Aakunuri Manjula* and G. Narsimha

Using an Efficient Optimal Classifier for Soil Classification in Spatial Data Mining Over Big Data

https://doi.org/10.1515/jisys-2017-0209
Received May 13, 2017; previously published online January 10, 2018.

Abstract: This article proposes an effectual process for soil classification. The input data of the proposed procedure is the Harmonized World Soil Database. Preprocessing aids to generate enhanced representation and will use minimum time. Then, the MapReduce framework divides the input dataset into a complimentary portion that is held by the map task. In the map task, principal component analysis is used to reduce the data and the outputs of the maps are then contributed to reduce the tasks. Lastly, the proposed process is employed to categorize the soil kind by means of an optimal neural network (NN) classifier. Here, the conventional NN is customized using the optimization procedure. In an NN, the weights are optimized using the grey wolf optimization (GWO) algorithm. Derived from the classifier, we categorize the soil category. The performance of the proposed procedure is assessed by means of sensitivity, specificity, accuracy, precision, recall, and F-measure. The analysis results illustrate that the recommended artificial NN-GWO process has an accuracy of 90.46%, but the conventional NN and *k*-nearest neighbor classifiers have an accuracy value of 75.3846% and 75.38%, respectively, which is the least value compared to the proposed procedure. The execution is made by Java within the MapReduce framework using Hadoop.

Keywords: MapReduce framework, principal component analysis, neural network, grey wolf optimization, accuracy, precision, recall, F-measure.

1 Introduction

In agriculture, the soil is the major resource of invention, which offers an intermediate for plant organization, seed germination, root expansion, and enlargement [1]. The soil is the extra species-rich habitation of the earth's ecosystem and its task comprises biomass creation, preservation of nutrient balance, chemical recycling, and water storage [12]. The soil and water can be effortlessly tainted by metals/metalloids because of direct contact (and interference) through the chemical procedure of metal discharge and recruitment that normally are results of mining [6]. Therefore, in the soil, data mining is significant. Soil spatial data mining (SDM) is a position region obtained by data mining for the quick investigation of spatial data. The objective of soil SDM is data investigation and knowledge discovery in a big quantity of spatial data gathered in a spatial scheme such as geographic information systems (GIS) [13]. The study endeavors to compare the performance of data mining algorithms among soil possessions such as particle allocation, clay content, and plasticity, which are employed to allocate soils a categorization and addition-

***Corresponding author: Aakunuri Manjula,** Jawaharlal Nehru Technological University, Hyderabad, India, e-mail: manjula30303@gmail.com
G. Narsimha: Department of IT, JNTUH College of Engineering, Karimnagar District, Telangana State, India

Open Access. © 2020 Walter de Gruyter GmbH, Berlin/Boston. This work is licensed under the Creative Commons Attribution 4.0 Public License.

ally with respect to soil restrictions and soil situation with respect to the subsequent uniqueness: acidity, alkalinity, sodicity, salinity, low cation replaceability, phosphorus fixation, cracking and swelling, depth, soil density, color, and nutrient content [14, 25].

Soil color is a complete pointer of the chemical composition and physical uniqueness of soils that a huge amount of soil information can be efficiently attained by the results of soil color. Therefore, the process of soil categorization and qualitative recognition rooted in soil colors is the major frequent process [10]. This process of soil categorization was an idea loaded, data reduced association, which instigates in the late 1950s, afterward the innovation of the digital computer, which permitted the complete computation of taxonomic detachment – detachment among position in multidimensional spaces, the axes definite by the typeset or features of the soil outline [11]. This soil categorization scheme offers a frequent metric for passing on the uniqueness of comparable soil kinds and for comparison among soils. A typical metric permits scientists and engineers to compare efforts or constantly pertain intend strategy transversely changeable soil situation [9]. Also, diverse soil categorization schemes were used and are discussed as pursue. In the middle of them, the existing Hungarian Soil Classification System (HSCS) was enhanced between the 1950s and the 1960s, founded on the genetic standard of Dokuchaev. The scheme is somewhat expressive; the taxonomic components are discriminated derived from the identification of a group of soil figuring procedure using a soil geographic method and restricted laboratory data [8, 20].

Afterward, the spatial disaggregation of a multicomponent soil categorization scheme was improved, as map polygons into entity element soil module have been established in an endeavor to commonly update soil maps and to generate class dissimilarity surrounded by the boundaries of unique review map elements [22]. Lastly, the Unified Soil Classification System (USCS) was enhanced, which is a broadly employed soil categorization scheme that uses the above-mentioned soil possessions to allocate a soil categorization. Although many soil categorization schemes have been enhanced, the procedures of assembling soil data and mapping soils, additionally the soil categorization scheme employed, considerably vary between the Alpine countries [3]. Thus, to conquer those concerns, classification and regression trees (CART) is a data mining representation procedure effectual for managing classification (or regression) tasks to offer robust analytical representation [27]. However, as soil possessions offer a range in their spatial dissimilarity, it is hard to classify soil models devoid of initiate fault or oversimplifications. Consequently, class limitations are typically elected subjectively by granting (1) an uncertainty about the accuracy of the grave threshold or range employed to identify association in a definite group and (2) an uncertainty about the eminence of the input maps [16]. The main contributions of the proposed technique are described as follows:

1. Efficient classifier is used for soil classification in SDM over big data.
2. The MapReduce framework is used to split the input dataset into free pieces, which are handled by the map tasks.
3. Principal component analysis (PCA) is used to reduce the data and the outputs of the maps are then contributed to reduce the tasks.
4. To classify the soil type, the method used optimal neural network (ONN) classification.
5. To improve the classification accuracy, weights are optimized by the grey wolf optimization (GWO) algorithm.

