Hybrid Regression Model for Soil Moisture Quantity Prediction

Sankhadeep Chatterjee
A.K.Choudhury School of Information Technology
University of Calcutta
Kolkata, India
chatterjeesankhadeep.cu@gmail.com

Sanmitra Kumar
Department of Information Technology
Academy of Technology
Aedconagar, India
sanmitra.kumar@aot.edu.in

Jeet Saha
Department of Information Technology
Academy of Technology
Aedconagar, India
jeet.saha@aot.edu.in

Soumya Sen
A.K.Choudhury School of Information Technology
University of Calcutta
Kolkata, India
iamsoumyasen@gmail.com

Abstract— **Predicting quantity of soil moisture would help the farmers who are involved in agriculture. Recently researchers have used various machine learning algorithms to predict the quantity of soil moisture. In this study a hybrid method of different types of regression algorithms have been employed for prediction. The proposed method first clusters the dataset using K-Means Clustering, then for each clusters an individual regression model is trained. The proposed method is compared with other regression models like DT (Decision Tree), MLP (Multilayer Perceptron)-Regressor, LR (Linear Regression), KNN-Regressor and SVM (Support Vector Machine Regression). Experimental results have shown that Hybrid model using Decision Tree Regression achieved an average RMSE of 0.002924 and outperformed other models.**

Keywords—soil moisture, hybrid regression model, k-means

I. INTRODUCTION

Good quality of raw materials is an important aspect on which all the agricultural industries are depended. A major part of these industries main resources are crop and vegetables. These also plays an important role in economy associated with it. With the growing up of more and more agriculture based industries and with the increasing demand of food, the total yield of crops and vegetables needs to increase simultaneously. There are different factors that affect the production of good quality and quantity of crops. So it is important to understand those factors properly in order to maximise the production. Mainly, there are two types of factors; one of which is genetic factors that refers to the genes, the chromosomes affecting the plants growth and development. In other words, plants displays unique traits that they inherits from their parent plants. The second factor is the environmental factors like climate, soil moisture, radiating energies, etc. Many researches have shown that for a healthy production, amount of soil moisture is one the most important factors [1]. Growth of plant cells are affected by the scarcity soil moisture, which reduces the plant growth. Conversely excessive amount of soil moisture cause reduction in growth at certain stages of plant life. So, it is clear from these statements that proper prediction of soil moisture is one of the important parameter affecting the agricultural production. The key challenge in agriculture is to optimise the uses of natural resources in order to meet the increasing demand of crops and vegetables. Best solution for optimising would be to integrate technological approaches for efficient use of both renewable and non-renewable resources. At the same time the system must be easy to implement and use [2]. With the advancement of technologies in agricultural resources, it is assumed that the use of pests and fertilizers, different irrigation techniques, machineries will increase over time. Along with these, efficient use of many non-renewable resources such as water, is important [3]. Recently, an automated aerial device has been developed for water stress management system, which is being deployed in agricultural fields. The system uses spatial and temporal sensing devices to monitor water stress remotely [4]. High water stress at certain stage of growth may have a harmful effect on various aspects of certain crops and vegetables, which later on cannot be compensated by a low water stress [5]. The above mentioned discussion reveals that soil moisture plays a vital role in agricultural fields. So, soil quality and prediction of soil moisture has become a topic of research.

Majority of the literature relies upon simple regression models to predict soil moisture quantity. In the current study, we propose a hybrid regression model framework in order to enhance the performance of existing regression models for soil moisture quantity. Initially, it is proposed as a hypothesis. Experimental studies have revealed that the proposed hybrid model is capable of enhancing the performance of regression models to a greater extent.

The rest of the paper is organised as follows: Section II reports related works. Next in section III the propsed method is explained. In section IV, experimental setup is reported along with dataset description. Finally, in section V, the results are reported.

II. RELATED WORK

Several ecological and statistical models have been proposed. Gill et al. [6] proposed Support Vector Machine technique for predicting soil moisture. The study reveals the inability of backpropagation based ANNs by comparing it with the proposed SVM based model in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Results are in favour of SVM, which could be a suitable choice over ANNs in predicting soil moisture quantity. Song et al. [7] proposed Feature Selection based ANN for soil moisture quantity prediction. The study revealed that the previously proposed methods considered all types of possible attributes, those are less important towards prediction of soil moisture quantity. Correlation measure supported Environmental Distribution Map was deployed to find the redundant features in order to overcome the problem. By measuring the RMSE for different set of features, the optimal features were selected. Then a multilayer feedforward network based on back-propagation algorithm was trained and tested to predict soil moisture quantity. Esmaeelnejad et al. [8] in an article proposed a pedotransfers function to predict soil moisture using ANN. Oliviu Matei [9] in study of real time soil moisture prediction used all types of Regression algorithms such as k-NN, SVM, ANN, Linear Regression, Decision Tree. The study revealed that model reached an average accuracy of 68.65% and RMSE is 0.033857. Recently, R. Pugazendi [10] proposed a hybrid model k-means clustering coupled with multilayer perceptron to predict rainfall. The literature survey revealed that most of the previously reported models have used basic forms of ANNs and other regression techniques which are not that fit to tackle with the problem of predicting soil moisture. To improve the performance of regression models hybrid models can be applied.

