Annexure - I:

Reprint of Journal Publications and Conference Proceedings
A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining

Dr. D. Ashok Kumar\(^1\), N. Kannathasan\(^2\)
\(^1\)Government Arts College, Tiruchirapalli- 620 022, Tamil Nadu, India
\(^2\)Mahatma Gandhi Government Arts College, Mahe, P.O. New Mahe-673 311, U. T of Puducherry, India

Abstract— Data mining has emerged as one of the major research domain in the recent decades in order to extract implicit and useful knowledge. This knowledge can be comprehended by humans easily. Initially, this knowledge extraction was computed and evaluated manually using statistical techniques. Subsequently, semi-automated data mining techniques emerged because of the advancement in the technology. Such advancement was also in the form of storage which increases the demands of analysis. In such case, semi-automated techniques have become inefficient. Therefore, automated data mining techniques were introduced to synthesis knowledge efficiently. A survey of the available literature on data mining and pattern recognition for soil data mining is presented in this paper. Data mining in Agricultural soil datasets is a relatively novel research field. Efficient techniques can be developed and tailored for solving complex soil datasets using data mining.

Keywords— Data Mining, Pattern Recognition, Soil Data Mining

I. INTRODUCTION

This Data mining software applications includes various methodologies that have been developed by both commercial and research centers. These techniques have been used for industrial, commercial and scientific purposes. For example, data mining has been used to analyze large datasets and establish useful classification and patterns in the datasets. Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods [16]. This paper outlines research which may establish if new data mining techniques will improve the effectiveness and accuracy of the Classification of large soil datasets. In particularly, this research work aims to compare the performance of the data mining algorithms with soil limitations and soil conditions in respect of the following characteristics: Acidity, Alkalinity and sodicity, Salinity, Low cation exchange capacity, Phosphorus fixation, Cracking and swelling properties, Depth, Soil density and Nutrient content. The use of standard statistical analysis techniques is both time consuming and expensive. If alternative techniques can be found to improve this process, an improvement in the classification of soils may result.

In many developing countries, hunger is forcing people to cultivate land that is unsuitable for agriculture and which can only be converted to agricultural use through enormous efforts and costs, such as those involved in the construction of terraces. Each country is known for its core competence. India's is agriculture. Yet, it only accounts for 17 per cent of the total Gross Domestic Product. With the pressure of urbanization, it is going to be a challenge to produce food for more people with less land and water.

Agriculture or farming forms the backbone of any country economy, since a large population lives in rural areas and is directly or indirectly dependent on agriculture for a living. Income from farming forms the main source for the farming community. The essential requirements for crop harvesting are water resources and capital to buy seeds, fertilizers, pesticides, labor etc. Most farmers raise the required capital by compromising on other necessary expenditures, and when it is still insufficient they resort to credit from sources like banks and private financial institutions. In such a situation, the repayment is dependent on the success of the crop. If the crop fails even once due to several factors, like bad weather pattern; soil type; improper, excessive, and untimely application of both fertilizers and pesticides; adulterated seeds and pesticides etc. then he is pushed into an acute crisis causing severe stress [58]. In addition, the plant growth depends on multiple factors such as soil type, crop type, and weather. Due to lack of plant growth information and expert advice, most of the farmers fail to get a good yield.

Most knowledge of soil in nature comes from soil survey efforts. Soil survey, or soil mapping, is the process of determining the soil types or other properties of the soil cover over a landscape, and mapping them for others to understand and use. Primary data for the soil survey are acquired by field sampling and supported by remote sensing.

The test dataset using for this research work collected from World Soil Information – ISRIC (International Soil Reference and Information Centre). Version 3.1 of the ISRIC-WISE database (WISE3-World Inventory of Soil Emission Potentials) was compiled from a wide range of soil profile data collected by many soil professionals world wide. All profiles have been harmonized with respect to the original Legend (1974) and Revised Legend (1988) of FAO-Unesco. Thereby the primary soil data and any secondary data derived from them can be linked using GIS to the spatial units of the soil map of the world as well as more recent Soil and Terrain (SOTER) databases through the soil legend code.
WISE3 is a relational database, compiled using MS-ACCESS. It can handle data on: (a) soil classification; (b) soil horizon data; (c) source of data; and methods used for determining analytical data. Profile data in WISE3 originate from over 260 different sources, both analogue and digital. Some 40% of the profiles were extracted from auxiliary datasets, including various Soil and Terrain (SOTER) databases and the FAO Soil Database (FAO-SDB), which, in turn, hold data collated from a wide range of sources.

WISE3 holds selected attribute data for 10,253 soil profiles, with some 47,800 horizons, from 149 countries. Individual profiles have been sampled, described, and analyzed according to methods and standards in use in the originating countries. There is no uniform set of properties for which all profiles have analytical data, generally because only selected measurements were planned during the original surveys. Methods used for laboratory determinations of specific soil properties vary between laboratories and over time. Some times, results for the same property cannot be compared directly. WISE3 will inevitably include gaps, being a compilation of legacy soil data derived from traditional soil surveys. These can be of a taxonomic, geographic, and soil analytical nature. As a result, the amount of data available for modeling is some times much less than expected. Adroit use of the data, however, will permit a wide range of agricultural and environmental applications at a global and continental scale (1:500000 and broader) [44].

The analysis of these datasets with various data mining techniques may yield outcomes useful to researchers in future.

II. MATERIALS AND METHODS

The rapid growth of interest in data mining is due to the (i) falling cost of large storage devices and increasing ease of collecting data over networks, (ii) development of robust and efficient machine learning algorithms to process this data, and (iii) falling cost of computational power, enabling use of computationally intensive methods for data analysis [37].

Data Mining (DM) represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data [18]. The DM process has developed due to the immense volume of data that must be handled easier in areas such as: business, medical industry, astronomy, genetics or banking field. Also, the success and the extraordinary development of hardware technologies led to the big capacity of storage on hard--disks, fact that challenged the appearance of many problems in manipulating immense volumes of data. Of course the most important aspect here is the fast growth of the Internet.

The core of the DM process lies in applying methods and algorithms in order to discover and extract patterns from stored data but before this step data must be pre--processed. It is well known that simple use of DM algorithms does not produce good results. Thus, the overall process of finding useful knowledge in raw data involves the sequential adhribution of the following steps: developing an understanding of the application domain, creating a target dataset based on an intelligent way of selecting data by focusing on a subset of variables or data samples, data cleaning and pre--processing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, the data mining step, interpreting mined patterns with possible return to any of the previous steps and consolidating discovered knowledge.

The DM contains many study areas such as machine--learning, pattern recognition in data, databases, statistics, artificial intelligence, data acquisition for expert systems and data visualization. The most important goal here is to extract patterns from data and to bring useful knowledge into an understandable form to the human observer. It is recommended that obtained information to be facile to interpret for the easiness of use. The entire process aims to obtain high--level data from low level data.

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Typically, a data mining algorithm constitutes some combination of the following three components.

- The model: The function of the model (e.g., classification, clustering) and its representational form (e.g. linear discriminants, neural networks). A model contains parameters that are to be determined from the data.
- The preference criterion: A basis for preference of one model or set of parameters over another, depending on the given data.
- The search algorithm: The specification of an algorithm for finding particular models and parameters, given the data, model(s), and a preference criterion.

A particular data mining algorithm is usually an instantiation of the model/preference/search components. The more common model functions in current data mining practice include:

1. Classification [41], [38], [42], [6], [39]: classifies a data item into one of several predefined categorical classes.
2. Regression [19], [12], [64], [45]: maps a data item to a real valued prediction variable.
3. Clustering [61], [50], [47], [52], [29], [31], [62], and [21]: maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models.
4. Rule generation [60], [35], [40], [43], [23], [55], [53], [67]: extracts classification rules from the data.
5. Discovering association rules [2], [63], [5], and [34]: describes association relationship among different attributes.
6. Summarization [32], [65], [25], [20]: provides a compact description for a subset of data.
7. Dependency modeling [22], [7]: describes significant dependencies among variables.

8. Sequence analysis [10], [33]: models sequential patterns, like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time.

Though, there are lots of techniques available in the data mining, few methodologies such as Artificial Neural Networks, K nearest neighbor, K means approach, are popular currently depends on the nature of the data.

Artificial Neural Network: Artificial Neural Networks (ANN) is systems inspired by the research on human brain (Hammerstrom, 1993). Artificial Neural Networks (ANN) networks in which each node represents a neuron and each link represents the way two neurons interact. Each neuron performs very simple tasks, while the network representing of the work of all its neurons is able to perform the more complex task. A neural network is an interconnected set of input/output units where each connection has a weight associated with it. The network learns by fine tuning the weights so as able to predict the call label of input samples during testing phase. Artificial neural network is a new techniques used in flood forecast. The advantage of ANN approach in modeling the rain fall and run off relationship over the conventional techniques flood forecast. Neural network has several advantages over conventional method in computing. Any problem having more time for getting solution, ANN is highly suitable states that the neural network method successfully predicts the pest attack incidences for one week in advance.

Pedotransfer functions (PTFs) provide an alternative by estimating soil parameters from more readily available soil data. The two common methods used to develop PTFs are multiple-linear regression method and ANN. Multiple linear regression and neural network model (feed-forward back propagation network) were employed to develop a pedotransfer function for predicting soil parameters using easily measurable characteristics of clay, sand, silt, SP, BD and organic carbon[51].

Artificial Neural Networks have been successful in the classification of other soil properties, such as dry land salinity (Spencer et al. 2004). Due to their ability to solve complex or noisy problems, Artificial Neural Networks are considered to be a suitable tool for a difficult problem such as the estimation of organic carbon in soil.

Support Vector Machines: Support Vector Machines (SVM) is binary classifiers (Burgos, 1998; Cortes and Vapnik, 1995). SVM is able to classify data samples in two disjoint classes. The basic idea behind is classifying the sample data into linearly separable. Support Vector Machines (SVM) are a set of related supervised learning methods used for classification and regression. In simple words given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

SVM is used to assess the spatiotemporal characteristics of the soil moisture products [4].

Decision trees: The decision tree is one of the popular classification algorithms in current use in Data Mining and Machine Learning. Decision tree is a new field of machine learning which is involving the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees and discrimination nets or production rules. Application of data mining techniques on drought related data for drought risk management shows the success on Advanced Geospatial Decision Support System (GDSS). Leisa J Armstrong states that data mining approach is one of the approaches used for crop decision making.

Research has been conducted in Australia to estimate a range of soil properties, including organic carbon (Henderson et al. 2001). The nation-wide database had 11,483 soil points available to predict organic carbon in the soil. An enhanced decision trees tool (Cubist), catering for continuous outputs was used for this study. A correlation of up to 0.64 was obtained between the predicted and actual organic carbon levels.

K nearest neighbor: K nearest neighbor techniques is one of the classification techniques in data mining. It does not have any learning phase because it uses the training set every time a classification performed. Nearest Neighbor search (NN) also known as proximity search, similarity search or closest point search is an optimization problem for finding closest points in metric spaces.

K nearest neighbor is applied for simulating daily precipitation and other weather variables (Rajigopalan and Lall, 1999).

