Clustering Tree Algorithm for Biodiversity Modeling of Diatoms

Andreja Naumoski, Georgina Mirceva, Kosta Mitreski

Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Skopje, North Macedonia
{andreja.naumoski,georgina.mirceva,
kosta.mitreski}@finki.ukim.mk

Abstract. Knowledge discovery from environmental data aims to understand the measured ecological data underlying patterns and to provide possible prediction of future events. Besides the environmental data at hand, this process must include a machine learning algorithm that will produce an accurate and human reasonable model. Both ecologists and decision-makers agree on this. That's why in this paper, we investigate the influence of the parameter tuining on a predictive tree decision-making algorithm, the predictive clustering trees, in the process of obtaining biodiversity models when using different number of discretization levels for the target attribute. Here, the biodiversity index is the target attribute, which is calculated from the diatoms' abundances and it is discretized. For building model, we use a decision tree machine learning algorithm that produces relatively accurate models and most important the models are very easy to interpret by the biologists, who do not need to be fimiliar with the inner working of the algorithm itself. Besides the experimental evaluation of the models' performance and statistical significance of the results, some of the obtained models are also presented and discussed.

Keywords: Ecological modeling · Biodiversity indices · Diatoms · Machine learning algorithm

1 Introduction

The biodivessity indces are computer from measured parameters. These mathematical measures can give ecologists a different view on the consistency and the dynamic of the diatom water community, therefore they give much richer information, besides the simple info for the diatoms' abundance. Biodiversity indices can correlate the diatoms' community (type of algaes) dynamics with the environmental stress factors well, if the sample's data are measured at the same time, and a well-suited algorithm can produce satisfactory informative models [1]. The same goal is focus of the habitat suitability modeling, where the models are focused on revealing the relationship between the abiotic factors and the dynamic of the diatom community [2]. Since diatoms are organisms that react very fast in the changing environment, combined with the biodiversity indi-

ces, that can be a very effective tool for recovery and prevention of the ecological degradation of the water ecosystem. As always, this modeling is not straight forward, since many ecological relationships are not linear in nature, and therefore, it is not easy to express the relationship between the biodiversity indices and the environmental parameters. One direction of research, of course, could be to investigate a machine learning algorithm, used usually for clustering or classification [3, 4, 5, 6, 7, 8, 9], that are efficient in such modeling of non-linear relationship. But we took a different path in this paper, and we focus on investigating a particular machine learning algorithm, the predictive clustering trees, which were used previously for modeling the relationship between the diatoms and the environment [10].

The biodiversity indices have been adapted from mathematical science for many research disciplines including ecological studies [11, 12]. They have been applied in studies to investigate the biodiversity in different ecological environments like oceans, lakes, ponds, mountains etc. [13, 14, 15]. There are also studies that investigate the importance of the differences and the common goals of the biodiversity indices, in order to better understand the impact of these indices that have on the outcome [16, 17, 18, 19]. In order to get the biodiversity information from the water sample, the researchers use the raw abundance data for each specie in the community, and then by using software that will calculate the biodiversity indices, we can obtain each index separately. If in this data, we incorporate the measurements regarding the environmental conditions, then we can try to find the relationship between these two elements of the equation.

As we pointed out earlier, in this paper we focus on one decision tree algorithm. This type of machine learning algorithm was also used in a study regarding the environmental influence on the diatoms' biodiversity indices in Lake Prespa [20, 21]. The results of these studies shed new light on the underlying ecological pattern in the relationship between the lake diatoms and the environmental factors. Besides this, we must look also on the downsides of these models, that many of them obtained medium accuracy and for many of the known biodiversity indices, the algorithms didn't produce any model. As the authors in the paper [21], concluded, further investigation of the biodiversity categorization of the target attributes is needed. Following this recommendation, in this paper, besides the investigation of the algorithm's parameter fitting, we investigate the influence of the different number of discretization levels.

