Diatom Ecological Modelling with Weighted Pattern Tree Algorithm by Using Polygonal and Gaussian Membership Functions

Abstract. Weighted Pattern Tree (WPT) algorithm is as an extension of the Pattern Tree (PT) algorithm, which could be used for fuzzy modelling. This algorithm utilizes the similarity between two fuzzy sets in order to quantify how much a particular tree model is confident to predict a given class. The Membership Functions (MFs) play an important role in model induction and thus on the model's performance. Therefore, this paper aims to investigate the influence of different MFs, not only by analyzing different mathematical distributions, but also to investigate the influence of the number of MFs per attribute used for fuzzification of the datasets, as well as the different settings of the algorithm. The experimental results show that WPTs with depth 10 using polygonal MFs with high number of MFs per attribute are excellent for describing the training data, while the models that are built with low number of MFs are excellent for making predictions for unseen data.

Keywords: Weighted Pattern Trees, Diatoms, Membership Functions, Statistical Significance.

1 Introduction

In many water related studies, the aquatic biologists try to find different ways to understand and reveal the relationship between the environment and how that environment influences on the aquatic organism's life cycle. Typically, a water quality class is defined in a certain range of the environmental stress factors in which the organism belongs. In such studies, the ecological experts usually relay on the data that is collected over period of time, and then they apply statistical methods to reveal any hidden pattern that can be found in the data. Usually classical statistical approaches such as canonical correspondence analysis, detrended correspondence analysis and principal component analysis are used as modelling techniques in this research area [1]. Although these techniques provide useful insights in the data, they are limited in terms of interpretability, and in most cases they are influenced by the subjective opinion of the expert.

With the advancements in computer science, particularly in the area of machine learning and data mining, where algorithms gain new knowledge from the data at hand, the algorithms provide better understanding of the data compared to the traditional statistical approaches. Many machine learning algorithms have provided better inside look in the data. They also provide models with better interpretability, and at the same time with improved accuracy. One such group of algorithms are decision trees [2], which partition the space in order to find the relationship between the input and output attributes. This could be done using various metrics to measure the benefit of making a given split, which has impact on the algorithm's performance. Also, there is a sub-class of

decision tree algorithms that are based on the concepts of fuzzy theory, which are introduced to further improve the performance.

There is an extensive research effort in developing algorithms based on fuzzy set-theory, and many of these advancements are inspired by crisp decision trees [2]. For example, Yuan and Shaw [3] proposed fuzzy decision trees induction using fuzzy entropy. Janikow [4], Olaru and Wehenkel [5] have introduced other algorithms for fuzzy decision tree induction. Suárez and Lutsko [6], and Wang and Chen [7] have presented some optimizations of fuzzy decision trees, while Nikravesh [8] presented evolutionary computation based multi-aggregator fuzzy decision trees. As successors, the Pattern Tree (PT) [9] and fuzzy operator tree [10] algorithms offer several improvements over the previous fuzzy based algorithms. The PT algorithm is able to retain the traditional tree like hierarchical structure. This algorithm provides opportunity to use different aggregation operators and similarity metrics with the aim to obtain better results. This is very important in fuzzy modelling because decisions could be made on basis on multiple-criteria group decision making. There were efforts, mainly in the area of rule fuzzy induction, to build such algorithms, like in the research made by Kóczy, Vámos and Biró [11], where they have proposed fuzzy theory concepts to model complex datasets with different aggregation operators including triangular intersections and average aggregation operators. The aggregation operators provide opportunity for multiple-criteria group decision making. This can be also achieved with the Weighted Pattern Tree (WPT) algorithm [12], which additionally gives weights how well a given pattern tree represent the corresponding class. Beside the use of fuzzy aggregation operators and similarity metrics, the algorithm could be applied by using various MFs.

Therefore, in this paper we experimentally evaluate the influence of the Trapezoidal, Triangular and Gaussian MFs on the WPT algorithm using different number of MFs per attribute over three ecological datasets. These datasets contain 10 input attributes detailing the abundances of the diatoms found in lake and one output attribute that describes the ecological status of the lake using certain parameters (in our case: Conductivity, pH and Saturated Oxygen). Furthermore, we investigate how the complexity and resistance to over-fitting influence on the model's performance. Models with different depths (5 and 10) are used for this purpose, as well as the four variants of the PT algorithm that are described in [9]. These four model variants are obtained by using various settings in the algorithm. We evaluate the resistance to over-fitting of the WPT algorithm by calculating the Root-mean squared error (RMSE) between the two evaluation procedures (in the first experiments the entire data set is used for both training and testing, while in the second experiments a cross validation is made). In order to ensure that the improvement in the performance is statically significant, we use statistical significance test. For this purpose, we use a two-stage procedure that combines the Aligned Friedman [13] test and post-hoc Hommel test [14], as described in [15].

