Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 2 (2010) 1124–1138

Procedia Environmental Sciences

International Society for Environmental Information Sciences 2010 Annual Conference (ISEIS)

Classifying diatoms into trophic state index classes with novel classification algorithm

Andreja Naumoski*, Kosta Mitreski

University "Ss. Cyril and Methodius" – Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia, P.O.Box 574, Karpos 2 bb, 1000 Skopje, R. Macedonia

Abstract

Diatoms are ideal bio-indicators of water ecosystem health and can be classified into one of the trophic state indexes (TSI) according to the nutrient level. Thus, the diatoms can be used to indicate the relationship between the organisms and the environmental parameters. In order to find the correct diatom- indicator connection, we can use a certain classification algorithm directly from measure data. This process of diatom classification can be significantly improved using information technology, especially data mining tools. In this direction, this paper work present several classification models with the novel method called aggregation trees based on evenly sigmoid shaped membership function (MF). Earlier, numerous statistical approaches have been used for this purpose, which provide very useful data inside information, but they are limited to interpretation. Further improvement is made by using decision trees, which increases interpretability, but remains not resistant to over fitting and robustness on data change. The proposed method in this paper synthesizes these advantages, in terms of interpretability, resistance of over-fitting and high classification accuracy compared with classical classification algorithms. This is confirmed by the experimental evaluation. Based on these evaluation results, one model for each TSI is presented and discussed. From ecological point of view, the described method improves the water quality and sustaining bio diversity understandings of this ecosystem. The method added new ecological knowledge about the ecological indicators for certain diatoms, which have been recently discovered.

Keywords: Aggregation trees; Lake Prespa; Diatoms; Trophic State Index; Sigmoid distribution; Classification models

1. Introduction

Lake Ecosystem classification according to his eutrophication status is a very important issue in today fast growing world. Demands of clear water, not just for drinking, but also for the survival and maintaining the

* Corresponding author.
E-mail address: andrejna@feit.ukim.edu.mk.

1878-0296 © 2010 Published by Elsevier Open access under CC BY-NC-ND license.
doi:10.1016/j.proenv.2010.10.122

organism's habitat, becomes ever important. In this direction, it is vital to manage these resources as much intelligence as we can with advance methods of new information technologies.

Focusing on this, the TSI classes define in the traditional way can be interpreted as a classification problem in the terms of data mining point of view. This property is used to discover the appropriate environment conditions for newly found diatom, which has been a subject of environmental informatics area of research very recently. Considering this, we deal with the typical classification problem, when we try to build a model that predicts (classifies) the correct diatom into TSI according to certain physico-chemical parameter. These objectives are usually very difficult to achieve by extracting knowledge directly from data without help from the information technologies. The eutrophication status measure is one of the key factors detecting the health of the ecosystem.

In this domain, classical statistical approach, such as canonical correspondence analysis (CCA), detrended correspondence analysis (DCA) and principal component analysis (PCA), are most widely used as modelling techniques [1]. Although these techniques provide useful insights in the data, they are limited in terms of interpretability. Obvious progress in this research area in a direction of interpretability, have been made using data mining techniques, particularly decision trees. These data mining methods, improves the problem of interpretability and increases the prediction power of the model trees. First attempt to model diatom-environment relationship for Lake Prespa, have been made by [2,3]. Various settings were applied to the datasets and thus different models were obtained, which later have been discussed with the biological expert. Several of the model produced, knowledge about the newly discovered the diatom's relationships with the environment for the first time [3].

After successful modelling the Lake Prespa diatoms, new class of multi-target decision trees was used, in order to reveal the dynamic nature of the entire set of physical-chemical parameters of this lake ecosystem [4]. These methods were more precise and also have greater interpretability than the previous methods. Nevertheless, these methods were not robust on data change, because the structure of the algorithm implies that. This is important because the environmental condition inside of the lake changes over small periods of time.

The robustness of data change and resistant to over-fitting of the fuzzy based concept is the main reason of extensive research on the fuzzy set based machine learning. Wang and Mendel [5] have presented an algorithm for generating fuzzy rules by learning from examples. Inspired by the classic decision tree induction by [6], there are substantial works on fuzzy decision trees. For example, [7] have proposed fuzzy decision trees induction using fuzzy entropy. [8] have presented different fuzzy decision tree inductions. [9,10] have presented optimizations of fuzzy decision trees. Most of the existing fuzzy rule induction methods including fuzzy decision trees [7] focus on searching for rules, which only use t-norm operators [11] such as the MIN and algebraic MIN. Research has been conducted to resolve this problem. [12] have proposed fuzzy signatures to model the complex structures of data points using different aggregation operators including MIN, MAX, and average, etc. [13] have investigated different aggregations in fuzzy signatures. [14] has presented evolutionary computation (EC) based multiple aggregator fuzzy decision trees. Recently, new method; pattern trees was introduced by [15], which satisfy the requirements stated above. The proposed method in this paper is very similar to the pattern trees method, but uses different membership functions and different similarity metrics.

The main question is: why use aggregation trees (AT) in the process of diatom classification? They are several reasons for this, and this entire concept is proofed in this paper. First of all, the proposed method is robust to over fitting, which is not the case with the classical methods and decision trees. Secondly, they obtain a compact structure, which is essential in the process of representation of the knowledge gain from the biological data. This is vital because later, the rules produced from the tree can be easily evaluated easily by the biological expert. And third, these models can achieve high classification accuracy. One of the reasons, why this method is better compared with the previous ones, is the use of different fuzzy membership functions.