Bayesian networks: A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over fitting of data Development of a data mining application for agriculture based on Bayesian networks were studied by Huang et al. (2008). According to him, Bayesian network is a powerful tool for dealing uncertainties and widely used in agriculture datasets. He developed the model for agriculture application based on the Bayesian network learning method. The results indicate that Bayesian Networks are a feasible and efficient.
Bayesian approach improves hydrogeological site characterization even when using low-resolution resistivity surveys [52].

**K means approach:** K means method is one of the most used clustering techniques in the data mining. The idea behind the K means algorithms is very simple that certain partition of the data in K clusters, the centers of the cluster can be computed as the mean of the all sample belonging to a cluster. The center of the cluster can be considered as the representative of the cluster. The center is quite close to all samples in the cluster.

K Means approach was used to classify the soil and plants (Camps-Valls et al., 2003).

**Fuzzy logic:** Fuzzy logic is a form of multi valued logic derived from Fuzzy set theory to deal with reasoning that is approximate rather than accurate. In contrast with "crisp logic", where binary sets have binary logic, fuzzy logic variables may have a truth value that ranges between 0 and 1 and is not constrained to the two truth values of classic propositional logic [46]. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions. Fuzzy logic emerged as a consequence of the 1965 proposal of Fuzzy set theory by Lotfi zadeh [1] [66]. Though fuzzy logic has been applied to many fields, from control theory to artificial intelligence, it still remains controversial among most statisticians, who prefer Bayesian logic, and some control engineers, who prefer traditional two-valued logic.

Fuzzy logic is used to the prediction of soil erosion in a large watershed (B.Mitra et al., ScienceDirect, Nov.1998).

**Genetic Algorithm:** The Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of Evolutionary Algorithm (EA), which generates solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover.

Soil liquefaction is a type of ground failure related to earthquakes. It takes place when the effective stress within soil reaches zero as a result of an increase in pore water pressure during earthquake vibration (Youd, 1992). Soil liquefaction can cause major damage to buildings, roads, bridges, dams and lifeline systems, like the earthquakes.

Genetic Algorithm approach is used for assessing the liquefaction potential of sandy soils (G. Sen et al. Nat. Hazards Earth Syst. Sci., 2010).

**Ant Colony Optimization:** The Ant Colony Optimization (ACO) algorithm is probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of ant colony algorithms family, in swarm intelligence methods, and it constitutes some Meta heuristic optimizations. Initially proposed by Marco Dorigo in 1992 in his Ph.D. thesis [13] [17], the first algorithm was aiming to search for an optimal path in a graph, based on the behavior of ants seeking a path between their colony and a source of food. The original idea has since diversified to solve a wider class of numerical problems, and as a result, several problems have emerged, drawing on various aspects of the behavior of ants.

Ant Colony Optimization is applied for estimating unsaturated soil hydraulic parameters (K.C.Abbaspour et al., ELSEVIER, 2001).

**Particle Swarm Optimization:** Particle Swarm Optimization (PSO) is a method for performing numerical optimization without explicit knowledge of the gradient of the problem to be optimized. PSO is originally attributed to Kennedy, Eberhart, and Shi [28] [54] and was first intended for simulating social behavior. The algorithm was simplified and it was observed to be performing optimization. The book by Kennedy and Eberhart [27] describes many philosophical aspects of PSO and swarm intelligence. An extensive survey of PSO applications is made by Poli [48] [49].

Particle Swarm Optimization is used for analysis of Soil erosion characteristics (Li Yunkai et al, Springer, Sep.2009).

**Simulated Annealing:** Simulated Annealing (SA) is a generic probabilistic Meta heuristic for the global optimization problem of applied mathematics, namely locating a good approximation to the global optimum of a given function in a large search space. It is often used when the search space is discrete (e.g., all tours that visit a given set of cities). For certain problems, simulated annealing may be more effective than exhaustive enumeration provided that the goal is merely to find an acceptably good solution in a fixed amount of time, rather than the best possible solution. The method was independently described by Scott Kirkpatrick, C. Daniel Gelatt and Mario P. Vecchi in 1983 [30] and by Vlado Cerny in 1985 [9]. The method is an adaptation of the Metropolis Hastings algorithm, a Monte Carlo method to generate sample states of a thermodynamic system, invented by N. Metropolis et al. in 1953 [36].

Simulated Annealing is used for analyzing Soil Properties (R.M. Lark et al., ScienceDirect, March, 2003).

### III. RESULTS AND DISCUSSION

The purpose of the study is to examine the most effective techniques to extract new knowledge and information from existing soil profile data contained within ISRIC-WISE soil data set. Several data mining techniques are in agriculture and allied area. Few of techniques are discussed here. K means method is used to forecast the pollution in the atmosphere (Jorquera et al., 2001). Different possible changes of weather are analyzed using SVM (Tripathi et al., 2006). K means approach is used for classifying soil in combination with GPS readings (Verheyen et al., 2001). Wine Fermentation process monitored using data mining techniques. Taste sensors are used to obtain data from the fermentation process to be classified using ANNs (Rud et al., 2004).

A brief survey of the related work in the area of soil mining is that the data involved here are high dimensional data and...
dimensionality reduction was addressed in classical methods such as Principal Component Analysis (PCA) [24]. There is a growing literature demonstrating the predictive capacity of the soil landscape paradigm using digital data and empirical numerical modeling techniques as specified by Christopher et al., [11]. The Eigen decomposition of empirical covariance matrix is performed and the data points are linearly projected. When the information relevant for classification is present in eigenvectors associated with small eigenvalues are removed, then this could lead to degradation in classification accuracy. Examples of spatial prediction have been provided, across a range of physiographical range of environment and spatial extents, for a number of soil properties by Gessler et al., [21] Tenenbaum et al.,[59] introduced the concept of Isomap, a global dimensionality reduction algorithm. The CCDR (classification constrained dimensionality reduction) algorithm [15] was only demonstrated for two classes and the performance was analyzed for simulated data. Bui et al., [8] demonstrated the potential for the discovery of knowledge embedded in survey of landscape model using rule induction techniques based on decision trees. It has the ability to mimic soil map using samples taken from it, and by implication it also captures the embedded knowledge. Related to agriculture, many countries are still facing a multitude of problems to maximize productivity [26]. Another concept of CCDR plots the classification error probability and its confidence interval using K nearest neighbor classifier [14]. Normally there is a decrease in error probability as dimension increases, and the optimal value is reached when dimension value varies between 12 - 14, which has been proved using entropic graph algorithm. However the food production has improved significantly during last two decades by providing it with good seeds, fertilizers, and pesticides and modern farming equipment [57]. The agriculture sector has seen a tremendous improvement.

IV. CONCLUSIONS

In this research survey, data mining and pattern recognition techniques for soil data mining studied. The survey aims to come out of the techniques being used in the agricultural soil science and its allied area.

The recommendations arising from this research survey are: A comparison of different data mining techniques could produce an efficient algorithm for soil classification for multiple classes. The benefits of a greater understanding of soils could improve productivity in farming, maintain biodiversity, reduce reliance on fertilizers and create a better integrated soil management system for both the private and public sectors.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions.

**Authors Biography**

D. Ashok Kumar did his Master degree in Mathematics and Computer Applications in 1995 and completed Ph.D., on Intelligent Partitional Clustering Algorithm’s in 2008, from Gandhigram Rural Institute–Deemed University, Gandhigram, Tamil Nadu, INDIA. He is currently working as Senior Grade Assistant Professor and Head in the Department of Computer Science, Government Arts College, Tiruchirappalli-620 022, Tamil Nadu, INDIA. His research interest includes Pattern Recognition and Data Mining by various soft computing approaches viz., Neural Networks, Genetic Algorithms, Fuzzy Logic, Rough set, etc.

N. Kannathasan is a Senior Grade Assistant Professor of Computer Science at the Mahatma Gandhi Government Arts College, Mahe, U.T. of Puducherry, INDIA. Prior to joining MGGAC, Mahe, he served at the Bharathidasan Government College for Women, Puducherry, Velammal College of Management and Computer Studies, Chennai, SRM Arts and Science College, Chennai, and AVC College, Mayiladuthurai. He received his M.Phil. Computer Science from the Bharathidasan University, Tiruchirappalli and M.Sc. Computer Science from Ayya Nadar Janaki Ammal College, Sivakasi.
Proceedings of the
International Conference on
MATHEMATICAL MODELING AND
APPLIED SOFT COMPUTING (MMASC 2012)

(Towards High Performance & Knowledge optimization)

11-13, JULY 2012

Volume – II

EDITORS
Dr. S.R.K. PRASAD
Dr. R. PRABHAKAR
Dr. V. SELLADURAI
Dr. K.S. RAMASWAMI

Coimbatore Institute of Technology
(Government Aided Autonomous Institution)
Coimbatore – 641 014, India
MMASC 2012 conference promoted and organized by Coimbatore Institute of Technology (Government Aided Autonomous Institution) Coimbatore – 641 014, India

Sponsored by

Department of Science and Technology
Ministry of Science and Technology
New Delhi

IEEE section
Chennai

Council of Scientific and Industrial Research
New Delhi

Department of Information Technology
Ministry of Communications and Information Technology, New Delhi

Indian National Science Academy
New Delhi

Catalyzed and supported by Tamilnadu State Council for Science and Technology, Government of Tamilnadu
“Proceedings of the international conference on
Mathematical Modeling and Applied Soft Computing”
Volume – II
1256 pages
Price: Rs 1640

Editors
S.R.K. Prasad
R. Prabhakar
V. Selladurai
K.S. Ramaswami
Coimbatore Institute of Technology
(Government Aided Autonomous Institution)
Coimbatore, India

Copyright © 2012, SHANGA VERLAG

SHANGA VERLAG

174, Nethaji Road, Pappanaicekenpalayam,
Coimbatore – 641 037.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or
transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise,
without prior written permission of the publisher.

All export rights for this book vest exclusively with
SHANGA VERLAG
Unauthorised export is a violation of terms of sale
and is subject to legal action.

Printed from camera-ready copy provided by Editors
in digital printer on 75 gsm maplitho paper

Published by GANESH SHANKAR D.,
on behalf of SHANGA VERLAG
and printed at SIVASALAPATHY PRINTERS,
174, Nethaji Road, Pappanaicekenpalayam,
Coimbatore – 641 037.
ANALYSIS OF LINEAR AND SEGMENTED LINEAR REGRESSION OF FRUIT YIELD ON SOIL SALINITY

D. Ashok Kumar
Government Arts College, Tiruchirapalli- 620 022, Tamil Nadu, India akudaiyar@yahoo.com
N. Kannathasan
Mahatma Gandhi Government Arts College, Mahe, P.O. New Mahe-673 311, U. T of Puducherry, India
nkathasan@gmail.com

ABSTRACT

This research paper aims to evaluate and compare the performance of the linear two-variable least square statistical data mining techniques Regression of y upon x and Segmented regression on the dataset of fruit yield and soil water salinity for bell peppers which is collected from the Arava Valley of Israel. Segmented Linear Regression gives a better performance of soil salinity on fruit yield than does the un-segmented regression technique Regression of y upon x. The segmented linear regression sum of the squares of the deviations ($\sum \hat{e}^2$) is lower than the un-segmented linear regression sum of the squares of the deviations.

