Chapter 1

Ensemble Methods in Environmental Data Mining

Goksu Tuysuzoglu, Derya Birant and Aysegul Pala

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74393

Abstract

Environmental data mining is the nontrivial process of identifying valid, novel, and potentially useful patterns in data from environmental sciences. This chapter proposes ensemble methods in environmental data mining that combines the outputs from multiple classification models to obtain better results than the outputs that could be obtained by an individual model. The study presented in this chapter focuses on several ensemble strategies in addition to the standard single classifiers such as decision tree, naive Bayes, support vector machine, and k-nearest neighbor (KNN), popularly used in literature. This is the first study that compares four ensemble strategies for environmental data mining: (i) *bagging*, (ii) bagging combined with random feature subset selection (the *random forest* algorithm), (iii) *boosting* (the AdaBoost algorithm), and (iv) *voting* of different algorithms. In the experimental studies, ensemble methods are tested on different real-world environmental datasets in various subjects such as air, ecology, rainfall, and soil.

Keywords: data mining, classification, ensemble learning, environmental data, bagging, random forest, AdaBoost

1. Introduction

Environmental data mining is defined as extracting knowledge from huge sets of environmental data. It is an interdisciplinary area of both computer and environmental sciences, including but not limited to environmental information management systems, decision support systems, recommender systems, environmental data analytics, and so on.

Environmental data mining based on ensemble learning is a rather young research area where a set of learners are trained sequentially on the dataset to better analyze and understand

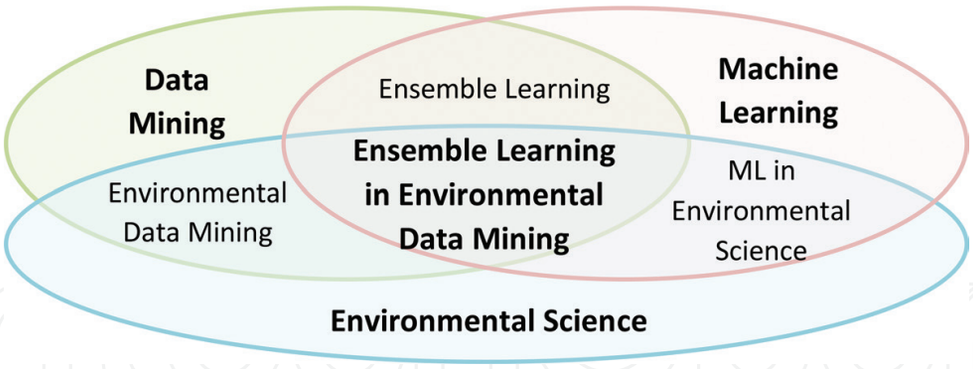

Figure 1. Interdisciplinary structure of ensemble learning in environmental data mining (ELEDM).

environmental processes and systems. However, it is not well-known yet how ensemble methodology can be utilized in order to improve the performance of a single method. For this purpose, this chapter presents the findings of a systematic survey of what is currently done in the area and aims to investigate the ability of different ensemble strategies for environmental data mining.

Ensemble learning in environmental data mining (ELEDM) can be drawn as a combination of three main areas: data mining (DM), machine learning (ML), and environmental science (**Figure 1**). ML in environmental science is learning-driven, meaning that machines teach themselves to recognize patterns by analyzing environmental data, whereas in contrast, DM is discovery-driven, meaning that patterns are automatically discovered from environmental data. DM uses many ML methods, including ensemble learning methods.

The novelty and main contributions of this chapter are as follows. First, it provides a brief survey of ensemble learning used in environmental data mining. Second, it presents how an ensemble of classifiers can be applied on environmental data in order to improve the performance of a single classifier. Third, it is the first study that compares different ensemble strategies on different environmental datasets in terms of classification accuracy.

2. Related work

Data mining techniques have been recently utilized in environmental studies for processing environmental data and converting it to useful patterns to obtain valuable knowledge and make right decisions when dealing with environmental problems. Many of the developed techniques in data mining can often be tailored to fit environmental data.

Recently, ensemble learning has been one of the active research fields in machine learning. Thus, it has been utilized in a very broad range of areas such as marketing, banking, insurance, health, telecommunication, and manufacturing. In contrast to these studies, our work proposes ensemble learning approach that combines several models to produce a result to solve environmental problems.

2.1. Ensemble-based environmental data mining

Ensemble classifiers have been applied to different environmental subjects, such as air [1–6], water [7–9], soil [10–12], plant [13], forests [14, 15], climate [16–18], noise [19], rainfall [20], energy [21–23], as well as living organisms [18, 24, 25]. Some of the ensemble-based environmental data mining studies have been compared in **Table 1**. In this table, the scopes of the studies, the year they were performed, the algorithms that were used in the studies, the type of data mining task, the success rate with the validation method, and the ensemble strategy are listed. In addition, if more than one algorithm is presented and compared with each other, the proposed one (the most successful one) is also indicated. As given in the table, ensemble of models for classification or prediction has higher interest than ensemble clustering and anomaly detection [2, 22] in environmental science. Although ensemble clustering has been used in many areas, especially in bioinformatics, only a few studies [4, 25] have been conducted so far in the environmental science.