The data at hand in our biodiversity classification quest consists of several physicochemical parameters that are measured during the EU monitoring program of the Lake Prespa, combined with the 10 biodiversity indices that are calculated from the diatoms' abundance data of the 116 different diatoms in the water sample. For evaluation purposes of the decision tree algorithm used in this paper, we have taken into consideration two different biodiversity models and using two different discretization levels - 3 and 5 of each biodiversity index. These two sets of models are also tested for accuracy and statistical significance. Furthermore, several settings are considered for the algorithm, like maximum tree size, maximum weight and maximum depth. The statistical significance test is done by using the two-stage procedure described in [22], combined with the Aligned Friedman test [23] and Hommel test [24]. By using the two Aligned Fried-

man and Hommel tests, the two-stage procedure forms a non-parametric statistical comparison, which will be used to estimate the statistical significance of the obtained set of models. Using this procedure, the models obtained from the datasets and applied algorithm are examined and ranked based on their performance statistics. The models that are obtained by the best performing algorithm in the experiments are taken as control model and then the rest of the models obtained with the other algorithms are tested with respect to the control model.

The rest of the paper is organized as follows: Section II presents the dataset description and the experimental setup, while in section III the experimental results are presented, as well as two decision tree models with their respective discussion. Finally, Section IV concludes the paper and the research direction is outlined.

2 Dataset Description and Experimental Setup

As we noted in the previous section, several physico-chemical parameters were measured, as well as the biological information about the diatoms that live in the lake, counted under microscope [25]. The ecological dataset used in this paper was gathered during the 16 months monitoring campaign of the Lake Prespa as part of the TRABOREMA Project [25]. Using this data, we can derive ecological knowledge regarding the underlying pattern that shapes the life of the diatoms' community in a direct correlation with the environmental conditions, and from that information, the ecologists and decision-makers can set new environmental directives and policies. The 116 different diatom species that were part of the sampling campaign, for the purpose of this paper are converted into 10 biodiversity indices. The set of 10 biodiversity indices that consist our set are: Brilouni [26], E1-D [27], E1/D [27], E-ln(D) [27], Heip [28], Hill [29], Hurlbert [30], Mcintosh [31], Simpson [32] and Shannon [33].

Due to the nature of the measurements, it is very difficult as well as time and technologically demanding to measure the exact value of each parameter on the lake surface and inside the lake. It is much harder to know the exact number of diatoms in each space and time in the lake, the ecological science uses categories or defined ranges where the given parameter can be expected or can be used to indicate some other phenomena. Therefore, it is much more useful to calculate the biodiversity index and to express it by using categories. For this purpose, as we previously stated, we transformed each biodiversity index into three categories (low, medium and high) and five categories (very low, low, medium, high, very high). We have done this by dividing the biodiversity indices into three and five equal intervals between the lowest and the highest value for each index. By doing this, we directly affect the accuracy of the obtained biodiversity models produced by the decision tree machine learning algorithm. The dataset, besides the 10 biodiversity indices, consists of the 21 measured input physico-chemical parameters, together form the evaluation dataset.

The model performance is evaluated by using the AUC-ROC evaluation measure. Two AUC-ROC values are obtained for each model. First, the descriptive performance is obtained by using the entire dataset for training the model and testing the model. On the other hand, the predictive performance of the biodiversity model is obtained using

the standard 10-fold cross-validation procedure. For the purpose of evaluation, the decision tree algorithm [34] implemented in the CLUS system is used, and by fitting its parameters to obtain the best models, we approach with three strategies. First, we constrain the tress with their size, and we set the maximum size for Garofalakis [35] pruning to 7 (MS7), 9 (MS9), 11 (MS11) and 13 (MS13). Second, we constrain the maximum depth of the tree which can grow to 3 (MD3), 4 (MD4) and 5 (MD5). And third, we constrain the minimal weight (minimal number of examples in each subtree to 2 (MW2), 4 (MW4), 8 (MW8), 16 (MW16) and 32 (MW32). Furthermore, we also compare these evaluation results with the No Constrained tree model (NoCo), to estimate the improvement we gain.

