PATTERN TREE SPATIAL MODELS FOR ECOLOGICAL CLASSIFICATION

Andreja Naumoski,
Faculty of Computer Science and Engineering
Skopje, R. Macedonia,
andreja.naumoski@finki.ukim.mk

Kosta Mitresski
Faculty of Computer Science and Engineering
Skopje, R. Macedonia,
kosta.mitreski@finki.ukim.mk

ABSTRACT

This paper further extends pattern trees membership functions, by implementing a modified sigmoid distribution. In this work we use this algorithm to extract knowledge for ecological classification task from the diatoms community measured dataset, which according the biological experts are used as bio-indicators in many water ecosystem environments. The first part of the algorithm transforms the input set from crisp values into fuzzy values, and then continues the induction of the tree. The transformation is achieved by using different membership functions, which have different shape and mathematical description. This is very important because later in the induction phase this will have effect on the classification accuracy and complexity of the obtained model. The modified sigmoid function that is put on test, have several advantage over the triangular and trapezoidal functions. The experiments on diatoms classification datasets showed that sigmoid shaped function algorithm models outperform the pattern tree models build based on the trapezoidal, triangular or Gaussian MF in terms of prediction accuracy. The diatom models based on this method produced valid and useful knowledge that later in the paper is interpreted. Finally, evaluation performance analyses of the build pattern trees with classical classification algorithms is presented and discussed.

I. INTRODUCTION

The water quality class (WQC) define in the traditional way can be interpreted as classification problem from data mining point of view. This property is used for finding the proper organism - environment relationship that has been a subject of eco-informatics area of research very recently. Considering this, we deal with the typical classification problem.

The main question is: why to use this method in process of knowledge discovery? First of all, the pattern trees (PTs) are robust to over fitting that is not the case with the fuzzy decision trees and ordinary decision trees, concluded [1]. Secondly, they obtain compact structure, which is essential in the process of representation of the knowledge gain from the biological data. In this domain, classical statistical approaches, such as canonical correspondence analysis (CCA), detrended correspondence analysis (DCA) and principal component analysis (PCA), are most widely used as modelling techniques [2]. Although these techniques provide useful insights in the data, the interpretation of the results is limited in terms of fitting the diatom in the one of the WQCs.

In order to improve the process, we use decision trees. Decision trees, fuzzy decision trees (FDTs) and PTs are effective tool for classification approach [1]. As sub-class of decision trees, the fuzzy and PTs have several advantages over classical decision trees. As successor of the benefits from the fuzzy decision trees, the PT can obtain high accuracy and it is robust of over-fitting, according to [1].

Because the fuzzy set based machine learning can overcome some problems of the classical learning, this is very active area of research. Wang and Mendel [3] have presented an algorithm for generating fuzzy rules. Inspired by the classic decision tree induction by Quinlan [4], there is great work on FDT. On other side, the Yuan and Shaw [5] have proposed FDTs induction using fuzzy entropy, while the Janikow [6], Olaru and Wehenkel [7] have presented different fuzzy decision tree inductions. Suárez and Lutsko [8], and Wang and Chen [9] have presented optimizations of fuzzy decision trees. Most of the methods include fuzzy decision trees [10] that focus on searching for rules which only use t-norm operators [11] such as the MIN or MIN. Research has been conducted to resolve this problem. Kóczy, Vámos and Biró [12] have proposed fuzzy signatures to model the complex structures of data points using different aggregation operators including MIN, MAX, and average etc. Mendis, Gedeon and Kóczy [13] have investigated different aggregations in fuzzy signatures. Nikravesh [14] has presented evolutionary computation (EC) based multiple aggregator fuzzy decision trees. The new PT method was proposed by [15] and they use simple evenly distributed trapezoidal, triangular and Gaussian membership functions (MFs). In this paper we propose new distributed MF; modified sigmoid MF that later in the paper is described together with the experiment evaluation over the diatom community datasets. From this dataset using PTs we will use to extract valuable knowledge about the diatom-environment relationship. This is vital because later, the rules and models produced from the tree can be easily evaluated by the known ecological references found in the literature. And last, the PTs will be tested for classification accuracy with other classification algorithms, which is the second criterion in the process of extracting knowledge. Also, we studied the influence of the number of MF per attribute, which leads to more precision range of the diatoms abundance.

