MODELLING THE RELATIONSHIP BETWEEN SATURATED OXYGEN AND DISTOMS‘ ABUNDANCE USING WEIGTHED PATTERN TREES WITH ALGEBRAIC OPERATORS

Prof. Dr. Naumoski A., Prof. Dr. Mirceva G., Prof. Dr. Mitreski K.
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje, Skopje, R. Macedonia

andrejanaumoski@mail.com

Abstract: *Machine learning has been used in many disciplines to reveal important patterns in data. One of the research disciplines that benefits from using these methods is eco-informatics. This branch of applied computer science to solve environmental problems uses computer algorithms to discover the impact of the environmental stress factors on the organisms' abundance. Decision tree type of machine learning methods are particularly interesting for the computer scientists as well as ecologists, because they provide very easy interpretable structure without any practical knowledge in mathematics or the inner working of the algorithm. These methods do not rely only on classical sets, but many of them are using fuzzy set theory to overcome some problems like overfitting, robustness to data change and improved prediction accuracy. In this direction, this paper aims to discover the influence of one particular environmental stress factor (Saturated Oxygen) on real measured data containing information about the diatoms' abundance in Lake Prespa, Macedonia, using weighted pattern tree (WPT) algorithm. WPT is a decision tree method variant that combines fuzzy set theory concepts, like similarity metrics, fuzzy membership functions and aggregation operators, to achieve better prediction accuracy, improve interpretability and increase the resistance to overfitting compared to the classical decision trees. In this study, we use Algebraic operators for aggregation. One WPT model is presented in this paper to relate the saturated oxygen parameter with the diatoms' abundance and reveal which diatoms can be used to indicate certain water quality class (WQC). The obtained results are verified with the existing knowledge found in literature.*

Keywords: ECOLOGICAL MODELING, ENVIRONMENTAL DATA, DIATOMS, FUZZY LOGIC, WEIGHTED PATTERN TREE (WPT), ALGEBRAIC OPERATORS

1. Introduction

The machine learning algorithms are more frequently used to discover underling drivers of the environmental stress factors that affects organisms' abundance. This is very important in ecological science that studies ecosystems that are threatened by pollution or they are under recovery program. In this case, monitoring is very important, as well as the data analysis. Data analytics usually is done by statistical algorithms, like canonical correspondence, detrended correspondence or principal component statistical analysis. These techniques provide useful understandings of the underling ecological processes. However, they are limited in terms of interpretability and in most cases, they suffer from subjective opinion of the domain expert. This is because the results are plotted on graph, from which the biological expert draws interpretations of the distances between groups and clusters of diatoms related with the environmental factors based on his previous knowledge. That gives the final model interpretation a degree of expert self-opinion, which should not be the case. That's why more and more experts use modelling techniques where the expert opinion is reduced to minimum, instead, an underling statistical procedure takes over to confirm the prediction model accuracy.

In this direction, the well-known ID3 [1], C4.5 [2] or CART [3] decision tree algorithms leave no room for expert influence, only in the stage when the model results needs to be verified with the existing knowledge found in literature [4]. Moreover, the decision tree algorithms produce easy interpretable model for which no mathematical knowledge is required, unlike the neural networks. There are various subgroups of decision tree algorithms, some of them differ on the type of heuristics used, some of them are different on how the model is learnt with different data partitioning strategies, and some of them on what type of sets is used to extract the knowledge (classical or fuzzy set theory). Fuzzy decision tree algorithm [5] is in a subgroup of decision tree algorithms that uses fuzzy set theory to improve descriptive and predictive performance of the models, as well as improve the interpretability of the models using fuzzy linguistic terms. Fuzzy linguistic terms play important role in interpretation, due to their similarity with the human language when transforming the model tree into rules. Beside this, the fact that in many research studies the fuzzy decision trees reported to have better performance than the crisp decision tree algorithms [6] at the expense of small increase of the time needed to build the model.

