International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
345 – 349

Improved Soil Data Prediction Model Base Bioinspired K-Nearest Neighbor Techniques for Spatial Data Analysis in Coimbatore Region

B. Murugesa Kumar, Dr. K. Ananda Kumar, Dr. A. Bharathi
Research Scholar, Bharathiar University, Coimbatore, Tamilnadu
Assistant Professor(Sl.Grade), Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu
Professor, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu

Abstract: In this research paper, agricultural Data Mining data are summarized. An improved Soil Data Prediction Model is developed to estimate the above parameters at locations for Coimbatore city. 142 locations were investigated for the development of the model. The model involves multiple regression equation, chi-square test and Bio inspired k-nearest neighbor classification. The GIS was used to manage the database and to develop thematic maps for depth, N value, free swell, liquid limit, plastic limit, plastic index, percentage gravel, percentage sand and percentage slit and clay. Field and laboratory studies were conducted in four locations and are compared with the predicted values.

Keywords: *Data mining, Soil Data Analysis, K-nearest neighbor Algorithm, Dataset*

*****

I. INTRODUCTION

Data Mining (DM) becomes popular in the field of agriculture for soil classification, wasteland management and crop and pest management. The variety of association techniques in DM and applied into the database of soil science to predict the meaningful relationships and provided association rules for different soil types in agriculture. Similarly, agriculture prediction, disease detection and optimizing the pesticides are analyzed with the use of various data mining techniques earlier. DM techniques for knowledge discovery in agriculture sector and introduced different exhibits for knowledge discovery in the form of Association Rules, Clustering, Classification and Correlation. In predicted the soil fertility classes using with classification techniques were Naïve Bayes, J48 and K-Nearest Neighbor algorithms. In6 used Adopted data mining techniques to estimate crop yield analysis. Multiple Linear Regression (MLR) method was used to find the linear relationship between dependent and independent variables. K-Means clustering approach was also use to form four clusters considering Rainfall as key parameter. Decision tree, Bayesian Network data mining techniques and the non-linear approaches were implemented. Optimization based Bayesian Network approach was considered as better than non-linear[1]. Due to expensive and time consuming process of subsoil investigation works, the present research was undertaken to develop a model in order to predict the important geotechnical related parameters such as N value, differential, free swell and the depth of foundation taking advantage of the spatial continuity and data mining technique using Weka.

Study area location

The Corporation of Coimbatore is the study area selected which is situated in the Southern part of the Peninsula at a longitude of $77^0$04’00” and $76^0$54’00” and latitude of $11^0$03’30” and 10□57’30”. The total geographical area of the city is 105.5 sq.km.A total of 138 bore hole locations were identified for the purpose of this study. Soil samplings were taken at each layer to analyse the soil for its grain size distribution, plasticity characteristics and differential free swell in the laboratory. The bore hole locations were selected such that it covers the entire area of Coimbatore City Corporation. It was also ensured that the four zones of the corporation namely Coimbatore North, Coimbatore South, Coimbatore West and Coimbatore East were also individually well covered. The zone wise number of bore hole investigation points are shown in Table-1.

Sl.No	Description of Zone	Number of bore hole locations
1	Coimbatore North	47
2	Coimbatore South	25
3	Coimbatore West	44
4	Coimbatore East	24

The lesser number of bore hole locations in Coimbatore East and Coimbatore south when compared to Coimbatore West and Coimbatore North is due to the presence of water bodies in these regions. The locations of bore hole are shown in Fig.1