The rest of the paper is organized as follows: Section II provides the definitions for similarity and aggregations metrics. In Section III we briefly introduce the modified sigmoid MFs. Section IV presents the diatoms abundance water quality datasets and the experimental setup, while the section V gives the experimental results and the interpretation for some of the model trees generated by the PT and the spatial fuzzy models. Finally, Section VI concludes the paper and research direction is outlined.

II. SIMILARITY AND FUZZY AGGREGATION METRICS

The PT diatom model is obtained by using different similarity measures and fuzzy aggregation, which are

presented in this section. The root mean square error (RMSE) of fuzzy sets A and B can be computed as:

$$Sim(A;B) = 1 - RMSE(A;B) = \sqrt{\frac{\sum_{i=1}^{n}\left(\mu_A(x_i) - \mu_B(x_i)\right)^2}{n}} \quad (1)$$

,where x_i, $i = 1, \ldots, n$, are the crisp discretized values, and $\mu_A(x_i)$ and $\mu_B(x_i)$ are the fuzzy membership values of x_i for A and B, that are two fuzzy sets defined on the universe of discourse U. The larger the value of Sim (A; B), the more similar A and B are. As $\mu_A(x_i)$, $\mu_B(x_i)$ [0, 1], $0 \leq$ Sim (A; B) ≤ 1 holds according to (1).

According fuzzy logic theory, the fuzzy aggregation are logic operators applied to fuzzy membership values or fuzzy sets. They have three sub-categories, namely t-norm, t-conorms, and averaging operators such as weighted averaging (WA) and ordered weighted averaging (OWA). In our experimental setup, we use the basic operators (Algebaric AND/OR) which operate on two fuzzy membership values a and b, where a, $b \in$ [0, 1] used in [10]. No weighted approach is studied in this paper.

III. Proposed sigmoid membership function

In the previous work [15], the process of generating fuzzy terms was achieved using three evenly distributed MFs: trapezoidal, triangular and Gaussian. Depending on the nature of the dataset, different MF will have different effect, meaning that the accuracy of the model will be changed.

The straight line MFs (triangular and trapezoidal) has the advantage of simplicity. They are simple, and in some case in process of building models can gain more precision power. Yet, many of the datasets have smoothed values and nonzero points, which apply to use more different MFs, such as sigmoid distribution curve.

In this paper we introduce modified evenly sigmoid MF, which in general is specified by three parameters, represented with eq. 2.

$$f(x;a;b) = \frac{1}{1 + e^{-a*(x-b)}} \quad (2)$$

In eq. 2 the parameter are constants, while the b parameter is located at the centre of the curve. In this paper, we propose that the eq. 2 be modified, by taking the mean values of the given data range into account. In this way, each fuzzy term will reflect the very nature of the tested dataset and evenly distributed sigmoid MF in the entire range. And finally when all this changes are taken into account eq. 3 mathematically represents the modified sigmoid MF as:

$$f(x;\mu;\sigma) = \frac{1}{1 + e^{-a*(x-\mu)}} \quad (3)$$

In eq. 3 the parameter a will get two values {1 and -1}, which will be intensively studied in this paper. Because of the smoothness and concise notation, modified sigmoid MFs, can be used for specifying fuzzy sets and ecological knowledge discovery and be used for wide range of different type of datasets.

IV. Data description and experimental setup

The datasets used in the experiments consist from 13 input parameters representing the TOP10 diatoms species (diatoms species that exist in Lake Prespa [16]) with their relative abundance per sample, plus the three WQCs for conductivity, pH and Saturated Oxygen. These measurements were made as a part of the TRABOREMA project [17]. The water quality classes are defined according the three physical-chemical parameters: Saturated Oxygen [18], Conductivity [19] and pH [18, 19] which are given in Table 1.