The rest of the paper is organized as follows: Section II provides description of the WPT algorithm and its main building blocks: definitions of the similarity metric, aggregations operators, as well as the three MFs that are used. In section III, the dataset description as well as the experimental setup procedures are described. Section IV presents the experimental results as well as the conclusions and discussions from this experimental evaluation. Section V concludes the paper and outlines our future work.

2 Weighted Pattern Tree Algorithm

The WPT algorithm uses the same concepts as the pattern tree algorithm. The similarity metric is also used to weight the confidence of a particular tree to predict a given class. Before this can be done, the WPT algorithm builds a tree model using aggregation operators and fuzzy terms from the input dataset. The input values for these fuzzy terms are obtained by fuzzification as every algorithm based on fuzzy theory requires, which could be done using different mathematical distributions. In this section, first we will describe the main concepts that are used in the algorithm, and then we will present how the WPT is induced.

In fuzzy theory, MFs are used to transform the crisp values or to transfer from the classical domain to fuzzy domain by a process known as fuzzification. This is an important part of the WPT algorithm, because the fuzzification has influence on the performance of the algorithm. The polygonal MFs (triangular and trapezoidal MFs) have advantages of simplicity, but in many cases datasets consist of smoothed and nonzero values, which are handled much better by Gaussian MF and thus more accurate models are produced. Besides the impact of the MF on the model's performance, also it is important how many MF are used per attribute. Namely, each attribute produce a predefined number of fuzzy terms that could be linguistically labelled. Therefore, in this paper we don't just experimentally evaluate the different types of MFs, but also we make experiments by using different number of MFs per attribute.

Another parameter that plays an important role in increasing the performance of the algorithm is the similarity metric. Therefore, it is important to consider metrics that will reflect the dataset properties. In this paper, we use the similarity metric proposed in [9], which could be used to calculate the similarity between two fuzzy sets A and B defined in the universe of discourse U as a complement of the root mean squared error metric. The similarity is computed as

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

where $\mu_A(x_i)$ and $\mu_B(x_i)$ are the membership degrees of an element x_i in the fuzzy sets A and B, respectively. The larger the similarity is, the more similar the fuzzy sets are.

The similarity metric could be used to evaluate the degree of similarity between a given fuzzy set and the class attribute. These fuzzy sets could be the initial sets that are obtained with fuzzification. However, they often do not provide best description of a given class. For that purpose, fuzzy aggregation operators are used in order to combine the fuzzy sets. There are several categories of aggregation operators, but mainly triangular norms are used as in many other fuzzy algorithms. In this research paper, we use the Algebraic AND and Algebraic OR aggregation operators. However, in future our focus can shift towards additional aggregation operators, since they also play an important role in producing more accurate models.

The induction of WPT starts by fuzzification of the dataset by using a particular MF. Each attribute is presented by a given number of fuzzy terms. For each of the fuzzy terms, a pattern tree model is generated. These pattern trees are called as primitive trees, and also we will refer to them as trees at level 0. However, they are too simple and could not provide accurate results in making predictions. Thus, by using aggregation operators, these primitive trees could be aggregated, thus obtaining more complex models. For that purpose, first, for each primitive tree the similarity between the fuzzy set that corresponds to the particular tree and the fuzzy set for the class attribute is calculated. The best primitive tree, which is the tree for which highest similarity is obtained, is further aggregated with the remaining primitive trees. In this way, the candidate trees are obtained. We will refer to these trees as trees at level 1. The candidate tree for which highest similarity is calculated is further aggregated with the remaining primitive tress, thus obtaining the candidate trees at level 2, and this is repeated until the tree reaches the predefined depth. In this way, for each class we obtain separate model tree, and for each of these trees a corresponding weight is assigned based on the similarity between the fuzzy set that corresponds to the tree and the fuzzy set for the class attribute. The model trees that are obtained in this way are simple trees. Besides aggregation of the best candidate tree with the primitive trees that are not used for making that candidate tree, also there is other possibility where besides the primitive trees (trees at level 0), also the trees at the remaining levels (levels 1, 2, etc.) could be considered in the aggregation. In this way, we can distinguish simple models, where in the aggregation only the trees at level 0 are aggregated with the best candidate tree, and general models, where the trees at all levels are considered in the aggregation. The difference between the PT and WPT algorithm is that in WPT besides generation of a model tree for each class, also a corresponding weigh is associated to that models. More details about the induction of the PT and WPT can be found in [9] and [12], respectively.