Ref.	Year	Type	Description	Data mining task	Ensemble strategy	Algorithms	Validation
[22]	2017	Energy	Identification of anomalous consumption patterns in building energy consumption	Anomaly detection	2, 4	RF, SVR, CCAD-SW using autoencoder and PCA, EAD	TPR = 98.10% FPR = 1.98% (for EAD model)
[18]	2016	Climate	Determination of the impact of climate change on the habitat suitability for large brown trout	Prediction	1, 2	Generalized additive models, MLP with bagging ensembles, RF, SVM, and fuzzy rule-based systems (TSK)	Threefold cross validation Weighted MSE = 0.18 (MLP with bagging ensembles) Overall true skill statistics (TSS) = 0.69 (RF)
[11]	2015	Soil	Classification of complex land use/land cover categories of desert landscapes using remotely sensed data	Classification	2, 3	RF and boosted ANNs	Mean class user's accuracy = 86.7% (for boosted ANN) and 86.6% (for RF ANN)
[26]	2015	Soil	Solve the problem of rare classes' classification on dust storm forecasting	Classification	2, 3	SMOTE with AdaBoost and RF (SARF), SVM, fuzzy ANN	Tenfold cross validation Accuracy = 96.51% (SARF)
[4]	2015	Air	Forecasting of air pollutant values for the Attica area	Clustering	2, 4	SOM for clustering, FFANN and RF ANN for regression, FIS to obtain fuzzy values	Tenfold cross validation RMSE and R^2

Ref.	Year	Type	Description	Data mining task	Ensemble strategy	Algorithms	Validation
[9]	2014	Water	Predictive modeling of groundwater nitrate pollution	Prediction	2	RF regression, LR	ROC = 0.923 (for model RF-A) AUC = 0.911 (for model RF-B)
[25]	2013	Living organisms	Construction of habitat models for living species in the Lake Prespa, Macedonia; in the soils of Denmark; and in the Slovenian rivers	Clustering	1, 2	RF and bagged multitarget predictive clustering tree (PCT) and single-target DT	Tenfold cross validation RRMSE
[5]	2012	Air	Prediction of the Macau's air pollution index	Prediction	1	Bootstrap sampling with replacement and random sampling without replacement using ANFIS method as base learner	RMSE = 12.21 (ANFIS with random sampling)
[2]	2011	Air energy	Detect overconsumption of fuel in aircrafts	Anomaly detection	1	Bootstrap sampling on each of the regression tree (tree), elastic network, GP, and stable GP regression methods	ROC = 0.90 NRMSE varied consistently between 85 and 90%

ANN, artificial neural network; SVR, support vector regression; PCA, principal component analysis; MLP, multilayer perceptron; SOM, self-organizing maps; EAD, ensemble anomaly detection; FIS, fuzzy inference system; GP, Gaussian process; MSE, mean squared error; RMSE, root-mean-square error; TPR, true positive rate; ROC, receiver operating characteristic.

Table 1. Comparison of ensemble-based environmental data mining studies.

The idea of using an ensemble of classifiers rather than the single best classifier has been proposed in several environmental data mining studies [5, 11, 26]. It is apparent that ensemble learners boost the performance of the single classifiers. Different models pick up different patterns in data. By pooling all these predictions together, as long as they are reasonably independent, informed, and diverse, the outcomes tend to be better.

One of the most popular ensemble learning strategies, *bagging*, is also well adapted to develop models for solving environmental problems. For example, it has been utilized to the forecast air pollution level of a region [5] and to establish habitat models for living species [25].

The second type of ensemble learning strategy, the *random forest* (RF) algorithm, has also been applied for classifying environmental data. It has been applied to predict pollutant occurrences in groundwater [9] and determination of the impact of climate change on the habitat suitability for a fish species [18] and to predict dust storm accurately [26].

Another ensemble learning strategy (*boosting*), the AdaBoost algorithm, has been used in various types of environmental applications such as for the classification of complex land use/land

cover categories of desert landscapes using remotely sensed data [11], to solve the problem of rare classes' classification on dust storm forecasting [26] and discovering plant species for automatic weed control [27].

Training with different algorithms in each ensemble (*voting*) is another commonly used ensemble strategy in environmental science. Some of the examples are for the identification of anomalous consumption patterns in building energy consumption [22] and forecasting air pollutant values of a region [4].

Differently from existing studies, the study presented in this chapter focuses on applying four distinct ensemble strategies to environmental datasets using (i) different training sets formed by random sampling with replacement (bagging), (ii) different training sets obtained by random instance and feature subset selection (random forest), (iii) different training sets using random sampling with replacement over weighted data (AdaBoost), and (iv) different algorithms (voting).