2 Literature Survey

Ref.	Topic	Goal	Technique(s)	Parameter(s) used	Result(s)
[20]	Using SDM to analyze area-diversity patterns on soil, vegetation, and climate: a case study from Almería, Spain	To analyze the various relationships such as area-pedodiversity, area-vegetation diversity, area-bioclimatic belts, diversity and pedodiversity, vegetation diversity, and bioclimatic diversity	Data mining technique and GIS technology	Map of soil associations, map of potential vegetation, map of bioclimatic belts	The use of similarity matrices allowed us to quantify the correlations among the factors that determine the spatial patterns. These results support and encourage further research in modeling pedodiversity and vegetation diversity relationships at different spatial scales with respect to climatic gradients and human impact. This knowledge is required for land management and conservation planning
[26]	Mining geostatistics to quantify the spatial variability of certain soil flow properties	To quantify the spatial variability of certain soil flow properties	Mining geostatistical approach	Darcy's permeability coefficient, spatial distribution of the α-parameter (relating the matrix potential to the unsaturated hydraulic conductivity)	Majority of precision agriculture strategies rely on statistical analyses (or image processing) of indirect measurements of the soil conditions obtained. This forecasting will be facilitated and informed by in situ measurements of water content obtained with spatially distributed autonomous and automated sensors on an IoT approach
[24]	Improved spatial resolution in soil moisture retrieval at arid mining area using apparent thermal inertia (ATI)	To calculate the soil moisture	A surface soil moisture model with improved spatial resolution was developed using remotely sensed ATI	The model integrates the surface temperature derived from TM/ETM+ image and the mean surface temperature from MODIS images to improve the spatial resolution of soil temperature difference based on the heat conduction equation	The improved soil moisture model is not only suitable for TM/ETM+ images but also for other higher spatial resolution images with thermal infrared band. The improved ATI has a stronger correlation with soil moisture than the wetness index
[4]	Factors causing spatial heterogeneity in soil properties, plant cover, and soil fauna in a nonreclaimed postmining site	To study the processes responsible for environmental heterogeneity at a manmade postmining landscape formed by the heaping of homogeneous overburden in longitudinal "waves"		Distribution of shrubs between waves affect between-waves distribution of other parameters, namely, soil Cox, as waves with more shrubs have larger litter input and more soil carbon	Spatial heterogeneity may help organisms deal with temporal fluctuations in environmental factors (animals can migrate between habitats; plants can spread their roots into several microhabitats). It follows that maintaining the spatial heterogeneity created by the heaping may enhance the spontaneous re-vegetation of postmining sites

2 Literature Survey (continued)

Ref.	Topic	Goal	Technique(s)	Parameter(s) used	Result(s)
[5]	Comparisons of spatial and nonspatial models for predicting soil carbon content based on visible and near-infrared spectral technology	To predict the soil properties	Partial least squares (PLS) regression (PLSR), PLS-geographically weighted regression (GWR), PLSR kriging, and PLS-GWR kriging were constructed to predict soil organic matter (SOM) based on soil spectral reflectance	The prediction capabilities of the models were evaluated using coefficient of determination (R^2), root mean square error, and ratio of performance to interquartile range	We discussed the spatial characteristics of SOM and its spectra and the importance of the spatial dependence of model residuals in predicting SOM
[2]	Effect of postmining land use on the spatial distribution of metal(loid)s and their transport in agricultural soils: analysis of a case study of Chungyang, South Korea	To analyze the spatial distribution of metal(loid)s and their transport in agricultural soils	Pollution sources were investigated using various statistical techniques, including multivariate statistical analysis and GIS, and the spatial distribution of metals and metalloids was determined	In this study, the concentrations of As, Cd, Cu, Ni, Pb, and Zn in agricultural soil near abandoned gold mine sites in a region of South Korea (Chungyang County) were analyzed	The distribution patterns of the metals in the soil of agricultural fields show that As and Cd, pollutants that originate from abandoned mines, have also been transported to fields that are 4 km or more away from their source and can have an effect on the surrounding environment. Agricultural fields polluted by diverse metal(loid) and their pollution characteristics and behaviors in soils must be understood
[15]	Evaluating large-extent spatial modeling approaches: a case study for soil depth (SD) for France	This work was to compare three spatially explicit modeling approaches for SD	The straightforward digital soil mapping approach based on regression tree modeling (RTM), gradient boosting modeling, and multiresolution kriging for large datasets	SD is the input parameter	The RTM approach used in this paper is a flexible method that can be applied – and easily improved – to map any other soil properties

3 Problem Definition

Soil SDM is a complicated area in data mining. The unpredictable development of spatial data and the standard practice of spatial databases designate the necessity for the mechanical recognition of spatial knowledge. The restrictions connected by obtainable soil SDM are as follows:

1. GWR in Ref. [5] leads to the prediction errors of soil; hence, they are not flexible in soil classification.
2. Traditional soil SDM methods in Ref. [2] are incapable of analyzing the behavior of soils and hence clearly leads to the incorrect classification of soil.
3. Conventional soil SDM methods are difficult to implement, as they increase the computational costs.
4. Using conventional soil SDM technologies reduce the performance rate and lack of accuracy.
5. Decreased efficiency in soil data mining techniques reduces the productivity of agriculture.
6. The present soil classification strategies still need the drying process of soil samples, which causes time complexity issues.

All these limitations motivated us to move on to the proposed method where all the problems are eliminated.

4 Proposed Methodology

SDM can be used to understand spatial data, discover the relation between space and nonspace data, set up the spatial knowledge base, excel the query, reorganize spatial database, and obtain concise total characteristics [19]. SDM can be used to analyze many social aspects through geographical data. One could find an interesting fact that the regional economy has strong spatial correlation [17]. In this article, we aim to suggest

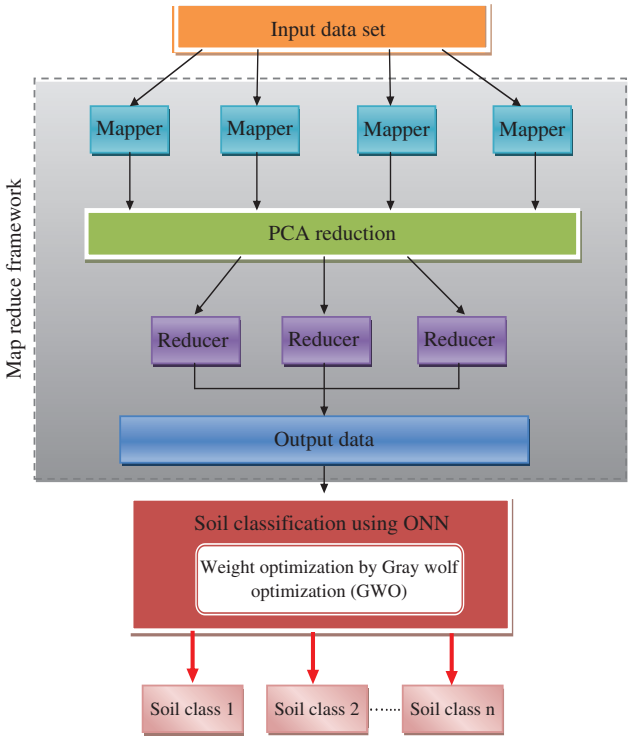

Figure 1: Semantic Structure of the Recommended Technique.