We examine the statistical significance of the model results with the two-stage procedure, described in details in [22]. As input into this procedure, we take the results of the descriptive and predictive power of the models, for each biodiversity index and particular parameter setting for the decision tree algorithm. Then, by using the non-parametric Aligned Friedman procedure [23], we rank the model performance in which we set the statistical significance *p*-value to 0.05. This procedure is used since the number of both biodiversity indices models and decision tree variants are small. If the outcome of the test shows rejection of the null hypothesis by the Aligned Friedman test, this means that there is a significant difference among the biodiversity index datasets, not the performance of the models obtained from these datasets. In case when the null hypothesis is rejected, then we proceed with the second part of the procedure, with the post hoc Hommel test [24]. It is important to note that the Hommel test changes the level of statistical significance in each iteration, therefore the results that we have to consider in our results are the adjusted *p*-values to make a fair comparison to the previously set *p*-value. In the next section, we present the outcomes of the Aligned Friedman test (Average rank out of each experiment) and the Hommel test (Control model and the models that have statistical significance). The reader should carefully consider all of this, and for further in-depth information on how this two-stage procedure works, we kindly ask the reader to read [23].

3 Experimental Results

This section consists from the results of the experimental evaluation of the biodiversity index datasets by using 3 and 5 categories (denoted as C3 and C5). The results for both descriptive and predictive part of the experiment are presented in this section. Furthermore, the results of the statistical significance experiments and two biodiversity models that are discussed at the end of the section.

3.1 Performance Analysis

The evaluation results for the descriptive performance, represented through the AUC-ROC metric, when using the 3 categories of the biodiversity target are presented in Table 1. Analyzing the results, it is clear that two decision tree algorithm parameters setting stand out: maximum depth 3 (MD3) and minimal weight 4 (MW4), and only

one biodiversity index stands out – Brilouni index, for which different algorithm settings achieved the highest performance. Besides the high AUC-ROC values above 0.93 for the best settings for the decision tree algorithm, also the rest of the models have relatively high AUC-ROC values between 0.7 and 0.9.

Table 1. The AUC-ROC values obtained in the evaluation of the descriptive performances of the decision tree algorithm by using 3 categories for the biodiversity indices. The highest value per algorithm's experimental setup is bolded

	Biodiversity Indices									
	Brilouni	E_{1-D}	$E_{1/D}$	$E_{-ln(D)}$	Heip	Hill	Hurlbert	Mcin-tosh	Simp-son	Shan-non
MD3	0.86	**0.96**	**0.97**	**0.95**	**0.94**	**0.96**	**0.94**	**0.96**	**0.96**	**0.97**
MD4	0.88	0.74	0.92	0.73	0.77	0.72	0.79	0.83	0.70	0.88
MD5	0.89	0.77	0.95	0.75	0.80	0.76	0.83	0.86	0.70	0.89
MS11	0.88	0.76	0.93	0.73	0.77	0.72	0.77	0.84	0.70	0.88
MS13	0.88	0.78	0.94	0.74	0.79	0.75	0.78	0.85	0.70	0.89
MS7	0.86	0.80	0.95	0.80	0.80	0.77	0.83	0.87	0.70	0.89
MS9	0.87	0.74	0.92	0.72	0.77	0.70	0.75	0.83	0.70	0.87
MW16	0.86	0.87	0.92	0.82	0.78	0.83	0.81	0.90	0.84	0.88
MW2	**0.93**	0.80	0.95	0.75	0.79	0.76	0.78	0.87	0.71	0.89
MW32	0.82	0.81	0.89	0.76	0.75	0.73	0.75	0.85	0.75	0.85
MW4	0.91	**0.96**	**0.97**	**0.95**	**0.94**	**0.96**	**0.94**	**0.96**	**0.96**	**0.97**
MW8	0.86	0.92	0.95	0.88	0.89	0.94	0.92	0.92	0.93	0.90
NoCo	**0.93**	0.79	0.89	0.71	0.75	0.70	0.70	0.83	0.73	0.83

The second table, Table 2, presents the obtained AUC-ROC values for the descriptive performance of the experimental evaluation for the biodiversity indices datasets with 5 category classes. As in Table 1, here again the algorithm's settings MD3 and MW4 are the optimal ones, except for the Brilouni index. The results in Table 2 present a similar picture, but slightly lower overall AUC-ROC values that range between 0.6 and 0.8.