The rest of the paper is organized as follows: Section II provides the definitions for similarity metrics and aggregation operators are presented. In Section III a novel evenly sigmoid membership functions are proposed. Section IV presents the diatom's abundance trophic state index classes, dataset description and the experimental setup. In section V present one prediction model for each TSI class in Lake Prespa. Section VI shows the prediction performance and experimental comparisons. Finally, Section VII concludes the paper and research direction is outlined.

2. Similarity metrics and fuzzy aggregation operators

The aggregation tree method described in this section is induced by using different similarity measures and fuzzy aggregation operators.

2.1. Similarity metrics

Let assume that A and B are two fuzzy sets [8] which are defined on the universe of discourse U. The root mean square error (RMSE) of fuzzy sets A and B can be computed as:

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

where x_i, i = 1,…,n, are the crisp values discretized in the variable domain, and $\mu_A(x_i)$ and μ_B=(x_i) are the fuzzy membership values of x_i for A and B. The RMSE based fuzzy set similarity can thus be defined as:

$$Sim(A;B)=1-RMSE(A;B) \quad (2)$$

The larger the value of *Sim*(A,B), the more similar A and B are. As $\mu_A(x_i)$, $\mu_B(x_i) \in [0,1]$, $0 \le Sim(A;B) \le 1$ holds according to (1) and (2). Note that the proposed method induction follows the same principle if alternative fuzzy set similarity definitions such as Jaccard are used [15]. In our experiments, we use only RMSE similarity metrics. Nevertheless other similarity metrics and membership functions are in focus for our further research.

2.2. Fuzzy aggregation operators

A fuzzy set operation is an operation in fuzzy sets. These operations are a generalization of crisp set operations. There are three sub-categories, namely t-norm, t-conorms, and averaging operators such as weighted averaging (WA) and ordered weighted averaging (OWA) [16]. In our experimental setup, we use the basic operators (Algebaric AND/OR) which operate on two fuzzy membership values a and b, where a, $b \in [0,1]$. (See equations 3). No weighted approach is studied in this paper.

$$\begin{aligned} &\text{MIN / MAX} \\ &\text{T-Norm}: \text{Min}\{a,b\} = a \wedge b \\ &\text{T-Conorm}: \text{Max}\{a,b\} = a \vee b \end{aligned} \quad (3)$$

Aggregation tree can be generated using different fuzzy aggregation operator sub-categories, which we plan to be a subject for our future research.

2.3. Aggregation trees algorithm

An aggregation tree is a tree which propagates fuzzy terms using different fuzzy aggregations, in this paper sigmoidal MF. Each aggregation tree represents a structure for an output class in the sense that how the fuzzy terms aggregate to predict such a class.

The extension of the simple aggregation tree, the general aggregation tree's induction, considers aggregating not only fuzzy terms, but also other aggregation trees. Subject to the particular demands (comprehensibility or performance), simple aggregation trees and general aggregation trees provide a highly effective methodology for real world applications, in our case, extracting knowledge from diatoms dataset. In this paper, we induce simple and general aggregation trees, because we want to find more general knowledge that fits better for knowledge discovery of correct diatom-environment classification. To our knowledge this is for the first time, to use this classification algorithm for this purpose.

3. Proposed membership function for diatoms classification

The straight line membership functions (triangular and trapezoidal) have the advantage of simplicity. They are simple, and in some case in the process of building aggregation trees gain relatively good precision power. Yet, many of the datasets, including the diatom-indicator relationship (see Fig. 1) have smoothed values and nonzero points. This conclusion applies to use more different membership functions for generate different fuzzy sets. Such functions are the Gaussian distribution curve: a simple Gaussian curve and signum membership function.

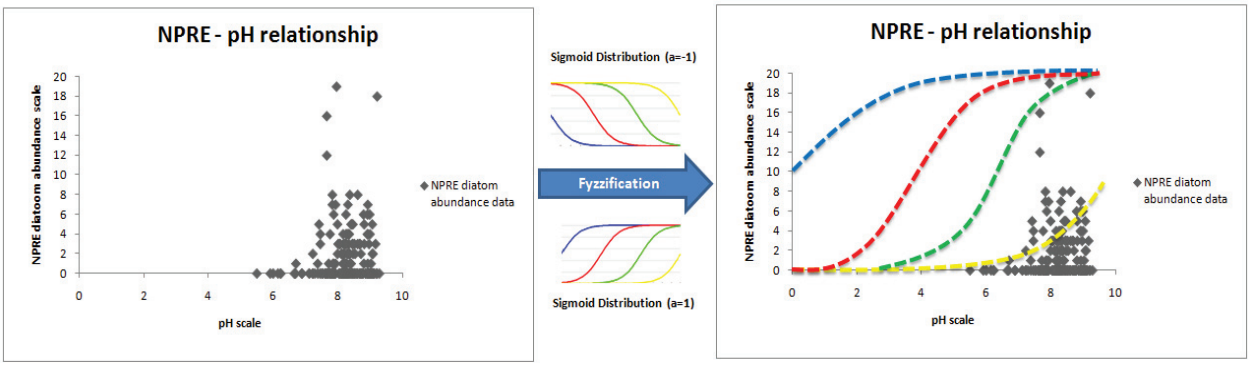

Fig. 1. NROT diatom and pH relationship within the given diatom dataset under sigmoid distribution area.