Keywords
Segmented linear regression, Fruit yield, Soil salinity, Regression of y upon x

1. Introduction

The Data Mining contains many study areas such as machine learning, pattern recognition in data, databases, statistics, artificial intelligence, data acquisition for expert systems and data visualization. The most important goal here is to extract patterns from data and to bring use-
ful knowledge into an understandable form to the human observer. The entire process aims to obtain high-level data from low-level data [1].

This research paper aims to evaluate and compare the performance of the linear two-variable 'least square' statistical data mining technique 'Regression of y upon x' and 'Segmented regression' on the dataset of fruit yield and soil salinity for bell peppers, collected from the Arava Valley of Israel [3].

Regression of y upon x is made if y is casually influenced by x, or to predict the value of y from a given value of x. In this regression, the sums of the squared deviations of y to the regression line, i.e. in the y-direction, are minimized. The relation between y and x need not be linear. It can be curved. To detect a non-linear relation, it is common practice to transform the values of y and x [4].

Segmented regression is a method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is fit to each interval. Segmented regression is useful when the independent variables, clustered into different groups, exhibit different relationships between the variables in these regions. The boundaries between the segments are breakpoints. The breakpoint can be important in decision making. Segmented linear regression is segmented regression whereby the relations in the intervals are obtained by linear regression. A segmented regression analysis is based on the presence of a set of (y, x) data, in which y is the dependent variable and x is the independent variable [2].

2. Material and Methods

2.1 Linear Regression y upon x

The linear regression of y upon x is designed to detect a relation like the following [2]

\[ y = ax + b + \epsilon, \text{ or } \hat{y} = ax + b \]  

(1)

Where

- \(a\) = the linear regression coefficient, giving the slope of the regression line
- \(b\) = the regression constant, giving the intercept of the regression line on the y axis
- \(\epsilon\) = a random deviation of the y value from the regression line
- \(\hat{y}\) = the expected value of y according to the regression \((\hat{y} = y - \epsilon)\).

This regression is used when the \(\epsilon\) values are independent of the values of y and x. It is used to predict the value of y from a value of x, regardless of whether they have a causal relation.

A regression line always passes through the central of the data \((x_-, y)\). A straight line through \((x, y)\) can be represented by

\[(y-y) = a (x-x)\]  

(2)

Where ‘a’ is the tangent of the angle \(a\).

Normally, the data \((x, y)\) do not coincide with the line, so a correct representation of the regression is

\[(y-y) = a (x-x) + \epsilon\]  

(3)

Where \(\epsilon\) is a vertical distance of the point \((x, y)\) to the regression line. The sum of all the line
\( \epsilon \) values equal to zero. The difference \( y - \epsilon \) gives a \( y \) value on the regression line, \( \hat{y} \). Substitution of \( \hat{y} = y - \epsilon \) in Eq. 3 gives

\[
(\hat{y} - y) = a(x - x)
\]

(4)

Where 'a' is called the regression coefficient of \( y \) upon \( x \).

This equation can also be written as

\[
\hat{y} = ax + y - ax
\]

(5)

By substituting \( b = y - ax \), gives Eq. 1.

To determine the best possible regression coefficient, one must minimize the \( \Sigma \epsilon^2 \) (the least squares method). In other words the choice of the slope of the line and the intercept must fit the points as well as possible. To meet this condition we must take

\[
a = \frac{\Sigma yx}{\Sigma x^2}
\]

(6)

Where,

\[
\Sigma yx = \Sigma (y - y) (x - x)
\]

(7)

\[
\Sigma x^2 = \Sigma (x - x)^2
\]

(8)

\[
\Sigma \epsilon^2 = \Sigma (y - y)^2
\]

(9)

in which the symbol \( \Sigma \) means 'reduced sum'.

Table 1. Fruit yield in the case of bell pepper in the Arava Valley of Israel [3]

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Soil Salinity EC (dS/m)</th>
<th>Number of Plants</th>
<th>Fruit Total yield</th>
<th>Average yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>19</td>
<td>1392</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>27</td>
<td>3776</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>28</td>
<td>3406</td>
<td>122</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>26</td>
<td>3571</td>
<td>137</td>
</tr>
<tr>
<td>5</td>
<td>0.5</td>
<td>30</td>
<td>4050</td>
<td>135</td>
</tr>
<tr>
<td>6</td>
<td>0.5</td>
<td>21</td>
<td>3589</td>
<td>171</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>22</td>
<td>3522</td>
<td>160</td>
</tr>
<tr>
<td>8</td>
<td>1.5</td>
<td>19</td>
<td>3672</td>
<td>193</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>24</td>
<td>3683</td>
<td>153</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>10</td>
<td>1207</td>
<td>121</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>21</td>
<td>2601</td>
<td>124</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>17</td>
<td>2686</td>
<td>158</td>
</tr>
</tbody>
</table>
The coefficient 'a' can be directly calculated from the (x, y) pairs of data. If 'a' is positive, the regression line slopes upward, and an increase in x causes an increase in y, and vice versa. If 'a' is negative, the regression line slopes downward. If the regression coefficient 'a' is zero, then there is no linear relation between y and x, and the line is horizontal.

In linear regression, the coefficient R equals the absolute value of the correlation coefficient r. In addition, \( r^2 \)\( \Sigma y^2 \) equals the linearly explained variation and

\( (1 - r^2)\Sigma y^2 \) is the residual variation, \( \Sigma e^2 \). The value of r can be calculated from

\[
r = \sqrt{\frac{(\Sigma x^2)(\Sigma y^2)}{\Sigma xy}}
\]

The correlation coefficient r can assume values of between -1 and +1. If r > 0, the coefficient 'a' is also positive. If r = 1 there is a perfect match of the regression line with the (x, y) data. If r < 0, the coefficient 'a' is also negative, and if r = -1, there is also a perfect match, although y increases as x decreases and vice versa. If r = 0, the coefficients 'a' is also zero, the regression line is parallel to the x-axis, i.e., horizontal, and the y variable has no linear relation with x.

Linear regression y upon x calculations for (y, x) data with y=AVERAGE fruit yield and x= Soil salinity EC\(_{\infty}\) (dS/m) of Table 1

\[
\begin{align*}
\Sigma x &= 78.50 & \Sigma x^2 &= 390.25 & \Sigma xy &= 9260.50 \\
\Sigma y &= 3093 & \Sigma y^2 &= 421763.00 & n &= 24 \\
x &= \frac{\Sigma x}{n} = 78.50/24 = 3.27 \\
y &= \frac{\Sigma y}{n} = 3093/24 = 128.88 \\
Equation 8: & \Sigma x^2 = 133.49 \\
Equation 9: & \Sigma y^2 = 23152.64 \\
Equation 7: & \Sigma xy = -856.19 \\
Equation 6: & a = -856.19/133.49 = -6.41 \\
Equation 5: & b = 149.84 \\
So, & \hat{y} = -6.41x + 149.84 \\
Equation 10: & r = -0.487 \\
\Sigma e^2 = (1-r^2)\Sigma y^2 = (1-0.237)*23152.64 \\
& = 17,665.46
\end{align*}
\]

278
2.2 Segmented Linear Regression

In agriculture, crops will often react to a production factor \( x \) within a certain range of \( x \), but not outside this range. Segmented linear regression applies linear regression to \( (x, y) \) data that do not have a linear relation. It introduces one or more breakpoints, whereupon separate linear regression is made for the resulting segments. Thus, the non-linear relation is approximated by linear segments [2].

Table 3  Segmented linear regression calculations for \((y, x)\) data with \(y=\text{Average fruit yield and } x=\text{Soil salinity EC}_{\text{a}} \text{(dS/m)} \) of Table 1

1) Segment with \( x < 3.5 \)
\[
\begin{align*}
\Sigma x &= 25.50 \quad \Sigma x^2 = 62.75 \\
\Sigma y &= 2131 \quad \Sigma y^2 = 313041 \\
\text{n} &= 15 \\
x &= \Sigma x / n = 25.50 / 15 = 1.70 \\
y &= \Sigma y / n = 2131 / 15 = 142.07 \\
\text{Equation 8: } \Sigma x^2 &= 19.40 \\
\text{Equation 9: } \Sigma y^2 &= 10296.93 \\
\text{Equation 7: } \Sigma xy &= 62.80 \\
\text{Equation 6: } a &= \Sigma xy / \Sigma x = 62.80 / 19.40 = 3.24 \\
\text{Equation 5: } b &= 136.56 \\
\text{So, } \hat{y} &= 3.24x + 136.56 \\
\text{Equation 10: } r &= 0.14 \\
\text{\( \Sigma e^2 \) } &= (1-r^2) \Sigma y^2 = (1-0.0196) \times 10296.93 = 10095.11
\end{align*}
\]

2) Segment with \( x > 3.5 \)
\[
\begin{align*}
\Sigma x &= 58 \quad \Sigma x^2 = 327.50 \\
\Sigma y &= 962 \quad \Sigma y^2 = 108722 \\
\text{n} &= 9 \\
x &= \Sigma x / n = 58 / 9 = 6.49 \\
y &= \Sigma y / n = 962 / 9 = 106.89 \\
\text{Equation 8: } \Sigma x^2 &= 15.39 \\
\text{Equation 9: } \Sigma y^2 &= 5894.89 \\
\text{Equation 7: } \Sigma xy &= -90.11 \\
\text{Equation 6: } a &= -90.11 / 15.39 = -5.86 \\
\text{Equation 5: } b &= 141.41 \\
\text{So, } \hat{y} &= -5.86x + 141.41 \\
\text{Equation 10: } r &= -0.3 \\
\text{\( \Sigma e^2 \) } &= (1-r^2) \Sigma y^2 = (1-0.09) \times 5894.89 = 5364.3
\end{align*}
\]

3. Results and Discussion

Table 4  The results of the linear regression and segmented linear regression of Table 1

<table>
<thead>
<tr>
<th>Statistical Data Mining Techniques</th>
<th>( \hat{y} )</th>
<th>( R^2 )</th>
<th>( \Sigma e^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression ( y ) upon ( x )</td>
<td>-6.41x + 149.84</td>
<td>0.237</td>
<td>17,665.46</td>
</tr>
</tbody>
</table>
The total $\sum y^2 = 10,095.11 + 5,364.3 = 15,459.41$ in Table 3 is lower than the $\sum y^2 = 17,665.46$ of Table 2, which represents the linear regression using all the data without a breakpoint. This means that the segmented regression gives a better analysis of the soil salinity on fruit yield than does the un-segmented linear regression upon $x$.

![Figure 1](scatter_diagram.png)

*Figure 1* Scatter diagram with regression for $(y, x)$ data with $y$=Average fruit yield and $x$=Soil salinity $\text{EC}_\sigma$ (dS/m) of Table 1

![Figure 2](scatter_diagram_segmented.png)

*Figure 2* Scatter diagram with segmented regression $x < 3.5 \text{ EC}_\sigma$ (dS/m) for $(y, x)$ data of Figure 1.
Segmented linearization will be illustrated with (y, x) data with y=Average fruit yield and x=Soil salinity ECₖ (dS/m) from Figure 1. The optimum breakpoint was at x=3.5.

Figure 4 clearly shows that the fruit yield is independent of soil salinity up to a salinity value of EC = 3.5 dS/m (the break-point, Bp, or threshold). Here, the graph shows a horizontal line, and the regression coefficient is statistically insignificant. Beyond the breakpoint (Bp=3.5dS/m), the relation has a statistically significant downward sloping trend.