Further research is done in improving the fuzzy decision tree algorithms. In that direction, the authors in [7] give detailed description of the pattern tree algorithm, which further improves the accuracy of the fuzzy decision tree algorithm. This is done by combining different types of membership functions and aggregation operators with various similarity metrics to achieve not only multi criteria decision-making, but also improving predictive accuracy and producing a model that is more resistant to overfitting. The pattern tree algorithm produces fuzzy rules with linguistic terms that can be obtained from the traditional hierarchical tree like structure. Furthermore, for each branch, a similarity between the target attribute and the input attribute is obtained, which evaluate the level of confidence of predicting the output attribute with that input attribute. Since each dataset may contains multi-class target attributes, one pattern tree model is built for each class of the target attribute. In this way a forest of pattern tree models is obtained, without knowing which tree holds the highest confidence of predicting the target class from the descriptive attributes. In this direction, the authors in [8] have presented the weighted pattern tree algorithm (WPT), which weights each pattern tree model to a class of the target attribute. In this way, the decision-making expert can select which output model is confident of predicting a given class. The WPT algorithm uses the similarity value between the target class attribute placed at the root of the model tree and the fuzzy term of the input attribute on the root branch. The membership degree of each input attribute is obtained using the process of fuzzification, which transforms the crisp attributes into fuzzy attributes by using different mathematical functions (triangular, trapezoidal, Gaussian, Bell etc.). These membership functions are widely used for fuzzification, depending on the nature of the input dataset and the purpose of the fuzzy system. Consequently, many researchers found out [9] that the fuzzification process has high influence on the accuracy of the classifier. Beside the different membership functions that WPT algorithm uses to build the final model, the WPT also uses different similarity metrics to find the most informative attribute related to the target class. Additionally, the WPT algorithm uses aggregation operators that relate each input attribute to the output attribute as operation between two fuzzy sets to narrow the search space. In both cases (similarity metrics and aggregation operators) there are many metrics and operators that may fit the modeler's needs to obtain high predicting accuracy. We use the recommendations that we suggested in [10], which are based on the experimental evaluation with different membership functions.

In this paper, we obtain one WPT model that consist from four sub-models to predict the relationship between the diatoms' abundances as indicators of water quality classes based on saturated oxygen parameter. The mathematical modelling is performed on ecological dataset that is comprised with ten input attributes and one output attribute. The ten input attributes represent the ten most abundant diatoms found in Lake Prespa water ecosystem [11] and one output target attribute that describes the ecological water quality class based on saturated oxygen parameter [12]. In this way, the obtained WPT model relates the diatoms' indicator status with the certain water quality class based on saturated oxygen parameter.

The rest of the paper is organized as follows: Section 2 provides description of the WPT building blocks: membership functions, similarity metrics and aggregations operators, as well as dataset description. In Section 3, we present the WPT model, we discuss the results and verify the obtained knowledge with the existing knowledge found in literature. The main conclusions and our future work are outlined in the Section 4.

2. Algorithm Concepts and Data Description

The WPT algorithm relies on the fuzzy theory concepts like membership functions, similarity metrics and aggregation operators, same as the pattern tree algorithm. However, the WPT uses additional information from the tree root similarity value to assign a degree of confidence or weight each model. The performance of the model depends on the membership function that is used, as well as the type of the similarity metric and aggregation operator. Following the recommendation that we gave in [10], we use the Bell membership function for building model.

The Bell membership function is defined with three parameters *a*, *b* and *c*. In order to achieve complete evenness between the fuzzy terms, we replace the parameter *a* with 10, *b* is replaced by σ calculated using (1), while *c* is replaced with μ (mean value of the range of each fuzzy term).

$$\sigma = \sqrt{\frac{Ln[10]}{2 * Ln[0.5 * r]}} \tag{1}$$

In equation (1), *r* stands for the length of the fuzzy range. After all these changes take place, the calculation of the Bell membership function fuzzy terms is made by

$$f(x;\mu;\sigma) = \frac{1}{1 + \left| \frac{x-\mu}{10} \right|^{2\sigma}}. \tag{2}$$

Additionally, very important factor that influence the model accuracy is the number of fuzzy terms used per attribute. This has effect on the interpretability of the models as well as on the type of analyses that is conducted on the obtained knowledge. That is why selecting the number of fuzzy terms can be done on basis of the needs of domain expert or based on experimental evaluation of the model. For this purpose, we use five fuzzy terms which corresponds with the number of classes used in the European Water Framework directive (Directive 2000/60/EC of the European Parliament) for water quality classification, where there are five categories (poor, bad, moderate, good and high).

The next component of the WPT model building process is selection of the appropriate similarity metric. The similarity metric can greatly affect the accuracy of the model as well as the selection process of the most confident model for prediction. For this purpose, we consider the same metric used in [10]. However, we want to note that maybe other similarity metrics, like Jacquard, Cosine or Squared Euclidean, could be more adequate for this purpose. The similarity metric Sim_{RMSE} used in [10] is based on the RMSE (Root mean squared error) distance metric and calculates the similarity between two fuzzy sets *A* and *B* as

$$Sim_{RMSE} = 1 - \sqrt{\frac{\sum_{i=1}^{n} \left(\mu_A(x_i) - \mu_B(x_i) \right)^2}{n}}. \tag{3}$$

