Published online 2017 July 16. **Research Article**

Using Data Mining to Predict the Concentration of Cadmium in Khuzestan Paddies, Iran

Ali Chamannejadian,[1,*] Mohammad Feizian,[1] and Abdol Amir Moezzi [2]

[1]Department of Soil Science Engineering, School of Agronomy Engineering and Technology College of Agriculture and Natural Resources, University of Lurestan, Khorramabad, IR Iran
[2]Department of Soil Science Engineering, School of Agronomy Engineering and Technology College of Agriculture and Natural Resources, University of Shahid Chamran, Ahvaz, IR Iran

[*]*Corresponding author*: Ali Chamannejadian, Department of Soil Science Engineering School of Agronomy Engineering & Technology College of Agriculture and Natural Resources, University of Lurestan, Khorramabad, IR Iran. Tel: +98-52675457, E-mail: chamannejadian@gmail.com

Received 2016 March 29; **Revised** 2016 November 29; **Accepted** 2016 December 18.

Abstract

Background: Rice is the second highly consumed foodstuff among Iranian people. However, high levels of cadmium (Cd) are reported in some paddy fields in Khuzestan province, Iran.
Objectives: The current study aimed at investigating the Cd concentration in rice grains by the decision tree using J48 algorithm. The current study also used WEKA software to implement the algorithm.
Methods: A total of 630 samples (9 attributes in 70 sampling areas) were taken from each paddy field (5 regions); hence, seed and soil samples were analyzed according to the standard laboratory procedures and finally, the data mining technique was used for the classification of trees by J48 algorithm to predict the concentration of Cd in rice seed.
Results: The results showed that the average concentrations of Cd in rice seed and soil were 81.4 and 273.6 μg/kg, respectively; it was also shown that J48 gives 95.71% accuracy, 0.899 Kappa coefficient, and less error (RMSE = 0.179), which make a good predictive model. A significant correlation was observed between soil characteristics and the concentration of Cd in rice seeds.
Conclusions: The data mining technology can be used to predicate Cd concentration in rice seeds, and also J48 algorithm is a simple designer to construct a decision tree; nevertheless, offers good results in experiments.

Keywords: Rice, Data Mining, Cadmium, J48

1. Background

Heavy metals are distributed in the environment because of human manipulations and natural chemical reactions (1). For example, application of metal-contaminated fertilizers, animal manures, and sewage sludge can result in high concentration of cadmium (Cd) in agricultural soils (all cases occurring in Khuzestan province, Iran). A food chain contaminated with such heavy metals is a major source for human poisoning. Plants play an important role in heavy metals transfer from contaminated soils to human body (2). This event is remarkable in highly consumed crops such as rice and wheat. Rice is the predominant food crop in the developing countries such as Iran; therefore, 96% of the world's rice is produced and consumed in such countries (3). Rice ranks second in the food chain of Iranian people. It is the most common crop grown in agricultural lands in the North of Iran (4). Data mining technique aimed at discovering useful information in a data set. Data mining is an advanced information processing technology, which discovers laws over data to obtain useful information. In a broad sense, any method that extracts information from data can be regarded as data mining, which includes a variety of information processing methods (5). Data mining is widely developed in various applications such as agriculture (6), analysis of organic matter (7), medical diagnosis (8), product design (9), marketing (10), credit card fraud detection (11), financial forecasting (12), and automatic abstraction (13).

2. Objectives

The current study pursued data classification and performance measure of the classifier algorithms based on true positive (TP) and false positive (FP) rates generated by J48 algorithm, when applied to the data set. The current study aimed at investigating the Cd concentration in rice grains by a decision tree using J48 algorithm. The study used WEKA software to implement the algorithm.

3. Methods

3.1. Study Area and Sampling Analysis

The study area was about 300 km2 in Khuzestan province of Iran (Figure 1). The study area included 5 sub-regions (Ahvaz, Dashte-azadegan, Baghmalek, Ramhurmoz, and Shushtar). A total of 70 soil samples were collected from paddy fields (Figure 1). Seed and soil samples were analyzed according to the standard laboratory procedures. All measurements and explanations in the current study were provided by the same authors (14). Data sets were analyzed with different software packages. Statistical analysis was conducted with SPSS version 17. Descriptive statistic variables such as mean, variance, maximum and minimum of Cd concentrations in soil and rice seed, and measured soil parameters were calculated. The correlation analysis was used to evaluate the relationship between soil properties and seed Cd concentrations.

3.2. Decision tree Algorithm J48

The J48 algorithm is an open source Java performance of the C4.5 algorithm in the weak data mining tool. The algorithm was developed by Ross Quinlan (15). Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. The decision trees made by C4.5 can be applied to different categories. Accordingly, C4.5 is frequently referential as a statistical arranger (16).