2.2. Advantages of ensemble-based environmental data mining

Some of the advantages of environmental data mining are given below:

- Prediction of parameters expected based on other parameters or under different cases in environmental studies, for example, prediction of rainfall [20], climate change [16–18] species richness/diversity [24, 25], and atmospheric parameters [28].
- Construction of models to reduce the consumption of energy [21–23] and raw materials [2] such as wood, grass, metal, steel, plastics, glass, paper, fuel, and natural gas.
- Clustering the items in environmental data to describe the current situation more clearly and to plan different activities for different clusters [4, 25].
- Classification of environmental audio and environmental noise [19].
- Processing ecological data for better modeling ecological systems [24, 25].
- Analyzing environmental data toward a better quality control such as air quality [1, 5, 6] and water quality [7–9].
- Identifying unexpected patterns from an environmental data using a data mining algorithm and detection of anomalies in environmental data [2, 22] to identify bad values, changes, errors, noises, frauds, and abnormal activities to realize the purpose of giving an alarm.
- Determination of the most important factor that affects the environment using a data mining technique such as decision tree and random forest [29].
- Development of a model to manage resources effectively [2, 21, 23], including environmental resources such as air, water, and soil; flow resources such as solar power [30] and wind energy; and natural resources such as coal, gas, and forests.
- Discovering patterns that can be used for better waste management and recycling.
- Analyzing the records of financial transactions related to environmental economics for better decision-making, i.e., investigating the financial impacts of environmental policies.

- Using ensemble methods as a preprocessing step before performing the essential environmental study.
- Clustering environmental documents according to their topics and main contents.
- Usage of process mining to improve work management in the environmental science.

3. Background information

3.1. Ensemble learning

Ensemble learning is a machine learning technique where multiple learners are trained to solve the same problem and their predictions are combined with a single output that probably has better performance on average than any individual ensemble member. The fundamental idea behind ensemble learning is to combine weak learners into one, a strong learner, who has a better generalization error and is less sensitive to overfitting in the presence of noise or small sample size. This is because different classifiers can sometimes misclassify different patterns and accuracy can be improved by combining the decisions of complementary classifiers.

3.2. Elements of an ensemble classifier

A typical ensemble framework for classification tasks contains four fundamental components descripted as follows:

- *Training set*: a training set is a special set of labeled examples providing known information that are used for training.
- *Base inducer*(s) or *base classifier*(s): an inducer is a learning algorithm that is used to learn from a training set. A base inducer obtains a training set and constructs a classifier that generalizes relationship between the input features and the target outcome.
- *Diversity generator*: it is clear that nothing is gained from an ensemble model if all ensemble members are identical. The diversity generator is responsible for generating the diverse classifiers and decides the type of every base classifier that differs from each other. Diversity can be realized in different ways depending on the accuracy of individual classifiers for the improved classification performance. Common diversity creation approaches are (i) using different training sets, (ii) combining different inducers, and (iii) using different parameters for a single inducer.
- *Combiner*: the task of the combiner is to produce the final decision by combining all classification results of the various base inducers. There are two main methods of combining: weighting methods and meta-learning methods. *Weighting methods* give each classifier a weight proportional to its strength and combine their votes based on these weights. The weights can be fixed or dynamically determined when classifying an instance. Common weighting methods are majority voting, performance weighting, Bayesian combination, and vogging. *Meta-learning methods* learn from new training data created from the predictions of a set of base classifiers. The most well-known meta-learning methods are stacking

and grading. While weighting methods are useful when combining classifiers built from a single learning algorithm and they have comparable success, meta-learning is a good choice for cases in which base classifiers consistently classify correctly or consistently misclassify.

4. Ensemble strategies

In order to construct an ensemble model, any of the following strategies can be performed:

4.1. Strategy 1: different training sets using random sampling with replacement

One ensemble strategy is to train different base learners by different subsets of the training set. This can be done by random resampling of a dataset (i.e., *bagging*; **Figure 2a**). When we train multiple base learners with different training sets, it is possible to reduce variance and therefore error.

4.2. Strategy 2: different training sets obtained by random instance and feature subset selection

The combination of bagged decision trees is constructed similar to Strategy 1 using one significant adjustment that random feature subsets are used (i.e., *random forest*; **Figure 2b**). When we have enough trees in the forest, random forest classifier is less likely overfit the model. It is also useful to reduce the variance of low-bias models, besides handling missing values easily.

4.3. Strategy 3: different training sets using random sampling with replacement over weighted data

This ensemble strategy can be implemented by weighted resampling of the dataset serially by focusing on difficult examples which are not correctly classified in the previous steps (i.e., *boosting*; **Figure 2c**). Boosting helps to decrease the bias of otherwise stable learners such as linear classifiers or univariate decision trees also known as decision stumps.

4.4. Strategy 4: different algorithms

The other ensemble strategy (i.e., *voting*; **Figure 2d**) is to use different learning algorithms to train different base learners on the same dataset. So, the ensemble includes diverse algorithms that each takes a completely different approach. The main idea behind this kind of ensemble learning is taking advantage of classification algorithms' diversity to face complex data.

4.5. Characteristic of different ensemble classifiers

Although ensemble classifiers have a common goal to construct multiple, diverse and predictive models and finally to combine their outputs, each strategy is carried out in different ways using different training sets, combiner or inducer. **Table 2** summarizes the properties of different ensemble strategies, the popular algorithms under each category and pros and cons of each ensemble classifier.