The results of the calculations using (3) reside in range between 0 and 1. To calculate this, the Sim_{RMSE} similarity metric uses the membership degrees for a given crisp value x_i in two fuzzy sets *A* and *B*, which are denoted as $\mu_A(x_i)$ and $\mu_B(x_i)$. The value of the Sim_{RMSE} similarity is propagated in each branch all the way up to the root of the model tree. Using only the similarity metric for estimating the relationship between the input and output attributes leads to low performances of the models. Therefore, the WPT algorithm also uses fuzzy aggregation operators, which are operations over two fuzzy sets to learn better models. Typically, triangular norms and conorms are used in fuzzy induction algorithms, and also in fuzzy decision tree algorithms. In [10], we used two fuzzy aggregation operators: Algebraic AND (T-norm) and Algebraic OR (T-conorm), which are defined as

Algebraic AND = $\mu_A(x_i) * \mu_B(x_i)$

Algebraic OR = $\mu_A(x_i) + \mu_B(x_i) - \mu_A(x_i) * \mu_B(x_i)$.

It is important to note that also some other Triangular T-norm or T-conorm aggregation operators can be used here, that may have influence on the performance of the model and its interpretation of the model.

The dataset used for modelling the relationship between saturated oxygen and diatoms' abundances is obtained from the EU funded TRABOREMA [11] project. The goal of this project was to assess the ecological status of the Lake Prespa. During the monitoring stage, valuable data about the physico-chemical parameters as well as organisms' relative abundance was collected. Overall, sixteen parameters were measured and 116 different diatom species are counted. For each sample, the relative abundance for all 116-diatom species is obtained. Using this type of monitoring, the relationship between the influencing factors and diatoms' relative abundance can be discovered. In the ecological literature [4], the relationship between environmental stress factors and diatoms as bio-indicators is well established, but for many diatoms, this relationship remains unknown. The environmental stress factors in the established ecological literature are represented with water quality classes (WQCs) based on a certain physico-chemical parameter. These classification systems can be found in the ecological literature, like the classification systems for Conductivity [13], pH [13], [12] and Saturated Oxygen [12]. The saturated oxygen classification system defines five classes for the target attribute, which are given in Table 1.

Table 1: *Water quality classes for the saturated oxygen physico-chemical parameter.*

Name of the water quality class	Parameter range
Oligosaprobous	> 85 %
β-mesosaprobous	70% - 85%
α-mesosaprobous	25% - 70%
α-meso / polysaprobous	10% - 25%
Polysaprobous	< 10%

Since WQC *Polysaprobous* doesn't contain any values in the measured dataset, this class was removed. Therefore, the final dataset contains 4 WQCs based on the saturated oxygen parameter. Considering this classification system, it is obvious that using machine learning algorithms we face with typical classification problems, where the saturated oxygen WQCs are the possible values for the target or predictive attribute, while the relative abundances of the top ten most abundant diatoms are the input or descriptive attributes. By using the parameters' settings described in the previously, in the next section we present the WPT model in order to describe the relationship between Saturated Oxygen and the diatoms as bio-indicators.

3. Results and Verification

Four models are obtained, one for each WQC, and for each model the highest similarity is given in Table 2.

Table 2: *The highest similarity that is obtained from the WPT models for the WQCs..*

Name of the water quality class	Highest similarity
Oligosaprobous	0.5249
β-mesosaprobous	0.5620
α-mesosaprobous	0.6897
α-meso / polysaprobous	0.8253

Based on the results in Table 2, the highest similarity is obtained for the WPT model for *α-meso / polysaprobous* WQC. The model for this class is presented on Fig. 1. The highest similarity obtained with this model shows that it has highest confidence to predict the corresponding WQC. The model depicts the diatoms that can be used to indicate this WQC.

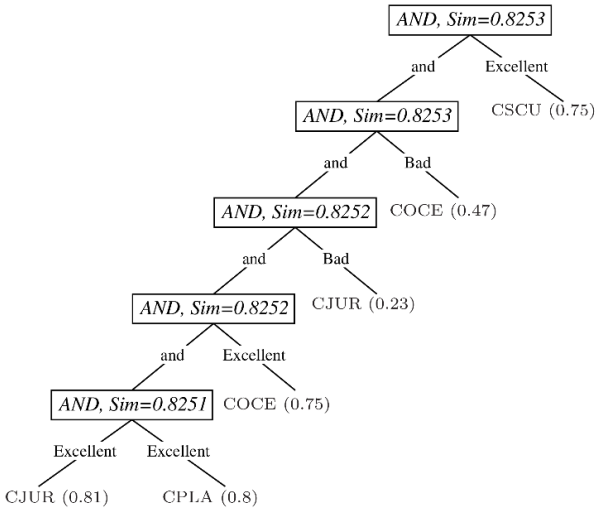

Fig. 1 *WPT model for the α-meso / polysaprobous WQC for Saturated Oxygen. In brackets, the similarity between the membership degrees for a given fuzzy term for a diatom and the WQC is given.*

According to the model, the *Cavinula scutelloides* (CSCU) and *Cocconeis placentula* (CPLA) diatoms can be used as excellent indicators for *α-meso / polysaprobous* WQC. The two other diatoms *Cyclotella juriljii* (CJUR) and *Cyclotella ocellata* (COCE) can be also used as excellent indicators since the similarity value between these two diatoms and the *α-meso / polysaprobous* WQC is higher compared to the similarity values for bad indicating properties.