3.3. Data Collection Cleaning and Checking

The dataset required for the current study was collected from the private soil testing laboratory in Shahid Chamran University of Ahvaz, Iran. The dataset contain various attributes and the respective values of soil samples taken from 5 regions of Khuzestan province. The dataset has 10 attributes and a total of 980 instances of soil samples. Table 1 shows the attribute description.

Table 1. Attribute Description

Attribute	Area	ECe	Sand	Silt	Clay	pH	TNV	OM	Cd seed	Cd DTPA
Description	Sampling site	Electrical conductivity decisiemen per meter	Sand value	Silt value	Clay value	pH value of soil	Total neutralizing value	Organic matter	Cd in rice seed	Extractable soil Cd

4. Result and Discussion

4.1. Descriptive Statistic Parameters

The descriptive statistics of the studied contents in 70 samples of rice seed from 5 sub-regions are shown in Table 2. The average of soil and plant Cd concentrations in the area were 81.4 and 273.6 μg/kg, respectively, which were lower than that of soil Cd and greater than that of plant Cd, based on the ISIRI (the Institute of Standards and Industrial Research of Iran) permitted limits for Cd in rice seed (ie, 0.2 mg/kg), and soil (ie, 3 mg/kg) (17-19). After examining the areas separately, it was observed that in some areas the amount of Cd in rice seed exceeded the permissible limit (Baghmalek) that should be considered (Table 3). The results showed a close relationship between Cd in the seed with ECe, TNV, Cd DTPA, as well as the relationship between CdDTPA, and pH and OM (Table 4). Similar researches were also conducted by the same technique such as Gholap and Sanap et al. but some recent researches reported different parameters such as the high percentage of lime that can play an important role in the behavior of Cd. Further researches by the same authors should address the relationship between the concentrations of Cd in rice seeds and soil (3, 14).

Table 2. Summary of Soil Characteristics in the Studied Areas

Soil Characteristics	N	Min	Max	Mean ± SD	CV (%)
pH	70	6.8	7.7	7.2 ± 0.22	3.0
ECe, dSm^{-1}	70	1.2	40.5	7.6 ± 6.76	8.9
Sand, %	70	2.0	48.0	17.0 ± 10.37	60.9
Silt	70	30	58.0	49.5 ± 5.48	11.1
Clay, %	70	16	52.0	33.4 ± 9.04	2. 7
TNV	70	22.4	49.9	48.5 ± 3.55	7.3
OM, %	70	0.3	1.7	0.8 ± 0.25	3.1
Cd seed, μg/kg	70	8.9	266.2	81.4 ± 53.69	65.9
Cd DTPA, μg/kg	70	63.3	521	273.6 ± 111.75	40.9

Abbreviations: CV, coefficient of variation; SD, standard deviation; TNV, total neutralizing value.

Table 3. The Mean Cd Concentrations in Rice Seed Grown in Different Areas

Element	Area	N	Mean ± SD	Min	Max
Cd	Ahvaz	12	270.9 ± 128	120	521
	Baghmalek	9	296.7 ± 135	127	465
	Dashte-azadegan	24	275.5 ± 118	63.3	493
	Ramhurmoz	5	249.6 ± 24.0	219	283
	Shushtar	20	269.6 ± 123	97	515
	Total	70	273.6 ± 117	63	521

4.2. Tuning Performance of J48 Algorithm and Decision Tree

The data mining technique of classification, using the J48 algorithm, was performed on the dataset including 21 samples of concern, to determine a relationship between the results obtained from the classification trees by algorithm J48 methods. The decision tree was provided for the prediction (Figure 2). Some of the characters of the decision tree are shown in Table 5. The quality of the predictions made by applying the J48 model which is presented