a process for soil SDM using a competent classifier. The proposed soil classification comprises preprocessing, MapReduce framework, and soil classification procedure. The proposed process employs input data such as the Harmonized World Soil Database. Actual data is habitually partial, conflicting, and/or deficient indefinite activity or movement and is possible to include numerous faults. Data preprocessing is a confirmed process of determining such concern. A superior data preprocessing assists to generate enhanced representation and will use minimum time. The next section of the proposed process is the MapReduce framework. A MapReduce framework typically divides the input dataset into a complimentary portion that is held by the map tasks in a completely equivalent manner. The scheme reduces the input data by PCA and the productivity of the maps, which are then contributed to the reduce phase. Lastly, the proposed process is used to categorize the soil kind by means of the classifier. Here, the categorization is derived from the ONN classifier. The conventional NN is customized using the optimization procedure. In an NN, the weights are optimized using the GWO algorithm. Derived from the classifier, we categorize the soil category. The general semantic structure of the recommended technique is shown in Figure 1.

In spatial data, the soil classification procedure flows in the course of the subsequent two sections: MapReduce framework and soil classification. In the MapReduce framework, the input soil data is condensed by means of PCA. Then, the condensed data is efficiently categorized by the best NN. The entire progression of the two major sections is discussed as follows.

4.1 MapReduce Framework

MapReduce is a structure for the competent distribution of the study of huge data on an outsized amount of servers. It is a similar and disseminated large-scale data processing model that has been widely investigated and extensively approved for huge data function recently. Incorporated among communications possessions provisioned by cloud scheme, MapReduce turns out to be much more dominant, flexible, and commercial because of the outstanding uniqueness of cloud computing. The MapReduce framework comprises two phases: map step and reduce step. Figure 2 illustrates the general arrangement of the MapReduce framework. There are five sections in MapReduce: Input (input involved data), Map (filtering and sorting the data), Reduction (redistribute the mapped data using PSA), Reduce (process each group of the redistricted data), and Output (collect all the reduce output). MapReduce courses are normally used to practice big files.

A MapReduce framework consists of two major phases: (1) Map phase: This phase inputs input soil data and divides it into M map task; every map task executes differently. (2) Reduce phase: In this phase, the separated soil data from the previous phase is condensed by PCA. The specific clarification of PCA is discussed in the next section.

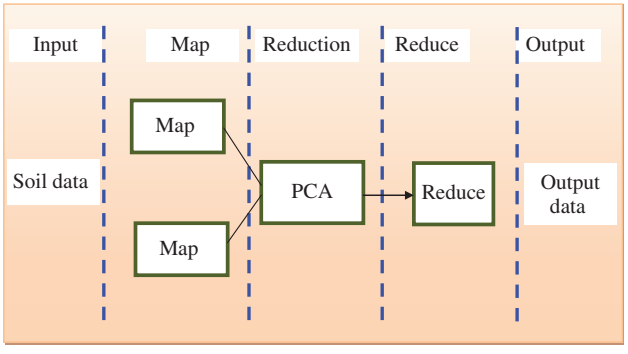

Figure 2: Flow Diagram of the MapReduce Framework.

4.2 PCA

PCA has been described as the majority of the important results from functional linear algebra. PCA is used often in every one of the figures of study, from neuroscience to computer graphics, as it is an easy, nonparametric process of removing appropriate information from confounds of data sets. By a smallest supplementary endeavor, PCA supplies a direction for how to reduce a compound dataset to a lesser measurement to expose the occasionally concealed, shortened dynamics that frequently motivates it. The foremost principles of PCA are the study of data to recognize models and discovery models to reduce the measurement of the dataset among the smallest failure of information.

Our preferred conclusion of the PCA is to offer an attribute space onto a slighter subspace that signifies our data "well". A potential function would be a model categorization task, where we want to reduce the computational expenses and the fault of limitation assessment by dropping the amount of measurement of our attribute space by removing a subspace that depicts our data "best".

4.3 Summary of the PCA Approach

Listed below are the six general steps for performing a PCA:

1. Take the whole dataset consisting of d-dimensional samples, ignoring the class labels.
2. Compute the d-dimensional mean vector (i.e. the means for every dimension of the whole dataset).
3. Compute the covariance matrix of the whole data set.
4. Compute eigenvectors ($e_1, e_2,\ldots,e_d$) and corresponding eigenvalues ($\lambda_1, \lambda_2,\ldots,\lambda_d$).
5. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d^*k-dimensional matrix M (where every column represents an eigenvector).
6. Use this d^*k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the mathematical equation: $y = M^T{}^*x$.

From the above procedure, the input soil data is reduced and then the resultant output is fed to the next phase of the MapReduce framework.

In the reduce phase, obtain every one of the condensed data from PCA and construct the productivity of the MapReduce scheme.

Lastly, the condensed productivity from the MapReduce framework is provided for to the categorization procedure. Agriculture is the field that provides food to the entire world [18]. There would be no existence of humans without agriculture. Hence, it is essential to increase the yield of agriculture. The yield of agriculture could be increased only via proper classification of soil [23]. Hence, in our suggestion scheme, the best NN is used for the soil categorization function. Each phase of the system of the best NN is discussed next.