A different story could be told if we look at the predictive performance obtained from the experimental evaluation on both datasets. The AUC-ROC results from the biodiversity indices datasets with 3 category classes, presented in Table 3, have the highest values if we select the maximum depth 5 (MD5) and maximum size of the tree model of 11 (MS11) as algorithm setting. The value of the range of AUC-ROC for these settings is around 0.9, while the rest of the models have obtained AUC-ROC values between 0.5 and 0.9.

Table 2. The AUC-ROC values obtained in the evaluation of the descriptive performances of the decision tree algorithm by using 5 categories for the biodiversity indices. The highest value per algorithm's experimental setup is bolded

	Biodiversity indices									
	Brilouni	E_{1-D}	$E_{1/D}$	$E_{ln(D)}$	Heip	Hill	Hurlbert	Mcin-tosh	Simp-son	Shan-non
MD3	0.73	**0.93**	**0.93**	**0.89**	**0.83**	**0.96**	**0.76**	**0.94**	**0.93**	**0.95**
MD4	0.77	0.65	0.83	0.68	0.66	0.69	0.63	0.72	0.69	0.74
MD5	0.83	0.71	0.86	0.74	0.69	0.73	0.69	0.78	0.71	0.76
MS11	0.75	0.66	0.81	0.71	0.65	0.69	0.63	0.72	0.67	0.74
MS13	0.75	0.67	0.82	0.71	0.66	0.71	0.62	0.73	0.68	0.75
MS7	0.73	0.76	0.88	0.78	0.72	0.77	0.71	0.84	0.73	0.82
MS9	0.74	0.65	0.80	0.70	0.65	0.68	0.62	0.70	0.66	0.74
MW16	0.71	0.79	0.81	0.73	0.67	0.81	0.67	0.80	0.77	0.81
MW2	**0.94**	0.68	0.83	0.72	0.66	0.72	0.64	0.74	0.70	0.76
MW32	0.71	0.70	0.80	0.70	0.64	0.77	0.60	0.76	0.72	0.75
MW4	0.88	**0.93**	**0.93**	**0.89**	**0.83**	**0.96**	**0.76**	**0.94**	**0.93**	**0.95**
MW8	0.80	0.86	0.88	0.84	0.78	0.90	0.73	0.88	0.88	0.88
NoCo	**0.94**	0.65	0.76	0.64	0.63	0.69	0.59	0.70	0.63	0.72

Table 3. The AUC-ROC values obtained in the evaluation of the predictive performances of the decision tree algorithm by using 3 categories for the biodiversity indices. The highest value per algorithm's experimental setup is bolded

	Biodiversity Indices									
	Brilouni	E_{1-D}	$E_{1/D}$	$E_{-ln(D)}$	Heip	Hill	Hurlbert	Mcin-tosh	Simp-son	Shan-non
MD3	0.80	0.78	0.76	0.78	0.78	0.79	0.78	0.80	0.72	0.77
MD4	0.79	0.72	0.65	0.69	0.68	0.68	0.67	0.67	0.69	0.68
MD5	0.71	0.65	0.69	0.69	0.88	**0.89**	**0.90**	**0.90**	**0.90**	**0.89**
MS11	**0.88**	**0.90**	**0.88**	**0.88**	**0.90**	**0.89**	0.60	0.60	0.65	0.63
MS13	0.61	0.61	0.62	0.63	0.59	0.60	0.62	0.63	0.59	0.68
MS7	0.68	0.69	0.67	0.69	0.65	0.69	0.71	0.59	0.64	0.58
MS9	0.55	0.56	0.56	0.55	0.54	0.53	0.54	0.53	0.56	0.57
MW16	0.52	0.63	0.64	0.64	0.64	0.62	0.63	0.64	0.63	0.67
MW2	0.63	0.60	0.64	0.77	0.79	0.79	0.79	0.79	0.78	0.81
MW32	0.80	0.81	0.77	0.79	0.80	0.61	0.63	0.63	0.63	0.62
MW4	0.63	0.62	0.62	0.58	0.61	0.64	0.62	0.54	0.57	0.59
MW8	0.57	0.56	0.58	0.56	0.60	0.66	0.54	0.55	0.63	0.81
NoCo	0.77	0.80	0.78	0.77	0.81	0.79	0.78	0.79	0.74	0.83