3 Dataset Description and Experimental Setup

The ecological datasets that are used to experimentally evaluate the influence of the shape, the number of MFs per attribute and the four different model variants are obtained from a real measured dataset collected within an EU project for ecological assessment of Prespa Lake [16]. During the time period of 16 months within the project timeframe, several physico-chemical parameters are measured. This dataset also contains measurements regarding the biological aspects of the lake, through measurements of the diatoms' abundances. These diatoms live in certain conditions, which are defined with the range of the physico-chemical parameters. Using the measured data, it is possible to find the relationship between the diatoms and the physico-chemical parameters in the environment. In this way, we may find out what are the required conditions in the environment in which a given diatom can survive. In this study, from these physico-chemical parameters, we consider three environmental stress factors whose classification systems can be found in the ecological literature [17, 18, 19]. Conductivity [17], pH [17,18] and Saturated Oxygen [18] classification systems are directly related to

some measured diatom species in a similar manner like the examples found in the ecological literature [19]. Since we are using a single target classification algorithm, which could be used for prediction of a single attribute, while from the measured dataset we consider three physico-chemical parameters, therefore we make a separate dataset by considering each of these three parameters as a single target attribute. The diatom ecological datasets consists from 10 input numeric attributes that represent the relative abundances of the 10 most abundant diatoms in the samples, and one target attribute that may obtain from 4 to 6 nominal values (different number of values for each dataset based on the considered target attribute).

As we mentioned previously, in this paper we evaluate these three ecological datasets and the influence of different number of MFs per attribute (3, 4, 5, 10, 20, 30, 50 and 100). The complexity and resistance to over-fitting of the models are evaluated by using the four different variants of WPTs, which could be simple weighted pattern trees (SWPT) and general weighted pattern trees (GWPT) with depth 5 and 10. The four variants of the WPTs are denoted with: SWPT5 (SWPT with depth 5), SWPT10 (SWPT with depth 10), GWPT5 (GWPT with depth 5) and GWPT10 (GWPT with depth 10). These four WPT variants lead to different properties and performances of the obtained models. For example, trees with higher depth (depth 10) would lead to more complex models but may have higher predictive power, while trees with lower depth (depth 5) are less complex but may have lower predictive power.

The descriptive classification accuracy, denoted as "Train" in the experimental results, shows how well the model tree suits to the training data, where the entire data set is used for both training the model and testing its accuracy. However, in this way we will obtain high results for the models that are over-fitted. Therefore, it is needed to estimate how accurate the model is for unseen data. Therefore, we also use the predictive classification accuracy, denoted as "Test" in the experimental results, by making 10-fold cross validation.

In order to confirm the statistical significance of the obtained results, the two-stage procedure proposed in [15] is employed, which combines the Aligned Friedman test [13] and post-hoc Hommel test [14]. In the first stage, non-parametric Aligned Friedman test [13] is used with so called aligned ranks, which is recommended when the number of experiments is not too large. The average rank according Aligned Friedman test is calculated by using Eq. 2, where k represents the number of different settings variants that could be used, while n is the number of datasets used in the experiments.

$$Rank = \frac{(k-1)[\sum_{j=1}^{k} X_j^2 - (kn^2/4)(kn+1)^2]}{\{[kn(kn+1)(2kn+1)]/6\} - (1/k)\sum_{i=1}^{n} X_i^2} \tag{2}$$

The average rank also considers the total rank for each of these parameters (k and n), where X_i is the total rank for the i-th dataset, while X_j is the total rank for j-th settings variant, respectively. For more information regarding the total rank, the reader is referred to [13]. In ER1 experiment, the number of MFs per attribute is fixed, and the average rank is calculated by using the three types of MFs and the four model variants,

so k = 12. Similarly, in ER2 experiment, the type of MFs is fixed so k = 32, while in ER3 experiment the model variant is fixed so k = 24. In this way, the average rank for each dataset is calculated, and also the average rank over all dataset is calculated by using Eq. 2.

Next, the obtained rank is compared for significance with a chi-squared distribution for k - 1 degrees of freedom. The p-value is computed using normal approximations [20], and if the null hypothesis is rejected, usually with high level of significance, we can proceed with the post hoc Hommel test [14]. Since the Aligned Friedman test does not examine the difference among the settings variants, it only shows differences among datasets. For this purpose, pairwise comparison is performed with a post-hoc procedure and a control variant is selected, which is the variant with highest rank, as it is indicated in [15]. The calculated p-value is compared with an appropriate level of significance, usually 0.05, to compensate for multiple comparisons. Because the post-hoc procedure adjusts the level of significance for each comparison, adjusted p-values are recommended to be used in order to make fair comparison among p-values.