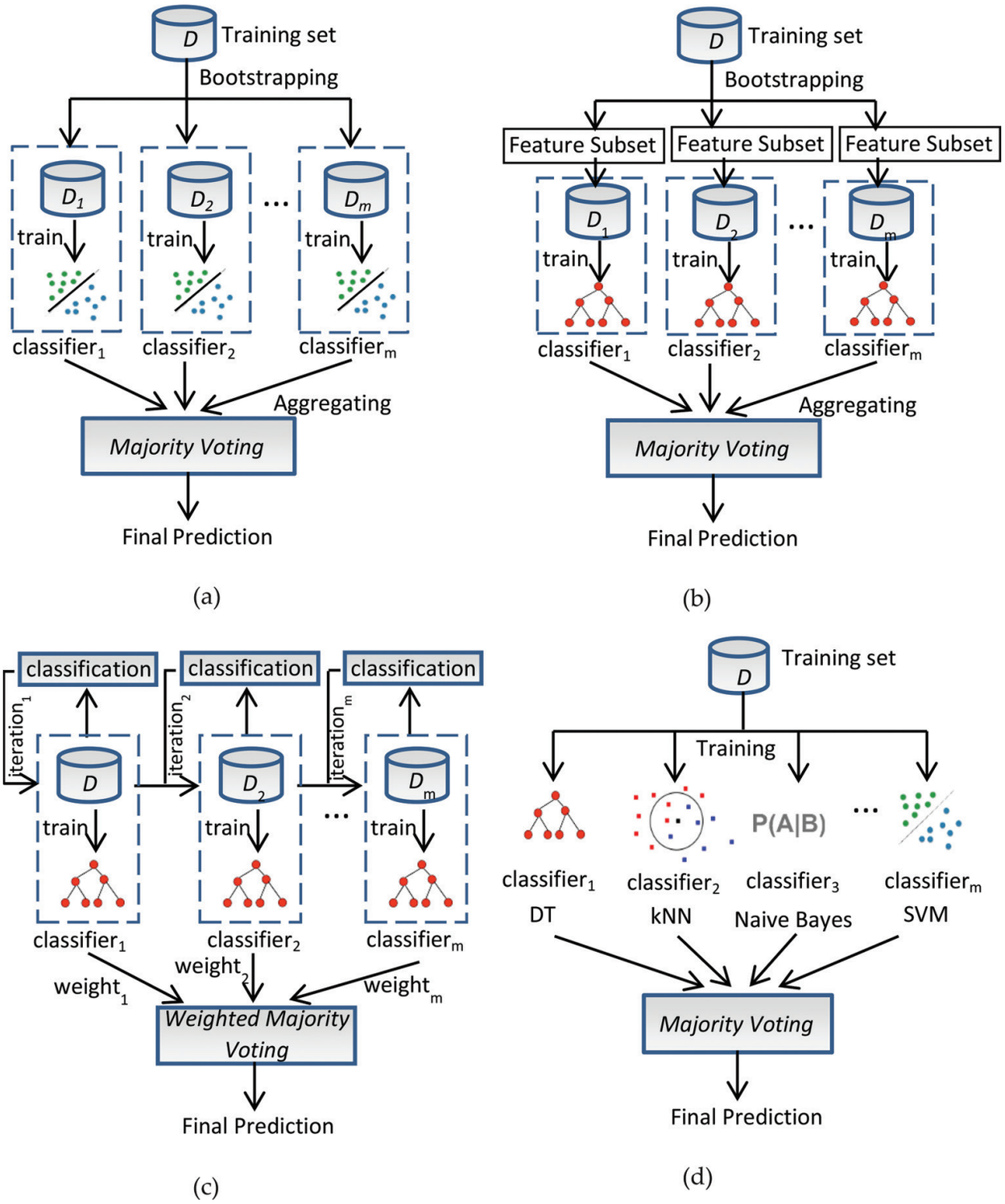

Figure 2. Different ensemble strategies: (a) bagging, (b) random forest, (c) AdaBoost, and (d) voting.

4.6. Challenges of ensemble learning in environmental data mining

Even ensemble-based environmental data mining is helpful based on the advantages indicated in Section 3; there are also challenges that could be overcome when you are aware. Challenges can be grouped under five main titles: selecting ensemble strategy, determining

Algorithm	Training set	Classifiers	Combiner	Inducer	Ensemble strategy	Advantage	Weakness
Bagging	Random resampling	Inducer independent	Majority voting	Single inducer	1	Minimizes variance	A relatively large ensemble size—loss of cooperation with each other
Random forest	Random resampling + feature subset	Inducer dependent (decision tree)	Majority voting	Single inducer	2		
Boosting	Weighted resampling	Inducer independent	Weighted majority voting	Single inducer	3	Boosts the performance of the weak learners	Degrades with noise
AdaBoost	Weighted resampling	Inducer independent	Weighted majority voting	Single inducer	3		
Stacking	Resampling and k-folding	Inducer independent	Meta-learning	Multiinducer	1, 4	Good performance	Storage and time complexity
Grading	Resampling and k-folding	Inducer independent	Meta-learning	Multiinducer	1, 4	Predictions are graded	Storage and time complexity
Voting	Same dataset	Inducer independent	Majority voting	Multiinducer	4	Increase predictive accuracy	How classifiers are selected
Voting	Same dataset	Inducer independent	Majority voting	Single inducer	4	Simple to understand and implement	Limited to a single algorithm performance