On the other hand, the highest AUC-ROC results from the biodiversity indices datasets with 5 category classes (see Table 4) have the same decision tree algorithm settings: maximum depth 5 (MD5) and maximum size of the tree model of 11 (MS11), but with lower values around 0.75 compared to Table 3. Similarly, the rest of the AUC-ROC values ranges between 0.5 and 0.7.

Table 4. The AUC-ROC values obtained in the evaluation of the predictive performances of the decision tree algorithm by using 5 categories for the biodiversity indices. The highest value per algorithm's experimental setup is bolded

	Biodiversity Indices									
	Brilouni	E_{1-D}	$E_{1/D}$	$E_{-ln(D)}$	Heip	Hill	Hurlbert	Mcin-tosh	Simp-son	Shan-non
MD3	0.66	0.66	0.68	0.67	0.67	0.66	0.67	0.65	0.66	0.64
MD4	0.64	0.66	0.58	0.61	0.60	0.59	0.60	0.61	0.60	0.59
MD5	0.59	0.58	0.58	0.58	0.68	**0.76**	**0.74**	**0.76**	**0.76**	**0.76**
MS11	**0.71**	**0.74**	**0.75**	**0.68**	**0.71**	**0.76**	0.56	0.62	0.61	0.62
MS13	0.60	0.64	0.63	0.61	0.64	0.56	0.59	0.60	0.53	0.52
MS7	0.55	0.53	0.53	0.56	0.50	0.53	0.53	0.53	0.53	0.58
MS9	0.58	0.53	0.58	0.56	0.59	0.62	0.60	0.56	0.63	0.56
MW16	0.60	0.55	0.56	0.57	0.56	0.56	0.56	0.56	0.54	0.56
MW2	0.55	0.56	0.55	0.62	0.64	0.61	0.62	0.65	0.62	0.65
MW32	0.63	0.63	0.62	0.65	0.65	0.56	0.60	0.59	0.60	0.59
MW4	0.59	0.63	0.59	0.62	0.56	0.58	0.59	0.54	0.59	0.60
MW8	0.58	0.54	0.60	0.57	0.60	0.59	0.54	0.55	0.60	0.70
NoCo	0.68	0.69	0.66	0.65	0.70	0.62	0.67	0.66	0.62	0.68

If we compare the experimental results between the biodiversity datasets with 3 categorization classes (Table 1 and Table 2) and 5 categorization classes (Table 3 and Table 4), there is a clear pattern that the models with 5 categorization classes have lower descriptive as well as predictive power. This is a general conclusion not only for the models with the highest value of the AUC-ROC metric, but also for all the models in general. Secondly, as in most of the experiments done with machine learning algorithms, the predictive performance of the models on unseen data is lower than the model on training data. But biodiversity models with 3 categories have better overfitting resistance, compared to the biodiversity models with 5 categories of the target attribute. And thirdly, if we compare each of the biodiversity models and their descriptive or predictive performance, there is a little difference between their highest AUC-ROC values.