4 Experimental Results

In this section, the experimental results are presented for the four model variants (SWPT5, SWPT10, GWPT5, GWPT10), as well as three types of MFs by using different number of MFs per attribute. The descriptive and predictive classification accuracies are determined, and also the RMSE between these two measures is calculated in order to estimate the resistance to over-fitting.

4.1 Performance Analysis

In this section, first, we present the experimental results obtained for the Conductivity dataset (see Table 1). The experiments with the triangular MF are characterized with highest peak of descriptive performance when we are using thirty MFs per attribute, while best predictive performance is obtained with ten MFs per attribute. If we investigate the different WPTs variants that we have employed, we may found out that GWPT5 and GWPT10 have highest predictive and descriptive classification accuracy, which is confirmed by the average accuracy.

When we are using the trapezoidal MF, the models with highest descriptive and predictive performance are achieved with hundred MFs per attributes, except for the descriptive analysis for GWPT5 variant when the best descriptive accuracy is achieved with twenty MFs per attribute. But, the average accuracy is the same as in the previous experiment.

For Gaussian MF, the highest value for the model's descriptive accuracy is when thirty MFs per attribute are used, while the predictive performance settle the highest peak between four and five MFs per attribute. In the case of Gaussian MF, the simple and general WPT variants are similar between each other, but it is interesting that the average accuracy, again, is highest for the descriptive performance for the GWPT10

variant, while the predictive performance is best with GWPT5. If we examine the resistance to over-fitting, we can note that best value is achieve with GWPT5 variant using triangular MF with value of almost 6% error. In the other experiments, various models obtain 2 to 4% higher error when it comes to resistance to over-fitting.

Table 1. Evaluation results for the Conductivity dataset by using different number of MFs per attribute. Train denotes descriptive classification accuracy, while Test denotes predictive classification accuracy. Underlined results show the models with highest descriptive classification accuracy, while bolded results show the models with highest predictive classification accuracy.

Triangular Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	71.56	72.94	75.23	78.44	77.06	76.61	75.69	72.48	75.00	6.82
Test[1]	69.09	70.04	**71.86**	68.61	66.75	67.25	69.05	69.09	68.97	
Train[2]	73.39	72.94	75.23	79.36	78.90	79.82	76.15	74.31	76.26	8.73
Test[2]	69.09	69.59	**71.41**	68.18	65.41	66.77	70.43	66.82	68.46	
Train[3]	71.10	74.77	75.23	77.98	76.15	76.61	74.31	72.48	74.83	5.93*
Test[3]	69.09	70.95	71.84	**72.36**	66.30	66.32	71.36	69.55	**69.72**	
Train[4]	72.02	74.77	75.23	78.90	77.98	81.19	76.15	74.77	76.38	8.12
Test[4]	69.07	69.13	70.93	71.90	66.77	65.84	**72.29**	68.18	69.27	
Trapezoidal Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	74.31	73.85	72.94	77.06	77.52	77.98	76.15	78.90	76.09	8.10
Test[1]	69.98	67.73	65.41	71.86	63.53	68.59	68.59	**72.27**	68.50	
Train[2]	73.85	73.85	72.94	79.82	79.36	81.19	79.82	82.11	77.87	10.31
Test[2]	69.07	66.36	64.48	70.95	64.46	67.68	68.14	**73.18**	68.04	
Train[3]	73.39	74.31	76.61	78.90	80.28	77.98	76.15	78.90	77.06	8.53
Test[3]	69.98	69.09	68.20	71.43	64.03	69.05	69.03	**74.13**	**69.37**	
Train[4]	73.85	74.77	76.15	80.28	81.19	81.19	80.28	81.19	78.61	11.03
Test[4]	69.98	67.73	67.29	70.06	63.12	67.68	68.59	**72.27**	68.34	
Gaussian Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	71.56	74.77	74.31	77.52	76.61	79.36	75.69	73.85	75.46	8.03
Test[1]	68.14	70.95	**73.27**	68.64	67.29	65.45	68.64	64.89	68.41	
Train[2]	72.48	74.77	74.31	77.06	77.98	81.19	76.61	76.61	76.38	8.55
Test[2]	69.09	**71.41**	71.39	67.73	64.91	68.18	70.45	67.21	68.80	
Train[3]	73.39	74.77	75.23	77.52	75.69	79.36	75.69	73.85	75.69	7.38
Test[3]	69.55	69.52	**72.79**	68.64	69.13	66.36	70.00	65.80	**68.97**	
Train[4]	73.39	74.77	75.23	77.98	78.44	81.19	76.61	76.61	76.78	8.77
Test[4]	69.07	70.00	**71.86**	67.25	66.32	69.55	70.45	65.37	68.73	