This relationship can be seen all over the diatom dataset between the biological aspect and the abiotic factors. For example, the relationship between the NROT diatom and the pH values in the diatom dataset can be covered with one of the several fuzzy sigmoid distributions were (a = -1) membership function for almost 99% of the data (see Fig. 1). It can be clearly seen that the red (doted) line covers (belong) the entire NROT (input fuzzy term) data within the given pH class (output class) range above 5 units. This is the main reason why we use the proposed fuzzy membership functions. We can make changes in order to fit more precisely into the given diatom range, the both proposed membership functions to increase the prediction power. All fuzzy sets have values from 0 to 1, because the A and B two fuzzy sets [8] are defined on the universe of discourse U [0, 1].

3.1. Evenly sigmoid distribution

Because the relationship between the diatoms and the TSI classes in many cases has evenly distributed distribution, we have modified previous equation (4), so that interception between two sigmoid functions has equal area and fit to the property of the relationship. We also propose that the equation (4) to be modify, by taking only the mean values (μ) of the given data range into account. In this way, each fuzzy MF per attribute will fallow the increasing and decreasing of the diatom's abundance and reflect the very nature of the tested dataset.

$$f(x;a;b)=\frac{1}{1+e^{-a*(x-b)}},\ a,b>0 \tag{5}$$

In equation 5, a and b parameters are positive constants. And finally when all this change is taken into account, the equation (5) mathematically represents the modified evenly sigmoid distributed membership function as:

$$f(x;\mu;a)=\frac{1}{1+e^{-a*(x-\mu)}}\ , \tag{6}$$

where the parameter a will get two values {1 and -1}, which will be intensively studied in this paper. It is expected, that the evenly sigmoid distribution better fallow the diatom-indicator relationship. The result of the fuzzification process for the proposed membership function is presented with Table 1.

Table 1. Fuzzy terms of the TOP10 diatoms after fuzzification

TOP10 Diatoms	Fuzzy Term - BAD	Fuzzy Term - WEAK	Fuzzy Term – GOOD	Fuzzy Term – VERY GOOD	Fuzzy Term - EXCELLENT
APED	0	3.25	6.5	9.75	13
CJUR	0	6.75	13.5	20.25	27
COCE	0	20.25	40.5	60.75	81
CPLA	0	8.5	17	25.5	34
CSCU	0	10.25	20.5	30.75	41
DMAU	0	3	6	9	12
NPRE	0	4.75	9.5	14.25	19
NROT	0	6	12	18	24
NSROT	0	5.75	11.5	17.25	23
STPNN	0	5.25	10.5	15.75	21

4. Data description and experimental setup

Lake Prespa is located at the border intersection of Macedonia, Albania and Greece (see Fig. 2). It covers an area of 301 km^2 at 850 m above sea level. The whole region that surrounds the lake was recently proclaimed a transboundary park (Prespa Park). The Prespa Park is well known for its great biodiversity, natural beauty and populations of rare water birds. However, the ecological integrity of the region is threatened by the increasing exploitation of the natural resources (inappropriate water management, forest destruction leading to erosion, overgrazing), inappropriate land-use practices, ecologically unsound irrigation practices, water and soil contamination from uncontrolled use of pesticides, lake siltation and uncontrolled urban development.

Monitoring of the state of Lake Prespa was performed during the EU project TRABOREMA. The measurements cover one and a half year period (from March 2005 to September 2006). Samples for analysis were taken from the surface water of the lake at 14 locations. The lake sampling locations are distributed in three countries (see Fig. 1) as follows: 8 in Macedonia, 3 in Albania and 3 in Greece. The selected sampling locations are representative for determining the eutrophication impact [17]. Through the lake measurements, a total of 218 water samples were collected. On these water samples, both physicochemical and biological analyses were performed.

The following physicochemical properties of the water samples were measured: temperature, dissolved oxygen, Secchi depth, conductivity, alkalinity (pH), nitrogen compounds (NO_2, NO_3, NH_4, inorganic nitrogen), sulphur oxide ions SO_4, and Sodium (Na), Potassium (K), Magnesium (Mg), Copper (Cu), Manganese (Mn) and Zinc (Zn).

The biological variables were the relative abundances of 116 different diatom taxa (for a complete list of diatom names and acronyms see [18]. Diatom cells were collected with a planktonic net or as attached growth on submerged objects (plants, rocks or sand and mud). This is the usual approach in studies for environmental monitoring and screening of diatom abundance. The sample, afterwards, is preserved and the cell content is cleaned. The sample is examined with a microscope, and the diatom taxa and abundance in the samples are obtained by counting 200 cells per sample. The specific taxon abundance is then given as the percent of the total diatom count per sampling site [18].