Table 2. Characteristic of different ensemble classifiers.

a satisfactory architecture, computational cost, complex nature of environmental data, and finally post processing:

- *Selecting ensemble strategy*: it is a difficult work to determine the best ensemble strategy in terms of accuracy, scalability, computational cost, usability, compactness, and speed of classification. Environmental researchers should know how to construct an ensemble model and be aware of alternative strategies and advantages/disadvantages of them. To overcome this problem, environmental data mining is mostly addressed to computer and environmental scientists working together.
- *Determining a satisfactory architecture*: there are two levels of problems in designing ensemble architecture. First, it is necessary to determine the optimal ensemble size. There are three approaches for determining the ensemble size: (i) preselection of the ensemble size, (ii) selection of the ensemble size while training, and (iii) postselection of the ensemble size (pruning). Second, how are learning algorithms and their respective parameters selected to construct the best ensemble? The best values for the input parameters of the algorithms should be determined through a number of tries. These problems are fundamentally different and should be solved separately to improve classification accuracy. Furthermore, it is necessary to update the model when new environmental data is acquired, allowing the up-to-date model to change over time.

- *Computational cost*: increasing the number of classifiers usually increases computational cost. To overcome this problem, users may predefine a suitable ensemble size limit, or classifiers can be trained in parallel.
- *Complex nature of environmental data*: it is necessary to deal with high dimensionality and complexity of environmental data. To reduce the dimensionality of the feature vector, feature selection techniques can be used such as principal component analysis, information gain, and ReliefF. Another problem is to deal with heterogeneous data by adding problem-specific science algorithms to the solution.
- *Post processing*: another critical issue is determining what the best voting mechanism (majority, weighted, average, etc.) for combining the outputs of base classifiers is. Furthermore, the final results should be presented in an appropriate form to help users understand and interpret easily.

5. Experimental study

In this study, different ensemble learning strategies were compared in terms of classification accuracy, precision (PRE), recall (REC), and f-measure (F-MEA). Four ensemble learning strategies were tested on six different real-world environmental datasets. The application was developed by using Weka open source data mining library.

5.1. Dataset description

In this experimental study, six different datasets that are available for public use were selected to determine the best ensemble strategy. Basic characteristics of the investigated environmental datasets are given in **Table 3**.

ID	Dataset name	Year	Attributes	Instances	Type	Link
1	Ozone (1 h)	2008	74	2536	Air	http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection
2	Ozone (8 h)	2008	74	2534		
3	Leaf	2014	17	340	Ecology	http://archive.ics.uci.edu/ml/datasets/Leaf
4	Eucalyptus	1991	20	736	Soil	https://weka.wikispaces.com/Datasets
5	Forest type	2015	28	523	Ecology	https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping
6	Cloud	1971	8	108	Rainfall	https://github.com/renatopp/arff-datasets/blob/master/statlib/nominal/cloud.arff

Table 3. Environmental datasets and their characteristics.

5.2. Comparison of ensemble strategies

Classification accuracies, precision, recall, and f-measure values for the applied algorithms were obtained using tenfold cross validation. Comparison of the classification accuracies of the applied algorithms for each dataset is displayed in **Figure 3**. Four weak learners (support vector machine (SVM), naive Bayes (NB), decision tree (DT) applied with C4.5 algorithm, and K-nearest neighbor (KNN)) and four ensemble learners (bagging, random forest (RF), AdaBoost, and voting) were used to construct classification models from environmental data. The base classifiers for the ensemble learners were selected as the one which gave the best classification accuracy among the applied weak learners for the respective dataset.

The experimental results were obtained with optimum parameters (given in **Table 4**) using grid search. The best parameters of SVM were found for the complexity parameter, *C* for the exponent value, *E* for polykernel parameters in the interval [10^k for $k \in \{-3, \ldots, 3\}$], and [1–10], respectively. To model DT, confidence factor, *C*, for pruning and the minimum number of objects, *M*, for leaf were obtained in the intervals of [0.05–0.95] and [1–10]. The number of neighbors, *N* for KNN classifier, was selected in the range of [1, 25]. For RF classifier, the number of randomly chosen attributes, *K*, and the number of iterations to be performed, *I*, were found in the intervals [0–15] and [10–100], respectively. The number of ensemble classifiers for bagging is 10 for each dataset. Weight threshold for weight pruning, *P*, and the number of iterations to be performed, *I*, were selected in the interval [10–100] for AdaBoost classifier. Voting was performed using the optimum parameters of SVM, NB, DT, KNN, and RF classifiers.

The objective of this experiment is to remark the success of the ensemble strategies in terms of classification accuracy concerning environmental data. According to the experimental results, it is apparent that the number of correctly classified instances is increased if ensemble strategies are applied. Especially, AdaBoost classifier provides significant performance gain compared to other models. SVM has superiority over other single learners; hence, most of the ensemble models selected it as the base learner.