Table 1: WQCs for the physical-chemical parameters

Physical-chemical parameters	Name of the WQC	Parameter range	Name of the WQC	Parameter range
Saturated Oxygen	*oligosaprobous*	SatO > 85	*α-mesosaprobous*	25-70
	β-mesosaprobous	70-85	*α-meso / polysaprobous*	10-25
pH	*acidobiontic*	pH < 5.5	*alkaliphilous*	pH > 7.5
	acidophilous	pH > 5.5	*alkalibiontic*	pH > 8
	circumneutral	pH > 6.5	*Indifferent*	pH > 9
Conductivity	*fresh*	Conduc < 20	*brackish fresh*	90 – 180
	fresh brackish	Conduc < 90	*brackish*	180 - 900

Among the input parameters, 10 are numerical parameters and the rest 3 are nominal with number of possible classes from 4 up to 6. We have made two variants of the method. First, so called; simple PT trees, which consist from 0 candidate trees and one low level tree with two different depths 5 – SPT5 and 10 - SPT10. And secondly, we induce models that consist from 2 candidate trees, 3 low level trees and depth equal to 5 – PT5 and 10 – PT10. For similarity definition we use RMSE similarity and only (Algebaric AND and OR) as fuzzy aggregation operator. Later, comparison with other crisp classifiers is done with simple and general PT models with different depth (5 and 10).

The configuration of the experiments is set up as follows. 1) A simple fuzzification method based on three evenly

distributed MFs including the modified sigmoid MF for each input variable is used to transform the crisp values into fuzzy values (Train); 2). Two experiments are carried out, with the first (Exp2 – odd-even) using odd labelled data as training set and even labelled data as test set, and the second (Exp3 – even-odd) using even labelled data as training set and odd labelled data as test set (2-fold cross validation) and 3) Standard 10-fold cross validation is used for testing the prediction performance accuracy of the built models (xVal). Table 2 shows results of the conducted experiments.

Table 2. Average prediction accuracy per WQC (in %)

Conductivity WQC – Average Prediction Accuracy (in %)					
Type of experiment	Triangular	Trapezoidal	Gaussian	Sigmoid (a= 1)	Sigmoid (a= -1)
Train	73.80	**74.66**	74.03	72.19	73.80
Exp2	69.15	**71.22**	68.12	68.46	69.15
Exp3	69.61	69.50	69.27	**70.87**	69.78
Classical Algo.	C 4.5	kNN	Bagging C4.5		Boosted C4.5
xVal	65.60	66.51	63.30		63.76
PT Algorithm	SPT5	SPT10	PT5		PT10
xVal (a= 1/ a= -1)	**71.83/71.83**	**71.36/71.36**	**73.16/73.16**		**71.32/71.32**
pH WQC – Average Prediction Accuracy (in %)					
Type of experiment	Triangular	Trapezoidal	Gaussian	Sigmoid (a= 1)	Sigmoid (a= -1)
Train	59.40	**60.95**	58.66	60.72	60.44
Exp2	46.30	41.90	47.92	**59.03**	54.86
Exp3	46.53	45.83	48.61	**52.31**	49.07
Classical Algo.	C 4.5	kNN	Bagging C4.5		Boosted C4.5
xVal	54.73	47.26	53.23		56.22
PT Algorithm	SPT5	SPT10	PT5		PT10
xVal (a= 1/ a= -1)	**58.10/59.41**	**58.55/58.93**	**58.10/59.00**		**58.55/58.25**
Saturated Oxygen – Average WQC Prediction Accuracy (in %)					
Type of experiment	Triangular	Trapezoidal	Gaussian	Sigmoid (a= 1)	Sigmoid (a= -1)
Train	**62.13**	61.13	62.00	61.26	59.83
Exp2	52.75	56.00	50.50	**57.63**	56.38
Exp3	50.37	52.35	51.36	53.96	**54.46**
Classical Algo.	C 4.5	kNN	Bagging C4.5		Boosted C4.5
xVal	**62.13**	**61.13**	**62.00**		**62.00**
PT Algorithm	SPT5	SPT10	PT5		PT10
xVal (a= 1/ a= -1)	58.50/57.00	58.50/56.50	58.00/57.00		58.50/55.51

V. EXPERIMENTAL RESULTS

A. Performance Analysis

The modified sigmoid shaped MF outperforms 2 of the 3 diatoms WQCs compare with other MFs. Both trapezoidal and combination of the two modified sigmoid MFs have obtained higher prediction accuracy than triangular and Gaussian MF in *odd-even* and *even-odd* experimental setup. Except for the Saturated Oxygen class, the Saturated Oxygen WQC and pH WQC, the method has obtained better prediction accuracy for sigmoid MF in Experiment 2 and 3. According to Table 2, the trapezoidal MF outperforms other fuzzy functions for pH WQC in train experiments, but sigmoid shaped MF outperformed other MF.