III. Proposed Method

The current study proposes a hybrid regression analysis method in order to improve the performance of regression models. The proposed hypothesis is motivated by the fact that instead of performing the regression analysis on a given dataset, it would be better if the dataset is clustered in to natural groups and then for each of those clusters a separate regression model is applied. Algorithm 1 explains the aforementioned concept. '*D*' denotes the initial dataset. clustering_algorithm() depicts any clustering algorithm. After applying the unsupervised clustering algorithm, it returns a set of cluster centers (C) where $|C|$ is at least 2 and set of set of data points P where each member P_i is the set of data points of ith cluster. In step 1, the clustering algorithm is applied on the initial dataset. In step 2, each set of data (P_i) is used to train a separate regression model (R_i). The train_regressor() can be any regression model training phase. Figure 1 depicts the working principle of the algorithm.

Algorithm 1: Hybrid Regression Model Training
Input: Initial Dataset (D)
Output: $C = \bigcup_i C_i, i \geq 2$ where C_i is cluster center of ith cluster, $P = \bigcup_i P_i$, where $P_i$is the set of data points in ith cluster, $R = \bigcup_i R_i, i \geq 2$ where R_i is a regressor trained with data points of ith cluster.

```
1. $P, C$ = clustering_algorithm($D$) ⊳ clustering_algorithm() is any unsupervised clustering algorithm
2. for each $P_i$ in $P$ do
     i. $R_i$ = train_regressor($P_i$) ⊳ train_regressor() trains a regression model using data points in $P_i$
3. End
```

Algorithm 2 explains the testing phase of the hybrid regression model. In step (i) of 1, we calculate the nearest cluster center of the test data. $\|T_i - C_x\|$ depicts the distance between T_i and xth cluster center C_x. Next, in step (ii), the corresponding regression model is used to calculate the predicted value of the target. Finally, in step 2, a performance metric is calculated. In the current study, Root Mean Square Error (RMSE) is used as performance measure. Figure 2 depicts the working principle of the whole proposed model.

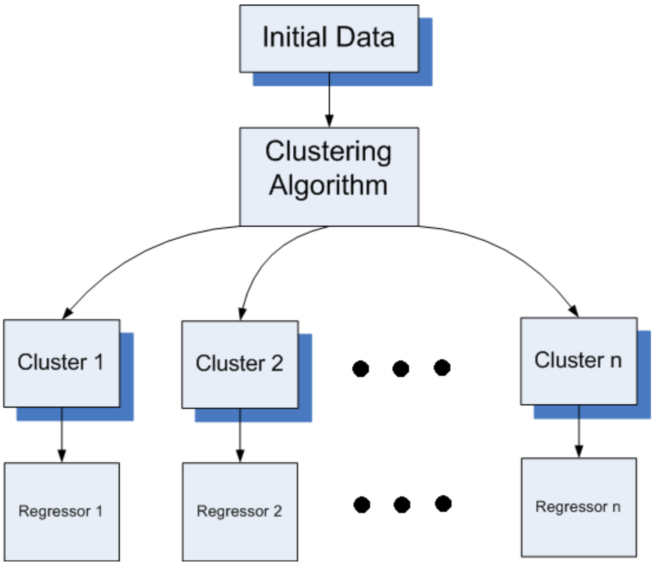

Fig. 1. Hybrid Regression Model training phase

Algorithm 2: Hybrid Regression Model Testing
Input: Testing Dataset (T)
Output: Performance metric

```
1. for each $T_i$ in $T$ do
     i. $\kappa = \arg\min_x(\|T_i - C_x\|)$
     ii. Predict target value using $R_\kappa$
2. Calculate performance metric
3. End
```