4. Conclusion

The relatively sum of the squares of the deviations (Σₑ) of the fruit yield parameter against soil salinity tested leads to the conclusion that the salinity effects on production can be accurately assessed.

It is also concluded that the segmented linear regression with breakpoint is an effective method for the assessment of salinity effects than does the un-segmented linear regression.
Acknowledgment

References


4) Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann, 2011.

Authors Biography

and completed Ph.D., on Intelligent Partitional Clustering Algorithm’s in 2008, from Gandhi gram Rural Institute–Deemed University, Gandhi gram, Tamil Nadu, INDIA. He is currently working as Senior Grade Assistant Professor and Head in the Department of Computer Science, Government Arts College, Tiruchirappalli–620 022, Tamil Nadu, INDIA. His research interest includes Pattern Recognition and Data Mining by various soft computing approaches viz., Neural Networks, Genetic Algorithms, Fuzzy Logic, Rough set, etc.,

N. Kannathasan is a Senior Grade Assistant Professor of Computer Science at the Mahatma Gandhi Government Arts College, Mahe, U.T. of Puducherry, INDIA. Prior to joining MGGAC, Mahe, he served at the Bharathidasan Government College for Women, Puducherry, Velammal College of Management and Computer Studies, Chennai, SRM Arts and Science College, Chennai, and AVC College, Mayiladuthurai. He received his M.Phil. Computer Science from the Bharathidasan University, Tiruchirappalli and M.Sc. Computer Science from Ayya Nadar Janaki Ammal College, Sivakasi.
A Study and Characterization of Chemical Properties of Soil Surface Data Using K-means Algorithm

D. Ashok Kumar
Department of Computer Science
Government Arts College
Tiruchirappalli, India-620 022.
akudaiyar@yahoo.com

N Kannathasan
Department of Computer Science
Mahatma Gandhi Government Arts College
Puducherry, India-673 311.
kathan@gmail.com

Abstract— Soil is a vital natural resource on whose proper use depends the life supporting system of a country and the socioeconomic development of its people. Clustering in agricultural soil datasets is a relatively novel research field. This research paper aims to study the Characterization of Chemical Properties of Soil Surface Data of Bhanapur Micro watershed of Koppal District, Karnataka using K-means algorithm. This work computed average silhouette width which provides an evaluation of clustering validity, and might be used to select an appropriate number of clusters in the soil dataset. Soil dataset clustered by using K-means two clustering with Euclidean distance which provides average silhouette value is 0.7736. And this work also proved high intra-class similarity: cohesive within clusters and low inter-class similarity: distinctive between clusters in the soil dataset by K-means which reassign points among clusters to decrease the sum of point-to-centroid distances, and then recomputed cluster centroids for the new cluster assignments. K-means clustering with Euclidean distance total sum of distance is 1.14402 based on the number of reassign soil data. By default, K-means begins the clustering process using a randomly selected set of initial centroid locations. K-means repeats the clustering process starting from different randomly selected centroids. The sum of distances within each cluster for that best solution is 1.1440 with Euclidean distance. K-means three clustering solution with Euclidean distance which provides average silhouette value is 0.6052. K-means three clustering solution with Cosine distance average silhouette value is 0.6219. Hierarchical Clustering dendrogram with Euclidean distance measure is 0.7935. Hierarchical Clustering dendrogram with Cosine distance measure is 0.6678. The results from hierarchical clustering dendrogram with Cosine distance are qualitatively similar to results from K-Means using three clusters. K-means cluster analysis with Euclidean and Cosine distance measures compared with Hierarchical clustering dendrogram. Based on the study and characterization of chemical properties of soil surface data using K-means, Cosine might be a good choice of distance measure.

Keywords— K-means; Centroid Based; Cluster Analysis; Soil Surface Data; Euclidean Distance; Cosine Distance

I. INTRODUCTION

Soil provides food, fodder and fuel for meeting the basic needs of human beings and animals. With the growth in human and animal population, demand for more food production is on the increase. This demands systematic appraisal of our soil resources with respect to their extent, distribution, characteristic, behavior and use potential, which is very important for developing an effective land use system for augmenting agricultural production on sustainable basis.

A comparison of different data mining techniques could produce an efficient algorithm for soil classification for multiple classes. The benefits of a greater understanding of soils could improve productivity in farming, maintain biodiversity, reduce reliance on fertilizers and create a better integrated soil management system for both the private and public sectors [3], [4].

A. Soil Surface Data

This research paper aims to study the Characterization of Chemical Properties of Soil Surface Data of Bhanapur Micro watershed of Koppal District, Karnataka using by K-means algorithm. The selected area of Bhanapur micro-watershed covers an area of 581.5 ha. It comprises of both red (75%) and black (25%) soils differing in their physical, chemical and agronomic characteristics [20]. Their characterization and classification is crucial for productive and sustainable management of soils. Knowledge about chemical properties of soils helps in better crop planning. Used on chemical properties of soil surface data, it will find the natural groupings among soil surface data, based on their pH, EC (ds/m) and OC (%).

B. K-means Algorithm

Clusters are simple and compact representation of a dataset and are useful in different applications. K-means is a standard and landmark algorithm for clustering data [5]. A good clustering method will produce high quality clusters with high intra-class similarity: cohesive within clusters and low inter-class similarity: distinctive between clusters. The quality of a clustering method depends on the similarity measure used by the method, its implementation, and its ability to discover some or all of the hidden patterns. Clusters found by K-means algorithm have convex shapes and each cluster is represented by a center.

K-means clustering is a partitioning method that treats observations in the data as objects having locations and distances from each other. It partitions the objects into k mutually exclusive clusters, such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster is
characterized by its centroid, or center point. Of course, the distances used in clustering often do not represent spatial distances.

The Conventional K-means algorithm described in Table I, one of the most used clustering algorithms, was first described by Macqueen [14]. It was designed to cluster numerical data in which each cluster has a center called the mean. The K-means algorithm is classified as a partitional or nonhierarchical clustering method [9]. In this algorithm, the number of clusters k is assumed to be fixed. There is an error function in this algorithm. It proceeds, for a given initial k clusters, by allocating the remaining data to the nearest clusters and then repeatedly changing the membership of the clusters according to the error function does not change significantly or the membership of the clusters no longer changes [8].

K-means algorithm can be divided into two phases (Fig. 1): the initialization phase and the iteration phase. In the initialization phase, the algorithm randomly assigns the cases into k clusters.

<table>
<thead>
<tr>
<th>TABLE I. The Pseudo Code of K-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong> Data set D, Number of Clusters k, Dimensions d:</td>
</tr>
<tr>
<td>{C_i is the i^{th} cluster}</td>
</tr>
<tr>
<td>[1.] Initialization Phase</td>
</tr>
<tr>
<td>1: ((C_1, C_2, \ldots, C_k) = \text{Initial partition of D.})</td>
</tr>
<tr>
<td>[2.] Iteration Phase</td>
</tr>
<tr>
<td>2: \textbf{repeat}</td>
</tr>
<tr>
<td>3: (d_{ij} = \text{distance between case } i \text{ and cluster } j;)</td>
</tr>
<tr>
<td>4: (n_i = \arg \min_{1 \leq j \leq k} d_{ij};)</td>
</tr>
<tr>
<td>5: Assign the case i to cluster (n_i;)</td>
</tr>
<tr>
<td>6: Recompute the cluster means of any changed clusters above;</td>
</tr>
<tr>
<td>7: \textbf{until} no further changes of cluster membership occur in a complete iteration</td>
</tr>
<tr>
<td>8: Output results.</td>
</tr>
</tbody>
</table>

Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized [10].

\[
E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2 \tag{1}
\]

where \(c_i\) is the centroid or medoid of cluster \(C_i\)

Centroid: the “middle” of a cluster [10]

\[
C_m = \frac{\sum_{i=1}^{N}(c_i)}{N} \tag{2}
\]

Radius: square root of average distance from any point of the cluster to its centroid [9]

\[
R_m = \sqrt{\frac{\sum_{i=1}^{N}(p - c_m)^2}{N}} \tag{3}
\]

Diameter: square root of average means squared distance between all pairs of points in the cluster [10]

\[
D_m = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{N}(p_i - p_j)^2}{N(N-1)}} \tag{4}
\]

As one of the most often used clustering algorithms, the K-means algorithm has some important properties:
- It is efficient in clustering large data sets, since its computational complexity is linearly proportional to the size of the data sets.
- It often terminates at a local optimum [1], [17].
- The clusters have convex shapes, such as ball in three-dimensional space [1].
- It works on numerical data.
- The performance is dependent on the initialization of the centers.

In the Iteration phase, the algorithm computes the distance between each case and each cluster and assigns the case to the nearest cluster.
Most of the variants of the k-means (Fig. 2) which differ in
- Selection of the initial $k$ means
- Dissimilarity calculations
- Strategies to calculate cluster means

Strength of the K-Means algorithm is relatively efficient $O(tkn)$, where $n$ is number of objects, $k$ is number of clusters, and $t$ is number of iterations. Normally, $k, t << n$.

The K-means algorithm has some drawbacks [15]. In particular, the performance is dependent on the initialization of the centers, as mentioned above. As a result, some methods for selecting good initial centers are proposed, for example, in [6] and [7]. Penna et al. [15] provide a comparison of four initialization methods: a random method, Forgy’s approach [1], Macqueen’s approach [14], and Kaufman’s approach [11]. Other initialization methods are presented in [13].

In the iteration phase of the algorithm, the objects will be moved from one cluster to another in order to minimize the objective function. Tarantino [18] presents a computational study of the shortcomings and relative merits of 17 reallocation methods for the K-means algorithm. Another drawback of the K-means algorithm is that it does not work effectively on high-dimensional data [12]. Also, working only on numerical data restricts some applications of the K-means algorithm.

II. CLUSTER ANALYSIS USING K-MEANS

Data clustering also known as cluster analysis is to discover the natural grouping(s) of a set of patterns, points, or objects [2]. Clustering is a technique, which is used to find groups of clusters that are somehow similar in characteristic from the given dataset for which the real structure is unknown. It is often confused with classification but there are some differences. In classification, the data are assigned to predefined classes or clusters, whereas in clustering the classes or clusters are defined and the only data available are unlabelled [19]. Cluster analysis can be defined as a wide variety of procedures that can be used to create a classification. These procedures empirically form “clusters” of groups of highly similar entities [1]. The main aim of cluster analysis is the finding similarities between chemical properties of soil data according to the characteristics found in the dataset and grouping similar data objects into clusters. Cluster analysis is a way to examine similarities and dissimilarities of observations or objects [10].

Similarity:
- Numerical measure of how alike two data objects are
- Value is higher when objects are more alike
- Often falls in the range $[0,1]$

Dissimilarity (e.g., distance):
- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

In this paper, MATLAB was used to study and characterize of chemical properties of soil surface data using K-means with Euclidean and Cosine distance measures and also this experimental study was compared with Hierarchical Clustering dendrogram.