The comparison with the classical classification algorithms shows similar results. In this setup we use same number of MF equal to 5; except for the Saturated Oxygen WQC we set the number of MF to 10. The prediction accuracy of the pattern increases for the conductivity WQC and pH WQC even in some cases with 10%, while for the most of cases from 2% to 5%. Saturated Oxygen WQC gained low performance, because the shape of the MF is not suitable for this WQC. However, it remains in focus for our future research.

B. Interpretation of the diatom models

Based on the performance results, in this section we presented several PTs models and their rules. We have built many models for each WQC, but due to paper constrains we present one model tree for each WQC. All the induced

models, have define range of Fuzzy Terms, which later will be commented. The number of MFs per attribute is 5 and all the model trees are generated using Experimental Setup 3 - xVal.

The method generated a separate rule for each class, which describes the diatoms abundance in correlation with physic - chemical property of the water. The PT model shown in Fig. 1 can be converted into rule which is stated below.

Rule1: If [*Diploneis mauleri* (DMAU) has μ=0.0 and *Amphora pediculus* (APED) has μ=0.0) and [*Cocconeis placentula* (CPLA) has μ=0.0 or *Navicula rotunda* (NROT) has μ=2.67)] or DMAU has μ=8.00 then the Conductivity WQ class is *fresh*.

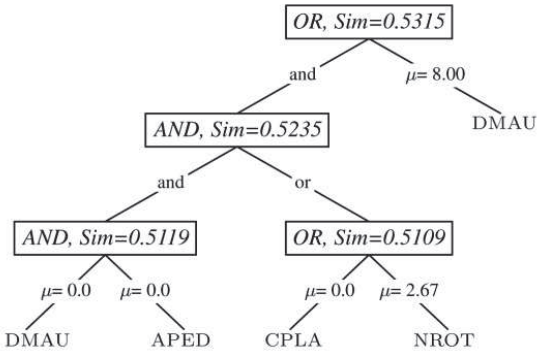

Figure 1: PT model obtained for the *fresh* - Conductivity WQC

From Rule1 can be easily seen that the two main diatoms NROT and DMAU, especially DMAU with higher abundant than the NROT diatom can be found in the water were Conductivity class is *fresh*. Using the generated rule we can immediately see what is the mean value of this diatoms in measure sample.

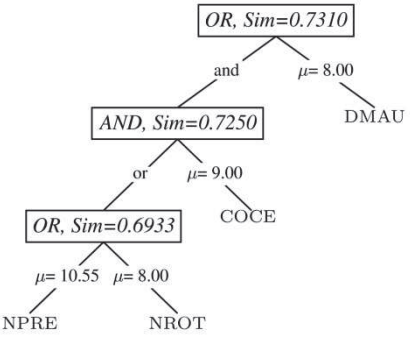

Figure 2: PT model obtained for the for the *circumneutral* pH WQC

According the model tree, the DMAU diatom can be found in the ecosystem with highest abundance that the rest of the diatoms in this model tree. The NROT diatom is less abundant than the DMAU diatom according the model with mean value of 2.67. Other diatoms have low level of abundance in the water according the model.

In the fallowing paragraph we present two more trees, one for the pH WQC – *circumneutral* and other one for Saturated Oxygen WQC – *β-mesosaprobous*. The rule induced from tree shown in Fig. 2 states:

Rule2: If [*Navicula prespanense* (NPRE) has μ=10.55 or NROT has μ=8.00] and *Cyclotella ocellata* (COCE) has μ=9.00 or DMAU has μ=8.00, then the pH WQ class is *circumneutral*.