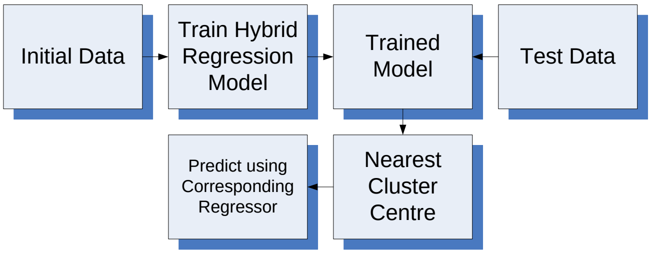
Initial Data
Train Hybrid Regression Model
Trained Model
Test Data
Predict using Corresponding Regressor
Nearest Cluster Centre

Fig. 2. Proposed Model working principle

IV. Experimental Setup

A. *Dataset Description*

UNIVERSITY OF TORONTO, Mississauga, Department Of Geography [11] surveyed for more or less 10 years on the nature of soil. The dataset from the survey represent measurements taken in 3 distinct environmental settings. The dataset is separated into 3 types soil such as field, forest, pond. The university employed HOBO U30 data logger coupled with sensors monitoring soil moisture, soil temperature, air temperature and relative humidity. The dataset of the month of February of 2010 has been used for this study.

The data set consists of 16 attributes. Among them, first four are the date & time on which the reading has been taken. Next twelve attributes are separated into three groups, four attributes in each group namely, field, forest and pond respectively. Four attributes in each group are a) Soil temperature at the field site in degrees Celsius from a sensor buried 30cm below surface, b) Air temperature at the field site in degrees Celsius, c) Relative humidity in percentage(%) at the field site, d) Soil water content in m3/m3 recorded at the field site by a sensor 30cm below soil surface.

V. Result and Discussion

Choosing the right value of K in k-means clustering is the most important aspect in getting the best cluster results. So, the foremost task in our study was to choose the k value. Starting with an initial k value of 2, it has been varied between 2 to 18 at interval of 4. For each k value the resultant RMSE value of the hybrid model has been calculated. The table 1 shows the variation in RMSE value with the increasing number of clusters i.e. with increment in the k value. It is clear from the result, that with increasing the number of clusters, the error i.e. the RMSE is minimized. But with further increment in k value in case of Decision Tree, the RMSE value increases. The dependency of RMSE with k value can be better understood from the Fig. 3, Fig. 4 and Fig. 5. For this study k value has been chosen as 10 as it gives the best result.

TABLE I. RMSE vs. Number Of Clusters

	2	6	10	14	18
MLP	1.826	1.708	1.512	1.661	1.445
DT	0.003017	0.002852	0.0025	0.003346	0.003798
LR	0.00678	0.005381	0.0045	0.00384	0.003699
KNN	0.006412	0.006007	0.0059	0.006162	0.005555
SVM	0.01956	0.016453	0.0126	0.011341	0.009735