A. Silhouette plot

Number of cluster specification is required in K-means clustering. Each cluster is represented by a so-called silhouette, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. The entire clustering is displayed by combining the silhouettes into a single plot, allowing an appreciation of the relative quality of the clusters and an overview of the data configuration. The average silhouette width provides an evaluation of clustering validity, and might be used to select an ‘appropriate’ number of clusters [16].

B. Two Clustering Solution ($k=2$) using Euclidean distance

The silhouette plot of a clustering with $k=2$ of the 75 numbers of chemical properties of red soil data (Fig. 3) displays a measure of how close each point in one cluster is to points in the neighboring clusters.

![Figure 3. Silhouette plot with (k=2) using Euclidean distance with average silhouette value 0.7736](image-url)
From the silhouette plot, most points in both clusters have a large silhouette value, greater than 0.8, indicating that those points are well-separated from neighboring clusters. However, each cluster also contains a few points with low silhouette values, indicating that they are nearby to points from other clusters.

In Fig. 4, the centroids of each cluster are plotted using circled X’s. Three of the points from the lower cluster, plotted with triangles, are very close to points from the upper cluster, plotted with squares. But, in fact, because the upper cluster is so spread out, those three points are closer to the centroid of the lower cluster than to that of the upper cluster, even though they are separated from the bulk of the points in their own cluster by a gap. Because K-means clustering only considers distances, and not density (local cluster criterion).

### C. Dissimilarity calculations

Similarity is expressed in terms of a distance function, typically metric: \( d(i, j) \). The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables. Distances are normally used to measure the similarity or dissimilarity between two data objects.

\[
d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2}
\]

Properties:

- \( d(i, j) \geq 0 \)
- \( d(i, i) = 0 \)
- \( d(i, j) = d(j, i) \)
- \( d(i, j) \leq d(i, k) + d(k, j) \)

At each iteration, the K-means algorithm with Euclidean distance reassigns points among clusters to decrease the sum of point-to-centroid distances, and then recomputed cluster centroids for the new cluster assignments (Table II).

#### D. Selection of the initial k means

By default, K-means begins the clustering process using a randomly selected set of initial centroid locations. Just as in many other types of numerical minimizations, the solution that K-means reaches sometimes depends on the starting points, and it is possible for it to reach a local minimum, where reassigning any one point to a new cluster would increase the total sum of point-to-centroid distances, but where a better solution does exist. K-means repeats the clustering process starting from different randomly selected centroids for each replicate (Table III).

### Table II. K-Means Clustering with Euclidean Distance

<table>
<thead>
<tr>
<th>( t ) (No. of Iterations)</th>
<th>( n ) (No. of reassign soil data)</th>
<th>Sum of distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>1.26091</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1.20517</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.17463</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.16694</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1.15289</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.14402</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1.14402</td>
</tr>
</tbody>
</table>

7 iterations, total sum of distances = 1.14402

### Table III. Sum of Distances for Different Randomly Selected Centroids

<table>
<thead>
<tr>
<th>( t ) (No. of Iterations)</th>
<th>Sum of distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.2406</td>
</tr>
<tr>
<td>5</td>
<td>1.14402</td>
</tr>
<tr>
<td>7</td>
<td>1.14402</td>
</tr>
<tr>
<td>7</td>
<td>1.14402</td>
</tr>
<tr>
<td>5</td>
<td>1.14402</td>
</tr>
<tr>
<td>4</td>
<td>1.30141</td>
</tr>
<tr>
<td>9</td>
<td>1.14402</td>
</tr>
<tr>
<td>10</td>
<td>1.14402</td>
</tr>
<tr>
<td>5</td>
<td>1.14402</td>
</tr>
<tr>
<td>5</td>
<td>1.14402</td>
</tr>
</tbody>
</table>

The output shows that, even for this relatively simple problem, non-global minima do exist. Each of these replicates began from a different set of initial centroids. Depending on where it started from, K-means reached one of two different solutions. However, the final solution that K-means returns is the one with the lowest total sum of distances, over all replicates. The sum of distances within each cluster for that best solution is 1.1440.
E. Three Clustering Solution (k=3) using Euclidean distance

A silhouette plot (Fig. 5) for this three-cluster solution indicates that there is one cluster that is well-separated, but that the other two clusters are not very distinct.

Again, the plot (Fig. 6) shows that how K-means has assigned the points to clusters. K-means has split the upper cluster from the two-cluster solution in Fig. 4, and that those two clusters are very close to each other. Depending on what intend to do with these data after clustering them, this three-cluster solution may be more or less useful than the previous, two-cluster solution. The average silhouette value was larger for the two-cluster solution, indicating that it is a better answer purely from the point of view of creating distinct clusters.


\[ \cos(d_1, d_2) = (d_1 \cdot d_2) / ||d_1|| ||d_2||, \]

where, indicates vector dot product, ||d|: the length of vector d.

From the silhouette plot (Fig. 7), these clusters appear to be only slightly better separated than those found using squared Euclidean distance in Fig. 3 and Fig. 5.

Notice that the order of the clusters is different than in the previous silhouette plot. This is because K-means chooses initial cluster assignments at random.

By plotting the raw data, the differences in the cluster shapes created using the two different distances. The two solutions are similar, but the two upper clusters are elongated in the direction of the origin when using cosine distance.

This plot (Fig. 8) does not include the cluster centroids, because a centroid with respect to the cosine distance corresponds to a half-line from the origin in the space of the raw data. However, a parallel coordinate plot of the normalized data points to visualize the differences between cluster centroids.

\[ \text{Figure 5. Silhouette plot with (k=3) using Euclidean distance with average silhouette value 0.6052} \]

\[ \text{Figure 6. Three clustering solution (k=3) of the 75 numbers of chemical properties of red soil data} \]

\[ \text{Figure 7. Silhouette plot with (k=3) using Cosine distance with average silhouette value 0.6219} \]
III. CLUSTER ANALYSIS USING DENDROGRAM

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram. A clustering of the data objects is obtained by cutting the dendrogram at the desired level, and then each connected component forms a cluster.

K-Means clustering produced a single partition of the chemical properties of soil surface data. Dendrogram creating a hierarchical tree of clusters on the chemical properties of soil surface dataset.

Cluster tree was created using distances between observations in the dataset using Euclidean distance. The cophenetic correlation is one way to verify that the cluster tree is consistent with the original distances. Large values indicate that the tree fits the distances well, in the sense that pairwise linkages between observations correlate with their actual pairwise distances. This tree (Fig. 9) seems to be a fairly good fit to the distance 0.7935.

The root node in this tree is much higher than the remaining nodes, confirmed with the K-Means clustering: there are two large, distinct groups of observations. Within each of those two groups, lower levels of groups emerge as smaller and smaller scales in distance. There are many different levels of groups, of different sizes, and at different degrees of distinctness. Based on the results from K-Means clustering, cosine might also be a good choice of distance measure.

The resulting hierarchical tree is quite different; suggesting a very different way to look at group structure in the dataset is 0.6678.

The highest level of this tree separates chemical properties of soil surface dataset into two very distinct groups. The dendrogram (Fig. 10) shows that, with respect to cosine distance, the within-group differences are much smaller relative to the between-group differences than was the case for Euclidean distance. This is exactly expected for these data, since the cosine distance computes a zero pairwise distance for objects that are in the same "direction" from the origin.

With 75 numbers of chemical properties of soil surface data, the plot is cluttered, but simplified dendrogram that does not display the very lowest levels of the tree.

The three highest nodes in this tree (Fig. 11) separate out four groups (in number of soil surface data): 20 18 34 3.
For many purposes, the dendrogram might be a sufficient result. However, used the cluster function to cut the tree and explicitly partition observations into specific clusters, as with K-Means. Using the hierarchy from the cosine distance to create clusters, specify a linkage height that will cut the tree below the three highest nodes, and create four clusters, then plot the clustered raw data.

The results from hierarchical clustering with cosine distance are qualitatively similar to results from K-Means, using three clusters. However, creating a hierarchical cluster tree allows visualizing, all at once, what would require considerable experimentation with different values for k in K-Means clustering.

Hierarchical clustering also allows to experiment with different linkages (Fig. 12). Clustering the chemical properties of soil surface dataset with single linkage, which tends to link together objects over larger distances than average distance does, gives a very different interpretation of the structure in the chemical properties of soil surface dataset.

IV. CONCLUSIONS

This work studied the characterization of chemical properties of soil surface data using K-means algorithm. Cluster analysis using K-means found similarities between chemical properties of soil data according to the characteristics found in the dataset and grouping similar data objects into clusters. And also K-means cluster analysis using Euclidean and Cosine distance measures compared with Hierarchical clustering Dendrogram. Based on the study and characterization of chemical properties of soil surface data using K-means, Cosine might be a good choice of distance measure.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions.

REFERENCES


[10] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann, 2011.


A Novel Artificial Neural Network Model for Prediction of Soil Total Porosity

Ashok Kumar D*, Kannathasan N#
*Department of Computer Science, Government Arts College, Tiruchirapalli-620 022, Tamil Nadu, India
akudaiyar@yahoo.com
#Department of Computer Science, Kanchi Mamunivar Centre for Post Graduate Studies, Puducherry-605008, India
nkthasan@gmail.com

Abstract: Background/Objectives: Many mapping of contaminated soil sites and the resulting cleanup are time consuming and expensive tasks, requiring extensive amounts of geology, hydrology, chemistry, GIS in environmental contamination and data mining and pattern recognition techniques. Artificial Neural Networks (ANN) has been inspired by biological neural networks, and are popular tools in the application of classification, prediction and recognition based problems. Methods/Statistical Analysis: This work proposed a novel Artificial Neural Network model with a Feed-forward Back-Propagation has been designed to predict soil Total Porosity. Soil samples were collected from International Soil Reference and Information Centre. The network structure used the default Levenberg–Marquardt algorithm (TRAINLM) for training, gradient descent with momentum weight/bias learning function (LEARNGDM) for learning and Mean Square Error (MSE) function for Performance estimation. Findings: The ANN model outputs for training, validation and testing in order to determine the relationship between the outputs of the network and the targets. After the network was trained, validated and tested, the generated model can be used to predict the parameter Total Porosity through new input pH. It was found that experimental analysis of Total Porosity was close to the predicted data calculated from the configuration and this confirms the developed ANN model is suitable for predicting the salinity concentrations. Applications/Improvements: Development of novel methods for classification in data mining and pattern recognition like genetic based, and hybrid based is required to find out the optimal number of clusters for soil sites based on soil texture with Total Porosity.

Keywords- Artificial Neural Network, Prediction, Total Porosity, Feed-forward Back-Propagation, Levenberg-Marquardt.

1. Introduction
An important factor influencing the productivity of the various ecosystems of the earth is the nature of their soils. Soils are vital for the existence of many forms of life that have evolved on our planet. Few feet of land from the surface is referred to as soil and it acts as a natural filter to sieve many substances that mix with the water.

But water is the medium that transports some contaminants into the groundwater from the soil surface (Fig.1).
color and usually light in texture and porous. The A horizon is commonly differentiated into a darker upper horizon or organic accumulation, and a lower horizon showing loss of material by eluviations. The B horizon is a third layer consists of mineral soil layer which is strongly influenced by illuviation. Consequently, this layer receives material eluviated from the A horizon. Due to enrichment of clay particles the B horizon also has a higher bulk density than the A horizon. The B horizon may be colored by iron oxides and aluminum oxides or by calcium carbonate illuviated from the A horizon. The fourth layer, the C horizon is composed of weathered parent material. The texture of this material can be quite variable with particles ranging in size from clay to boulders. The C horizon is also not significantly influenced by the pedogenic processes, translocation, and/or organic modifications. The lower most layers in a typical soil profile is called the R horizon. This soil layer simply consists of unweather bedrock (Fig.2).