3.2 Ranking the Experiments

As we gave some short discussion and conclusion regarding the results from the experimental evaluation for different experimental setups, we noticed that the best model

results can sometimes be a random chance. Therefore, as we stated at the begging of the paper, we will test the obtained results for statistical significance. And to that, we used the two-step procedure consisting of the Aligned Friedman test [23] as well as the post-hoc Hommel test [24] to determine the models with a statistical difference when the number of categories for the class attribute is 3 and 5, as well as a combination of both these datasets and the results from the statistical analyses. The results are presented in Table 5. According to the analysis for the descriptive performances for the biodiversity datasets with 3 and 5 categories for the class attribute (C3_Train and C5_Train), the models built by the MW4 decision tree algorithm variant has the best average rank (lowest value) and this parameter setting is taken as control model. The MW4 model compared to the rest of the models, has statistically significant results, except for the MD3 model variant.

Table 5. The average rank of the decision tree classification algorithm experimental setups. The model with the best average rank is taken as a control model in the second part of the statistical significance procedure. The model variants that are rejected by the Hommel test are underlined. These models have a statistically significant difference with the control method

	C3_Train	**C3_Test**	**C5_Train**	**C5_Test**	**C3C5_Train**	**C3C5_Test**
MD3	2.40	3.80	2.39	3.05	2.40	3.42
MD4	9.80	5.95	9.45	6.85	9.62	6.40
MD5	6.45	**3.25**	5.65	4.30	6.05	3.77
MS11	9.55	4.15	9.80	3.40	9.67	3.77
MS13	7.85	9.90	8.80	8.15	8.32	9.02
MS7	6.20	7.40	4.95	12.5	5.75	9.95
MS9	11.0	12.8	11.2	8.90	11.1	10.8
MW16	6.35	8.80	6.20	10.3	6.27	9.55
MW2	5.95	5.25	6.85	6.75	6.40	6.00
MW32	9.50	5.85	9.25	6.65	9.37	6.25
MW4	**1.65**	10.2	**1.65**	8.55	**1.65**	9.40
MW8	3.90	10.2	3.25	8.70	3.57	9.47
NoCo	10.4	3.35	11.5	**2.90**	10.9	**3.12**

A different situation is observed when we look at the evaluation of the predictive model performance, in which according to Table 5 the best average rank for the biodiversity dataset with 3 categories (C3_Test) is MD5, while the decision tree algorithm without any modifications (NoCo) is the best biodiversity model with 5 category classes. And most important is that these control models have statistically significant results compared to the rest of the models. The last analysis of the biodiversity combined datasets, which consist of both 3 and 5 categorizations per target attribute. The conducted analysis showed that the MW4 decision tree variant is best for the C3C5_Train dataset, while the NoCo variant for the predictive performance on the C3C5_Test dataset. And

finally, the results of the two-stage procedure, show that these control models are statistically significantly difference with most of the other models.

3.3 Model Results

In this section, we present two decision tree models for biodiversity modeling for the Simpson biodiversity index using 3 and 5 discretization levels (C3 and C5). On Fig. 1, we can note that the most influencing factor on the Simpson biodiversity index is the levem of potassium (K), followed by the levels of manganese (Mn) and magnesium (Mg). According to the model, higher values of potassium above 3.36 mg/l in the water sample would result in high diatom Simpson biodiversity, while the lower values will result in a general medium to low Simpson biodiversity. Further down the model, we can note that the low biodiversity according to the Simpson formula, will be achieved if we have Magnesium concentration higher than 1.73 mg/l.

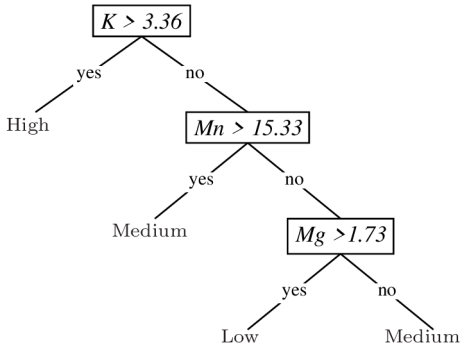

Fig. 1. Biodiversity tree models obtained using C3 discretization and MD5 model variant in the algorithm for the Simpson biodiversity index

The second model presented on Fig. 2 shows a quite different picture even with the fact that the three physico-chemical paramters that occurred in the previous model occur in this model too. According to this model, the levels of conductivity is the most important, followed by the levels of magnesium and the sulfates (SO_4). If we look down carefully in the leaves of the model, we can note that high value of the Simpson biodiversity index is achieved if potassium concentration is higher than 3.36 mg/l, pH is higher than 7.85 and total nitrogen concertation on the water sample is higher than 1.97 mg/l. On the other end of the scale, the low and bad values for the Simpson biodiversity index, according to the model is found when potassium concentrations are lower than 3.36 mg/l, SO_4 is higher than 24.04 mg/l, and total nitrogen concentration is lower than 1.97 mg/l and conductivity is less than 152 μS/cm.