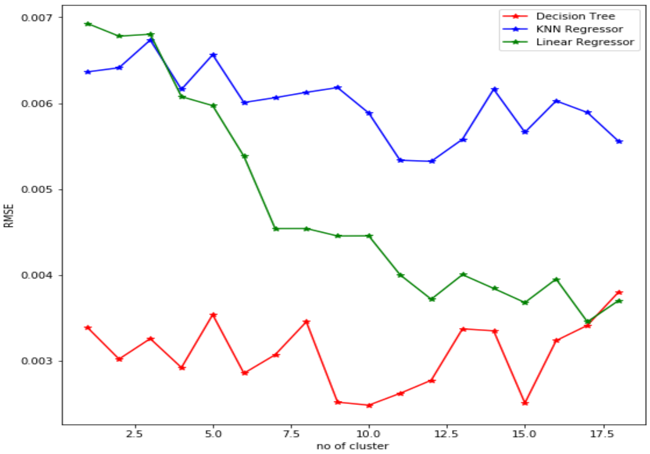

Fig. 3. RMSE vs. No. of clusters of KNN, Decision Tree, Linear Regression Model

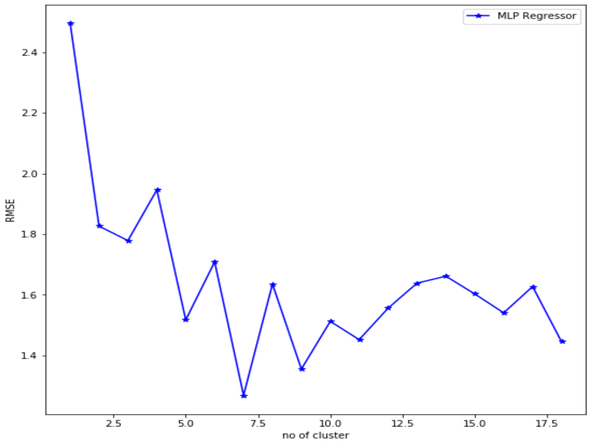

Fig. 4. RMSE vs. No. of clusters of MLP Regression Model

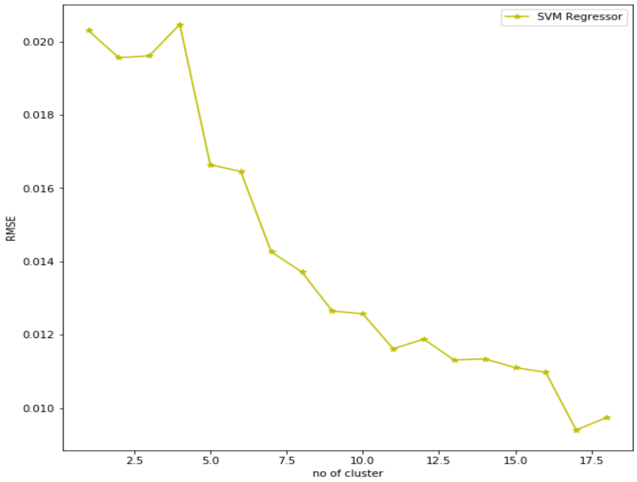

Fig. 5. RMSE vs. No. of clusters of SVM Regression Model

All the regression model has been compared has been compared with the hybrid approach of that regression model. Each regression model has been iterated for 10 times and finally the average RMSE of the 10 iterations has been considered. Similarly, the hybrid model of each regressor has been iterated 10 times to obtain the average RMSE. From Table I, it is clear that for every regression model, the hybrid model has outperformed its base model. A graphical representation of the comparison is shown in the Fig. 6.

TABLE I. RMSE Of Base And Hybrid Model

	Base	**Hybrid**
MLP	2.175	1.297
DT	0.003388	0.002617
LR	0.005498	0.00407
KNN	0.006365	0.005972
SVM	0.020303	0.013747

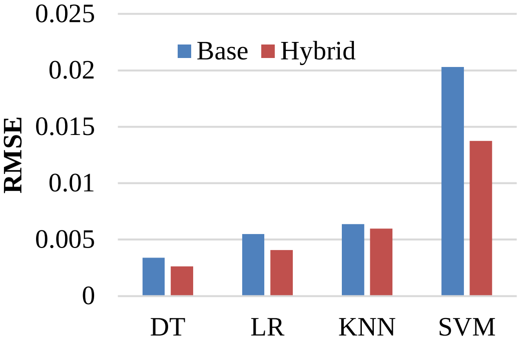

Fig. 6. Comparison between base and hybrid model of DT, LR, KNN, SVM

The current study reveals that decision tree regression model has outperformed all other regression model. So, decision tree regression has been coupled with the hybrid model to predict the soil moisture quantity. The hybrid model of decision tree outperforms the base model of decision tree and also the other regression models. The average RMSE of the hybrid model is calculated as 0.002924 which is than other models. The Table II shows the RMSE values of all the 10 iterations and the average RMSE of the hybrid decision tree model and other base models. The actual and predicted soil moisture quantity of both the normal decision tree and its hybrid model has been represented graphically in Fig. 7 and Fig. 8 respectively.