 ISSN: 2454-4248
345 – 349

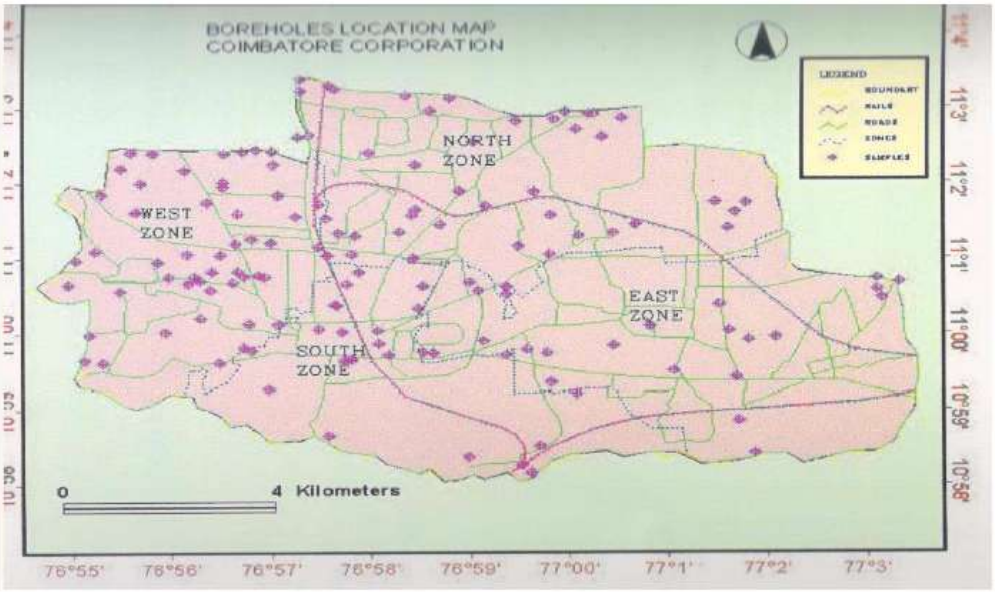

Fig.1 Bore-Holes location map Coimbatore Corporation

Industrialized Data Mining

Data Mining is essential to discover the agricultural related knowledge such as soil fertility, yield prediction and soil erosion. Soil prediction helps to for soil remedy and crop management. Classification algorithms involve finding rules that partition the data into disjoint groups. A set of classification rules are generated by such a classification process, which can be used to classify future data. Following section give explanation of classification algorithms such as Naive Bayesian classifier, J48 decision tree classifier and JRip classifier.

Naive Bayes

A Naive Bayes classifier is one of the classifiers in a family of simple probabilistic classification techniques in machine learning. It is based on the Bayes theorem with independence features. Each class labels are estimated through probability of given instance. It needs only small amount of training data to predict class label necessary for classification.

J48 (C4.5)

The J48 is one of the classification-decision tree algorithm and it slightly modified from C4.5 in Weka. It can select the test as best information gain. This algorithm was proposed by Ross Quinlan. C4.5 is also referred to as a statistical classifier. J48 predicts dependent variable from available data. It builds tree based on attributes values of training data. This classifies data with the help of feature of data instances that said to have information gain. The importance of error tolerance is developed using pruning concept.

JRip

IREP optimized version is Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen. This algorithm is a propositional guideline learner. J-Rip classifier is one of the decision tree pruning models based on association rules. It is an effective technique to reduce error pruning. In this algorithm, the training data is split into two sets and with the help of pruning operators the error is reduced on both the sets. Finally rules are formed from two sets such as Growing set and Pruning set.

Mechanized System

Soil classification system is essential for the identification of soil properties. Expert system can be a very powerful tool in identifying soils quickly and accurately .Traditional classification systems include use of tables, flow-charts. This type of manual approach takes a lot of time, hence quick, reliable automated system for soil classification is needed to make better utilization of technician's time [9]. We propose an automated system that has been developed for classifying soils based on fertility. Being rule-based system, it depends on facts, concepts, theories which are required for the implementation of this system. Rules for soil classification were collected from soil testing lab. The soil sample instances were classified into the fertility class labels as: Very High, High, Moderately High, Moderate, Low, and Very Low. These class labels for soil samples were obtained with the help of this system and they have been used further for comparative study of classification algorithms.

 ISSN: 2454-4248
345 – 349

II. RESEARCH METHODOLOGY

Dataset Collection

The dataset is part of surveys which are carried out regularly in Pune District. Primary data for the soil survey are acquired by field sampling. These samples are then sent for chemical and physical analysis at the soil testing laboratories; hence this dataset was collected from a private soil testing lab in Pune. It contains information about number of soil samples taken from 3 regions of Pune district (Khed, Bhor, and Velhe). Dataset has 9 attributes and a total 1988 instances of soil samples.