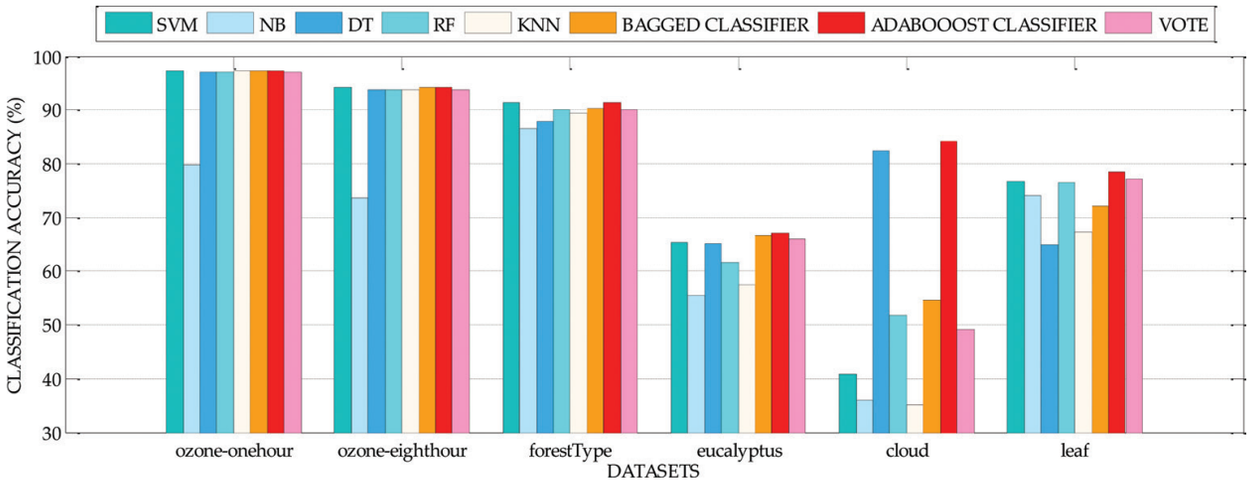

Figure 3. Comparison of single and ensemble classifiers in terms of classification accuracies.

Dataset	SVM		DT		KNN		RF		AdaBoost	
	C	E	C	M	N	Distance metric	I	K	I	P
Ozone (1 h)	10^3	2	0.05	1	17	Euclidean distance	10	5	10	10
Ozone (8 h)	10^3	5	0.55	1	5	Chebyshev distance	10	5	100	10
Leaf	10^0	1	0.05	2	1	Manhattan distance	60	0	100	10
Eucalyptus	10^1	1	0.15	2	9	Manhattan distance	50	2	80	40
Forest type	10^0	1	0.15	3	11	Manhattan distance	50	11	10	10
Cloud	10^0	1	0.05	1	17	Euclidean distance	100	0	100	40

Table 4. Optimum classifier parameters corresponding to each dataset.

There are a number of cases resulting in poor classification performance, such as the following:

- In case of the presence of either noisy or missing data
- If there is an insufficient number of instances available
- If there are too many number of classes
- If a complex relationship is inherent
- If the feature dependencies are ignored
- If the feature selection is not well performed
- If the algorithm parameters are not correctly determined
- If the class labels are imbalanced

For example, because the number of instances in "cloud" dataset is very few (due to the insufficient number of instances), inferior results are obtained for most of the applied algorithms as expected. However, even in such cases while some algorithms fail, some others manage to perform well (e.g., $C_{4.5}$ DT 82%). In this situation, the classifier's performance can also be enhanced by applying ensemble learning methods as in the case of AdaBoost with 84% classification accuracy for the same dataset. AdaBoost is a powerful ensemble learning algorithm because its distribution update step ensures that instances misclassified by the previous classifier are more likely to be included in the training data of the next classifier with the chance of further enhancement.

Due to the fact that classification accuracy as a performance metric is not just enough to decide whether a learner is considerably good or not, the precision, recall, and f-measure values were also calculated for each model (**Table 5**). It is also clear from the table values that applying ensemble strategies compared to single learners makes more sense in terms of classifier performance.