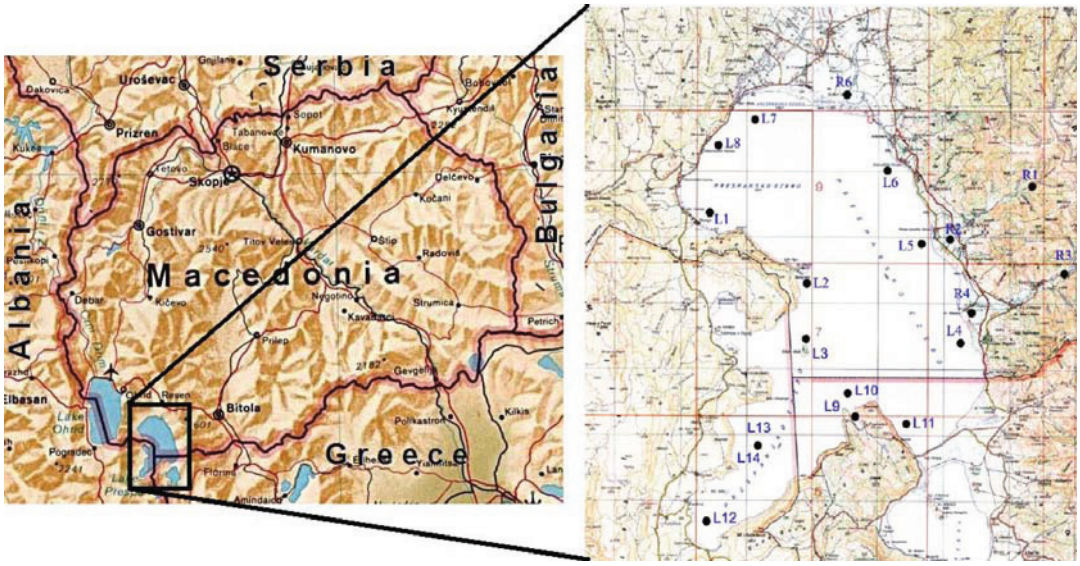

Fig. 2. Position of Lake Prespa (left) and the sampling locations (right).

The datasets used in this paper, as experimental dataset consist from 12 input parameters representing the TOP10 most abundant diatom taxa, with their abundance per sample, plus two trophic state indexes are according to concentration of Total Phosphorus and Secchi Disk. Nevertheless, any water quality or trophic state class could be

used as an input parameter defined by the need of the stake-holders and decision makers. The trophic stet index (eutrophication parameters) is calculated with Carlson's formula [19].

Table 2. Water Quality Classes for the Physical-chemical parameters

Physical-chemical parameters	Name of the WQC	Parameter range
Trophic State Index – Total Phosphorus (TSI_TP)	*Oligotrophic*	TP < 30 - 40
	Mesotrophic	40 – 50
	Eutrophic	50 – 70
	Hypereutrophic	70 - 100
Trophic State Index – Secchi Disk (TSI_SD)	*Oligotrophic*	SD >8m – 4m
	Mesotrophic	4m – 2m
	Eutrophic	2m - 0.5m
	Hypereutrophic	0.5m – 0.25m

4.1. Experimental Setup

We conducted three types of experiments, which are set up follows:

1) A fuzzification method based the novel membership function presented in this paper for each input variable are used to transform the crisp values into fuzzy values and the same subset used as a train-train (**Train**);

2) Two experiments are carried out. The first experiment (Exp2 – odd-even) is using odd; predecessor attributes counting from the first to the dataset, and even; follower attributes from the second attribute in the dataset as a test set. The second experiment (Exp3 – even-odd) is using even; counting from the first attribute in the dataset labeled data as training set an odd labeled data as a test set. This experimental setup is actually 2-fold cross validation analysis. (**Train/Test**)

3) Standard 10-fold cross validation is used for testing of the prediction performance accuracy of the algorithm with the classical classification algorithms (C4.5, kNN, SVM, NBTree, LADTree, etc.). (**Test**)

For similarity definition, we use RMSE similarity and Alegbaric AND and OR for fuzzy aggregation metric. For evaluation purpose, we induce simple aggregation trees (SAT) and general aggregation trees (AT). The simple trees consist from 1 candidate tree, 0(zero) low levels and two different depths; 5-(SAT5) and 10-(SAT10). While general aggregation trees consist from 2 candidate tree, 3 low levels trees and two different depths; 5-(AT5) and 10-(AT10). In the section 5 general aggregation tree which consists from 2 candidate trees, 3 low level trees and depth = 3, are discussed, based on the highest similarity value for each TSI.

5. Aggregation tree models for Lake Prespa

Based on the performance results, in this section we give an interpretation of several model trees and their rules derived from them. We have built many classification model trees for each TSI class, but due to a large number of build trees (almost 60 different model trees) we present one model tree for some of the TSI classes.

All the induced classification models have defined range of fuzzy terms, which later are commented. The number of MFs per attribute is m=5, according to Table 1. All the model trees were obtained using Experimental Setup 2 (Train/Test).

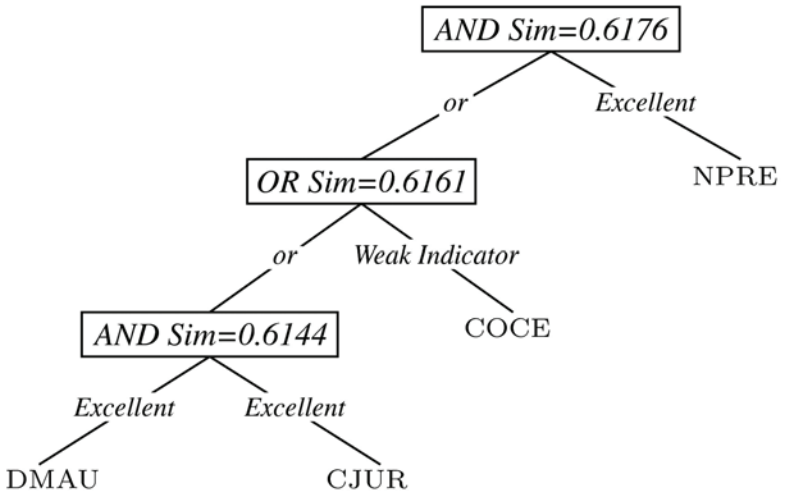

Fig. 3. Aggregation tree generated using proposed sigmoid (+1) MF for the *mesotrophic* class of the TSI_SD.