The soil properties such as texture, porosity, specific yield depend on the total volume of groundwater recharge, water storage and discharge, also the extent of groundwater contamination1.

Soil is a mixture of three soil parts: sand, silt, clay. Classification of these parts is based on grain size. The following Table.1 shows the soil part and corresponding diameter of the parts. The relative proportion of soil parts in a particular soil determines its soil texture.

<table>
<thead>
<tr>
<th>Name of separate</th>
<th>Diameter range (millimeters)</th>
<th>Porosity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>2.0 – 0.05</td>
<td>25-50</td>
</tr>
<tr>
<td>Silt</td>
<td>0.05 – 0.002</td>
<td>35-50</td>
</tr>
<tr>
<td>Clay</td>
<td>Less than 0.002</td>
<td>33-60</td>
</tr>
</tbody>
</table>

The ratio of different-sized mineral particles in soil is the basis for soil texture. Clay, silt, or sand particles are considered as mineral particles. Clay particles hold water molecules due to electric force on them and make it suitable for growing crops. Clay rich soil is resistant to erosion. Silt particles too are small enough to hold water in the soil but the silt which is exposed washes away easily, depleting nutrients in the soil. Water passes through sandy soil very quickly and does not hold enough moisture to support plant growth. Exposed sandy soils are much subjected to erosion if the angle of the land, i.e. slope, is high. The soil textural triangle gives the textural name, which based on the percentage of sand, silt, and clay within the soil sample (Fig.3). The triangle is divided into 10-percent portions of clay, silt and sand. The summation of the three percentages must total 100 percent5.

Soil system is composed of air, water, dead organic matter, and various types of living organisms (Fig.4). Various factors such as organisms living in that soil, climate, topography, parent material and time influence the formation of soil. The sum of mineral particles alone does not constitute a true soil but it is influenced, modified, and supplemented by organisms living in that soil.

Humus is the biochemical substance that makes the upper layers of the soil usually looks dark, colored dark brown to black. Organic activity is usually profuse in the upper layers of the soil. For instance, one cubic centimeter of soil can be the home of more than a million bacteria.

The Porosity of a soil is decided by shape and spatial arrangement of soil particles. It is the air space or void space between soil particles. Infiltration, ground water movement, and storage occur in these void spaces. Porosity of soil typically decreases as pH value increases and also particle size increases because of soil contamination or soil pollution (Fig.5).

Bulk density is an important indicator of soil compaction and can be calculated as the dry weight of soil divided by its volume and is expressed in g/cm³. This volume is the sum of volume of soil particles and the volume of pores among soil particles. Structural support, water and solute
movement, and soil aeration are dependent factors of bulk density. Bulk densities above threshold level indicate unsupportive for plant growth (Table 2). This parameter is also used to convert between volume and weight of soil; also it is used to express physical, chemical and biological measurements of soil on a volumetric basis for soil quality assessment. This increases the validity of comparisons by removing error associated with differences in soil density at time of sampling.

The increased porosity of soil is due to mixing and movement of air and water from the soil top to lower layer (Fig. 6) where roots of plants occupying (rhizosphere). Increased volume of air and water available to roots has a positive effect on plant productivity. Earthworms and larval and adult insect stages also play a crucial role to produce most of the humus found in a soil through the incomplete digestion of organic matter. Movement of water to lower layers of the soil causes both mechanical and chemical translocations of organic matter.

Table 2. General Relationship of soil bulk density to root growth based on soil textural

<table>
<thead>
<tr>
<th>Type of Mineral Particle</th>
<th>Ideal bulk densities for Plant Growth (g/cm³)</th>
<th>Bulk densities that restrict root growth (g/cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>&lt;1.60</td>
<td>&gt;1.80</td>
</tr>
<tr>
<td>Silt</td>
<td>&lt;1.40</td>
<td>&gt;1.65</td>
</tr>
<tr>
<td>Clay</td>
<td>&lt;1.10</td>
<td>&gt;1.47</td>
</tr>
</tbody>
</table>

Fig. 6. (a) Soil with good structure (pH <=7) (b) Soil with poor and dense structure (pH >7)

Soil is polluted by various anthropogenic activities such as addition of industrial wastes, pesticides, fertilizers, domestic sewage etc. to soil. Poly nuclear aromatic hydrocarbons and petroleum hydrocarbons and heavy metals are the most common contaminants of soil. Urbanization and industrialization are the major contributors of soil pollution. It leads to numerous health hazards and the chemicals reach human being through contaminated water and plants.

Artificial Neural Network (ANN) is a software system that imitates the neural networks of the brain of man. Neural networks are powerful tools that have the ability to identify underlying highly complex relationships from input-output data only. The study indicates that ANN is efficient in simulating the complicated phenomena during the training process and uses them to simulate the results for new inputs. The technique is capable of dealing with uncertainties in the inputs and can extract information from incomplete or contradictory data sets.

2. Literature Survey
In summary of related 348 papers and reports, ANNs have been applied in solving problems in food quality and safety (35.34%), crop (22.7%), soil and water (14.37%), precision agriculture (6.61%), food processing (2.3%), greenhouse control (2.01%), agriculture vehicle control (1.15%), agricultural pollution (1.15%), agricultural biology (1.15%) and others (2.3%) such as bio energy and agricultural facilities. These ANN applications have been created mainly through classification (45.11%), modeling and prediction (43.97%), control (4.02%) and simulation (2.59%), parameter estimation (2.01%), detection (1.15%), data clustering (0.57%), optimization (0.29%) and data fusion (0.29%) as well.

In summary of 136 related papers and reports, FL has been applied in solving problems in crop management (17%), soil and water (16%), food quality and safety (14%), animal health and behavior (10%), agricultural vehicle control (8%), precision agriculture (7%), greenhouse control(7%), agricultural machinery (4%), food processing (4%), air quality and pollution (3%), agricultural facilities (2%) and others (6%) such as natural resources management and agricultural product design. These applications have been created through FL mainly by control (28%), modeling and prediction (24%), classification (24%), fuzzy clustering (9%), rule-based Inference (7%), multi sensor data fusion (4%), optimization (1%) and others (3%) such as thresholding and pattern inference.

Based on a summary of 83 papers and reports, GAs have been applied in solving problems in crop management (31%), water management (27%), food quality and safety (11%), food processing (6%), precision agriculture (4%), agricultural biology (4%), agricultural machinery (2%), agricultural facilities (2%), animal behavior (2%), and others (11%) such as agricultural vehicle, robotics, and pollution. GAs are basically an optimization and search method. The applications for optimization take the largest portion of the total, 66%. GAs have been also used to assist with modeling and prediction (18%), classification (12%), control (2%), data clustering (1%) and value thresholding (1%).

3. Materials and Methods
Classification and Prediction are two modes of data analysis that can be used to extract the models that describe important data classes or to predict the future data trends. Classification predicts categorical labels. Prediction models continuous valued functions.

Data classification is a process of two-steps:

Model Construction: describing a set of predetermined classes.
- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute.
The set of tuples used for model construction is a training set. The model is represented as classification rules, decision trees, or mathematical formulae. Model Usage: for classifying future on unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is an independent of a training set. If the accuracy is acceptable, use the model to classify data tuples whose class labels are unknown.

Prediction: This data analysis task of numeric prediction, where the model constructed predicts a continuous-valued function, or ordered value, as opposed to categorical label. This model is a predictor. Regression analysis is a statistical methodology which is most often used for numeric prediction. Learning is broadly classified into two types: Supervised Learning and Unsupervised Learning (clustering). Supervised Learning: The training data (observations, measurements, etc.) are accompanied by the labels indicating the class of the observations. New data is classified based on the training set. Unsupervised Learning (Clustering): The class label of training data is unknown. Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data.

3.1 ISRIC-WISE Soil Profile Data

The test dataset consists of 705 samples of particular region as in Table 3. Using for this research work collected from World Soil Information – ISRIC (International Soil Reference and Information Centre). Version 3.1 of the ISRIC-WISE database (WISE3-World Inventory of Soil Emission Potentials) was compiled from a wide range of soil profile data collected by many soil professionals worldwide. All profiles have been harmonized with respect to the original Legend (1974) and Revised Legend (1988) of FAO-UNESCO.

Table 3. List of soil variables, their abbreviations and units of measurement

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHH2O</td>
<td>Soil reaction in water</td>
<td>pH units</td>
</tr>
<tr>
<td>SAND</td>
<td>Sand</td>
<td>% (mass)</td>
</tr>
<tr>
<td>SILT</td>
<td>Silt</td>
<td>% (mass)</td>
</tr>
<tr>
<td>CLAY</td>
<td>Clay</td>
<td>% (mass)</td>
</tr>
<tr>
<td>TOTPOR</td>
<td>Total Porosity</td>
<td>%</td>
</tr>
</tbody>
</table>

3.2 Classification Rules Model

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Major Textural Classes</th>
<th>Relative % of Sand, Silt, Clay</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sand</td>
<td>Must contain Sand &gt; 85% and % of silt plus 1.5 times the % of clay shall not exceed 15 and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>between 25 and 50 of Total Porosity and pH value between 5.8 to 7</td>
</tr>
<tr>
<td>2.</td>
<td>Loamy Sand</td>
<td>A) Upper Limit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B) Must contain 85 to 90 % of sand and the % of silt plus 1.5 times the % of clay is not less</td>
</tr>
<tr>
<td></td>
<td></td>
<td>than 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C) Lower Limit: Must contain 70 to 85% sand and the % of silt plus twice the % of clay does</td>
</tr>
<tr>
<td></td>
<td></td>
<td>not exceed 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D) pH value between 5.8 and 7 and Total Porosity between 25 and 50</td>
</tr>
<tr>
<td>3.</td>
<td>Sandy Loam</td>
<td>A) Contains 20% or less clay and Total Porosity between 25 to 50 and the pH value between 5.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B) The % of silt plus twice the % of clay exceeds 30 and has 52% or more sand, or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C) Contains &lt;7% clay, &lt;50% silt, and between 43 and 52% sand</td>
</tr>
<tr>
<td>4.</td>
<td>Loam</td>
<td>Contains 7 to 27% clay, 28 to 50% silt, &lt;53% sand and Total Porosity between 37.5 to 46.5 and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pH value between 6 to 6.5</td>
</tr>
<tr>
<td>5.</td>
<td>Silty Loam</td>
<td>A) Contains 50% or more silt and 12 to 27% clay, or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B) Contains 50% to 80% silt and &lt;12% of clay</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C) Contains 6 to 7 pH value and Total Porosity value between 35 to 50</td>
</tr>
<tr>
<td>6.</td>
<td>Silt</td>
<td>Contains 80% or more silt and &lt;12% clay and Total Porosity between 35 to 50 and pH value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>between 6 to 7</td>
</tr>
<tr>
<td>7.</td>
<td>Sandy Clay Loam</td>
<td>Contains 20 to 35% clay, &lt;28% silt, and 45% or more sand and Total Porosity between 37.5 to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46.5 and pH value between 6 to 6.5</td>
</tr>
<tr>
<td>8.</td>
<td>Clay Loam</td>
<td>Contains 27 to 40% clay and 20 to 45% sand and Total Porosity between 37.5 to 46.5 and pH value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>between 6 to 6.5</td>
</tr>
<tr>
<td>9.</td>
<td>Silty Clay Loam</td>
<td>Contains 27 to 40% clay and &lt;20% or more sand and Total Porosity between 37.5 to 46.5 and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pH value between 6 to 7</td>
</tr>
<tr>
<td>10.</td>
<td>Sandy Clay</td>
<td>Contains 35% or more clay and 45% or more sand and Total Porosity value between 37.5 to 42.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and pH value between 6 to 6.5</td>
</tr>
<tr>
<td>11.</td>
<td>Silty clay</td>
<td>Contains 40% or more clay and 40% or more silt and Total Porosity between 42.5 to 46.5 and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pH value between 6.1 to 6.5</td>
</tr>
<tr>
<td>12.</td>
<td>Clay</td>
<td>Contains 40% or more clay, &lt;45% sand, and &lt;40% silt and Total Porosity from 33 to 60 and pH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>value from 6 to 7</td>
</tr>
</tbody>
</table>
4. ANN for Prediction of Soil Total Porosity