4.4 ONN Classifier

The proposed procedure employs the best NN for soil categorization. The conventional NNs are customized by means of the GWO algorithm. The GWO algorithm is engaged to optimize the influence in the NN. The foremost intention of the ONN is to classify the input spatial soil data. For the preparation function, the backpropagation algorithm is used in our recommended procedure. NN encompasses a sequence of nodes (neurons) that obtain numerous associations among further nodes. Each one association includes an influence related to it, which can be diverse in force, in similarity through neurobiology synapses. The artificial NN (ANN) structure has an input layer and an output layer as well as hidden layers between these two layers. The amount of these layers is reliant on the difficulty we are difficult to explain, that is fundamentally on the consumer. There are two very significant sections in the NN system: the preparation section and the experiment section. The general progression of the best NN is shown next.

4.5 ONN Function Steps

1. Fix weights for every neuron, except the neurons in the input layer.
2. Develop the NN with the input soil data as the input units, H_u as the hidden units, and O_u as the output unit.
3. The computation of the proposed bias function for the input layer is

$$S = \alpha + \sum_{n=0}^{H_u - 1} w_1 S_{d1} + w_2 S_{d2} + w_3 S_{d3} + \ldots w_n S_{dn} \tag{1}$$

In the proposed best NN, the influences are optimized using the GWO algorithm. Each system of the GWO is demonstrated in the next section.

4.6 GWO Algorithm

The grey wolves adequately enclose *Canidae* species ancestors and are esteemed as the apex predators offering their location at the wherewithal's food sequence [21]. They habitually illustrate a partiality to construct suitable as a cluster. The heads represent a male and a female, marked as α, which is the majority division in incriminating of enchanting appropriate selection presenting different features; they are fundamentally supplementary wolves that effectively suggest a few support to the α in the option constructing or equivalent cluster task. The choices prepared by the α are permitted onto the group. The β conveys to the second grade in the outstanding arrange of the grey wolves. They are basically complementary wolves that adequately treaty a number of assist to the α in the option generating or corresponding group performance. The omega, which is the smallest division of the grey wolf group, by and great task as a replacement contribution into the further primary wolves very practically on each occasion and is acceptable to obtain just the diminutive leftovers enchanting following an enormous blowout by the organizer wolves. In the GWO process, the tracking (optimization) is directed by then α, β, δ, and ω. For picking the best influence, the proposed procedure employs the GWO algorithm. The pseudo code for the customized GWO algorithm is demonstrated below.

Pseudo code for GWO
Begin
Step 1: Initialize the random weights $w_i = (i = 1, 2, 3, \ldots, n)$
  Initialize a, A, and C are the coefficient vectors
Step 2: Find the fitness of the initial weight
  $Fit_i = \min MSE$
Step 3: Separate the solution based on the fitness
  w_α = first best search weight
  w_β = second best search weight
  w_δ = third best search weight
While ($t <$ max number of iteration)
  For each search weight
Step 4: Update the position of the current search weight

$$w_p(t+1) = \frac{w_{p1} + w_{p2} + w_{p3}}{3}$$

  End for
Step 5: Calculate the fitness of the new search weight
Step 6: Update w_α, w_β, and w_δ
Step 7: Store the best weight so far attained
Iteration = Iteration + 1
End while
Stop

The step-by-step process of the GWO algorithm is mentioned below.

Step 1: Initialization process
Initialize the input random weights and a, A, and C as coefficient vectors.

Step 2: Fitness evaluation
Evaluate the fitness performance based on Eq. (2) and then pick the best result.

$$Fit_i = \min \text{MSE} \tag{2}$$

Step 3: Separate the solution based on the fitness
Now, determine the different result on the foundation of the fitness value. Let the first best fitness results be w_α, the second best fitness results w_β and the third best fitness results w_δ.

Step 4: Update the position
We presume that the α, β, and δ obtain the enhanced facts about the probable position of the prey to replicate precisely the tracking activities of the grey wolves. Because of the result, we accumulate the primary three best influences accomplished until now and necessitate the further search influence (including the omegas) to modify their situation along with the situation of the best search influence. For replication, the innovative weight $w_p(t+1)$ is predictable by the formulas below:

$$\vec{K} = |\vec{C}.w_n(t+1) - w_n(t)| \tag{3}$$

where $\vec{K}^\alpha = |\vec{C}_1.w_{p\alpha} - w_p|$, $\vec{K}^\beta = |\vec{C}_2.w_{p\beta} - w_p|$, $\vec{K}^\delta = |\vec{C}_3.w_{p\delta} - w_p|$

$$w_p(t+1) = \frac{w_{p1} + w_{p2} + w_{p3}}{3} \tag{4}$$

where $w_{p1} = w_{p\alpha} - \vec{A}_1 \cdot (\vec{K}^\alpha), w_{p2} = w_{p\beta} - \vec{A}_2 \cdot (\vec{K}^\beta), w_{3p} = w_{p\delta} - \vec{A}_3 \cdot (\vec{K}^\delta)$

$$\vec{A} = 2\vec{a}r_1 - \vec{a} \text{ and } \vec{C} = 2r_2 \tag{5}$$

where t is the iteration number, $p(t)$ is the prey location, A and C are the coefficient vector, $\vec{a}$ is linearly reduced from 2 to 0, and r_1 and r_2 are the random vector [0, 1]. It can be distinguished that the final influence would be in an arbitrary position prearranged a circle, which is precise by the position of α, β, and δ in the investigate break. It also meant by α, β, and δ evaluate the location of the prey and supplementary wolves update their location randomly in the region of the prey. Examination and utilization are specific by means of the adaptive values of a and A. The adaptive values of limitation a and A allow GWO to effortlessly changeover in the middle of examination and utilization. By retreating A, half of the iterations are dedicated to the examination ($|A| < 1$) and the other half is devoted to the convention. Attaching the demeanor, the following equations are used keeping in mind the conclusion objective to provide arithmetical representation.

Step 5: Fitness calculation
Calculate the fitness of the new search weight using Eq. (2) and then store the best solution.

Step 6: Stopping criteria
Replicate Steps 3–5, awaiting an improved fitness or highest amount of iterations are collected. Derived from on top of the declared procedure, accomplish the best influence. After that, the best influence is used for the additional procedure.

