Predicting Groundwater Areas Using Data Mining Techniques: Groundwater in Jordan as Case Study

Faisal Aburub, Wael Hadi

Abstract**—**Data mining is the process of extracting useful or hidden information from a large database. Extracted information can be used to discover relationships among features, where data objects are grouped according to logical relationships; or to predict unseen objects to one of the predefined groups. In this paper, we aim to investigate four well-known data mining algorithms in order to predict groundwater areas in Jordan. These algorithms are Support Vector Machines (SVMs), Naïve Bayes (NB), K-Nearest Neighbor (kNN) and Classification Based on Association Rule (CBA). The experimental results indicate that the SVMs algorithm outperformed other algorithms in terms of classification accuracy, precision and F1 evaluation measures using the datasets of groundwater areas that were collected from Jordanian Ministry of Water and Irrigation.

Keywords**—**Classification, data mining, evaluation measures, groundwater.

I. Introduction

WATER is considered as one of the most important elements, not only for human life, but also for all types of life on the planet. The provision of water represents one of the main problems in many countries. Jordan is one of these countries and is recognized as one of the water-poorest countries in the world. According to [1], "Jordan suffers from water scarcity, which poses a threat that would affect all sectors that depend on the availability of water for the sustainability of their activities for their development and prosperity". Moreover, Jordanian citizens only have access to around 1,000 cubic meter per year, while American citizen use more than 9,000 cubic per year [2].

According to [1], groundwater is the main water resource in Jordan. Moreover, the groundwater resource is the only water supply available in many parts of the country. In Jordan, there are 12 main basins, which are comprised of a set of aquifers. The most important aquifers are the Amman/Wadi Elsir, Basal and Ram.

Countries like Jordan need to exploit every available water resource in order to provide water for their citizens. One of these methods is to use data mining techniques to predict and find groundwater. Many methods have been developed to predict groundwater areas. For example, [3] proposed a method to check if the ground water from selected areas are potable or not. As multivariate is present, Principal Component Analysis with data mining techniques using JRIP rules were employed for classifying the ground water. The authors conclude that data mining techniques can be employed for quicker classification of water potability. Reference [4] presents two models to predict groundwater levels in an unconfined shallow aquifer in the Searsville basin, part of the Jasper Ridge Biological Preserve. Linear regression does a good job of predicting groundwater levels in the summer, when water levels are low, while the neural network does a good job of predicting groundwater levels in winter, when water levels are high. This result supports the combination of linear regression and neural networks for predicting hydrologic response up to one year in advance. Moreover, [5]-[9] developed different methods with varying techniques that aim to detect groundwater areas.

According to the abovementioned paragraphs, it is important to discover new areas of groundwater with minimum resources and cost. Four well-known data mining techniques have been used to predict groundwater areas in Jordan. In order to discover new groundwater sites, Jordanian groundwater experts identified seven features that used as input parameters for the four data mining techniques. The groundwater features are: Elevation, Valleys, Slope of the earth's surface, the annual long-term average of rainfall, the annual long-term average of temperatures, Geological out-crop and the faults of earth surface.

The research aims to investigate four well-known data mining techniques: CBA, SVMs, NB and KNN, to determine which better data mining techniques that can predict groundwater sites in Jordan.

The rest of the paper is organized as follows: in Section II, the literature review is presented. The data mining techniques are described in Section III, followed by the experimental results in Section IV. Finally, conclusions and future works are presented in Section V.

II. Literature Review

This section presents related works of machine learning and data mining in groundwater applications.

Reference [3] proposed a method to test if the ground water of some areas are potable or not. Principal Component Analysis combined with the data mining technique using JRIP rules was employed for classifying the ground water. The authors conclude that data mining techniques can be employed for quicker classification of water potability. Reference [4] presents two models to predict groundwater levels in an unconfined shallow aquifer in the Searsville basin. Linear regression does a good job of predicting groundwater levels in the summer when water levels are low, while the neural network does a good job of predicting groundwater levels in the winter, when water levels are high. This result supports the

Faisal Aburub and Wael Hadi are with the University of Petra, MIS Department, Amman Jordan (e-mail: faburub@uop.edu.jo, whadi@uop.edu.jo).

combination of linear regression and neural networks for predicting hydrologic response up to one year in advance.