The classification model represented as aggregation tree for *mesotrophic* class of the TSI_SD according to the model is in correlation with seven diatoms (see Fig. 3). Each condition branch of the tree contains a measure for similarity between the diatoms and the output class. The classification model shown in Fig. 3 can be converted into a rule which is stated with Rule1.

Rule1: If TSI_SD class is *mesotrophic* **THEN** (*Diploneis mauleri* (DMAU) is **Excellent Indicator** AND *Cyclotella juriljii* (CJUR) is **Excellent Indicator**) OR *Cyclotella ocellata* COCE is **Weak Indicator** AND *Navicula prespanense* (NPRE) is **Excellent Indicator**. The rule has confidence of 61.76%.

From Rule1, it can be easily noted that the DMAU, CJUR and NPRE are excellent indicators of *mesotrophic* waters according to the mode tree. The model recognizes the COCE diatom as a weak indicator, or this diatom hardly can be found in these waters.

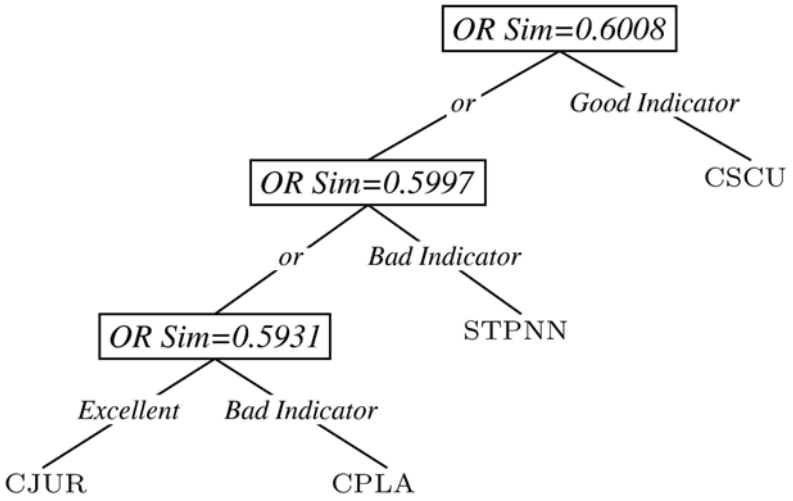

Figure 4. Aggregation tree generated using proposed sigmoid (-1) MF for the *mesotrophic* class of the TSI_SD.

Two more trees are presented in this section, one for the TSI_SD *mesotrophic* class with sigmoid (-1) MF and other one for TSI_TP – *eutrophic* class generated with sigmoid (-1) MF. The rule induced from the tree shown in Fig. 4 states:

Rule2: If TSI_SD class is *mesotrophic* **THEN** (CJUR is **Excellent Indicator** OR *Cocconeis placentula* (CPLA) is **Bad Indicator**) OR *Staurosirella pinnata* (STPNN) is Bad Indicator OR *Cavinula scutelloides* (CSCU) is **Good Indicator**. The rule has confidence of 60.08%.

This rule has a slight lower confidence factor than the Rule1. According to the classification model, CJUR is an excellent indicator of *mesotrophic* waters, which was conferred by the model tree given with Fig. 3. The model identifies the CSCU diatom as a good indicator of such waters. The rest of the four diatoms; STPNN and CPLA taxa cannot exist in such water, thus cannot be used as ecological indicators of mesotrophic waters.

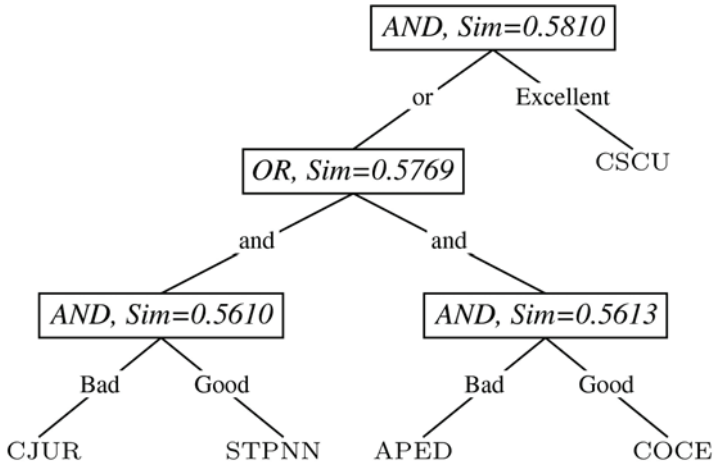

Fig. 5. Aggregation tree generated using proposed sigmoid (-1) MF for the *eutrophic* class of the TSI_TP.

On Fig. 5 the last model for *eutrophic* TSI class is presented. This model was obtained using the proposed algorithm with sigmoid (-1) membership function. A simple rule is derived from the tree, stated below:

Rule3: If TSI_TP class is *eutrophic* **THEN** ((CJUR is **Bad Indicator** OR STPNN is **Good Indicator**) OR (*Amphora pediculus* (APED) is **Bad Indicator** AND COCE is **Good Indicator**) AND CSCU is **Excellent Indicator**. The rule has confidence of 58.10%. The classification model identifies the COCE as a good indicator of eutrophic waters, while the CSCU diatom as an excellent indicator. Other diatoms such as CJUR, STPNN and APED diatoms are not eutrophic taxa according to the presented model.