Dataset	Algorithm	PRE	REC	F-MEA	Dataset	Algorithm	PRE	REC	F-MEA
Ozone (1-h)	SVM	0.97	0.97	0.95	Eeucalyptus	SVM	0.65	0.65	0.65
	NB	0.96	0.79	0.86		NB	0.62	0.55	0.55
	$C_{4.5}$ DT	0.94	0.97	0.95		$C_{4.5}$ DT	0.66	0.65	0.64
	RF	0.94	0.97	0.95		RF	0.61	0.61	0.61
	K-NN	0.97	0.97	0.95		K-NN	0.57	0.57	0.56
	$K\text{-}NN_{Bagged}$	0.97	0.97	0.95		SVM_{Bagged}	0.66	0.66	0.66
	$K\text{-}NN_{AdaBoost}$	0.97	0.97	0.95		$SVM_{AdaBoost}$	0.67	0.67	0.67
	Vote	0.94	0.97	0.95		Vote	0.67	0.65	0.65
Ozone (8-h)	SVM	0.93	0.94	0.93	Cloud	SVM	0.37	0.40	0.37
	NB	0.92	0.73	0.80		NB	0.49	0.36	0.32
	$C_{4.5}$ DT	0.87	0.93	0.90		$C_{4.5}$ DT	0.82	0.82	0.82
	RF	0.91	0.93	0.91		RF	0.51	0.51	0.51
	K-NN	0.87	0.93	0.90		K-NN	0.33	0.35	0.32
	SVM_{Bagged}	0.92	0.94	0.93		$C_{4.5}$ DT_{Bagged}	0.55	0.54	0.54
	$SVM_{AdaBoost}$	0.93	0.94	0.93		$C_{4.5}$ $DT_{AdaBoost}$	0.84	0.84	0.84
	Vote	0.93	0.93	0.91		Vote	0.47	0.49	0.46
Forest types	SVM	0.91	0.91	0.91	Leaf	SVM	0.78	0.76	0.76
	NB	0.86	0.86	0.86		NB	0.75	0.74	0.74
	$C_{4.5}$ DT	0.88	0.88	0.87		$C_{4.5}$ DT	0.66	0.65	0.64
	RF	0.90	0.90	0.90		RF	0.77	0.76	0.76
	K-NN	0.89	0.89	0.89		K-NN	0.69	0.67	0.67
	SVM_{Bagged}	0.90	0.90	0.90		SVM_{Bagged}	0.72	0.72	0.71
	$SVM_{AdaBoost}$	0.91	0.91	0.91		$SVM_{AdaBoost}$	0.79	0.78	0.78
	Vote	0.90	0.90	0.90		Vote	0.77	0.77	0.76

Table 5. Precision (PRE), recall (REC), and f-measure (F-MEA) results using tenfold cross validation for respective algorithms in each dataset.

6. Conclusion and future work

This study aims to provide helpful guidelines for future applications by presenting the advantages and challenges of ensemble-based environmental data mining and comparing alternative ensemble strategies through experimental studies. It compares four different ensemble

strategies for environmental data mining: (i) bagging, (ii) bagging combined with random feature subset selection, (iii) boosting, and (iv) voting. In the experimental studies, ensemble methods are tested on different real-world environmental datasets.

In the future, the following studies can be carried out:

- Multistrategy ensemble learning that combines several ensemble strategies can be addressed, instead of a single ensemble strategy.
- Text mining, web mining, and process mining have been used in many engineering fields. However, there is very limited usage of them in environmental engineering. Future research can focus on these subjects.
- Some ontologies can be developed for environmental domain. We believe that the future environmental data mining studies will be supported by the ontologies to extract semantic relationships, to improve accuracy, and to develop better decision support systems.

Author details

Goksu Tuysuzoglu[1], Derya Birant[2]* and Aysegul Pala[3]

*Address all correspondence to: derya@cs.deu.edu.tr

1 Graduate School of Natural and Applied Sciences, Dokuz Eylul University, Izmir, Turkey

2 Department of Computer Engineering, Dokuz Eylul University, Izmir, Turkey

3 Department of Environmental Engineering, Dokuz Eylul University, Izmir, Turkey

References

[1] Stojić A, Stojić SS, Reljin I, Čabarkapa M, Šoštarić A, Perišić M, Mijić Z. Comprehensive analysis of PM10 in Belgrade urban area on the basis of long-term measurements. Environmental Science and Pollution Research. 2016;**23**:10722-10732. DOI: 10.1007/s11356-016-6266-4

[2] Srivastava AN. Greener aviation with virtual sensors: A case study. Data Mining and Knowledge Discovery. 2012;**24**:443-471. DOI: 10.1007/s10618-011-0240-z

[3] Al Abri ES, Edirisinghe EA, Nawadha A. Modelling ground-level ozone concentration using ensemble learning algorithms. In: Proceedings of the International Conference on Data Mining (DMIN'15); 27-30 July 2015; Las Vegas. USA: The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2015. pp. 148-154

[4] Bougoudis I, Demertzis K, Iliadis L. HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens. Neural Computing and Applications. 2016;**27**:1191-1206. DOI: 10.1007/s00521-015-1927-7

[5] Lei KS, Wan F. Applying ensemble learning techniques to ANFIS for air pollution index prediction in Macau. In: International Symposium on Neural Networks (ISNN'12); 11-14 July 2012. Berlin, Heidelberg: Springer; 2012. pp. 509-516

[6] Singh KP, Gupta S, Rai P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment. 2013;**80**:426-437. DOI: 10.1016/j.atmosenv.2013.08.023

[7] Granata F, de Marinis G. Machine learning methods for wastewater hydraulics. Flow Measurement and Instrumentation. 2017;**57**:1-9. DOI: 10.1016/j.flowmeasinst.2017.08.004

[8] Budka M, Gabrys B, Ravagnan E. Robust predictive modelling of water pollution using biomarker data. Water Research. 2010;**44**:3294-3308. DOI: 10.1016/j.watres.2010.03.006