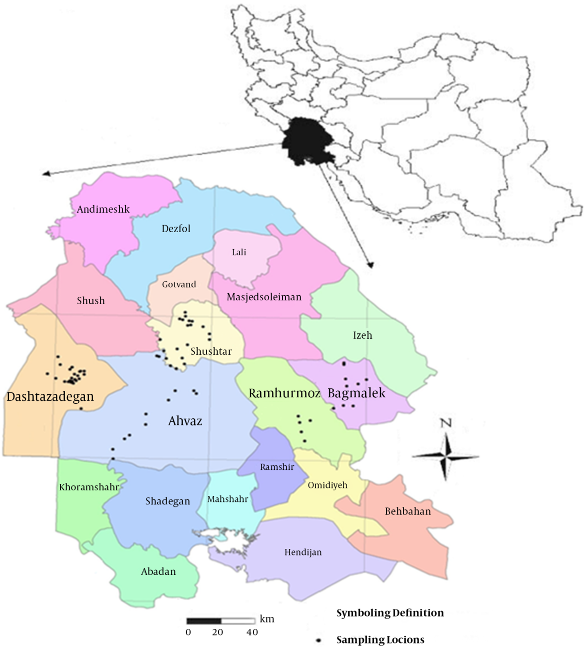


Figure 1. Distribution of Sampling Locations

Table 4. The Correlation Coefficient Between Heavy Metals and Soil Attributes

	ECe	Sand	Silt	Clay	pH	TNV	OM	Cd seed
Sand	0.030							
Silt	0.090	0.340						
Clay	0.127	0.750	0.170					
pH	0.183	0.138	0.220	0.056				
TNV	0.090	0.111	0.090	0.013	0.137			
OM	0.139	0.277[a]	0.230	0.375[b]	0.084	0.012		
Cd seed	0.435[b]	0.210	0.013	0.138	0.009	0.277[a]	0.111	
Cd DTPA	0.032	0.091	0.210	0.141	0.280[a]	0.091	0.376[b]	0. 271a

Abbreviations: Cd DTPA, extractable soil cadmium; Cd seed, cadmium in rice seed; OM, organic matter.

[a] $P < 0.05$.

[b] $P < 0.01$.

in Table 6, which indicates that the J48 normative model could predict the concentration of Cd in the rice seed accurately. The accuracy of decision tree in the current study was about 95.7%, which made a good predictive model (Table 5). This performance was confirmed by the MAE and RMSE values (Table 6). The Kappa coefficient was around 0.89, which was a great value for forecasting models (Table 7) (20).

Table 5. Detailed Accuracy by Class for J48 Algorithm

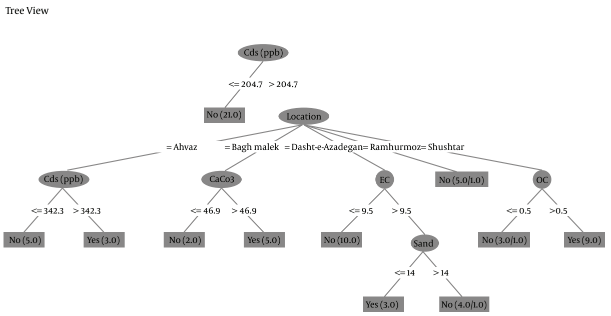
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.87	0	1	0.87	0.93	0.904	0.989	0.976	Yes
1	0.13	0.94	1	0.969	0.904	0.989	0.993	No
0.957	0.088	0.96	0.957	0.956	0.904	0.989	0.987	Weighted Avg.

Abbreviations: FP Rate, false positive rate; MCC, Matthews correlation coefficient; PRC Area, precision recall curve area; ROC Area, receiver operating characteristic area; TP Rate, true positive rate.

4.3. Conclusion

The current study used the algorithm J48 and prediction techniques to analyze Cd concentration in rice samples. It was demonstrated a comparative study of vari-

Figure 2. Prediction of the Cadmium Concentration in Khuzestan Paddies by the Decision Tree Using J48 Algorithm



Yes: More than guide value for cadmium, No: Less than guide value for cadmium (guide value for Cd: 0.2 mg/kg).

Table 6. Performance Estimation for J48 Algorithm by WEKA Tool

	Value
Correctly classified instances	95.71%
Incorrectly classified instances	4.29%
Kappa statistics	0.8995
Mean absolute errors	0.0633
Root mean squared errors	0.1780

Table 7. Interpretation of Kappa Coefficient

	Poor	Slight	Fair	Moderate	Substantial	Almost Perfect
Kappa	0	0.2	0.4	0.6	0.8	1
Kappa	Agreement					
< 0	Less than chance agreement					
0.01 - 0.20	Slight agreement					
0.21 - 0.40	Fair agreement					
0.41 - 0.60	Moderate agreement					
0.61 - 0.80	Substantial agreement					
0.81 - 0.99	Almost perfect agreement					

ous classification J48 algorithms (C4.5) with the help of data mining tool WEKA. J48 algorithm is a simple designer to construct a decision tree, but it had the best result in the experiments. Various decision tree algorithms can be used to predict the concentration of Cd in rice seed. According to authors' best knowledge, J48 had 95.71% accuracy, 0.899 Kappa coefficient, and less error (RMSE = 0.179), which made a good predictive model. In future, according to the results on less error and classification areas, it is recommended to build a fertilizer recommendation system, cropping pattern, and given soil.

Acknowledgments

The current study was financially supported by Shahid Chamran University of Ahvaz, Lurestan University of Khorramabad, and Food and Drugs Safety Evaluation Research Center, Ahvaz Jundishapur University of Medical Sciences, Iran.