[1]SWPT5, [2]SWPT10, [3]GWPT5, [4]GWPT10

In Table 2, the experimental results obtained for the pH dataset are presented. The first part of the table depicts the influence of the triangular MF for which best descriptive results are achieved using high number of MFs per attribute (fifty and hundred). On the other side, for predictive accuracy, the models with low number of MFs per attribute (between three and five) obtain best results.

Table 2. Evaluation results for the pH dataset by using different number of MFs per attribute. Train denotes descriptive classification accuracy, while Test denotes predictive classification accuracy. Underlined results show the models with highest descriptive classification accuracy, while bolded results show the models with highest predictive classification accuracy.

Triangular Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	58.26	61.01	62.84	62.84	62.39	62.84	65.60	65.60	62.67	14.78
Test[1]	56.30	**57.12**	56.73	48.92	41.49	45.71	48.42	43.85	49.82	
Train[2]	59.63	61.47	62.84	64.68	66.51	69.27	70.64	71.56	65.83	17.58
Test[2]	56.30	57.14	**57.16**	50.74	43.31	43.90	49.81	47.06	50.68	
Train[3]	60.09	61.93	61.01	63.30	62.39	62.84	66.97	65.60	63.02	13.65
Test[3]	**55.82**	53.90	54.44	48.44	44.22	50.80	48.87	47.10	50.45	
Train[4]	61.47	63.30	61.01	65.14	66.51	69.27	73.85	72.48	66.63	17.00
Test[4]	**55.82**	54.39	55.80	50.74	45.13	49.42	50.69	48.44	**51.30**	
Trapezoidal Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	59.17	60.55	61.93	62.39	65.60	67.43	66.06	65.14	63.53	14.26
Test[1]	53.01	**54.85**	54.46	46.58	51.13	50.24	48.42	44.29	50.37	
Train[2]	59.17	60.55	61.93	63.76	66.97	68.35	70.64	72.94	65.54	16.83
Test[2]	53.01	54.85	**56.73**	47.97	50.17	49.35	47.03	46.15	**50.66**	
Train[3]	60.55	60.09	63.30	62.39	65.60	65.60	66.06	65.14	63.59	14.82
Test[3]	52.55	55.30	**55.41**	46.10	48.42	47.03	50.30	43.79	49.86	
Train[4]	60.55	60.09	63.76	63.76	68.35	69.27	71.10	73.39	66.28	18.26
Test[4]	52.55	53.90	**58.14**	47.03	48.38	46.15	47.06	46.13	49.92	
Gaussian Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	58.72	57.80	61.93	63.30	62.39	59.63	66.97	65.60	62.04	13.16
Test[1]	**55.35**	53.01	54.89	53.92	47.94	42.51	49.81	45.17	50.32	
Train[2]	58.26	57.34	62.39	63.30	67.43	63.76	72.02	71.56	64.51	16.14
Test[2]	55.35	**55.80**	55.37	52.99	48.38	41.10	51.65	46.56	**50.90**	
Train[3]	59.17	57.80	55.50	63.30	61.01	59.63	66.06	64.68	60.89	12.22*
Test[3]	53.05	52.51	50.28	**53.46**	48.40	45.26	47.49	47.49	49.74	
Train[4]	59.63	57.80	56.42	63.76	66.06	65.60	72.02	71.56	64.11	15.82
Test[4]	53.05	**56.71**	52.10	52.53	48.38	43.44	50.24	48.40	50.61	

[1]SWPT5, [2]SWPT10, [3]GWPT5, [4]GWPT10

The simple WPTs obtain best predictive accuracy, while general WPTs are better in descriptive analysis. The average accuracy in the overall performance analysis for the triangular MF, placed the GWPT10 and GWPT5 among the best for building diatom ecological models. WPT with depth 10 are better in ecological modelling than the WPT with depth 5 based on the model's descriptive accuracy.