[9] Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L. Predictive modeling of groundwater nitrate pollution using random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of the Total Environment. 2014;**476**:189-206. DOI: 10.1016/j.scitotenv.2014.01.001

[10] Heung B, Hodúl M, Schmidt MG. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma. 2017;**290**: 51-68. DOI: 10.1016/j.geoderma.2016.12.001

[11] Halmy MWA, Gessler PE. The application of ensemble techniques for land-cover classification in arid lands. International Journal of Remote Sensing. 2015;**36**:5613-5636. DOI: 10.1080/01431161.2015.1103915

[12] Wang Q, Xie Z, Li F. Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. Environmental Pollution. 2015;**206**: 227-235. DOI: 10.1016/j.envpol.2015.06.040

[13] Crimmins SM, Dobrowski SZ, Mynsberge AR. Evaluating ensemble forecasts of plant species distributions under climate change. Ecological Modelling. 2013;**266**:126-130. DOI: 10.1016/j.ecolmodel.2013.07.006

[14] Engler R, Waser LT, Zimmermann NE, Schaub M, Berdos S, Ginzler C, Psomas A. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. Forest Ecology and Management. 2013;**310**:64-73. DOI: 10.1016/j.foreco.2013.07.059

[15] Healey SP, Cohen WB, Yang Z, Brewer CK, Brooks EB, Gorelick N, et al. Mapping forest change using stacked generalization: An ensemble approach. Remote Sensing of Environment. 2018;**204**:717-728. DOI: 10.1016/j.rse.2017.09.029

[16] Gaál M, Moriondo M, Bindi M. Modelling the impact of climate change on the Hungarian wine regions using random forest. Applied Ecology and Environmental Research. 2012;**10**:121-140. DOI: 10.15666/aeer/1002_121140

[17] Nelson TA, Coops NC, Wulder MA, Perez L, Fitterer J, Powers R, Fontana F. Predicting climate change impacts to the Canadian Boreal forest. Diversity. 2014;**6**:133-157. DOI: 10.3390/d6010133

[18] Muñoz-Mas R, Lopez-Nicolas A, Martínez-Capel F, Pulido-Velazquez M. Shifts in the suitable habitat available for brown trout (*Salmo trutta* L.) under short-term climate change scenarios. Science of the Total Environment. 2016;**544**:686-700. DOI: 10.1016/j.scitotenv.2015.11.147

[19] Bravo-Moncayo L, Naranjo JL, García IP, Mosquera R. Neural based contingent valuation of road traffic noise. Transportation Research Part D: Transport and Environment. 2017;**50**:26-39. DOI: 10.1016/j.trd.2016.10.020

[20] Kühnlein M, Appelhans T, Thies B, Nauss T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. Remote Sensing of Environment. 2014;**141**:129-143. DOI: 10.1016/j.rse.2013.10.026

[21] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy. 2014;**127**:1-10. DOI: 10.1016/j.apenergy.2014.04.016

[22] Araya DB, Grolinger K, ElYamany HF, Capretz MA, Bitsuamlak G. An ensemble learning framework for anomaly detection in building energy consumption. Energy and Buildings. 2017;**144**:191-206. DOI: 10.1016/j.enbuild.2017.02.058

[23] Jovanović RŽ, Sretenović AA, Živković BD. Ensemble of various neural networks for prediction of heating energy consumption. Energy and Buildings. 2015;**94**:189-199. DOI: 10.1016/j.enbuild.2015.02.052

[24] Knudby A, Brenning A, LeDrew E. New approaches to modelling fish–habitat relationships. Ecological Modelling. 2010;**221**:503-511. DOI: 10.1016/j.ecolmodel.2009.11.008

[25] Kocev D, Džeroski S. Habitat modeling with single-and multi-target trees and ensembles. Ecological Informatics. 2013;**18**:79-92. DOI: 10.1016/j.ecoinf.2013.06.003

[26] Zhang Z, Ma C, Xu J, Huang J, Li L. A novel combinational forecasting model of dust storms based on rare classes classification algorithm. In Geo-Informatics in Resource Management and Sustainable Ecosystem (GRMSE'14); October 2014. Berlin, Heidelberg: Springer; 2015. pp. 520-537

[27] Mathanker SK, Weckler PR, Taylor RK, Fan G. AdaBoost and support vector machine classifiers for automatic weed control: Canola and Wheat. In: 2010 Pittsburgh, Pennsylvania, 20-23 June 2010; American Society of Agricultural and Biological Engineers. 2010. p. 1

[28] Lima AR, Cannon AJ, Hsieh WW. Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy. Computers & Geosciences. 2013;**50**:136-144. DOI: 10.1016/j.cageo.2012.06.023

[29] Luo Q, Kathuria A. Modelling the response of wheat grain yield to climate change: A sensitivity analysis. Theoretical and Applied Climatology. 2013;**111**:173-182. DOI: 10.1007/s00704-012-0655-5

[30] Mohammed AA, Yaqub W, Aung Z. Probabilistic forecasting of solar power: An ensemble learning approach. Intelligent Decision Technologies. Smart Innovation, Systems and Technologies. 2015;**39**:449-458. DOI: 10.1007/978-3-319-19857-6_38