– The activation function for the output layer is estimated as

$$\text{Active } (S) = \frac{1}{1 + e^{-S}} \tag{6}$$

– Recognize the learning error as follows:

$$\text{Output}(O_u) = \frac{1}{2}\sum_{n=0}^{H_u-1}(\text{Desired}_n - \text{Actual}_{n'})^2 \tag{7}$$

where Desired_n is desired output and Actual_n is actual output.

In the best NN, the mistake should be in the least value. Subsequently, the qualified NN is well qualified for presenting the experiment stage. Derived from the qualified data, the NN arrangement efficiently categorized the soil category. The analysis result of the proposed procedure is discussed in Section 5.

5 Results and Discussion

This section offers the comprehensive results of the effect obtained from the proposed soil categorization in SDM, which is carried out in the functioning platform of Java within the MapReduce framework using Hadoop. The analysis outcome and the performance of the proposed process are discussed in the subsequent sections.

5.1 Dataset Description

The proposed process was investigated by the Harmonized World Soil Database version 1.2.

The Harmonized World Soil Database is a 30-arc second raster database with more than 15,000 different soil mapping components that merge local and countrywide updates of soil information worldwide (SOTER, ESD, Soil Map of China, WISE) from the information restricted within the 1:5,000,000 scale FAO-UNESCO Soil Map of the World. The resulting raster database consists of 21,600 rows and 43,200 columns, which are linked to harmonized soil property data. The dataset is available at http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/.

5.2 Evaluation Metrics

The assessment metrics for the performance of the proposed scheme are sensitivity, specificity, accuracy, precision, recall, and F-measure. The standard count values such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are presented.

5.2.1 Sensitivity

The ratio of the number of TPs to the sum of TP and FN is called sensitivity.

$$\text{Sensitivity} = \frac{\text{No of TP}}{\text{No of TP} + \text{No of FN}} \times 100 \tag{8}$$

5.2.2 Specificity

Specificity is defined as the ratio of the number of TNs to the sum of TNs and FPs.

$$\text{Specificity} = \frac{\text{No of TN}}{\text{No of TN} + \text{No of FP}} \times 100 \tag{9}$$

5.2.3 Accuracy

Accuracy can be calculated using the measures of sensitivity and specificity.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \tag{10}$$

5.2.4 Precision

Precision means the nearness of two or more dimensions to each other. It is defined as the ratio of TP to the sum of TP and FP.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{11}$$

5.2.5 Recall

Recall is defined as the ratio of TP to the sum of TP and FN.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{12}$$

5.2.6 F-measure

F-measure comprises the harmonic mean of the combinations of precision and recall.

$$\text{F-measure} = 2 * \frac{\text{Precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

5.3 Performance Analysis

The proposed procedure is discussed in the next section. Table 1 illustrates the performance analysis of the proposed procedure. The sensitivity, specificity, accuracy, precision, and recall, and F-measure values are presented. In our procedure, 90% of the input data is allowed for preparation and the outstanding 10% is allowed for the experiment. The performance analysis of the proposed procedure with varying data size is discussed below.

In Table 1, using the proposed process for the data dimension of 1000, the sensitivity value is 0.89676%, the specificity value is 0.78947368%, the accuracy value is 0.9046%, the precision value is 0.98979%, the

Table 1: Performance Analysis by Varying Data Size.

Data size	Sensitivity	Specificity	Accuracy	Precision	Recall	F-measure
1000	0.8967	0.7894	0.9046	0.9897	0.8967	0.9409
2000	0.8657	0.8181	0.9138	0.9489	0.8657	0.9054
3000	0.8950	0.75	0.9107	0.9797	0.8950	0.9354
4000	0.8355	0.784	0.8815	0.9447	0.8355	0.8867
5000	0.8416	0.7777	0.8639	0.9721	0.8416	0.9022

Table 2: Performance Analysis by the Cross-Validation Method.

Cross-validation	Testing percentage	Sensitivity	Specificity	Accuracy	Precision	Recall	F-measure
1	30	0.82265	0.75	0.854769	0.9638	0.826	0.886443
2	20	0.85604	0.775862	0.863999	0.98095	0.85265	0.91817
3	10	0.89676	0.78947368	0.9046	0.98979	0.89676	0.940986

recall value is 0.89676%, and the F-measure value is 0.940986%. For the data dimension of 2000, the sensitivity value is 0.865787%, the specificity value is 0.81818%, the accuracy value is 0.913846%, the precision value is 0.948979%, the recall value is 0.865787%, and the F-measure value is 0.905476%. For the data dimension of 3000, the sensitivity value is 0.89506%, the specificity value is 0.75%, the accuracy value is 0.9797%, the precision value is 0.9797%, the recall value is 0.89506%, and the F-measure value is 0.935484%. For the data dimension of 4000, the sensitivity value is 0.83553%, the specificity value is 0.78409%, the accuracy value is 0.881538%, the precision value is 0.944737%, the recall value is 0.8355%, and the F-measure value is 0.88678%. For the data dimension of 5000, the sensitivity value is 0.8416499%, the specificity value is 0.77777%, the accuracy value is 0.86399%, the precision value is 0.972163%, the recall value is 0.841649%, and the F-measure value is 0.902211%.

Table 2 presents the performance analysis by the cross-validation method with a steady data dimension of 1000. For cross-validation 1, the sensitivity value is 0.82265%, the specificity value is 0.75%, the accuracy value is 0.854769%, the precision value is 0.9638%, the recall value is 0.826%, and the F-measure value is 0.886443% for the preparation percentage of 70 and the experiment percentage of 30. For cross-validation 2, the sensitivity value is 0.85604%, the specificity value is 0.775862%, the accuracy value is 0.863999%, the precision value is 0.98095%, the recall value is 0.85265%, and the F-measure value is 0.91817% for the preparation percentage of 80 and the experiment percentage of 20. For cross-validation 3, the sensitivity value is 0.89676%, the specificity value is 0.78947368%, the accuracy value is 0.9046%, the precision value is 0.98979%, the recall value is 0.89676%, and the F-measure value is 0.940986% for the preparation percentage of 90 and the experiment percentage of 10. The average accuracy value for the cross-validation method is 87.44%.