Sources of data

A number of national and international sources have been used in compiling the database. For agricultural data, the main source was the data set created by James W. McKinsey, Jr. and Robert Evenson (Yale university, CT, USA). As discussed later in the report, a number of corrections have been made in the original data set by cross-checking the same with government publications. Some of the main government publications used in compiling the database are:

- Agricultural Situation in India
- Area and Production of Principal Crops in India
- Agricultural Prices in India
- Fertilizer Statistics (published by Fertilizer Association of India)
- Statistical Abstracts of India

Climate data from over 160 meteorological stations are from the Food and Agricultural Organization (FAO) of the United Nations.

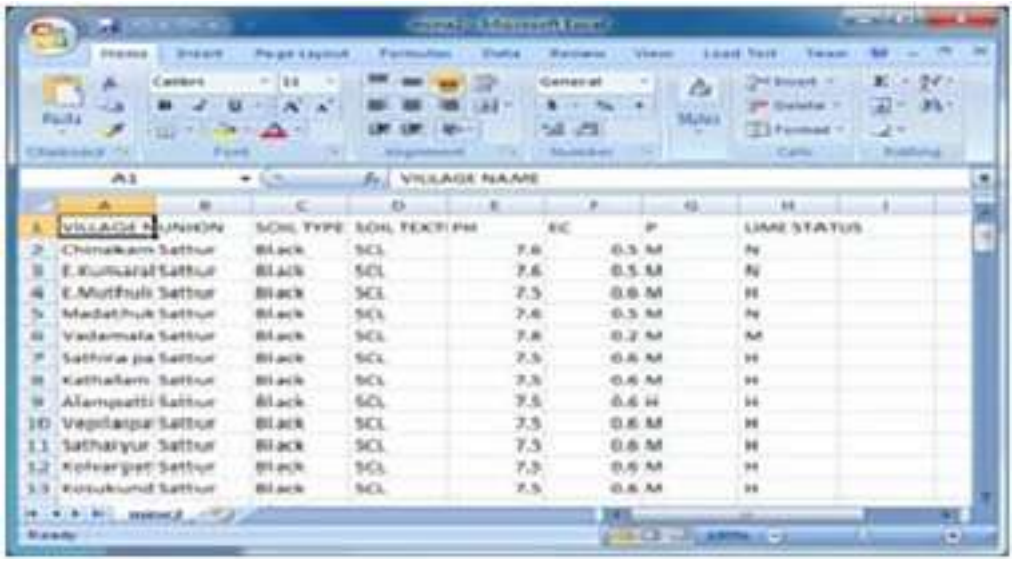

Fig 2: Soil Data set

IMPROVED SOIL DATA PREDICTION MODEL (ISDPM)

Statistical Modeling on the prediction of the soil characteristics was developed by using Weka and data mining technique called Bio-inspired k-nearest neighbour. This model constitute of multiple regression equation, chi-square test and k- nearest neighbour classification.

The correlation analysis measures the degree of association between two sets of quantitative data, while regression analysis explains the variation in one variable, based on the variation in one or more of these variables. The variable, of which the variation is explained, is called dependent variable while the variables which are used to explain the variation are called the independent variables. If there is only one dependent variable and only one independent variable is used to explain the variations in it, then the model is known as simple regression model. If the multiple independent variables are used to explain the variations in a dependent variable, it is called multiple regression models and the process of analysis is called multiple regression analysis[2-4].

The general regression model is of the type

$$Y = a + b1x1 + b2x2 + ……..+bn\ xn$$

Where, Y is the dependent variable and x1, x2, ….., xn are the independent variables expected to be related to Y and expected to explain or predict Y. b1, b2,…..,bn are the coefficients of the respective independent variables, which will be determined from the input data.