Artificial Neural Networks are generally presented as systems of interconnected nodes which can compute values from inputs. Simple artificial nodes are class of statistical models could be called “neural” if they possess the following characteristics: consist of sets of adaptive weights, i.e. numerical parameters that are tuned by a learning algorithm, and are capable of approximating non-linear functions of their inputs. The adaptive weights are conceptually connection strengths between neurons, which are activated during training and prediction. The network structure is composed of a set of neurons connected by links and organized in number of layers. Each layer is fully interconnected to the preceding layer by weights. Initial suggested weights are progressively adjusted during the training process by comparing predicted outputs with measured data (targets). The computation of network weights and biases is known as “training step”. The objective of the back-propagation training algorithm is to find the optimal weights by minimizing the Mean Square Error (MSE) of the output values. Simultaneously, learning functions are used to update the layer’s weight and bias. This procedure is completed in the “validation step”, where the network is improved to avoid data over-fitting. After that a set of data is randomly used to examine the network generalization, i.e. the “test step”.

Each scalar input (p) is multiplied by a scalar weight (W), and then added to a scalar bias (b), resulted in the net input \( n = Wp + b \). Finally, the result is passed through the transfer function \( f \), which gets the neuron’s output \( a \), where \( a = f(Wp+b) \) (Fig.7). Those functions can be (i) linear transfer functions: It is purely a linear function \( y = mx+c \) and it is also called as PURELIN transfer function. It is used in the final layer to find a linear approximation to a nonlinear function, or (ii) sigmoid transfer functions: used in the hidden layers to generate the output between 0 and 1 and also known as LOGSIG transfer function and between -1 and 1 for Tan-Sigmoid or TANSIG, even if the input data have infinity values.

During training, each neuron in the layer adjusts its weight vector toward the closest group of input vectors (Fig.9). The magnitude of the gradient and the number of validation checks were used to terminate the network training. At epoch: 6 iterations, the gradient was equal to 3.758 (i.e. at gradient less than 1e-005, the training will stop). The number of validation checks was equal to 6; which is the appropriate value to stop training. The performance plot (Fig.11) shows the value of the function, in basis of training, validation, and test behaviors, versus the iteration number. The best validation performance, based on the mean square
error, was 0.0035334 at epoch 0. Since the validation and test curves are very similar, therefore no major problems or over-fitting occurred with the training (Fig. 10).

The final weights and biases were:

Weight to layer 1 from input 1 \((w_{1;1}) = [8.1898; -7.5562; -7.0081; 7.0008; 7.6404]\)

Weight to layer 2 from layer 1 \((w_{2;1}) = [1.9045 -1.9357 0.95875 -0.59307 -0.67968]\)

Bias to layer 1 \((b_{1}) = [-9.1555; 3.8046; -0.013523; 3.4992; 6.3596]\)

Bias to layer 2 \((b_{2}) = [1.6103]\).

Linear regression plots the network outputs for training, validation and testing in order to determine the relationship between the outputs of the network and the targets (Fig. 12). In each plot, the dashed line represents the perfect result, i.e. outputs = targets, whereas the solid line corresponds to the best fit linear regression. As the R-value approaches to one, then there is an exact linear relationship. The regression results (R-value) were 0.36427, 0.31759 and 0.26087 for training, validation and test, respectively. Those results were corresponding to a total response of 0.33523. The lower regression results can be attributed to fewer training data (accounting for 78 points) and/or the ANN configuration, in terms of number of hidden layers and neurons, might not being optimal. After the network was trained, validated and tested, the generated model can be used to predict the parameter Total Porosity through new input pH data.

It was found that experimental analysis of Total Porosity was close to the predicted data calculated from the configuration and this confirms the validity of this model.
5. Conclusion

Many mapping of contaminated soil sites and the resulting cleanup are time consuming and expensive tasks, requiring extensive amounts of geology, hydrology, chemistry, GIS in environmental contamination and data mining and pattern recognition techniques. The soil salinity based on alkalinity was predicted and ANN with a structure of 1–5–1 was proposed. The network showed an acceptable ability to capture the interrelationship between input: pH and output: Total Porosity concentrations. Values of correlation coefficient (R) of training, validation and test were 0.36427, 0.31759 and 0.26087, respectively. It is concluded that the developed ANN model is suitable for predicting the salinity concentrations. Development of novel methods for classification in data mining and pattern recognition like genetic based, and hybrid based is required to find out the optimal number of clusters for soil sites based on soil texture with Total Porosity.

References

A Novel Soil Profile Feature Reduction Model using Principal Component Analysis

D. Ashok Kumar1* and N. Kannathasan2

1Department of Computer Science, Government Arts College, Tiruchirapalli - 620022, Tamil Nadu, India; akudaiyar@yahoo.com
2Department of Computer Science, Kanchi Mamunivar Centre for Post Graduate Studies, Puducherry - 605008, U.T. of Puducherry, India; nkthasan@gmail.com

Abstract
Background/Objectives: Data Mining has been used to analyze large datasets and establish useful classification and patterns in the datasets. The efficient analysis of data in different format becomes a challenging work. Methods/Statistical Analysis: This work proposed a novel Soil Profile Feature Reduction Model using Principal Component Analysis for data reduction. The proposed model uses the method of k-Means clustering and PC A approach for feature reduction which initially applies PCA to acquire reduced uncorrelated attributes showing maximal eigenvalues in the dataset with minimal loss of information. Again proposed model uses k-Means on the PCA reduced dataset to find out discriminative features that will be the most sufficient ones for classification. Findings: The weight by PCA generates attribute weights of the soil profile dataset using a component created by the PCA. The component is mentioned by the component number parameter. The normalize weight parameter is usually set to true to spread the weights between 0 and 1. The attribute weights reflect the relevance of the attributes with respect to the class attribute. The higher weight of an attribute is more relevant, it is considered. This is a combination of clustering approach with feature reduction to get a minimal set attributes relating a suitably high accuracy in describing the original features. The result of clustering is same after reducing the attributes using PCA. The experimental results prove that proposed model is reducing number of initial attributes, reducing computational complexity and improving predictive accuracy in High Dimensional Datasets. Applications/Improvements: The same soil profile feature is implemented by using the other techniques instead of PCA algorithm in future.

Keywords: Clustering, Feature Reduction, k-Means, Principal Component Analysis, Soil Profile

1. Introduction

The nature of soil varies with the ecosystem and so the productivity of that ecosystem. Soils support all life forms on the planet and play vital role for their existence. The top layer of the soil is a natural filter for many contaminants is shown in Figure 1. Soil has five layers and the property of each layer may vary from region to region. The property is dependent on a variety of environmental parameters of that region is depicted in Figure 2.

The soil properties such as texture, porosity, specific yield depend on the total volume of groundwater recharge, water storage and discharge, also the extent of groundwater contamination1.

Soil is a mixture of sand, silt and clay and each of them vary in their particle size (Table 1). Soil is made of small mineral particles that differ in size based on the type of soil. Clay, silt and sand are common soils which have an increasing order of particle size respectively. The particles of clay are charged and can attract water molecules.

Figure 1. Soil water contamination.
A Novel Soil Profile Feature Reduction Model using Principal Component Analysis

Soil system is composed of air, water, dead organic matter and various types of living organisms (Figure 4). It is rich in a number of microorganisms and living matter. The dead and decaying matter of the soil plays a crucial role in formation of humus.

Shape and spatial arrangement of soil particles decide the porosity of that soil. It is the air space or void space between soil particles. Infiltration, ground water movement and storage occur in these void spaces. Porosity of soil typically decreases as pH value increases and also particle size increases because of soil contamination or soil pollution (Figure 5).

Clay is one of the suitable soil for cultivation of crops. It is resistant to erosion, can hold water better than the other soils but the silt of the clay can easily be washed away and so the nutrients. Whereas the sand allow the water to pass through easily and quickly and it cannot hold water like that of the clay and therefore it cannot support the growth of flora. The sand which is exposed is highly subjected to erosion if the angle and slope of land is too severe. The soil textural triangle gives the textural name, which based on the percentage of sand, silt and clay within the soil sample (Figure 3). The triangle is divided into 10-percent portions of clay, silt and sand. The summation of the three percentages must total 100 percent.

![Figure 2. Soil profile.](image1)

Table 1. USDA particle size and porosity ranges for sand, silt and clay

<table>
<thead>
<tr>
<th>Name of separate</th>
<th>Diameter range (millimeters)</th>
<th>Porosity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>2.0 – 0.05</td>
<td>25-50</td>
</tr>
<tr>
<td>Silt</td>
<td>0.05 – 0.002</td>
<td>35-50</td>
</tr>
<tr>
<td>Clay</td>
<td>Less than 0.002</td>
<td>33-60</td>
</tr>
</tbody>
</table>

Bulk density is an important indicator of soil compaction and is given in Table 2. The porosity of soil is governed by soil biota and increases the water and air movement to rhizosphere shown in Figure 6.

![Figure 3. USDA 12 basic soil textural triangle.](image2)

Table 2. General relationship of soil bulk density to root growth based on soil textural

<table>
<thead>
<tr>
<th>Type of Mineral Particle</th>
<th>Ideal bulk densities for Plant Growth (g/cm³)</th>
<th>Bulk densities that restrict root growth (g/cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>&lt;1.60</td>
<td>&gt;1.80</td>
</tr>
<tr>
<td>Silt</td>
<td>&lt;1.40</td>
<td>&gt;1.65</td>
</tr>
<tr>
<td>Clay</td>
<td>&lt;1.10</td>
<td>&gt;1.47</td>
</tr>
</tbody>
</table>
Figure 6. (a) Soil with good structure (pH\leq7). (b) Soil with poor and dense structure (pH >7).

Soil is polluted by various anthropogenic activities such as addition of industrial wastes, pesticides, fertilizers, domestic sewage etc. to soil. Polynuclear aromatic hydrocarbons and petroleum hydrocarbons and heavy metals are the most common contaminants of soil. Urbanization and industrialization are the major contributors of soil pollution. It leads to numerous health hazards and the chemicals reach human being through contaminated water and plants.