This model tree shows that several diatoms can be found to exist in *circumneutral* waters. According to the tree model, the NPRE diatom can be more likely to be found with COCE diatom in *circumneutral* waters. Nevertheless, NROT and DMAU diatoms exist in these waters and they are important habitats, but less abundant than the previous ones, according the model tree.

Using the PT method with modified sigmoid MF the induced tree presented in Fig. 3 shows the suitable habitat of Saturated Oxygen WQC - *β-mesosaprobous* on the four diatoms (APED, COCE, NROT and STPNN). These diatoms and their combination can exist in this WQC.

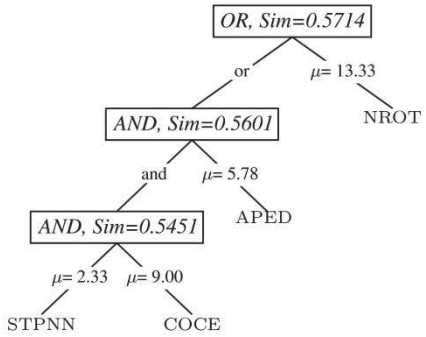

Figure 3: PT model obtained for the *β-mesosaprobous* Saturated Oxygen WQC

Rule3: If [*Staurosirella pinnata* (STPNN) has μ=2.33 and COCE has μ=9.00] and APED has μ=5.78 or NROT has μ=13.33, then the Saturated Oxygen WQ class is *β-mesosaprobous*.

This model tree indicated that the NROT diatom is mostly likely to be found in the *β-mesosaprobous* waters, or COCE and APED diatoms together, which are less abundant that the NROT diatom. STPNN diatoms according the model tree is the less abundant in these waters.

C. Spatial fuzzy diatom models

The diatom model given in Fig. 3, the leaf 3 (APED has μ=5.78 relative abundance) is represented as spatial fuzzy model (see Fig. 4) in order to investigate the spatial information of the given diatom. According to the model, the APED diatom can be relative good indicator for low level concentration of metals in locations L_8 and L_1, while other location medium concentration of metals (Cu, Mg, Mn and Zn) is suitable for this diatom. Eutrophication parameter (nitrogen and phosphorus) for almost all measuring station is low, except for the L_4 location. Saturated Oxygen and Secchi Disk has no measurements for some location, so further investigation is needed to confirm the APED diatom suitable habitat. Each leaf from the diatom model can be represented

with GIS, in this way for all the diatoms can be found the indicating properties.

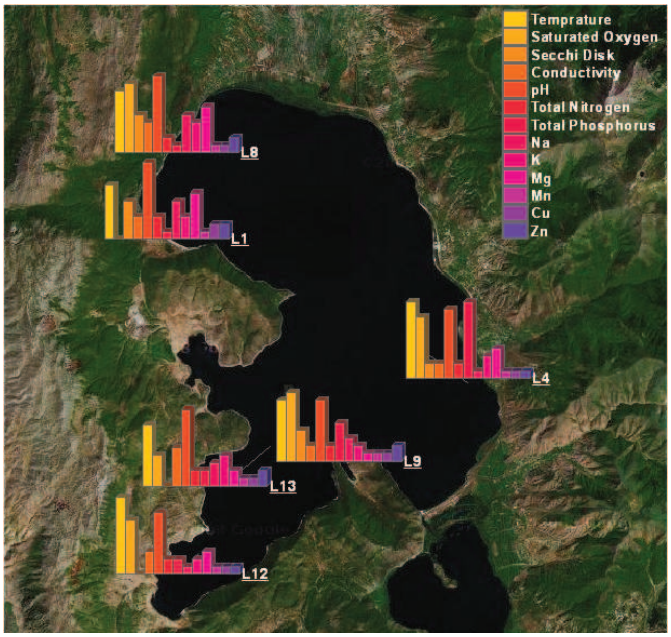

Figure 4: Spatial fuzzy model for leaf 4 – APED with μ=5.78 for *β –mesosaprobous* class