Y- Variable
X1-Sand X6 - Plasticity Index
X2- Gravel X7 -Depth
X3- Silt and Clay X8- N value
X4 – Liquid limit X9 – Free Swell
X5- Plastic limit

Predication of N value is given by the equation,

 ISSN: 2454-4248
345 – 349

N Value = 1625.0577*X1 - 545.0223* X2 + 0.5367* X3+ 0.4972* X4+0.4309* X5+ 0.7287* X6 -0.5967* X7 - 1.2586* X8 - 0.1178* X9+ 2284.5463

Prediction of Depth is given by the equation,

Depth= -208.6702* X1 - 43.6708* X2 - 0.0031* X3 - 0.0035* X4 - 0.006* X5 -0.0054* X6+ 0.0035* X7 - 0.0238* X8 + 0.0032* X9 +292.0515

Prediction of Free Swell is given by the equation,

Free Swell = 7162.1547* X1 -1648.0475* X2 - 2.0628* X3- 2.3616* X4-1.8845* X5 + 2.0963* X6 - 0.6052* X7 - 2.1195* X8 + 4.0621* X9- 9068.4318

The regression output gives the coefficients of the independent variables. The coefficients of independent variables are substituted in the above model[5].

Validation

A very powerful test for testing the significance of discrepancy between theory and experiment is known as Chi-square Test of Goodness of Fit. It enables us to find the deviation of the experiment from theory is just by chance or is it really due to inadequacy of the theory to fit the observed data. The value of the test-statistic is

$$X^2 = \sum_{I=1}^{n} \left\{ \frac{(O_i - E_i)^2}{E_i} \right\}$$

where

$X2$ = the test statistic that asymptotically approaches a $\chi 2$ distribution.

Oi = an observed frequency;

Ei = an expected (theoretical) frequency, asserted by the null hypothesis;

n = the number of possible outcomes of each event.

The chi-square fitness test shows that, the calculated value for N value as 29.9934, for depth as 8.55 and for free swell as 124.78 and are less than standard tabulated value (140.2 for more than 100 degree of freedom). Hence the model developed is adequate.