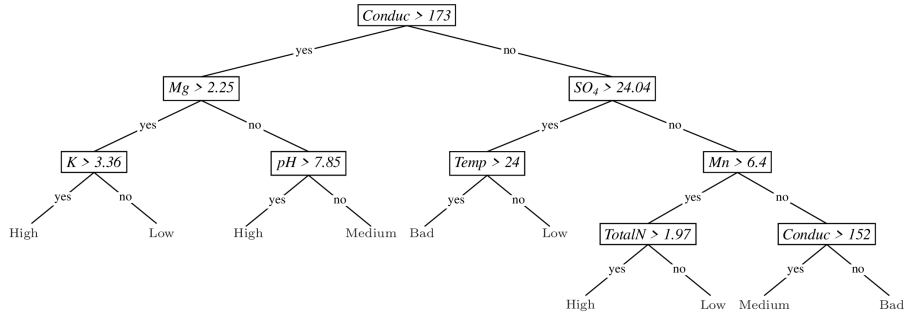

Fig. 2. Biodiversity tree models obtained using C5 discretization and MD5 model variant in the algorithm for the Simpson biodiversity index

4 Conclusion

In this paper, we experimentally evaluated the statistical significance of the diatom biodiversity models in a relationship with the environmental condition by using a state-of-the-art decision tree machine learning algorithm. Furthermore, we tried different strategies by fitting different parameters of the algorithm, and more important we tested the influence of the number of discretization levels used for the class attribute. The obtained performance results for both descriptive and predictive performance of the models were measured using the AUC-ROC metric and then they were used as input into the two-stage procedure to estimate the statistical significance of the results. Based on the results presented in this paper we can draw several conclusions.

The best descriptive performance is obtained using decision tree algorithm settings with minimum weight 4 (MW4) followed by the maximum depth 3 (MD3), while the models with maximum depth 5 (MD5) and No Constraints (NoCo) are the ones with best predictive power, according to the statistical significance tests. If we compare the results from the models with 3 and 5 biodiversity categorization levels, according to the results there is a clear pattern that the models with 5 categorization classes have lower descriptive as well as predictive power than the models with 3 categorization classes. Based on these results two biodiversity models were later presented and then discussed in terms of the environmental condition influence on the Simpson biodiversity index.

As we draw into the conclusion of the results from this paper, we can note that there is still work to be done in the research of diatoms' biodiversity modeling. As future work, we plan to examine other biodiversity indices, as well as different classification algorithms that use fuzzy or fuzzy-rough modeling.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