2. Materials and Methods

Regression analysis was used for numeric prediction\(^7\). Learning is broadly classified into two types: Supervised Learning and Unsupervised Learning. Supervised Learning: The training data are accompanied by the labels representing the class of the observations. New data is classified based on the training set. Unsupervised Learning: The class label of training data is unknown. Given a set of measurements, observations with the aim of establishing the existence of classes or clusters in the data\(^8-10\).

Table 3. List of soil variables their abbreviation and units of measurement

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHH(_O)</td>
<td>Soil reaction in Water</td>
<td>pH units</td>
</tr>
<tr>
<td>ECE</td>
<td>Electrical Conductivity of Saturated Paste</td>
<td>dS m(^{-1})</td>
</tr>
<tr>
<td>ORGC</td>
<td>Organic Carbon</td>
<td>% (mass)</td>
</tr>
<tr>
<td>CECC</td>
<td>Cation Exchange Capacity of clay size fraction</td>
<td>Cmol kg(^{-1})</td>
</tr>
<tr>
<td>TEB</td>
<td>Total Extractable Bases</td>
<td>Cmol kg(^{-1})</td>
</tr>
<tr>
<td>CACO</td>
<td>Calcium Carbon content</td>
<td>% (mass)</td>
</tr>
<tr>
<td>GYPS</td>
<td>Gypsum content</td>
<td>% (mass)</td>
</tr>
<tr>
<td>SAND</td>
<td>Sand</td>
<td>% (mass)</td>
</tr>
<tr>
<td>SILT</td>
<td>Silt</td>
<td>% (mass)</td>
</tr>
<tr>
<td>CLAY</td>
<td>Clay</td>
<td>% (mass)</td>
</tr>
<tr>
<td>TOTPOR</td>
<td>Total Porosity</td>
<td>%</td>
</tr>
</tbody>
</table>

Principal Component Analysis (PCA), a mathematical procedure, uses an orthogonal transformation to convert a set of observations of possibly correlated attributes into a set of values of uncorrelated attributes called Principal Components\(^11\). The test dataset consists of 16383 samples of particular region as in Table 3 using for this research work collected from World Soil Information International Soil Reference and Information Centre\(^3\).

3. Proposed Model using PCA

The Proposed Model consists of three phases such as k-Means Clustering on Original Dataset, to reduce the unnecessary attributes using Principal Component Analysis Feature Reduction method and k-Means Clustering on Reduced Dataset which is shown in Figure 7. The proposed model uses the method of k-Means clustering and PCA approach for attribute reduction, which initially applies PCA to obtain reduced uncorrelated attributes mention maximal eigenvalues in the dataset with minimum loss of information. Then again proposed model uses k-Means on the PCA reduced dataset to find out discriminative features that will be the most sufficient ones for classification.

![Figure 7. Proposed model.](image-url)
A Novel Soil Profile Feature Reduction Model using Principal Component Analysis

**Figure 8.** Steps for PCA method.

PCA projects high dimensional data to a lower dimension and projects the data in the least square sense. It captures big principal variability in the data and ignores small variability and reduces the dimensionality of a data set by finding a new set of variables, small than the original set of variables. PCA can also retain most of the sample's information while it extracting relevant information from confusing data sets. It is a simple, non-parametric method for dimension reduction. For this reason it is useful for the compression and classification of data. Principal Component Analysis is used in image process and compression, pattern recognition and other data dimension reduction. Steps for PCA are given below in Figure 8.

Cluster analysis is the assignment of a set of observations into subsets so that observations in the same cluster are similar in some sense. The algorithm k-Means is the process of partitioning a group of data points into a small number of clusters. The steps for k-Means algorithm are shown in Figure 9.

**4. Experimental Evaluation and Result Analysis**

Feature reduction model using Principal Component Analysis is reducing number of initial attributes, reducing computational complexity and improving predictive accuracy. The number of principal components is lesser or equal to the number of original attributes.

**Table 4.** Cluster in original dataset

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Cluster_0</th>
<th>Cluster_1</th>
<th>Cluster_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHH</td>
<td>2</td>
<td>-98</td>
<td>6.479</td>
</tr>
<tr>
<td>ECE</td>
<td>-98</td>
<td>0.132</td>
<td>6.479</td>
</tr>
<tr>
<td>ORGC</td>
<td>-98</td>
<td>1.645</td>
<td>6.479</td>
</tr>
<tr>
<td>CECC</td>
<td>-98</td>
<td>53.252</td>
<td>6.479</td>
</tr>
<tr>
<td>TEB</td>
<td>-98</td>
<td>16.453</td>
<td>6.479</td>
</tr>
<tr>
<td>CACO</td>
<td>-98</td>
<td>1.206</td>
<td>6.479</td>
</tr>
<tr>
<td>GYPS</td>
<td>-98</td>
<td>0</td>
<td>6.479</td>
</tr>
<tr>
<td>SAND</td>
<td>-98</td>
<td>38.302</td>
<td>6.479</td>
</tr>
<tr>
<td>SILT</td>
<td>-98</td>
<td>36.937</td>
<td>6.479</td>
</tr>
<tr>
<td>CLAY</td>
<td>-98</td>
<td>24.761</td>
<td>6.479</td>
</tr>
<tr>
<td>TOTPOR</td>
<td>-98</td>
<td>49.533</td>
<td>6.479</td>
</tr>
</tbody>
</table>

**4.1 Phase-I**

Step 1: Collection of dataset: The soil profile data set that is collected is taken.

Step 2: Application of k-Means algorithm: Using the dataset and got the following cluster listing, this is shown in Table 4.

Step 3: Clustering using Centroid Plot View method: Using Centroid Plot View method obtained the 2D as shown in Figure 10.

**Figure 10.** Centroid plot view of original dataset.
4.2 Phase-II

Step 1: Collection of the dataset soil profile: The collected soil profile dataset.

Step 2: Finding the all covariance matrix, Eigen values and Eigen vectors of the dataset: which is shown in Table 5 and Table 6.

Step 3: Clustering using Centroid Plot View method: Using Centroid Plot View method obtained the 2D as shown in Figure 11.

Step 4: Determining the number of meaningful Principal Components: To eliminate the weaker components from the Principal Components set the corresponding Eigen value which is lesser than the other points in that dataset. Consider only the highest values in the dataset those values are taken into account.

Step 5: Finding the reduced data set using the reduced Principal Components: The transformation matrix with reduced Principal Components is applied to the dataset to produce the new reduced projected dataset which can be used for further data analysis, which is shown in Table 7.

Table 5. Eigen values of the covariance matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>Standard Derivation</th>
<th>Proportion of Variance</th>
<th>Cumulative Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC_1</td>
<td>56.547</td>
<td>0.797</td>
<td>0.797</td>
</tr>
<tr>
<td>PC_2</td>
<td>23.429</td>
<td>0.137</td>
<td>0.934</td>
</tr>
<tr>
<td>PC_3</td>
<td>14.314</td>
<td>0.051</td>
<td>0.985</td>
</tr>
<tr>
<td>PC_4</td>
<td>6.609</td>
<td>0.011</td>
<td>0.996</td>
</tr>
<tr>
<td>PC_5</td>
<td>3.088</td>
<td>0.002</td>
<td>0.998</td>
</tr>
<tr>
<td>PC_6</td>
<td>1.895</td>
<td>0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>PC_7</td>
<td>1.438</td>
<td>0.001</td>
<td>1.000</td>
</tr>
<tr>
<td>PC_8</td>
<td>0.519</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PC_9</td>
<td>0.418</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PC_10</td>
<td>0.179</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PC_11</td>
<td>0.033</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 6. Eigen vectors of the covariance matrix

<table>
<thead>
<tr>
<th>Attribute</th>
<th>PC_1</th>
<th>PC_2</th>
<th>PC_3</th>
<th>PC_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHH₂O</td>
<td>0.258</td>
<td>0.099</td>
<td>-0.022</td>
<td>-0.020</td>
</tr>
<tr>
<td>ECE</td>
<td>0.241</td>
<td>0.117</td>
<td>-0.033</td>
<td>-0.087</td>
</tr>
<tr>
<td>ORGC</td>
<td>0.245</td>
<td>0.101</td>
<td>-0.049</td>
<td>-0.095</td>
</tr>
<tr>
<td>CECC</td>
<td>0.405</td>
<td>-0.663</td>
<td>0.362</td>
<td>0.213</td>
</tr>
<tr>
<td>TEB</td>
<td>0.297</td>
<td>-0.206</td>
<td>-0.025</td>
<td>0.517</td>
</tr>
<tr>
<td>CACO</td>
<td>0.245</td>
<td>0.095</td>
<td>-0.032</td>
<td>0.006</td>
</tr>
<tr>
<td>GYPS</td>
<td>0.241</td>
<td>0.117</td>
<td>-0.032</td>
<td>-0.089</td>
</tr>
<tr>
<td>SAND</td>
<td>0.327</td>
<td>0.555</td>
<td>0.651</td>
<td>0.008</td>
</tr>
<tr>
<td>SILT</td>
<td>0.344</td>
<td>-0.292</td>
<td>-0.170</td>
<td>-0.726</td>
</tr>
<tr>
<td>CLAY</td>
<td>0.297</td>
<td>0.209</td>
<td>-0.610</td>
<td>0.359</td>
</tr>
<tr>
<td>TOTPOR</td>
<td>0.363</td>
<td>0.157</td>
<td>-0.195</td>
<td>-0.082</td>
</tr>
</tbody>
</table>

4.3 Phase-III

Step 1: k-Means cluster analysis using reduced dataset 2D by PCA: This is shown in Table 8.

The Weight by PCA operator produces attribute weights of the given dataset using a component created by the PCA. The component is mentioned by the component number parameter. The normalize weight parameter is usually set to true to spread the weights between 0 and 1. The attribute weights reflect the relevance of the attributes with respect to the class attribute. The higher weight of an attribute is more relevant, it is considered. Weighted for original dataset is shown in Table 9 and Weighted for reduced dataset is shown in Table 10.
Clustering was performed by proposed model on the soil profile directly using k-Means. Using PCA the features of the data set were then reduced and clustering was performed again. The clustering results of both the data sets are found to be same. Therefore the model has proved that the PCA is a good feature reduction technique, it can handle data without any loss.

5. Conclusion

The proposed model first finds the cluster of soil profile dataset using the k-Means clustering method and then the PCA approach of feature reduction technique has been implemented on that data set having continuous attribute values. As a result of which a number of uncorrelated and discriminative attributes, more adequate for classification has been obtained. These attributes also specifies the maximal variances among the dataset by retaining the original property of the dataset. Then model applied the k-Means clustering method on that reduced dataset. The same cluster listing obtained in both the methods. So, the experimental results prove that the PCA method is highly useful for getting the same cluster and also takes less time. Therefore, improving the effectiveness of the proposed model in terms of classification accuracy will be investigated in the future using some other feature reduction methods, like Wavelet Transform.

6. References

2. Study guide: Soil mechanics level 1, module 3, uniform soil classification system, National employee development staff, soil conservation services. United States Department of Agriculture; 1987.