5.4 Effectiveness of the Suggested Technique

In this section, the efficiency of the proposed procedure is compared to a further obtainable procedure. The detail clarification is demonstrated in Figure 3.

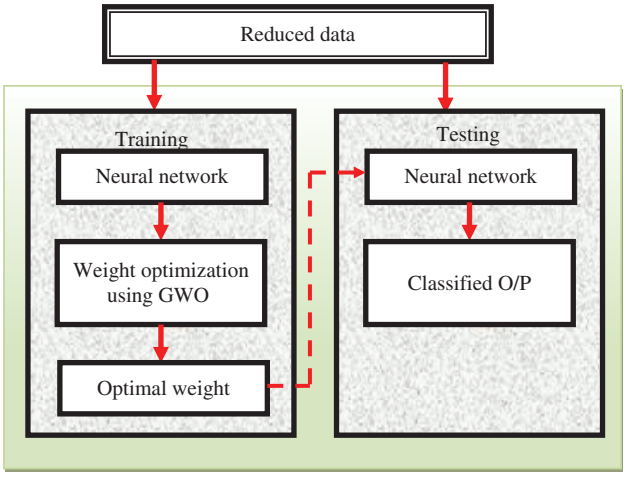

Figure 3: Overall Process of the ONN.

5.5 Comparison Analysis of Performance Metrics

The accuracy value is the foremost feature in the soil categorization scheme. It is significant for the process to offer the elevated accuracy value to provide the best process. Figure 4 illustrates the proportional study for the accuracy value by the obtainable categorization process. We allow the obtainable categorization scheme as the conventional NN and *k*-nearest neighbor (KNN) classifier.

Figure 4 illustrates the accuracy of the proposed ANN-GWO compared to the accuracy of the obtainable classifiers such as NN and KNN. The conventional NN classifier has 75.3846% accuracy and the KNN classifier has 75.38% accuracy, but the proposed process has 90.46% accuracy, which is the highest value compared to the obtainable procedure.

Figure 5 illustrates the sensitivity of the proposed ANN-GWO compared to the sensitivity of the obtainable classifiers such as NN and KNN. The conventional NN classifier has a sensitivity value of 0.75% and the KNN classifier has 0.7538% sensitivity value, but the proposed classifier ANN-GWO has 0.89676% sensitivity value, which is the highest value compared to the obtainable classifier procedure.

Figure 6 illustrates the specificity of the proposed ANN-GWO compared to the specificity of the obtainable classifiers such as NN and KNN. The NN classifier has a specificity value of 0.77166% and the KNN classifier has 0.811269% sensitivity value, but the proposed classifier ANN-GWO has 0.789474% sensitivity value. Compared to the obtainable procedure, the recommended procedure has the least specificity value. Although it has the least specificity value, the categorization accuracy is the highest value compared to the further categorization procedure.

Figure 7 illustrates the precision and recall of the proposed ANN-GWO compared to the precision and recall of the obtainable classifiers such as NN and KNN. The precision value of the NN classifier is 0.8229% and the KNN classifier has 0.81887% precision value, but the proposed classifier ANN-GWO has 0.98979% precision value. Compared to the obtainable process, the proposed categorization procedure has the highest

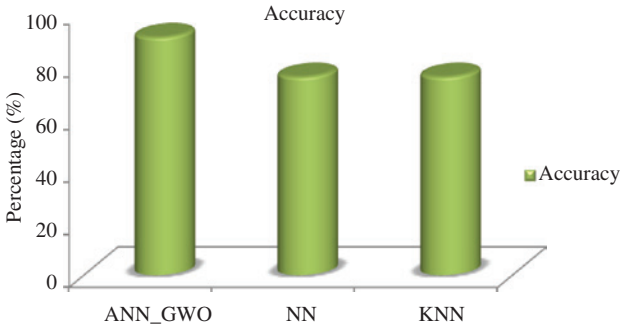

Figure 4: Accuracy Comparison for Different Classifiers.

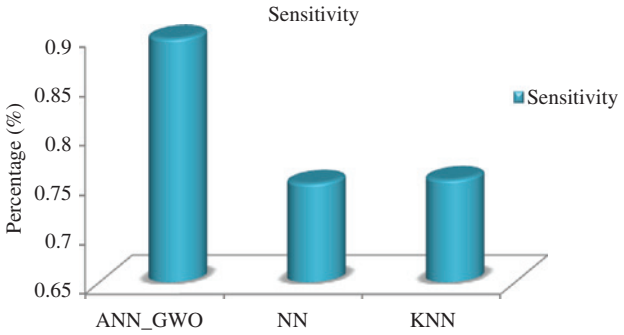

Figure 5: Sensitivity Comparison for Different Classifiers.

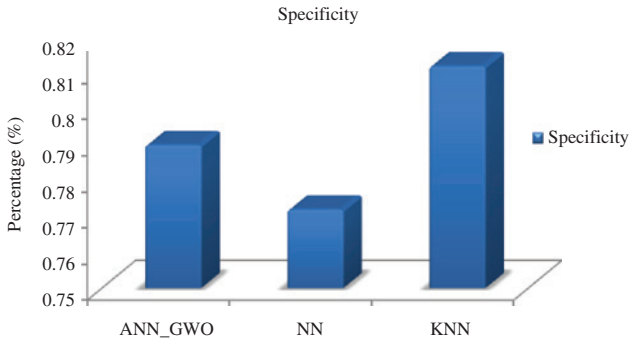

Figure 6: Specificity Comparison for Different Classifiers.

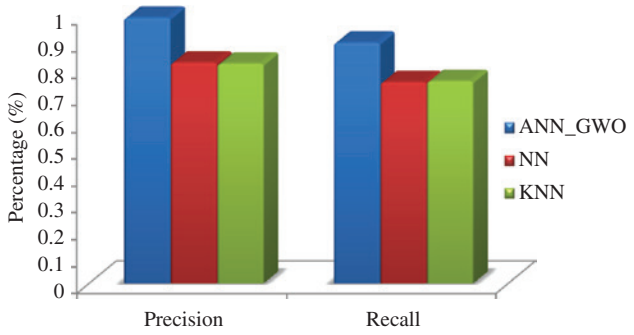

Figure 7: Precision and Recall Comparison for Different Classifiers.