References

1. Villéger SS., Mason NWH., Mouillot D.: New multidimensional functional diversity indices for a multifaceted framework in functional ecology. Ecology 89:2290–2301 (2008)
2. Recknagel F., Michener W.: Ecological Informatics. Data management and knowledge discovery. 3rd Edition. Springer International (2018)
3. Fielding AH:. Cluster and classification techniques for the biosciences. Cambridge: Cambridge University Press (2007)
4. Levatić J., Kocev D., Debeljak M., Džeroski S.: Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecological Modelling 306:294-304 (2015)
5. Kampichler C., Wieland R., Calmé S., Weissenberger H., Arriaga-Weiss S.: Classification in conservation biology: A comparison of five machine-learning methods. Ecological Informatics 5(6):441-450 (2010)
6. Drew CA., Wiersma YF., Huettmann F.: Predictive species and habitat modelling in landscape ecology. Springer (2011)
7. Oppel S., Meirinho A., Ramirez I., Gardner B., O'Connell AF., Miller PI., Louzao K.: Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation 156:94-104 (2012)
8. Prasad AM., Iverson LR., Liaw A.: Newer classification and regression tree techniques: bagging and random forests for ecological predictions. Ecosystems 9:181-199 (2006)
9. Naumoski A., Mitreski K.: Influence of climate change on diatoms diversity indices in Lake Prespa. Journal of E-Analytical - Energy and Climate Change - Southeast Europe in Focus 2(1) (2009)
10. Kocev D., Naumoski A., Mitreski K., Krstic S., Dzeroski S.: Learning habitat models for the diatom community in Lake Prespa. Ecological Modelling 221(2):330-337 (2010)
11. Pielou EC.: Shannon's formula as a measure of specific diversity. Its use and misuse, American Naturalist 100:463-465 (1966)
12. Pielou EC.: Species-diversity and pattern diversity in the study of ecological succession. Journal of Theoretical Biology 10:370-383 (1966)
13. Tang C., Yi Y,. Yang Z., Zhou Y., Zerizghi T., Wang X., Cui X., Duan P.: Planktonic indicators of trophic states for a shallow lake (Baiyangdian Lake, China). Limnologica 78 (2019)
14. Zimmerman RH.: The ecology of mosquito larvae (Diptera: Culicidae) in the subalpine zone of the eastern Sierra Nevada Mountains, California, USA. Journal of Vector Ecology 44(1) (2019)
15. Udaya NKP., Karthick NA., Poomagal D., Susithra M., Chandran M., Bhuvaneswari S.: Fungal endophytes of an aquatic weed Marsilea minuta Linn. Current Research in Environmental & Applied Mycology (Journal of Fungal Biology) 8(1):86-95 (2018)
16. Magurran E.: Ecological diversity and its measurement. Princeton University Press, Princeton. NJ (1988)
17. Pielou EC.: Ecological diversity. Wiley Interscience, New York (1975)
18. Magurran AE.: Measuring biological diversity. Blackwell, Oxford UK (2004)
19. Hill MO. Diversity and evenness: a unifying notation and its consequences. Ecology 54:427-431 (1973)
20. Naumoski A.: Multi-target modelling of diatoms diversity indeces in Lake Prespa. Applied Ecology and Environmental Research 10(4):521-529 (2012)
21. Naumoski A.: Learning models of abiotic influence on the biodiversity indices in Lake Prespa. Proceedings of the 12th International Conference for Informatics and Information Technologies (CIIT 2015), Bitola, Macedonia:257-261 (2015)

22. García S., Fernández A., Luengo J., Herrera F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180:2044-2064 (2010)
23. Friedman M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32:674–701 (1937)
24. Hommel G.: A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75:383–386 (1988)
25. Krstić S.: Description of sampling sites. TRABOREMA Project, FP6 (2006)
26. Brillouin L.: Science and information theory. Academic Press. New York. 320 (1956)
27. Smith B., Wilson JB.: A consumer's guide to evenness indices. Oikos 76:70–82 (1996)
28. Heip C.: A new index measuring evenness. J. Mar. Biol. Assoc. 54:555–557 (1974)
29. Alatalo RV (1981) Problems in the measurement of evenness in ecology. Oikos 37:199-204.
30. Hurlbert SH.: The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577–586 (1971)
31. Mcintosh RP.: An index of diversity and the relation of certain concepts to diversity. Ecology 48: 392–404 (1967)
32. Simpson EH.: Measurement of diversity. Nature 163:688 (1949)
33. Shannon CE., Weaver W.: The mathematical theory of communication. University Ilinois Press, Urbana (1963)
34. Blockeel H., Struyf J.: Efficient algorithms for decision tree cross-validation. Journal of Machine Learning Research 3:621–650 (2002)
35. Garofalakis M., Hyun D., Rastogi R., Shim K.: Building decision trees with constraints. Data Mining and Knowledge Discovery 7(2):187-214 (2003)