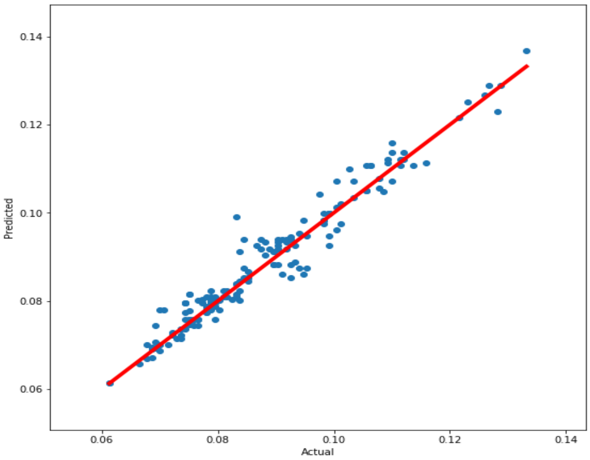

Fig. 7. Hybrid Decision Tree Model

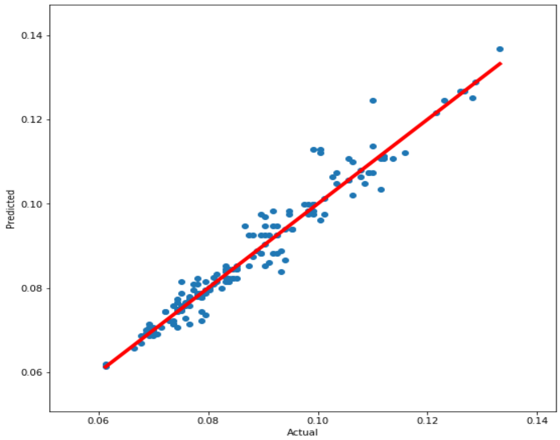

Fig. 8. Normal Decision Tree Model

TABLE II. A COMPARATIVE STUDY OF DIFFERENT REGRESSION MODELS IN TERMS OF RMSE

ITERATIONS	**RMSE OF REGRESSION MODELS**					
	HYBRID (k=10)	**DT**	**MLP**	**LR**	**KNN**	**SVM**
1	0.002480	0.003433	2.60	0.006926	0.006365	0.020303
2	0.002672	0.003170	2.48	0.006926	0.006365	0.020303
3	0.003256	0.003415	2.52	0.006926	0.006365	0.020303
4	0.002818	0.003217	2.25	0.006926	0.006365	0.020303
5	0.003021	0.003422	2.23	0.006926	0.006365	0.020303
6	0.003209	0.003260	2.27	0.006926	0.006365	0.020303
7	0.002824	0.003375	1.81	0.006926	0.006365	0.020303
8	0.002949	0.003382	0.86	0.006926	0.006365	0.020303
9	0.002951	0.003446	0.08	0.006926	0.006365	0.020303
10	0.003064	0.003422	0.04	0.006926	0.006365	0.020303
AVERAGE	**0.002924**	**0.003354**	**1.71**	**0.006926**	**0.006365**	**0.020303**

CONCLUSION

The current work proposed a hybrid regression model to predict the soil moisture quantity. The hybrid model involves an initial phase of data clustering. An extensive analysis is done where the proposed hybrid model frame work is compared with the base model on which the model is applied. The results have indicated that the hybrid model is highly capable of improving the performance of regression models to a greater extent in predicting soil moisture quantity prediction.

REFERENCES

[1] Liang, Y., S. Kang, and C. Zhang. "The effects of soil moisture and nutrients on cropland productivity in the highland area of the Loess Plateau." ACIAR MONOGRAPH SERIES 84 (2002): 187-194.

[2] Rockström, Johan, et al. "Sustainable intensification of agriculture for human prosperity and global sustainability." Ambio 46.1 (2017): 4-17

[3] Valipour, Mohammad, et al. "Agricultural water management in the world during past half century." Archives of Agronomy and Soil Science 61.5 (2015): 657-678

[4] Gago, Jorge, et al. "UAVs challenge to assess water stress for sustainable agriculture." Agricultural water management 153 (2015): 9-19

[5] J.F. Bierhuizen, and N.M. De Vos, “The effect of soil moisture on the growth and yield of vegetable crops.”

[6] Gill, M. K., Asefa, T., Kemblowski, M. W., & McKee, M. (2006). Soil moisture prediction using support vector machines. JAWRA Journal of the American Water Resources Association, 42(4), 1033-1046

[7] Song, J., Wang, D., Liu, N., Cheng, L., Du, L., & Zhang, K. (2008, December). Soil moisture prediction with feature selection using a neural network. In Computing: Techniques and Applications, 2008. DICTA'08. Digital Image (pp. 130-136). IEEE.

[8] Esmaeelnejad, L., Ramezanpour, H., Seyedmohammadi, J., &Shabanpour, M. (2015). Selection of a suitable model for the prediction of soil water content in north of Iran. Spanish Journal of Agricultural Research, 13(1), 1202.

[9] Oliviu Matei, Teodor Rusu, Adrian Petrovan, Gabriel Mihut. A data mining system for real time soil moisture prediction. 181 (2017) 837-844. Procedia Engineering.

[10] R. Pugazendi, P. Usha. A hybrid model of k-means clustering and multilayer perceptron for rainfall. J. Computing & Int System (2017) 36-40.

[11] UNIVERSITY OF TORONTO, Mississauga, Department Of Geography, https://www.utm.utoronto.ca/geography/resources/environmental-datasets

[12] M. A. Figueiredo, “Adaptive Sparseness for Supervised Learning”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 1150-1159, 2003.

[13] Weber M., Welling M., Perona P. (2000) Unsupervised Learning of Models for Recognition. In: Computer Vision - ECCV 2000. vol 1842. Springer, Berlin, Heidelberg

[14] Algorithm AS 136: A K-Means Clustering Algorithm, J. A. Hartigan and M. A. Wong, Journal of the Royal Statistical Society. Series C (Applied Statistics).Vol. 28, No. 1 (1979), pp. 100-108