VI. Conclusion

The experiments on diatoms datasets WQC dataset showed that modified sigmoid MF outperform the models obtained with the method which uses trapezoidal, triangular or Gaussian in terms of prediction accuracy. The two modifications of the sigmoid shaped MF for a parameter of 1 and -1, achieved better results and in some cases the sigmoid MF with a=1 is even better in terms of prediction accuracy. This is very important for different types of datasets. Also, in this paper we have compared the prediction accuracy between the proposed method and the ordinary crisp classification algorithms and showed improvement of the classification accuracy for some of the WQC dataset. The mixed datasets *odd-even* and *even-odd* (2-fold cross validation) performed better, which means that the generalization of the proposed method is greater. Conducted experiments on the diatoms datasets show that the average prediction accuracy for the modified sigmoid MF is greater than the classical crisp classifiers. In terms of interpretability, previously used methods has low level score for this important property of the ecological model, which puts the PT in the highest place in the methods that should be used for diatom classification.

Further research on developing more MF in process of building pattern tress is necessary. The current version also needed to be updated in term of new similarity definitions. More sophisticated fuzzy aggregations and similarity definitions may be more suitable for diatoms community dataset and can therefore lead to higher accuracy. From ecological point of view, other physico - chemical parameters can be used for diatom classification and contribute in the generalization of the method.

References

[1] D. Kocev, A. Naumoski, K. Mitreski, S. Krstić, S. Džeroski, "Learning habitat models for the diatom community in Lake Prespa," Journal *of Ecological Modelling,* vol. 221, no. 2. pp. 330-337, 2010.

[2] E. F. Stroemer, and J. P. Smol, *The diatoms: Applications for the Environmental and Earth Sciences*. Cambridge University Press, Cambridge, 2004.

[3] L.X. Wang, and J.M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 22, no. 6, pp.1414-1427, 1992.

[4] R.J. Quinlan, "Decision trees and decision making," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 20, no. 2, pp.339-346, 1990.

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

[6] C.Z. Janikow, "Fuzzy decision trees: issues and methods," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 28, no. 1, pp.1-14, 1998.

[7] C. Olaru, and L. Wehenkel, "A complete fuzzy decision tree technique," *Fuzzy Sets and Systems*. vol. 138, pp. 221-254, 2003.

[8] A. Suárez, and Lutsko, J.F. "Globally optimal fuzzy decision trees for classification and regression, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no.12, pp. 1297-1311, 1999.

[9] X. Wang, B. Chen, G. Olan, and F. Ye, "On the optimization of fuzzy decision trees," *Fuzzy Sets and Systems,* vol. 112, pp. 117-125, 2000.

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

[11] B. Schweizer, and A. Sklar, "Associative functions and abstract semigroups," *Publ. Math. Debrecen.* vol. 10, pp. 69-81, 1963.

[12] L.T. Kóczy, T. Vámos, G. and Biró, "Fuzzy signatures". *EUROFUSE-SIC*. pp. 210-217, 1999.

[13] B.S.U. Mendis, T.D. Gedeon, and L.T. Kóc, "Investigation of aggregation in fuzzy signatures," *3rd International Conference on Computational Intelligence, Robotics and Autonomous Systems, Singapore*, 2005.

[14] M. Nikravesh, "Soft computing for perception-based decision processing and analysis: web-based BISC-DSS," *Studies in Fuzziness and Soft Computing.* vol. 164, pp. 93-188, 2005.

[15] Z.H. Huang, T.D. Gedeon, and M, Nikravesh, M, "Pattern Trees Induction: A New Machine Learning Method." *IEEE Transaction on Fuzzy Systems.* vol. 16, no. 3, pp. 958-970, 2008.

[16] Z. Levkov, S. Krstič, D. Metzeltin, and T. Nakov, "Diatoms of Lakes Prespa and Ohrid (Macedonia). *Iconographia Diatomologica*. vol. 16, pp. 603, 2006

[17] TRABOREMA Project - WP3.: EC FP6-INCO project no. INCO-CT-2004-509177, 2005-2007.

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

[19] A. Van Der Werff, and H. Huls, *Diatomeanflora van Nederland*. Abcoude - De Hoef, 1957, 1974.