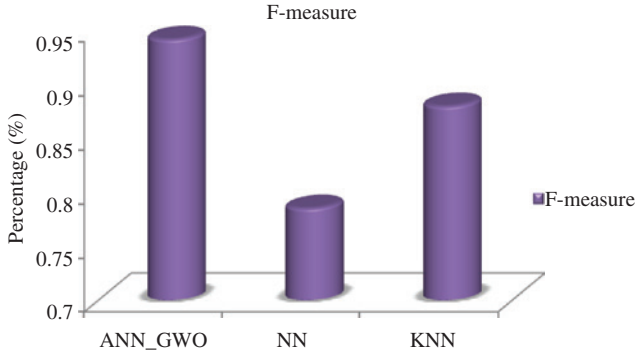

Figure 8: F-Measure Comparison for Different Classifiers.

precision value. The recall value of the NN classifier is 0.75% and the recall value of the KNN classifier is 0.7538%, whereas the proposed classifier has 0.89676% recall value, which is better than the obtainable classifiers.

Figure 8 illustrates the F-measure value of the proposed ANN-GWO compared to the F-measure value of the obtainable classifiers such as NN and KNN. The NN classifier has 0.784768% F-measure value and the KNN classifier has 0.8792% F-measure value, but the proposed classifier has 0.940986% F-measure value, which is better than the obtainable classifiers.

In Figure 9, the proposed method compares the result for PCA with and without eigenvector k value. The performance is evaluated by accuracy, sensitivity, and specificity values. Here, the proposed PCA with eigenvector has an accuracy value of 90.46%, the sensitivity value is 0.89676%, and the specificity value is

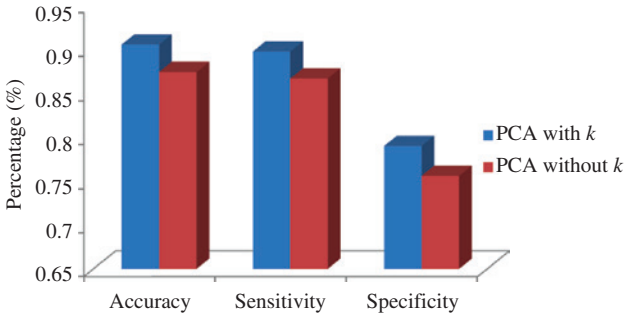

Figure 9: Comparison Results for PCA with and without *k* Value.

Table 3: Comparison Results for Soil Classification with and without Optimization.

Metrics	Soil classification with optimization	Soil classification without optimization
Accuracy	90.46	88.72
Sensitivity	0.8967	0.7958
Specificity	0.8112	0.7643
Precision	0.9897	0.8564
Recall	0.8967	0.7958
F-measure	0.9409	0.8451

0.789%, but the proposed system PCA without eigenvector has an accuracy value of 87.35%, the sensitivity value is 0.8661%, and the specificity value is 0.7556%, which is the minimum value compared to PCA with eigenvector. In the graph, PCA with *k* value improves the proposed performance compared to PCA without *k* value. The comparison result for with and without optimization is shown in Table 3.

Table 3 shows the comparison result for soil classification performance with and without optimization. It is clearly shown that the proposed technique with the optimization method has a better result compared to that without the optimization method. Here, the accuracy value of the proposed soil classification with optimization is 90.46% but that without the optimization method is 88.72%, which is minimum value compared to the proposed technique. From the result, the proposed method concludes that classification with the optimization method outperforms the classification without optimization.

The proposed performance is compared to the existing technique [7] in Figure 10. For comparison, the proposed technique considers the existing support vector machine (SVM) and PCA with SVM.

Figure 10 illustrates the accuracy value of the proposed ANN-GWO compared to the accuracy value of the obtainable classifiers such as SVM and PCA + SVM.

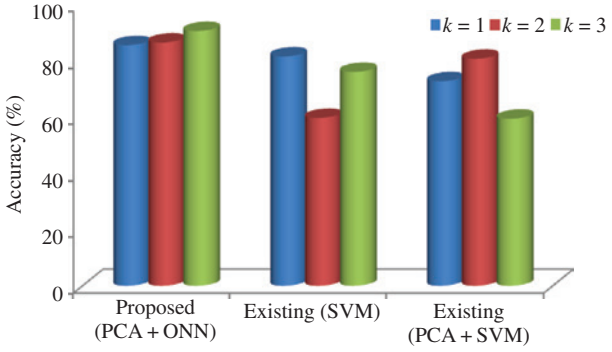

Figure 10: Comparison Results for Various Existing Methods.

For cross-validation 1, the accuracy value of the SVM classifier is 81.44% and PCA + SVM classifiers have 72.63% accuracy value, but the proposed classifier ANN-GWO has 85.47% accuracy value.

For cross-validation 2, the accuracy value of the SVM classifier is 59.61% and PCA + SVM classifiers have 80.61% accuracy value, but the proposed classifier ANN-GWO has 86.399% accuracy value.

For cross-validation 3, the proposed classifier ANN-GWO has 90.46% accuracy value, but the existing classifier has 75.94% for the SVM classifier and 59.31% for the PCA + SVM classifier. Compared to the obtainable process, the proposed categorization procedure has the highest accuracy value. As a result, the recommended procedure has the highest result compared to the obtainable classifier for spatial data soil categorization.

6 Conclusion

An effectual soil categorization in SDM in excess of huge data is proposed in this article. The execution is made by Java within the MapReduce framework using Hadoop. Originally, the input soil data is preprocessed and then the resulting productivity is supplied to the MapReduce framework. Finally, the condensed soil data is categorized using the best NN. The performance of the proposed procedure is assessed by means of sensitivity, specificity, accuracy, precision, recall, and F-measure. The analysis results illustrate that the recommended ANN-GWO process has an accuracy of 90.46%, but the conventional NN and KNN classifiers have an accuracy value of 75.3846% and 75.38%, respectively, which is the least value compared to the proposed procedure. In the future, investigators can use a variety of categorization procedures to accomplish superior quality in performance.