Reference [5] developed a map to display groundwater sites and where they may be most vulnerable to contamination using the SINTACS model. Some environmental factors have been considered in the model such as hydrogeology, hydrology, topography and pedology. In order to model groundwater sites using this model, special knowledge and an understanding of the mutual relationships of environmental factors are required. Reference [6] used GIS methods to apply geostatistical analysis on groundwater levels for 95 well sites in northeast Libya. Normality of the groundwater level was investigated using spatial data analysis (ESDA) tools.

Reference [7] developed a model to predict the level of groundwater using neural network technique. This model uses average rainfall as an input parameter. While reference [8] developed a decision support system (DSS) based on data mining results to complement the deterministic models to detect new ground water sites. A DSS is a powerful, easy-to-use package that combines data, analytical results, predictive models, and supporting graphics that allows resource managers and stakeholders to evaluate alternative management strategies.

Reference [9] developed a method to detect the quality of ground water for Thanjavur District, Tamilnadu, India and classify the suitability of water for drinking. The chemical parameters of water such as Ph, EC, HCo3, Cl, SO4 and TDS were used for classifying using data mining techniques. The data mining classifiers support for faster identification of water quality.

III. Data Mining Techniques

This section discuses four well-known data mining techniques: CBA, SVMs, NB and KNN. CBA is the first approach that integrates the association rule task with the classification data mining task. Also, SVMs is considered as one of the most accurate data mining algorithms that performs classification by building an N-dimensional hyperplane that optimally splits the data into two classes. Moreover, NB is a simple probabilistic algorithm based on Baye's theorem. Finally, KNN is a statistical data mining algorithm, which has been intensively studied in pattern recognition over five decades. The next four subsections discuss the four algorithms that we consider.

A. CBA

Reference [10] considered CBA as one of first method works that presented the utilization of association rule in classification. The CBA algorithm works through three steps: rule generation, model building and prediction. In the first step and according to [11], the algorithm employs the Apriori algorithm to find the frequent rules among training data features. Any rule is called a frequent rule if it has support greater than or equal to the inputted minimum support. For the second step, the frequent rules are sorted according to certain criteria. Then, some of the frequent rules are pruned because these rules may be redundant, conflict, or noise. The remaining rules are sorted according to confidence and support, and inserted into CBA model, this step is called, model building. Finally, to predict a new test instance with the suitable class, the first rule in the set of ranked rules that matches the test instance predicts it to the class label of the matched rule. According to [10]; [12], if no rule matches to this instance, it assigns the default class.

B. (SVMs

SVMs developed by [13], are a supervised data mining algorithm which can be used for either classification or regression challenges. However, it is frequently used in classification problems. SVMs have become the method of choice to solve difficult classification problems in a wide range of application domains. SVMs built on the fundamental of minimization of structural risk. In linear classification, SVMs produce a hyper plane that splits the training data into two groups with maximum-margin. A maximum-margin hyper plane is a hyper plane which separates two of points and is at equal distance from the two. Mathematically, SVMs learn the sign function f (x) = sng (wx + b), where w is a weighted vector in R^n. SVMs find the hyper plane y = wx + b by separating the space R^n into two half-spaces with the maximum-margin. Linear SVMs can be generalized for non-linear problems. To do so, the data is mapped into another space H and we perform the linear SVMs algorithm over this new space.

C. NB

The NB is a simple and effective algorithm that utilizes the likelihoods of each feature belonging to each class value to do a prediction step [14].

NB streamlines the computation of likelihoods by supposing that the likelihood of each feature belonging to a given class value is independent of all other features.