For verification of the results, we compared them with the ecological preferences found in [4]. For the CJUR and NPRE diatoms no records exist for their ecological preferences because they are newly described taxa. The CPLA diatom is eutrophic with medium oxygen demand according [4], while the model shows that the CPLA diatom is an excellent indicator for waters with low values of saturated oxygen. On the other hand, the CSCU diatom is *alkalibiontic*, freshwater to brackish water specie, being *oligosaprobic* indicator with eutrophic preferences according [4]. The model gives directions that this diatom is an excellent indicator for *α-meso / polysaprobous* class opposite from which the known ecological literature directs. And finally, according the model the COCE diatom is excellent indicator of *α-meso / polysaprobous* WQC, which if also obtained by other models that we generated with other experiments, thus this could add additional knowledge in the literature for this diatom. According [4], the trophic ecological preferences are the only one known for this diatom and since the trophic state index classification is out of the scope of this paper, adding new knowledge regarding the saturated oxygen demand for this diatom, could enrich its indicating properties.

4. Conclusion

In this paper, we presented a technique based on fuzzy set theory that could reveal the relationship between the saturated oxygen *α-meso / polysaprobous* WQC and the ten most abundant diatoms found in Lake Prespa. We built a model that presents which diatoms can be used for indicating the WQCs for saturated oxygen. The results from the WPT model are compared with the known ecological preferences found in the literature.

In future, we plan to investigate the influence of other similarity metrics and aggregation operators to further improve the accuracy of the models. Other type of membership functions could be more suitable for diatom modelling, thus revealing more valuable knowledge from the measured data. Also, modelling other water quality classes or trophic index classes could increase the applicability of the algorithm.

Acknowledgement

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

References

[1] J.R. Quinlan, "Induction of decision trees", Mach. Learn., vol. 1, pp. 81–106, 1986.

[2] "C4.5: Programs for Machine Learning", San Francisco, CA: Morgan Kaufmann, 1993.

[3] L. Breiman, J. Friedman, R. Olshen, C. Stone, "Classification and Regression Trees", Belmont, Wadsworth, 1984.

[4] H. Van Dam, A. Martens, J. Sinkeldam, "A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands", Netherlands Journal of Aquatic Ecology, vol. 28, no. 1, pp. 117–133, 1994.

[5] Y. Yuan, M.J. Shaw, "Induction of fuzzy decision trees", Fuzzy Sets and Systems, vol. 69, no. 2, pp. 125–139, 1995.

[6] Y.-l. Chen, T. Wang, B.-s. Wang, Z.-j. Li, "A survey of fuzzy decision tree classifier", Fuzzy Information and Engineering, vol. 1, no. 2, pp. 149–159, 2009.

[7] Z.H. Huang, T. D. Gedeon, "Pattern trees", in: Proc. of IEEE International Conference on Fuzzy Systems, pp. 1784–1791, 2006.

[8] Z. Huang, M. Nikravesh, B. Azvine, T.D. Gedeon, "Weighted pattern trees: a case study with customer satisfaction dataset", International Fuzzy Systems Association World Congress 2007, pp. 395–406, Springer, Berlin, Heidelberg, 2007.

[9] M. Zeinalkhani, M. Eftekhari, "Fuzzy partitioning of continuous attributes through discretization methods to construct fuzzy decision tree classifiers", Information Sciences, vol. 278, pp. 715–735, 2014.

[10] A. Naumoski, G. Mirceva, K. Mitreski. "Experimental Evaluation of Different Membership Functions on Weighted Pattern Trees for Diatom Modelling", 14th International Conference on Natural Computation; Fuzzy Systems and Knowledge Discovery, IEEE, 2018.

[11] "TRABOREMA Project" WP3, EC FP6-INCO project no. INCO-CT-2004-509177, 2005–2007.

[12] A. Van Der Werff, H. Huls, "Diatomeanflora van Nederland", Abcoude - De Hoef, 1957, 1974.

[13] K. Krammer, H. Lange-Bertalot, "Die Ssswasserflora von Mitteleuropa 2: Bacillariophyceae. 1 Teil", pp. 876, Stuttgart: Gustav Fischer-Verlag, 1986.