 100 neurons were considered in hidden layer. Also, the ANN model with 'train scg' training methodology converged with faster rate when 120 neurons were considered.

The other DM technique that has been explored is SVM. We have used the lib SVM tools for implementing the SVM model in this thesis. Lib SVM is also widely used tools in various real life applications these days. The parameter that have been experimented are (i) the kernel used in SVM, (ii) the number of folds in N-fold cross validation, (iii) tolerance level (Є) and (iv) scaling parameter. Hence we have considered three different kernels; three values of n, in n-fold cross validation, four values of tolerance and four different When we employed linear kernel for SVM modeling, we achieved an accuracy of 62.8 % with

3/4/5 fold cross validation. Also an accuracy of 59.3 % was achieved with RBF kernel when 3/4/5 – fold cross validation was used. As such the highest accuracy 642.8 % has been achieved in this work.

6.1 Future Work

The present studies use limited parameters of climate such as mean temperature, maximum temperature, rainfall and number of rain days (as input for prediction of the major forest type). The detailed classification of India forest type was done Champion and Seth (1968) adopting the number of criteria like climate, physiognomic traits, species composition, locality factor, altitude, aspect, soil etc. This study may be extended by increasing the parameters one by one. Example, First we may add latitude and then soil condition for predicting the result. One may also attempt to classify to a lower level of subtype/species level.

REFERENCES

[1]. FSI (2009) India State of Forest Report (2009): Forest Survey of India (Ministry of Environment and Forests) Dehradun.
[2]. FAO (2010) Global Forest Resources Assessment (2010); FAO Forestry Paper 163 Food and Agriculture Organization of United Nation, Rome, (2010).
[3]. FSI (1987) India State of Forest Report (1987): Forest Survey of India (Ministry of Environment and Forests) Dehradun.
[4]. Spur, S. H., and Barnes, B. V., (1980). Forest Ecology, 3rd Edition. John Wiley and Sons, Inc.
[5]. Hand *et al.*,(2001). Hand D., Mannila H, Smyth P., Principals of Data Mining, Inference and prediction, Springer, 2001.
[6]. Berry *et al.,* (2000), Berry, M.J.A. and Linoff., G.S. Mastering Data Mining, John Wiley & Sons, Inc 2000.
[7]. Han *et al.,* (2001) Han, J., Kamber, M. Data Mining Concepts and Technique, Morgan Kaufmann Publisher, 2001.
[8]. Fayyad *et al.,* (1996). Fayyad , U . Data Mining and Knowledge discovery: Making sense out of Data, IEEE ert, Oct. 20-25, 1996.
[9]. Agarwal *et al.,* (1993) R Agarwal, C. Faloutsos, and A Swami. Efficient similarity search in sequence database. In proc. of Fourth International Confrerence on Foundation of Data Organization nad Algorithms, Chicago, Oct 1993.
[10]. Alvarez, S., 1995. Generation of terrain textures using neural networks. Unpublished masters thesis. Department of Computer Science, Colorado State University, 44 pp.
[11]. Fowler, C.J., Clarke, B.J., 1996. Corporate distress prediction: a comparison of the classification power of a neural network and a multiple discriminant analysis model. Accounting Forum 20 (3–4), 251–269.