The likelihood of a class value given a value of a feature is named the conditional likelihood. By multiplying the conditional likelihoods together for each attribute for a given class value, we have a likelihood of a data object belonging to that class. To do a prediction we can compute the likelihoods of the object belonging to each class and select the class value with the highest likelihood.

D. kNN

The KNN algorithm [14] is simple. Based on training and test data, the KNN finds the kNNs of the training data, and uses classes of the k-neighbors to assign the class of the test instance. The scores of similarity of each neighbor instance to the test instance are used as a weight of the classes of the neighbor instance. When several kNNs share a class, then the pre-neighbor weights of that class should be added together, and the result of the added weights should be used as the likelihood score of that class with regard to the test instance. In order to find a ranked list for the test instance, the scores of the candidates' classes should be sorted.

IV. Experimental Results

A. Settings

In all experimentations, the 10-fold cross-validation evaluation method has been used. Moreover, four well-known data mining algorithms have been compared for predicting new groundwater areas. These algorithms are CBA, NB, SVMs and KNN.

The experimentations were conducted on an I7 machine with 16G main memory; the experiments of all algorithms were conducted using the WEKA software [15] environment. The bases of our experiments are three well-known evaluation measures (Accuracy, Precision and F1).

Finally, we have set the minimum support and minimum confidence thresholds for CBA to 1% and 50%, respectively for all experimentations. Previous studies, such as [16]-[19], suggested that the value of minimum support range from 1% to 5% and a minimum confidence threshold is 50%.

B. Datasets

Datasets of 900 groundwater areas have been investigated in our experiments that they are belonging to two classes ("Yes", "No"), containing seven features (attributes) to distinguish the groundwater areas. These features are elevation, faults, rainfall, slope, temperature, wadis and outcrop. Our groundwater datasets are collected from the Jordanian Ministry of Water and Irrigation.

C. Pre-Processing

One of important step in data mining is preparing the input data which may be unstructured, sparse, and may contain noise, such as incomplete transactions, records redundancy, missing values, poor image, etc. [20]. Thus, the quality of the produced output classification systems is significantly impacted by the quality of the input data set. There are various types of features that require different types of data or maps. For instance, drainage networks and sub-catchment areas can be obtained from topographic maps, rock type and soil type can be derived from geological maps and the precipitation rate and rainfall frequency can be extracted from weather maps provided by rainfall stations that are scattered around the country. Therefore, pre-processing the different kinds of data to extract the appropriate features for further analysis is a crucial task. There are different software systems that can be utilized to process the input data, though the usage of software relies on the type of data that requires processing. Overall, GIS related software like ARC-MAP, ARC-View and other mathematical modelling techniques are mainly used to extract important features before the analysis stage begins.

V. Results Analysis

After investigating the four data mining g techniques shown in Table I, we found that the SVMs algorithm outperformed all other data mining algorithms with respect to the classification, Accuracy, evaluation measure. In particular, the SVMs outperformed NB, CBA and KNN with 1.1%, 10.2% and 1.2%, respectively. Also, the SVMs algorithm outperformed all algorithms with regards to the Precision evaluation measure. The SVMs outperformed NB, CBA and KNN with 1.4%, 10.9% and 1.5%, respectively. Moreover, the SVMs dominated all algorithms with reference to the F1 evaluation measure. In particular, the SVMs outperformed NB, KNN and CBA with 1.2%, 1.6% and 17.2%, respectively.

Finally, the CBA algorithm produces the worst results because the groundwater areas datasets are unbalanced, the datasets contain 683 areas that belongs to the class "Yes" and 217 areas that belongs to the class "No". In general, all algorithms produce good results that indicate that data mining algorithms are a suitable and helpful tool for predicting new groundwater areas.