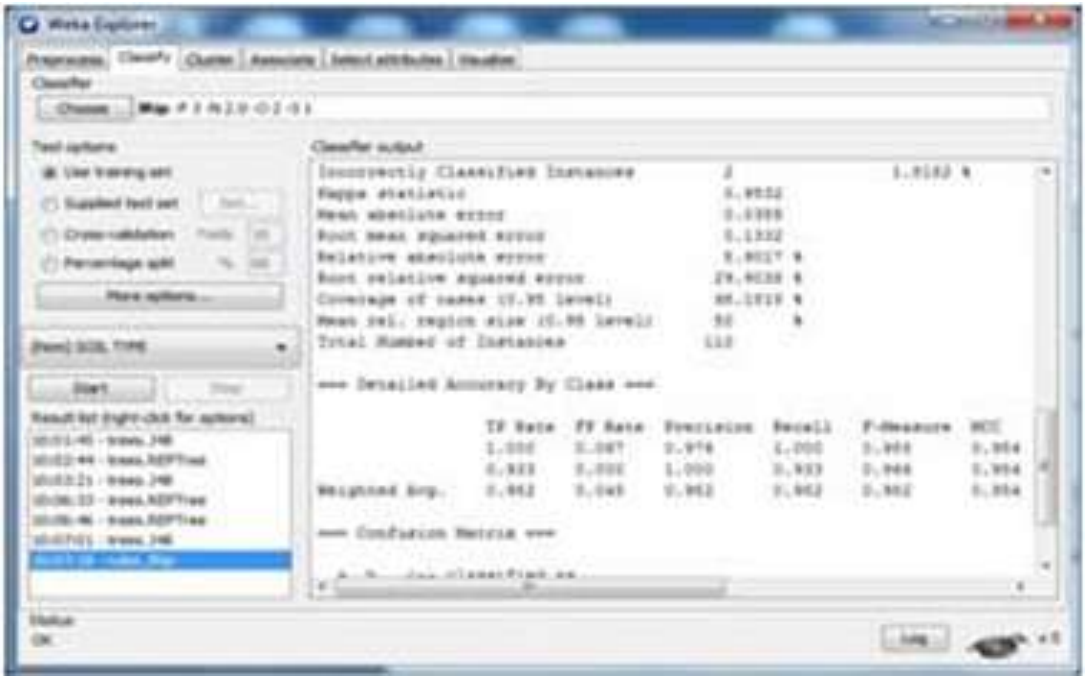

Fig 3: ISDPM data analysis using Weka

 ISSN: 2454-4248
345 – 349

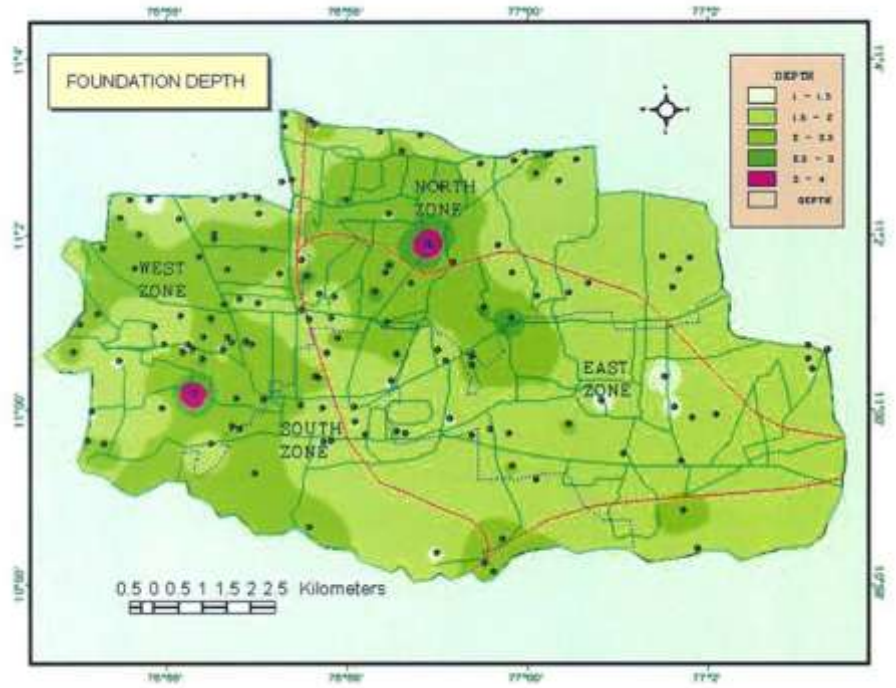

Fig.4 Spatial Variation of Foundation Depth

III. CONCLUSION

An improved Soil Data Prediction Model(ISDPM) is developed for the predication of the parameters required for foundation design. It gives the predicated values based on regression analysis and Bio-inspired K-nearest neighbour. The values predicated using ISDPM are found to have close agreement with actual values. The SCPM has flexibility to include additional parameters and additional data for any other specific purpose oriented studies. The good performance of ISDPM is confirmed by Chi-Square test for goodness of fit where the calculated values are well within the tabulated values.

REFERENCES

[1] Geetha MCS. Implementation of association rule mining for different soil types in agriculture. International Journal of Advanced Research in Computer and Communication Engineering. 2015 Apr; 4(4):520–2.

[2] Fayer M.J. et al (1995), "Estimating recharge rates for a groundwater model using GIS," Journal of Environmental quality V 25 I 3 May-June 1996. pp 510-518.

[3] Kothyari et al (1997), "Sediment yield estimation using GIS," Hydrological science journal V 42 n 6 Dec. 1997.pp 833-843.

[4] Meijerink et al (1996), "Comparison of approaches for erosion modeling using flow accumulation with GIS," Application of GIS system in Hydrology and water resource management. IAHS Publication n 235. IAHS press, Wallingford, Engl. pp 437-444.

[5] Gandhimathi et. al. / International Journal of Engineering Science and Technology,Vol. 2(7), 2010, 2982-2996

[6] Gholap J, Lngole A, Gohil J, Shailesh, Attar V. Soil data analysis using classification techniques and soil attribute prediction. 2012 Jun; 9(3):1–4.

[7] Anuradha C, Velmurugan T. A comparative analysis on the evaluation of classification algorithms in the prediction of student performance. Indian Journal of Science and Technology. 2015 Jul; 8(15):1–12.

[8] Narain B. Study for Data Mining techniques in classification of agricultural land soils. Journal of Advanced Research in Computer Engineering. 2011 Jan-Jun; 5(1):35–7.

[9] Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology. 2015 Nov; 8(29):1–8.

[10] Chandrakar PK, Kumar S, Mukherjee D, Applying classification techniques in Data Mining in agricultural land soil. International Journal