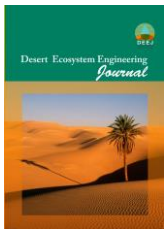
DEEJ (2018) 7 (19) 81-95

University of Kashan

Desert Ecosystem Engineering Journal

Journal homepage: http://deej.kashanu.ac.ir



Using Data Mining Algorithms in Separation of Sediment Sources in Nodeh Watershed, Gonabad

Mehdi bashiri1, Mahsa Ariapour[2], Ali Golkarian[3]

Received: 5/03/2018 Accepted: 2/07/2018

Extended Abstract

Introduction: Reduction of sediment supply requires the implementation of soil conservation and sediment control programs in the form of watershed management plans. Sediment control programs require identifying the relative importance of sediment sources, their quantitative *ascription* and identification of critical areas within the watersheds. The sediment source ascription is involves two main steps so that in the first, several diagnostic tracers are selected for obvious and significant separation of potential sources of sediment and in the second step selected tracers for potential sources of sediment are compared, with corresponding values extracted from the sediment samples taken in the watershed outlet. Also, due to the large amount and complexity of data available, nowadays in geo- and environmental sciences, we face the need to develop and incorporate more robust and efficient methods for their analysis and modelling. Therefore recent fundamental progress in data mining algorithms can considerably contribute to the development of the emerging field - environmental data science.

Methodology: According to what was said, in this research, the data mining algorithms used to separate sediment sources in the Nodeh watershed of Gonabad located in Razavi-Khorasan province by using the geochemical (includes the 21 elements of Mg, Sr, Mn, Ba, Zn, Y, V, Ti, Pb, P, Na, Li, K, Cu, Cr, Co, Ce, B, Ca, Al and Fe), granulometric (includes the D_{90}, D_{50}, D_{10}, percent of sand, percent of silt, percent of clay, skewness and kurtosis and the diameters less than 1, 2 and 4 millimeters and less than 500, 250, 125 and 63 microns) and lithological variables (includes the quartz, tuff, laterite, dacite, andesite, dolomite, calcite, andesitic tuff, lithic andesite and salt). A set of 11 classification algorithms includes the decision tree, random forest, regression methods, discriminant analysis, local linear model tree, nearest neighbor analysis, support vector machine, logistic regression, artificial neural network, pattern recognition and group method of data handling programmed in the MATLAB software and the results compared based on the coefficient of determination and mean squared error.

Results and Discussion: Study of geochemical element concentrations in 7 geological units showed that the Ca, Fe, Mg and Al elements have the highest and B and Co have the lowest concentrations within the soil samples. Overall evaluation of classification algorithms in training stage showed that the discriminant analysis, random forest, k nearest neighbor and support vector machines with linear, polynomial, multiple and RBF kernels with maximum values of the coefficient of determination (R^2=1) and minimum values of the mean squared error (RMSE=0) are the most accurate algorithms in sediment source separation but the regression trees method has the worst performance. Also, at testing stage, the support vector machines with RBF kernel was the most accurate

1. Assistant Professor University of torbat heydarieh - - Razavi-khorasan, University of torbat heydarieh - me.bashiri@yahoo.com
2 M.Sc. Student University of torbat heydarieh - - Razavi-khorasan, University of torbat heydarieh
3 Assistant Professor Ferdowsi University of mashhad - - Razavi-khorasan, Ferdowsi University of Mashhad
DOI: 10.22052/deej.2018.7.19.49

and the classification trees with maximum error rate was the most inaccurate algorithm. Also, entrance of geochemical and granulometric variables lead to the highest and lowest accuracy in the sediment source separation, respectively. Using the geochemical variables for the separation of sediment sources, types of support vector machines, nearest neighbor analysis, discriminant analysis and the random forest algorithm had the highest coefficients of determination and lowest error values in the training and testing stages. By entering the lithological variables, the random forest algorithm had the highest accuracy for the sediment sources classification in the training and testing stages and the discriminant analysis and support vector machines were located thereafter. Finally, fitting the classification algorithms using granulometric variables showed that the support vector machines had highest accuracy in the training and testing stages of models and the random forest and nearest neighbor analysis were ranked thereafter.

Conclusion: Totally, due to the proper accuracy and performance of data mining classifier algorithms, application of these methods in the natural sciences is suggested especially in the large amounts of data. These algorithms are used to find patterns in large sets of data and help classify new information. Especially, the support vector machines that are supervised classifier algorithms and besides that, in the natural sciences have successful results. In the watershed management considering the time and cost, sediment source ascriptions are difficult to obtain using monitoring techniques, but data mining procedures, have emerged as a potentially valuable alternative. Therefore, application and evaluation of these methods are suggested for further studies and natural sciences data.

Keywords: Classification algorithms, Element density, Nodeh watershed, Sediment source ascription.