TABLE I
Results Produced by Four Data Mining Algorithms on Groundwater Datasets

Algorithms	Classification Accuracy	Precision	F1
CBA	78.2	78.6	71.6
SVMs	88.4	89.5	88.8
NB	87.3	88.1	87.6
KNN	87.2	88.0	87.2

VI. Conclusion and Future works

This paper aims to compare four well-known supervised learning algorithms (SVMs, NB, KNN and CBA) with regards to classification accuracy, precision and F1 measures in relation to groundwater areas datasets. The test results show the SVMs algorithm outperformed the other three algorithms in relation to all used measures. The results also show that there is potential use for automated data mining algorithms in predicting groundwater areas.

References

[1] Jordan Ministry of Water and Irrigation – Reports 2013-2016. http://www.mwi.gov.jo/sites/enus/SitePages/MWI%20BGR/Reports.aspx

[2] Nortcliff A, Carr G, Potter RB, Darmame K. (2008) Jordan's Water Resources: Challenges for the Future. Geographical Paper No. 185, The University of Reading.

[3] Karthik, D., & Vijayarekha, K. (2014). Multivariate Data Mining Techniques for Assessing Water Potability. Rasayan Journal of Chemistry, 7 (3):256-259.

[4] Maatta, S. (2011). Predicting groundwater levels using linear regression and neural networks, CS229 final project, December 15, 2011.

[5] Al Kuisi, M., El-Naqa, A., & Hammouri, N. (2006). Vulnerability mapping of shallow groundwater aquifer using SINTACS model in the Jordan Valley area, Jordan. Environmental Geology, 50(5), 651-667.

[6] Salah, H., (2009). Geostatistical analysis of groundwater levels in the south Al Jabal Al Akhdar area using GIS. GIS Ostrava.

[7] Kumar, S., Dirmeyer, P. A., Merwade, V., DelSole, T., Adams, J. M., & Niyogi, D. (2013). Land use/cover change impacts in CMIP5 climate simulations: A new methodology and 21st century challenges. Journal of Geophysical Research: Atmospheres, 118(12), 6337-6353.

[8] Cook, J.B., Roehl, E.A. and Daamen, R.C., 2013. Predicting the Impact of Climate Change on Salinity Intrusions in Coastal South Carolina and Georgia. *Proceedings of the 2013 Georgia Water Resources Conference*, held April 10–11, 2013, at the University of Georgia.

[9] Karthik, D., Vijayarekha, K. & Abirami S. (2015). Classifying ground water quality using data mining technique for Thanjavur district, Tamilnadu, India. Journal of Chemical and Pharmaceutical Research, 7(3):1724-1727.

[10] Liu B., Hsu W. and Ma Y. (1998). Integrating classification and association rule mining. Proceedings of the KDD, (pp. 80-86). New York, NY.

[11] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

[12] Antonie M. and Zaiane O. (2002). Text Document Categorization by Term Association, Proceedings of the IEEE International Conference on Data Mining (ICDM '2002), (pp.19-26), Maebashi City, Japan.

[13] Vapnik V. (1995). The Nature of Statistical Learning Theory, chapter 5. Springer-Verlag, New York.

[14] Hadi, W., Thabtah, F., ALHawari, S., & Ababneh, J. (2008). Naive Bayesian and k-nearest neighbour to categorize Arabic text data. In Proceedings of the European Simulation and Modelling Conference. Le Havre, France (pp. 196-200).

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[16] Hadi, W. (2015). ECAR: A New Enhanced Class Association Rule. Advances in Computational Sciences and Technology, 8(1), 43-52.

[17] Abdelhamid, N., Ayesh, A., & Hadi, W. (2014). Multi-label rules algorithm based associative classification. Parallel Processing Letters, 24 (1), 1450001-1-1450001-21.

[18] Hadi, W. (2013). EMCAR: Expert Multi Class Based on Association Rule. International Journal of Modern Education and Computer Science, 5(3), 33-41.

[19] Thabtah, F., Hadi, W., Abdelhamid, N., & Issa, A. (2011). Prediction Phase in Associative Classification Mining. International Journal of Software Engineering and Knowledge Engineering, 21(06), 855-876.

[20] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.