5.1. Verification of the model results

Namely, out of the 10 top dominant diatoms in Lake Prespa, CJUR and NPRE are newly described taxa (diatom species) with no record for their ecological preferences in the literature. Also, DMAU, NROT and NSROT do not have any ecological reference in the literature. Based on this, the results from the models are the first known ecological reference for TSI classes. The ecological references are given according to latest diatom ecology publications [20] and databases (European Diatom Database - http://craticula.ncl.ac.uk/Eddi/jsp/index.jsp). Based on this, the APED is an eutrophic taxon tolerant to elevated N concentrations, CSCU is also eutrophic taxon-alkalibiont, CPLA is an eutrophic taxon with medium oxygen demand, COCE is a mesotrophic to eutrophic taxon, while STPNN is a hyper-eutrophic (oligo-eutrophic; indifferent) taxon frequently found on moist habitats [4,17,18,21].

If we compare the results from the classification models, we can make several remarks about the results from the models. According to the models, APED diatom is bad indicator of eutrophic waters, based on the TSI_TP, no other

relations have been found with the environment. The CSCU diatom is a good indicator of mesotrophic waters, and excellent indicator of eutrophic water. A CPLA diatom is not mesotrophic taxa, while COCE diatom is a good indicator of mesotrophic and eutrophic waters. This means that a COCE diatom is meso and eutrophic taxon, which is verified by the known ecological reference for this diatom. STPNN diatom according to the classification models presented with Fig. 5, it is a good indicator for eutrophic waters. It is important to notice that the classification models revealed that the CJUR diatom is an excellent indicator of mestrophic diatoms, but bad indicator of eutrophic indicator. DMAU and NPRE diatoms are also excellent indicators of mesotrophic waters. This statements should be tested with more models and data, before any conclusion is made for the newly discover taxa.

According to the models, we have proven that using the proposed method it is possible to extract valuable knowledge from the dataset. We have added several ecological preferences of these TOP10 diatoms for some of the TSI classes. Although the prediction model should be further improved by prediction accuracy, the proposed method and ecological preference have been found to be relevant for such a task. The proposed method confirmed some of the diatoms ecological preference, some of them need more work, and for the unknown diatoms we have added some new ecological knowledge.

Table 3. The average prediction accuracy (in %) per TSI for each fuzzy MF.

TSI according Secchi Disk (TSI_SD) or total phosphorus concentration (TSI_TP)		Triangular	Trapezoidal	Gaussian	Evenly Sigmoid (+1)	Evenly Sigmoid (-1)
TSI_SD	Train	84.41	84.21	84.34	84.67	**84.80**
TSI_SD	Exp2	38.42	83.42	**83.82**	83.16	80.79
TSI_SD	Exp3	38.42	84.61	84.74	**86.32**	**86.32**
TSI_TP	Train	45.58	46.16	**49.48**	45.70	41.00
TSI_TP	Exp2	38.42	36.58	41.51	**41.55**	38.53
TSI_TP	Exp3	38.42	36.12	37.04	**39.79**	36.93

6. Performance evaluation

The proposed evenly sigmoid shaped MFs outperformed 2 of the 3 diatoms TSI classes compared with other MFs. In Table 3 we present the highest prediction accuracy of proposed method for extracting knowledge from diatom data over different combinations of training-test sets compared with the membership functions in [15]. The evenly distributed sigmoid (-1) membership function has obtained higher prediction accuracy than the other MF in train experiment and experiment 3. The evenly sigmoid (+1) and sigmoid (-1) MF vs. other MF in experiment 2 have achieved lower prediction accuracy for the TSI_SD.

The evaluation performance's analysis in details for a different number of MF for the TSI_TP is given in Table 3. For experiments 2 and 3, the proposed method with evenly sigmoid membership function has achieved greater prediction accuracy.

6.1. Comparison with crisp classifiers

In order to improve the classification accuracy and maintaining the robustness of the data change which comes by using fuzzyfication of the input data, we use the aggregate trees to extract knowledge directly from the dataset. Most of the classic decision trees – classification algorithms, produce very strict interpretability of acquired knowledge

from the data. Also, these algorithms are not very robust on data change, which is not the case with the proposed method tree.

The experimental results confirm these findings, by comparing the AT used for the diatom's dataset with other algorithms. The results are presented in Table 4. The results were obtained with a number of MF equal to 5. Using the different change of the parameter {1 and -1} we evaluate the performance compared with classical crisp classifiers for evenly distributed sigmoid membership function. The classification algorithm performance, were tested by building four variants of the proposed method.

Table 4. 10-fold cross validation classification accuracy (in %) of crisp classification algorithms against proposed evenly distributed sigmoid function for L =2, M=3 and depth=3.

TSI according Secchi Disk (TSI_SD) or total phosphorus concentration (TSI_TP)		C 4.5	kNN	Bagging C4.5	REP Tree
TSI_SD	xVal-1	83.16	73.16	83.16	83.16
		SAT5	SAT10	AT5	AT10
TSI_SD	Evenly sigmoid (+1)	83.60	83.60	83.07	83.07
TSI_SD	Evenly sigmoid (-1)	**84.12**	**84.12**	**84.12**	**84.12**
TSI_TP	xVal-1	39.91	39.45	41.28	41.74
		SAT5	SAT10	AT5	AT10
TSI_TP	Evenly sigmoid (+1)	39.68	39.72	38.72	41.08
TSI_TP	Evenly sigmoid (-1)	39.18	39.70	**43.42**	39.72

Most of the cases we have 1% to 3% increase of prediction power, and the prediction accuracy of the aggregated trees increases for the both trophic state index classes. The proposed method has been proven to be excellent data mining technique for knowledge extraction for diatoms-indicator relationship with high classification accuracy.

