Data Mining Techniques in Soil Data Analysis for Effective Agriculture Data

T. Mathavi Parvathi[1] Dr. Paul Rodrigues[2]
[1]Research Scholar [2]Professor
[1,2]Department of Computer Science & Engineering
[1]Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu India [2]King Khalid University Saudi Arabia

Abstract— The advancement in computers provided large amount of data. The task is to analyze the input data and obtain the required data which can be done by various data mining techniques. Present work focusses on analysis of relationships in spatial datasets are regional and there is a great need for regional regression methods that derive regional reflects different spatial characteristics of different regions. Naive Bayes, J48 (C4.5) and JRip Algorithms were used to analyse the data JRip reported to be simple, efficient classifier of soil data. The selected soil attributes were Nitrogen, Phosphorus, Calcium, Magnesium, Sulphur, Iron, Zinc, Potassium, PH and Humus. The attributes were predicted by linear regression. Even though all regressions provided almost equal results least Median Square depicts better results. This paper proposes a regional regression technique for regions that are defined by a categorical attribute, in particular soil type. The result is a series of hierarchically grouped regions according to their similarity.

Key words: Soil Data, Attributes, JRip, Datamining, Spatial Mining

I. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used for industrial, commercial and scientific purposes. As such the process of data mining involves sorting through large amounts of data and discovering patterns in the data (Witten and Eibe, 2005). Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods (Cunningham and Holmes, 1999).The recent advances in data mining technologies are successfully applied to the management of natural resources also. Armstrong et al., (2007) have reported that data mining empowers farmers in selection of site specific crop varieties by studying soil fertility. The principle objective of classification of soil is for predicting the engineering properties and behaviour of soil finally dictating the choices for use. Laboratory and various statistical techniques are time consuming and highly expensive, efficient techniques can be developed for solving complex soil data sets using data mining to improve the effectiveness and accuracy of the classification of large soil data sets [Kumar and Kannathasan, 2011). Verheyen et al., (2001) have studied the soil characteristics by k-means approach and GPS based data mining techniques. Soil testing laboratories either government or private sector provides various protocols for soil analysis and literature regarding the soil characteristics. Based on the data describes the soil composition and also recommends suitable fertilizers based on the data. It also helps farmers to supply the suitable fertilizer for suitable crop for a particular season.

The present work has taken up to analyze the soil data by various data mining techniques and the outcome obtained has been used by researchers and also farmers for selection of suitable crop and fertilizer to that soil.

II. METHODOLOGY

A. Data Collection

The Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. There is a need for effective and efficient methods to extract unknown information from spatial data sets of large size and complexity [6]. Due to the widespread application of geographic information systems (GIS) and GPS technology, private industries and the general public also have more and more interest in both contributing and using geographic data. Spatial data mining is still at a very early development stage and its limits and potentials are yet to be defined. In spatial data mining, the data cannot tell stories unless we formulate appropriate questions to ask and use appropriate methods to seek the answers from the data.

Pei et al. focuses on the development of a new method for point pattern analysis for detecting feature from spatial point processes using collective nearest neighbor [7]. Establishing spatial clustering methods are often sensitive to the parameterization of the clustering algorithm, particularly to the scale at which one theorizes clustering occurs, as such an assumption often must be made a priori to the application of the clustering technique. Consequently, the results of clustering may be highly subjective.

Sub-path	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)	(1, 7)	(1, 8)	(1, 9)	(1, 10)	(1, 11)
SUM	7	1	2	1	6	11	15	12	17	22	12

Table 1: Look table of SUM function in the sample data

III. GIS MODELS

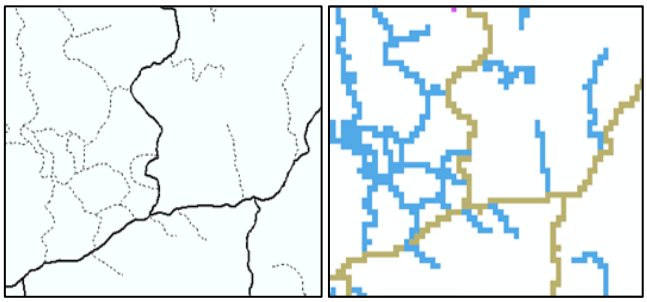

Fig. 1: Vector Data Model and Raster Data Model

Raster images come in the form of individual pixels, and each spatial location or resolution element has a pixel associated where the pixel value indicates the attribute, such as color,

elevation, or an ID number. Raster images are normally acquired by satellites, optical scanner, digital CCD camera and other raster imaging devices. Because a raster image has to have pixels for all spatial locations, it is strictly limited by how big a spatial area it can represent. When increasing the spatial resolution by 2 times, the total size of a two-dimensional raster image will increase by 4 times because the number of pixels is doubled in both X and Y dimensions [12]. The same is true when a larger area is to be covered when using same spatial resolution. Vector data can be easily converted to raster data. Figure 1, shows the graphical representation of raster and vector data models.

IV. JRIP

This algorithm implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP. In this paper, three classification techniques (naïve Bayes, J48 (C4.5) and JRip) in data mining were evaluated and compared on basis of time, accuracy, Error Rate, True Positive Rate and False Positive Rate. Tenfold cross-validation was used in the experiment. Our studies showed that J48 (C4.5) model turned out to be the best classifier for soil samples.