The next MF is the trapezoidal function, for which higher number of MFs per attribute (from thirty till hundred) give better descriptive performance, while the models that have low number of MFs per attribute have best predictive performance. Regarding the WPT variant, the models with depth 5 are worse in both descriptive and predictive analysis, while both simple and general models with depth 10 are better in the experimental evaluation. This is confirmed by the average performance analysis, which puts the SWPT10 as the best model with highest predictive accuracy, while GWPT10 is best in the descriptive analysis.

And finally, the Gaussian MF doesn't change the pattern that we found for this dataset. It attains high descriptive accuracy using high number of MFs per attributes (exactly fifty) and low number of MFs per attribute for predictive accuracy (between three and ten). When it comes to depth analysis, again, simple and general WPT with depth 10 are best for obtaining models with high descriptive and predictive accuracy. This is confirmed with the obtained average classification accuracy, but compared to the triangular MF, SWPT10 are best in the case of Gaussian MF. The analysis of the RMSE regarding the resistance to over-fitting over this dataset, shows lower resistance to over-fitting compared to the Conductivity dataset (twice much higher error of 12.22%), and the best model variant to build such models is using GWPT5. The other variants are having 3 to 5% lower resistance to over-fitting.

The experimental results for the last dataset that we used, the Saturated Oxygen dataset, are presented in Table 3. Here, the triangular MF continues the trend of separation between the models regarding their descriptive and predictive accuracy, when it comes to the number of MFs per attribute. In this case, the descriptive accuracy is higher with higher number of MFs per attribute (more than twenty), while the models obtained by using low number of MFs per attribute (between three and five) are having highest value for predictive accuracy. There is no big difference when using simple or general WPT, even with different depths. That's why there is not much difference in the average accuracy, where SWPT5 slightly overruns the other variants for predictive analysis, while GWPT10 is best in the descriptive analysis.

If we examine the trapezoidal MF, the trend of separating the train and test performance with different number of MFs per attribute stops here. Here, the best models are obtained when the number of MFs per attribute is higher than 5. When it comes to model variants, the results are very close, and there is not much difference between the models with depth 5 and 10. The average accuracy shows that the best model with highest descriptive accuracy is the same as for the triangular MF, while SWPT10 is best for obtaining models with high predictive performance.

The best model according to the descriptive performance is the same when using the Gaussian MF, while best average predictive performance is obtained by using the GWPT10 variant. Here, there is not much difference between the descriptive and predictive performance when it comes to number of MFs per attribute. It is noticeable that

for the modes obtained by using more than 5 MFs per attribute, better descriptive as well as predictive accuracy is achieved. Here, again, there is no noticeable difference regarding the WPT variants, both simple and general models with different depths can be used and they obtain very similar results.

Table 3. Evaluation results for the Saturated Oxygen dataset by using different number of MFs per attribute. Train denotes descriptive classification accuracy, while Test denotes predictive classification accuracy. Underlined results show the models with highest descriptive classification accuracy, while bolded results show the models with highest predictive classification accuracy.

Triangular Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	60.70	61.19	63.68	65.67	65.17	62.19	67.66	63.18	63.68	9.53*
Test[1]	**57.00**	56.50	55.50	53.50	52.50	53.50	56.00	53.00	**54.69**	
Train[2]	60.70	61.69	64.68	67.16	66.17	66.17	68.66	69.15	65.55	11.99
Test[2]	55.50	**56.50**	**56.50**	52.50	54.50	52.00	54.50	52.50	54.31	
Train[3]	60.70	59.70	60.20	64.68	68.66	62.69	67.66	63.18	63.43	11.76
Test[3]	**57.00**	54.50	53.00	51.00	52.00	48.00	52.50	53.00	52.63	
Train[4]	60.20	59.70	61.69	66.67	69.65	69.65	68.66	68.66	65.61	13.70
Test[4]	**56.00**	54.50	55.00	53.00	53.50	49.00	52.50	52.00	53.19	
Trapezoidal Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	60.20	60.70	62.19	65.17	66.17	61.69	66.67	67.66	63.81	11.04
Test[1]	53.00	53.00	55.00	55.00	**56.00**	46.00	**56.00**	52.00	53.25	
Train[2]	60.20	61.19	63.68	67.66	70.65	64.68	69.15	69.15	65.80	13.07
Test[2]	53.00	52.00	55.50	57.00	**58.00**	46.50	54.00	51.00	**53.38**	
Train[3]	60.20	61.19	62.19	66.17	66.17	61.69	67.16	67.66	64.05	12.01
Test[3]	53.00	53.00	**56.00**	54.50	55.00	43.50	54.50	52.00	52.69	
Train[4]	60.70	61.69	63.68	67.66	68.66	64.68	69.65	69.15	65.73	13.49
Test[4]	53.50	52.00	56.00	**56.50**	**56.50**	44.50	53.50	51.50	53.00	
Gaussian Membership Function										
	3	**4**	**5**	**10**	**20**	**30**	**50**	**100**	**Avg**	**RMSE**
Train[1]	56.22	57.71	57.71	63.18	65.67	65.67	65.67	65.67	62.19	10.95
Test[1]	52.00	52.00	49.00	56.00	50.00	**53.00**	51.00	**53.00**	52.00	
Train[2]	56.72	58.21	58.71	65.17	67.66	70.15	69.15	69.65	64.43	13.50
Test[2]	53.00	52.50	49.50	**54.00**	52.50	53.00	51.00	51.00	52.06	
Train[3]	58.71	60.20	61.19	61.69	68.16	65.67	65.17	65.67	63.31	15.19
Test[3]	52.00	52.50	**53.00**	51.50	50.50	41.50	42.50	**53.00**	49.56	
Train[4]	58.71	60.20	58.71	60.70	69.65	70.15	69.15	69.65	64.61	16.52
Test[4]	54.00	53.50	52.50	50.00	51.50	44.00	43.00	**56.00**	**54.31**	