Footnote

Authors' Contribution: Ali Chamannejadian: conception and design, generation of data, collection of data, assembly of data, analysis of data, interpretation of data, drafting of the manuscript. Mohammad Feizian: conception and design, interpretation of data, revision of the manuscript, approval of the manuscript. Abdolamir Moezzi: interpretation of data, revision of the manuscript, approval of the manuscript.

References

1. Harmanescu M, Alda LM, Bordean DM, Gogoasa I, Gergen I. Heavy metals health risk assessment for population via consumption of vegetables grown in old mining area; a case study: Banat County, Romania. *Chem Cent J.* 2011;**5**:64. doi: 10.1186/1752-153X-5-64. [PubMed: 22017878].

2. Khan MU, Malik RN, Muhammad S. Human health risk from Heavy metal via food crops consumption with wastewater irrigation practices in Pakistan. *Chemosphere.* 2013;**93**(10):2230–8. doi: 10.1016/j.chemosphere.2013.07.067.
3. Chamannejadian A, Moezzi AA, Sayyad G, Jahangiri A, Jafarnejadi A. Effect of soil characteristics on spatial distribution of cadmium in calcareous paddies. *Int J Agric.* 2013;**3**(1):139.
4. Khaniki GR, Zozali MA. Cadmium and lead contents in rice (Oryza sativa) in the North of Iran. *Int J Agric Biol.* 2005;**6**:1026–9.
5. Ji H, Songlin W, Qinglin W, Xiaonan C. Douhe Reservoir Flood Forecasting Model Based on Data Mining Technology. *Proc Environ Sci.* 2012;**12**:93–8. doi: 10.1016/j.proenv.2012.01.252.
6. Huang J, Yuan Y, Cui W, Zhan Y. Development of a Data Mining Application for Agriculture Based on Bayesian Networks. *Comput Comput Technol Agric.* 2008;**258**:645–52. doi: 10.1007/978-0-387-77251-6_70.
7. Wu S, Liu J, Hu Y, Wang J, Pellizzari E, editors. Using data mining techniques to identify volatile organic compounds associated with asthma attack. Proc. Joint Statistical Meetings. 2002; pp. 3809–12.
8. Kumar DS, Sathyadevi G, Sivanesh S. Decision support system for medical diagnosis using data mining. *Int J Comput Sci Issues.* 2011;**8**(3):147–53.
9. Dong J, Zhao Y, Peng T. A review of design pattern mining techniques. *Int J Software Engin Knowledge Engin.* 2009;**19**(06):823–55.
10. Sing'oei L, Wang J. Data mining framework for direct marketing: A case study of bank marketing. *Int J Comput Sci Issues.* 2013;**10**(2):198–203.
11. Brause R, Langsdorf T, Hepp M. Neural data mining for credit card fraud detection. 11th IEEE International Conference. 1999. pp. 103–6.
12. Mehzabin Shaikh P, Chhajed MGJ. Review on Financial Forecasting using Neural Network and Data Mining Technique. *Glob J Comput Sci Technol.* 2012;**5**(2):263–7.
13. Ghafoorian M, Taghizadeh N, Beigy H, editors. Automatic Abstraction in Reinforcement Learning Using Ant System Algorithm. AAAI Spring Symposium: Lifelong Machine Learning. 2013; .
14. Chamannejadian A, Sayyad G, Moezzi AA, Jahangiri A. Evaluation of estimated daily intake (EDI) of cadmium and lead for rice (Oryza sativa L.) in calcareous soils. *Iran J Environ Health Sci Engin.* 2013;**10**(1):28. doi: 10.1186/1735-2746-10-28.
15. Gholap J. Performance tuning of J48 Algorithm for prediction of soil fertility. *Asian J Comput Sci Inf Technol.* 2012;**2**(8):1–5.
16. Sanap SA, Nagori M, Kshirsagar V. Classification of Anemia Using Data Mining Techniques. *Swarm Evolutionar Memetic Comput.* 2011;**7077**:113–21. doi: 10.1007/978-3-642-27242-4_14.
17. Alloway B. Heavy metal in soils. Glasgow: Blackie and Academic and professional; 1995.
18. Afyoni M, Khoshgoftarmanesh AH, Dorostkar V, Moshiri R. Zinc and Cadmium content in fertilizers commonly used in Iran. International Conference of Zinc Crops. Istanbul. .
19. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. ; 2012.
20. Legrand G, Nicoloyannis N. Data Preprocessing and Kappa Coefficient. *Rough Sets Fuzzy Sets Data Mining Granular Comput.* 2005;**3641**:176–84. doi: 10.1007/11548669_19.