Classifier	Naïve Bayes	JRip	J48
Correctly Classified Instances	765	1794	1827
Incorrectly Classified Instances	1223	194	161
Accuracy	38.40%	90.24%	91.90%
Mean Absolute Error	0.229	0.0411	0.0299

Table 2: Comparison of different classifiers

V. HIERARCHICAL CLUSTERING

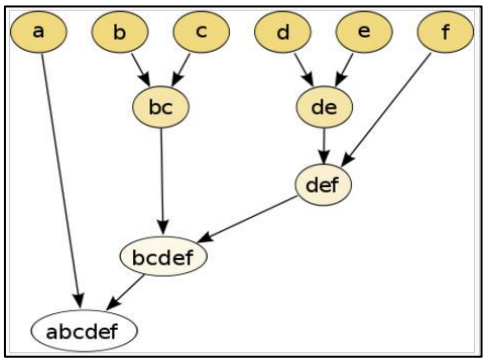

Fig. 2: An Example Hierarchical Dendrogram Representation

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of hierarchical clustering is:

1) Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item [22]. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2) Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3) Compute distances (similarities) between the new cluster and each of the old clusters.
4) Repeats steps 2 and 3 until all items are clustered into a single cluster of size N.

VI. DATA COLLECTION

Import the raster data in GRASS and calculate the NDVI values using GRASS. We have used a script to generate raster map with NDVI values. Import all the required data to ArcGIS for further processing. Convert both yield and NDVI data to point data to extract the point value, these point maps are then spatially joined with each other. From the given agriculture data, we could identify that all the soil types are loamy soils. Gardeners are advised that a loamy garden soil is best for just about all plants. Table 4, shows the abbreviated soil names for the soil types in the experimental agriculture data. By using RMSE technique we have calculated Root Mean Square Error (RMSE) for each soil type and for the entire field. This RMSE is used as measure to identity relationships among the regions in a field.

The script created for performing the above statistics is:

```
FD = read.csv("D:/R Scripts/All_Soil_Types.csv",
header = False)
Soil_Types = c("BBDL39slope", "BBL36slope",
"ABLL69slope", "BSL36slope",
"BSL03slope", "TSL01slope", "HWL03slope",
"SCL03slope", "FRL02slope", "SARC69slope",
"CFSL02slope",
"HTLFS06slope", "CFSL26slope")
# declare a few vector variables to hold the results
nn <- length(Soil_Types)
rse <- rep(1:nn, 0)
X2 <- rep(1:nn, 0)
radj2 <- rep(1:nn, 0)
intercept <- rep(1:nn, 0)
Pcoef <- rep(1:nn, 0)
rmse <- rep(1:nn, 0)
for (i in 1:nn)
{
sd <- subset(DD, DD$Soil_Name ==
Soil_Types[i], select = MDVI:Yield)
MDVI <- sd$MDVI
Yield <- sd$Yield
mytitle = Soil_Types[i]
plot(MDVI,Yield, main = mytitle)
fit <- lm(Yield ~ MDVI)
abline(fit)
sfit <- summary(fit)
eqn <- paste("Y = ", round(fit$coefficients[2],3),
"X + ", round(fit$coefficients[1],3), "(R^2 = ",
round(sfit$r.squared,3),")")
xtextpos <- 1.4*min(MDVI)
mtext(eqn, side = 3)
# text(xtextpos, min(Yield), labels = eqn)
invisible(readline(prompt="Press [enter] to
continue"))
(r2[i] <- sfit$r.squared)
(radj2[i] <- sfit$adj.r.squared)
```



```
        (rse[i] <- sfit$sigma) # residual standard error
        (rmse[i] <- sqrt(mean(sfit$residuals)^2))
        fit$coefficients # intercept and x-coof
        (intercept[i] <- fit$coefficients[1])
        (xcoef[i] <- fit$coefficients[2])
}
```

VII. CONCLUSION

In this paper, the comparative analysis of three algorithms like Naïve Bayes, JRip and J48 is projected. Based on the GIS and spatial database value soil can be classified according to the RMSE algorithm the accuracy and efficient is good. JRip classification algorithm gives better result of this dataset and is correctly classified into maximum number of instances comparing with the other two. JRip can be recommended to predict soil types.

REFERENCES

[1] S. Diplaris, G. Tsoumakas, P. Mitkas & I. Vlahavas, (2005), "Protein Classification with Multiple Algorithms", P. Bozanis and E.N. Houstis (Eds.): pp. 448 – 456
[2] A. Sharma, (2011), "A Comparative Study of Classification Algorithms for Spam Email Data Analysis"; International Journal on Computer Science and Engineering, ISSN: 0975-3397 Vol.2 PP:450-455.
[3] 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.
[4] . Solanki J, Mulge Y. Different techniques used in data mining in agriculture. International Journal of Advanced Research in Computer Science and Software Engineering. 2015 May; 5(5):1223–7.
[5] Bhuyar V. Comparative analysis of classification techniques on soil data to predict fertility rate for Aurangabad District. International Journal of Emerging Trends and Technology in Computer Science. 2014 Mar-Apr; 3(2):200–3
[6] Fathima NG, Geetha R. Agriculture crop pattern using data mining techniques. International Journal of Advanced Research in Computer Science and Software Engineering. 2014 May; 4(5):781–6.
[7] Suman, Naib BB. Soil classification and fertilizer recommendation using WEKA. International Journal of Computer Science and Management Studies. 2013 Jul; 13(5):142–6.
[8] Ramesh D, Vardhan VB. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering. 2013 Sep; 2(9):3477–80.