[1]SWPT5, [2]SWPT10, [3]GWPT5, [4]GWPT10

For the Saturated Oxygen dataset, the resistance to over-fitting is better than for the pH dataset, and it is worse compared to the Conductivity dataset, peaking lowest error of 9.53%.

The in-depth analysis found mixed patterns which cannot be distinguished. In order to test/report statistical significance of the results, next, we present the results from the two-stage procedure.

4.2 Ranking the MFs and WPT variants

In this section, we perform ranking by using different number of MFs per attribute, different types of MFs, and different WPTs variants. The ranking is done by using the two stage procedure presented in [15] in order to measure the statistical significance of the obtained results. This procedure combines the Aligned Friedman test [13] and the post-hoc Hommel test [14]. The results from the ranking are presented in Table 4.

Table 4. The average ranks over the three datasets by using different settings (different number of MFs per attribute, different types of MFs and WPTs variants). The best rank is bolded and the model with best average rank is taken as control model for the post-hoc Hommel test. The rejected null hypothesis based on the Hommel adjusted *p*–values are underlined, and they correspond to statistically significant differences between the examined and the control model.

Different settings	**Conductivity**		**pH**		**Saturated Oxygen**		**All Datasets**	
ER1	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
3	86.45	40.75	82.16	22.66	83.62	**36.21**	290.5	91.6
4	73.00	41.41	78.75	18.41	78.96	42.58	230.4	93.2
5	65.66	**33.92**	69.12	**16.92**	69.25	37.75	204.6	**84.4**
10	24.04	36.83	52.62	51.92	46.96	40.00	126.7	132.8
20	28.71	82.75	38.62	68.67	22.96	43.75	91.6	194.7
30	**14.16**	68.25	37.37	74.08	39.67	78.95	97.5	221.8
50	46.54	35.08	**13.38**	55.96	**21.79**	53.14	**72.6**	152.7
100	49.42	49.00	15.96	78.37	24.79	55.62	82.8	184.6
ER2	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Triangular	62.95	**43.53**	38.33	**44.17**	44.98	**35.75**	144.6	**121.4**
Trapezoidal	**28.62**	53.00	**37.87**	52.98	**38.98**	43.58	**103.8**	148.3
Gaussian	53.92	48.97	69.30	48.34	61.53	66.17	185.0	163.7
ER3	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
SWPT5	71.18	50.71	66.77	52.87	69.25	40.20	204.7	142.6
SWPT10	35.98	58.25	32.12	**39.12**	31.54	**37.75**	98.12	**131.1**
GWPT5	61.54	**33.10**	68.54	57.27	61.92	64.71	192.5	157.5
GWPT10	**25.29**	51.94	**26.56**	44.73	**31.29**	51.33	**88.6**	146.7

The results from this ranking confirmed what we have concluded in the previous section: higher number of MFs per attribute (thirty for Conductivity and fifty for pH and Saturated Oxygen datasets) are best for obtaining models with higher descriptive accuracy, and these results showed as statistically significant compared to the models built with lower number (three, four and five) of MFs per attribute. Completely different picture can be seen for the models' predictive accuracy, where lower number of MFs per attribute, especially five MFs per attribute, are statistically significant compared to the models that are built with higher number of MFs per attribute (higher than ten). The conducted experiments that include all the datasets confirms this. Another set of experiments that were conducted evaluated different types of MFs. In most cases, the results for the triangular and trapezoidal MFs are statically significant compared to the results for the Gaussian MF. And the final analysis was made to evaluate the different WPT variants, which confirmed some of the discussion points that we made in the previous section. In most of the cases simple and general WPT models with depth 10 achieve best predictive and descriptive performance, and they are statistically significant compared to the models with depth 5. GWPT10 obtains statically significant descriptive performance compared to the models with depth 5, while SWPT10 is the best model in predictive analysis but without statistical significance.