Bibliography

[1] F. R. Ajdadi, Y. A. Gilandeh, K. Mollazade and R. P. R. Hasanzadeh, Application of machine vision for classification of soil aggregate size, *Elsevier Soil Tillage Res.* **162** (2016), 8–17.

[2] D. Badía, C. Martí, J. M. Aznar and J. León, Influence of slope and parent rock on soil genesis and classification in semiarid mountainous environments, *Elsevier Geoderma* **193** (2013), 13–21.

[3] J. Baruck, O. Nestroy, G. Sartori, D. Baize, R. Traidl, B. Vrščaj, E. Bräm, F. E. Gruber, K. Heinrich and C. Geitner, Soil classification and mapping in the Alps: the current state and future challenges, *Elsevier Geoderma* **264** (2016), 312–331.

[4] E. Bol, The influence of pore pressure gradients in soil classification during piezocone penetration test, *Elsevier Eng. Geol.* **157** (2013), 69–78.

[5] E. C. Brevik, C. Calzolari, B. A. Miller, P. Pereira, C. Kabala, A. Baumgarten and A. Jordán, Soil mapping, classification, and pedologic modeling: history and future directions, *Elsevier Geoderma* **264** (2016), 256–274.

[6] C. Candeias, P. F. Ávila, E. F. Da Silva and J. P. Teixeira, Integrated approach to assess the environmental impact of mining activities: estimation of the spatial distribution of soil contamination (Panasqueira mining area, Central Portugal), *Elsevier Environ. Monit. Assess.* **187** (2015), 1–23.

[7] M. Fauvel, J. Chanussot and J. A. Benediktsson, Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas, *J. Adv. Signal Process.* **2009** (2009), 1–14.

[8] M. Fuchs, V. Láng, T. Szegi and E. Michéli, Traditional and pedometric approaches to justify the introduction of swelling clay soils as a new soil type in the modernized Hungarian Soil Classification System, *Elsevier Catena* **128** (2015), 80–94.

[9] D. R. Gambill, W. A. Wall, A. J. Fulton and H. R. Howard, Predicting USCS soil classification from soil property variables using random forest, *Elsevier J. Terramech.* **65** (2016), 85–92.

[10] P. Han, D. Dong, X. Zhao, L. Jiao and Y. Lang, A smartphone-based soil color sensor: for soil type classification, *Elsevier Comput. Electronics Agric.* **123** (2016), 232–241.

[11] P. A. Hughes, A. B. McBratney, B. Minasny and S. Campbell, End members, end points and extragrades in numerical soil classification, *Elsevier Geoderma* **226** (2014), 365–375.

[12] J. Ö. G. Jónsson and B. Davíðsdóttir, Classification and valuation of soil ecosystem services, *Elsevier Agric. Syst.* **145** (2016), 24–38.

[13] A. V. Krishna Prasad, S. Rama Krishna, D. Sravan Kumar, K. Suresh and I. S. Ravi Varma, Spatial data mining using novel neural networks for soil image classification and processing, *Int. J. Recent Trends Eng. Technol.* **3** (2010), 156–159.

[14] A. Kumar and N. Kannathasan, A survey on data mining and pattern recognition techniques for soil data mining, *Int. J. Comput. Sci. Issues* **8** (2011), 422–428.
[15] C. S. Lee, T. M. Sung, H. S. Kim and C. H. Jeon, Classification of forensic soil evidences by application of THM-PyGC/MS and multivariate analysis, *Elsevier J. Anal. Appl. Pyrol.* **96** (2012), 33–42.
[16] R. W. Lourenço, P. M. B. Landim, A. H. Rosa, J. A. F. Roveda, A. C. G. Martins and L. F. Fraceto, Mapping soil pollution by spatial analysis and fuzzy classification, *Elsevier Environ. Earth Sci.* **60** (2010), 495–504.
[17] A. Manjula and G. Narsimha, A review on spatial data mining methods and applications, *Int. J. Comput. Eng. Appl.* **7** (2014), 208–218.
[18] A. Manjula and G. Narsimha, Towards precision agriculture: a review of the present state-of-the-art, *Int. Conf. Rough Sets Knowl. Technol.* (2014), 74–78.
[19] A. Manjula, G. Narsimha and S. Katherapaka, Spatial data mining: a recent survey and new discussions, *Int. J. Comput. Sci. Inf. Technol.* **2** (2011), 1501–1504.
[20] E. Michéli, V. Láng, P. R. Owens, A. McBratney and J. Hempel, Testing the pedometric evaluation of taxonomic units on soil taxonomy – a step in advancing towards a universal soil classification system, *Elsevier Geoderma* **264** (2015), 340–349.
[21] S. Mirjalili, S. M. Mirjalili and A. Lewis, Grey wolf optimizer, *J. Adv. Eng. Softw.* **69** (2014), 46–61.
[22] T. W. Nauman and J. A. Thompson, Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees, *Elsevier Geoderma* **213** (2014), 385–399.
[23] V. Rajeswari and K. Arunesh, Analysing soil data using data mining classification techniques, *Indian J. Sci. Technol.* **9** (2016), 1–4.
[24] A. A. Shahri, A. Malehmir and C. Juhlin, Soil classification analysis based on piezocone penetration test data – a case study from a quick-clay landslide site in southwestern Sweden, *Elsevier Eng. Geol.* **189** (2015), 32–47.
[25] J. D. Sitton and B. A. Story, Estimating soil classification via quantitative and qualitative field testing for use in constructing compressed earth blocks, *Elsevier Proc. Eng.* **145** (2016), 860–867.
[26] G. M. Vasques, J. A. M. Demattê, R. A. V. Rossel, L. Ramírez-López and F. S. Terra, Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths, *Elsevier Geoderma* **223** (2014), 73–78.
[27] B. K. Waruru, K. D. Shepherd, G. M. Ndegwa and A. M. Sila, Estimation of wet aggregation indices using soil properties and diffuse reflectance near infrared spectroscopy: an application of classification and regression tree analysis, *Elsevier Biosyst. Eng.* **152** (2016), 148–164.