6.2. Over-fitting comparison with C4.5, KNN, SVM, Naive BayesNet, REPTRee, NB-Tree and LAD-Tree

Over-fitting refers to the phenomena that a classifier may fit well to the training data but is not generalized enough to classify unseen data. The 10-fold cross validation based experiments fairly present the normal behavior of classifiers, but it does not reveal which classifiers are prone to over-fitting.

In this section, the whole data (rather than the 10-fold cross validation data) are used to train and test all the classical classification algorithms and four variants of method using the same experimental setup. The results collectively are shown in Table 5.

The classical classifiers (C4.5, kNN, SVM, NB, REPTree, NBTree, LADTree) obtained from the Weka machine learning toolkit [22] from a crisp classifier group. The default settings of each classifier in the toolkit are used. For example, the minimal number of instances per leaf is set to 2 for C4.5 and the number of neighbours to use is set to 1 for KNN. In REP Trees number of Boosting Interaction are set to 15, while the LADTree the number of Boosting Interaction is set to 10. For each classifier, the root mean square error (RMSE) of the classification accuracy between the 10-fold cross-validation and whole data based is shown at the bottom of the table.

Table 5. Whole data based classification accuracy (in %) of C4.5, kNN, SVM, NB, REPTRee, NBTree, LADTree and four variants of AT over six datasets.

Sigmoid (+1)/(-1)		C4.5	kNN	SVM	NB	REP Tree	NB Tree	LAD Tree	SAT 5	SAT10	AT5	AT10
(+1)	TSI_SD	83.16	73.16	84.21	56.84	82.11	84.21	83.16	83.60	83.60	83.07	83.07
(+1)	TSI_TP	39.91	39.45	40.37	28.90	41.74	33.49	41.74	36.08	35.58	37.40	37.38
(+1)	RMSE	27.98	40.32	3.57	6.50	83.16	15.93	16.27	3.70	4.70	7.01	7.35
(- 1)	TSI_SD	83.16	73.16	84.21	56.84	82.11	84.21	83.16	84.12	84.12	83.07	83.60
(- 1)	TSI_TP	39.91	39.45	40.37	28.90	41.74	33.49	41.74	39.13	38.70	40.04	40.06
(- 1)	RMSE	27.98	40.32	3.57	6.50	83.16	15.93	16.27	2.82	2.80	**1.19**	4.45

The RMSE reveals how much improvement one classifier can gain based on the whole data experiment comparing to the 10-fold cross validation one. It is assumed that the more gain for one classifier, the more likely that the classifier is prone to over-fitting. The results of this test revealed that SVM maintains the best generality compared with the classical approaches (with RMSE being 3.57) and four variants of aggregation trees perform slightly worse (with RMSE being 3.70, 4.70, 7.01 and 7.35 respectively for sigmoid (+1), while RMSE begins from 1.19 to 4.45, gain much better performance for the sigmoid (-1) MF. Most important is the fact, that the four variants of aggregation trees have maintained the interpretability, which is the property of the C4.5 and also remain resistant to over-fitting, using sigmoid (-1) membership function.

Other crisp classifiers perform worse than SVM and variants of proposed method, but better than C4.5. It is not surprising that KNN performs the worst as it is totally biased to the nearest neighbour in classification and makes no attempt to find a general model. Even complex aggregation trees do not suffer from over-fitting.

7. Conclusion

Classifying the lake ecosystem using diatoms, can be greatly improved with the proposed method, not just for Lake Prespa, but for any lake ecosystem. The current involvement of the information technology in solving Lake Prespa and its inflow rivers ecological problems through environmental management are low. Keeping this in mind, the proposed method in this paper, further improves the methods used for such environmental management and compared with the previous used have been several advantaged. Not just improving the interpretability of the gain models, to make easy interpretably for biological experts, but producing models, fast and more accurate results.

The experiments on diatom dataset TSI dataset show that the two modified sigmoid MFs for aggregation trees outperformed previously used MFs in terms of prediction accuracy. This is very important for different types of datasets. In our case, the diatoms have very tight value range over the physical-chemical parameters, and if we want to define the abundance range of the diatoms, we have to increase the number of MFs per attribute. The mixed datasets odd-even and even-odd performed better, which means that the generalization of the proposed method is greater. 10-fold cross validation used to compare the performance of this algorithm with crisp algorithms, proof that we develop a membership function distribution which outperformed classical classification algorithms in terms of prediction power and maintained resistance of the over-fitting.

More important is the interpretation of the proposed method, outperforms the classical statistical methods such as: PCA, CCA, DCA and other methods, used previously. The obtained models have clearly stated prediction in terms of finding correct diatom-indicator relationship. For example, model tree presented with Fig. 5 for eutrophic TSI_TP using proposed evenly sigmoid (-1) MF for the clearly states that the CSCU diatom can be an indicator of these waters. Nevertheless, many of the models produce rules that include relationship between several diatoms at once with the physical-chemical parameters. The experiments showed that machine learning tools can extract some

valuable knowledge in a relatively comprehensible form, even when the application area is so extremely complex also for humans and the data are far from being perfect. In order to produce a precise prediction model mainly depends from the selection of the relevant forecasting attributes, which is driven by training data. In case were we use method with fuzzy set theory, we have tried to decrease the chance of any incorrectness or irrelevance in the data can distort the results.