5 Conclusion

In this paper work, we performed extensive experimental evaluation of the influence of the different MFs on the classification accuracy on diatom datasets using the WPT algorithm. Beside the influence of the different MFs on the descriptive and predictive performances of the models, we also investigated the influence of the number of MFs use per attribute and the different WPTs variants that have different complexities on the accuracy of the models. The results from the evaluation revealed some interesting patterns, like best ranked model with highest descriptive classification accuracy can be obtained using high number of MFs per attribute (fifty) in combination with triangular MF and GWPT models with depth 10. On the other hand, models built with lower number of MFs per attribute in combination with Triangular MF and SWPT with depth 5, have highest predictive classification accuracy. These conclusions are based on the outputs obtained from the two-stage procedure for testing the statistical significance of the results.

As future work, we plan to investigate other types of MFs, as well as different similarity metrics and aggregation operators that influence on the classification accuracy of the models.

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. Stroemer, E.F., Smol, J.P.: The diatoms: Applications for the Environmental and Earth Sciences, Cambridge University Press, Cambridge, (2004).
2. Quinlan, J.R.: Decision trees and decision making. IEEE Transactions on Systems, Man, and Cybernetics 20(2), 339–346, (1990).
3. Yuan, Y., Shaw, M.J.: Induction of fuzzy decision trees. Fuzzy Sets and Systems 69(2), 125–139, (1995).
4. Janikow, C.Z.: Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man, and Cybernetics 28(1), 1–14, (1998).
5. Olaru, C., Wehenkel, L.: A complete fuzzy decision tree technique. Fuzzy Sets and Systems 138, 221–254, (2003).
6. Suárez, A., Lutsko, J.F.: Globally optimal fuzzy decision trees for classification and regression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(12), 1297–1311, (1999).
7. Wang, X., Chen, B., Olan, G., Ye, F.: On the optimization of fuzzy decision trees. Fuzzy Sets and Systems 112, 117–125, (2000).
8. Nikravesh, M.: Soft computing for perception-based decision processing and analysis: web-based BISC-DSS, Studies in Fuzziness and Soft Computing, vol. 164, pp. 93–188, Springer Berlin/Heidelberg, (2005).
9. Huang, Z., Gedeon, T.D.: Pattern trees. In Fuzzy Systems, 2006 IEEE International Conference, pp. 1784–1791. IEEE, (2006).
10. Yi, Y., Fober, T., Hüllermeier, E.: Fuzzy operator trees for modeling rating functions. International. Journal of Computational Intelligence and Applications 8(04), 413–428, (2009).
11. Kóczy, L.T., Vámos, T., Biró, G.: Fuzzy signatures. EUROFUSE-SIC, pp. 210–217, (1999).
12. Huang, Z., Nikravesh, M., Azvine, B., Gedeon, T.D.: Weighted pattern trees: a case study with customer satisfaction dataset. In International Fuzzy Systems Association World Congress, pp. 395–406, Springer, Berlin, Heidelberg (2007).
13. Hodges, J., Lehmann, E.: Ranks methods for combination of independent experiments in analysis of variance. Annals of Mathematical Statistics 33, 482–497, (1962).
14. Hommel, G.: A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75, 383–386, (1988).
15. 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).
16. "TRABOREMA Project" WP3, EC FP6-INCO project no. INCO-CT-2004-509177, (2005–2007).
17. Krammer, K., Lange-Bertalot, H.: Die Ssswasserflora von Mitteleuropa 2: Bacillariophyceae. 1 Teil pp. 876, Stuttgart: Gustav Fischer-Verlag, (1986).
18. Van Der Werff, A., Huls, H.: Diatomeanflora van Nederland. Abcoude - De Hoef, (1957, 1974).
19. Van Dam, H., Martens, A., Sinkeldam, J.: A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology 28 (1), 117–133, (1994).
20. Abramowitz, M.: Handbook of Mathematical Functions. With Formulas, Graphs, and Mathematical Tables, Dover Publications, (1974).