From ecological point of view, it is very important that the proposed algorithm is resistant to data change, which in this case is true and we have also added several ecological references for the unknown diatoms. Data change is a property of any system due to the changeable environmental condition. We believe that studies like ours that combines the ecological, hydro-biological, together with information technologies, especially in the area of environmental informatics, are necessary to provide understanding of the physical, chemical and biological processes and their relationship to aquatic biota for predicting a certain effect. Using decision system support system with such implemented algorithms we can increase the chance to keep the ecosystem healthy and the organism survival at a high rate.

Further research needs focus on developing more MF in a process of building aggregation trees is necessary. More similarity metrics may be more suitable for this diatom community dataset and can therefore, lead to higher accuracy. In future we plan to test more datasets with a greater number of input parameters and use of weighted approach on the classification problem of the diatom's community.

References

[1] Stroemer EF, Smol JP. *The diatoms: Applications for the environmental and earth sciences*. Cambridge University Press, Cambridge. 2004; 192–198.

[2] Džeroski S, Mitreski K, Krstić S, Naumoski A. Constructing habitat models for diatoms in Lake Prespa using machine learning method of regression trees. In: *Proceedings of the 6th European conference on Ecological Modelling*, ECEM '07. Trieste, Italy: Challenges for ecological modelling in a changing world: global changes, sustainability and ecosystem based management: conference proceedings, [S.l.: s.n.]. 2007; 149–150.

[3] Naumoski A, Kocev D, Atanasova N, Mitreski K, Krstić S, Džeroski S. Predicting chemical parameters of water quality form diatoms abundance in lake Prespa and its tributaries. In: *4th International ICSC Symposium on Information Technologies in Environmental Engineering - ITEE* 2009. Thessaloniki, Greece, Springer Berlin Heidelberg press. 2009; 264–277, doi: 10.1007/978-3-540-88351-7.

[4] Kocev D, Naumoski A, Mitreski K, Krstić S, Džeroski S. Learning habitat models for the diatom community in Lake Prespa. Journal of *Ecol Model* 2010; **221**(2): 330–337.

[5] Wang LX, Mendel JM. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics* 1992; **22**(6): 1414–1427.

[6] Quinlan RJ. Decision trees and decision making. *IEEE Transactions on Systems, Man, and Cybernetics* 1990; **20**(2): 339–346.

[7] Yuan Y, Shaw MJ. Induction of fuzzy decision trees. *Fuzzy Sets and Systems* 1995; **69**(2): 125–139.

[8] Olaru C, Wehenkel L. A complete fuzzy decision tree technique. *Fuzzy Sets and Systems* 2003; 138: 221–254.

[9] Suárez A, Lutsko JF. Globally optimal fuzzy decision trees for classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1999; **21**(12): 1297–1311.

[10] Wang X, Chen B, Olan G, Ye F. On the optimization of fuzzy decision trees. *Fuzzy Sets and Systems* 2000; **112**: 117–125.

[11] Schweizer B, Sklar A. Associative functions and abstract semigroups. *Publication Mathematica Debrecen* 1963; **10**: 69–81.

[12] Kóczy LT, Vámos T, Biró G. Fuzzy signatures. *EUROFUSE-SIC* 1999; 210–217.

[13] Mendis BSU, Gedeon TD, Kóc LT. Investigation of aggregation in fuzzy signatures. In: *CD Proceeding of the 3rd International Conference on Computational Intelligence*, Robotics and Autonomous Systems, Singapore; 2005.

[14] Nikravesh M. Soft computing for perception-based decision processing and analysis: Web-based BISC-DSS. *Studies in Fuzziness and Soft Computing* 2005; **164**: 93–188.

[15] Huang ZH, Gedeon TD, Nikravesh M. Pattern Trees Induction: A New Machine Learning Method. *IEEE Transaction on Fuzzy Systems* 2008; **16**(3), 958–970.

[16] Peter KE, Radko M, Endre P. Triangular Norms. *Dordrecht: Kluwer*. 2000; 78–98.

[17] Krstić S. Description of sampling sites. *FP6-project TRABOREMA: Deliverable 2.2*; 2005.

[18] Levkov Z, Krstić S, Metzeltin D, Nakov T. Diatoms of Lakes Prespa and Ohrid (Macedonia). *Iconographia Diatomologica* 2006; **16**: 603.

[19] Carlson RE, Simpson J. A Coordinator's Guide to Volunteer Lake Monitoring Methods. *North American Lake Management Society* 1996; 96.

[20] Van Dam H, Mertens A, Sinkeldam J. A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. *Netherlands J Aq Ecol* 1994; **28**(1), 117–133.

[21] Džeroski S, Mitreski K, Krstić S, Naumoski A. Learning habitat models for the diatoms of Lake Prespa. In: *Proceedings of the 7-th National Conference with international participation ETAI* 2007. Ohrid, Macedonia: [S.l.]: Society for Electronics, Telecommunications, Automatics and Informatics of the Republic of Macedonia; 2007.

[22] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). *Morgan Kaufmann*; 2005.