ALGORITHM WITH EVENLY DISTRIBUTED GAUSSIAN FUNCTION FOR DIATOM CLASSIFICATION

Andreja Naumoski
Ss. Cyril and Methodius University,
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia

Kosta Mitreski
Ss. Cyril and Methodius University,
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia

ABSTRACT

The diatom organisms are good bio-indicators of certain ecosystem environments. According the national directive for water quality classification, each WQC represent a water quantity of certain physico-chemical parameters in certain range define by biological experts. The property of bio-indicator is used to characterize the environment and thus helping in process of classification of the diatoms in the correct water quality classes (WQCs). In this direction we use pattern trees; trees which have combined the advantages of the information theory and fuzzy theory to model (predict) in which WQC belongs the certain diatom. Because many of the newly discover diatoms does not have ecological preference, this algorithm significantly improves the process of fast and accuracy classification. In our approach we divide each diatom into three evenly ranges with Gaussian functions, which will be represented with fuzzy terms (low, medium and high) similar as the WQC range classes. Using this data mining techniques we can closely reflect the very nature of the diatoms dataset, which later the experiments will confirms this assumption, by taking into account the mean and the standard deviation of each diatom range. The experimental results have shown that the extract knowledge has high level of confidence factor in many cases and the trees obtained have high accuracy compared with other classification algorithms. As future work we intend to expand the number of fuzzy membership and inspect their influence, to implement more fuzzy aggregation functions and similarity definitions in process of pattern trees.

I. INTRODUCTION

The water quality classes define in the traditional way can be interpreted as classification problem in the terms of data mining point of view. This property is used for finding the proper diatom (organism)-environment relationship, which has been a subject of eco-informatics area of research very recently. Considering this, we deal with the typical classification problem, when we try to find the correct organism - environment relationship. 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 [18]. Although these techniques provide useful insights in the data, they are limited in terms of interpretability. In order to improve the problem of interpretability we use decision trees. As sub-class of decision trees, the fuzzy and pattern trees have several advantages over classical decision trees. Pattern trees are resistant to over-fitting and robust to dataset change. This is the main reason of extensive research on fuzzy set based machine learning. Wang and Mendel [12] have presented an algorithm for generating fuzzy rules by learning from examples. Inspired by the classic decision tree induction by Quinlan [10], there are substantial works on fuzzy decision trees. For example, Yuan and Shaw [15] have proposed fuzzy decision trees induction using fuzzy entropy. Janikow [11], Olaru and Wehenkel [8] have presented different fuzzy decision tree inductions. Suárez and Lutsko [14], and Wang and Chen, et al. [13] have presented optimizations of fuzzy decision trees. Most of the existing fuzzy rule induction methods including fuzzy decision trees [9] focus on searching for rules which only use t-norm operators [7] such as the MIN and algebraic MIN. Research has been conducted to resolve this problem. Kóczy, Vámos and Biró [3] 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 [4] have investigated different aggregations in fuzzy signatures. Nikravesh [5] has presented evolutionary computation (EC) based multiple aggregator fuzzy decision trees. As successor of the benefits from the fuzzy decision trees, the pattern trees can obtain high accuracy, robustness of over-fitting and etc.

Huang and Gedeon [1] have first introduced the concept of pattern trees and proposed a novel pattern tree induction method by means of similarity measures and different fuzzy aggregations. In their algorithm, they use simple evenly distributed trapezoidal, triangular and Gaussian membership function. In this paper we will use Gaussian evenly distributed membership function, to obtain correct diatoms-WQC relationship and to predict the quantity of diatoms in measured sample for each WQC. The quantities of diatoms are express through fuzzy terms. At the end of the paper extensive experiment evaluation over the diatom community datasets is made. From this dataset using pattern trees we have extracted valuable knowledge.

The rest of the paper is organized as follows: Section II provides the definitions for similarity, aggregations and pattern trees. In Section III the process of WQ classification with pattern trees for the diatoms is presented. Section IV describes the diatoms abundance water quality datasets and the experimental setup. Section V gives the experimental results and performance evaluation. Finally, Section VI concludes the paper and outlines the research direction.

II. DEFINITIONS FOR SIMILARITY AND FUZZY AGGREGATION METRICS

The pattern tree induction method is composed by using different similarity measures and fuzzy aggregation, which

©2010 Institute of Informatics.

are presented in this section. The process induction of pattern trees in great details by authors in explained in [1].

A. Similarity definitions

Let A and B be two fuzzy sets [11] defined on the universe of discourse U. The root mean square error (RMSE) of fuzzy sets A and B can be computed as:

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

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

$$\mathrm{Sim}(A;B)=1-\mathrm{RMSE}(A;B)\ . \qquad (2)$$

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], $0 \leq \mathrm{Sim}(A;B) \leq 1$ holds according to (1) and (2). Another similarity definition is given by equation (3) and is known as Jaccard similarity measure.

$$Jaccard\ similarity\ measure=\frac{A\cap B}{A\cup B}\ . \qquad (3)$$

Note that the pattern tree induction follows the same principle if alternative fuzzy set similarity definitions such as Jaccard or any other are used.

Table 1: Basic T-norms and T-conorms

Name	T-Norm	T-Conorm
MIN/MAX	$\mathrm{Min}\,\{a,b\} = a\wedge b$	$\mathrm{Max}\,\{a,b\} = a\vee b$
Algebaric AND/OR	ab	$a+b-ab$
Lukasiewicz	$\mathrm{Max}\,\{a+b-1,0\}$	$\mathrm{Min}\,\{a+b,1\}$
Einstein	$\frac{ab}{2-(a+b-ab)}$	$\frac{a+b}{1+ab}$
MIN/MAX	$\mathrm{Min}\,\{a,b\} = a\wedge b$	$\mathrm{Max}\,\{a,b\} = a\vee b$

B. Fuzzy aggregation definitions

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 only the basic operators which operate on two fuzzy membership values *a* and *b*, where *a*, *b* ∈ [0, 1] shown in Table 1 (Algebraic AND/OR, without WA and OWA).

In [1] the authors combine for the first time, all the three sub-categories of fuzzy aggregation. We will use this combination to improve the knowledge extraction procedure form the diatoms dataset. As can be seen, a pattern tree can be generated using different fuzzy aggregation functions.

III. Water quality classification with Pattern trees

As we pointed earlier before, the water quality class is in fact a classification problem which can be represented with fuzzy membership function. In this direction we will transform the crisp values into fuzzy values and then assign certain membership name to that particular range.

The results of the process are presented in Table 2. First we divide the data into two groups, but maintaining into single file, the TOP10 diatoms abundance data and three water quality classes from measured SatO, pH and Conductivity parameters. Then using automatic procedure each diatom is divide into three evenly ranges, which will be represented with fuzzy membership functions and names like (low, medium and high) shown in Table 2.

We use evenly distributed Gaussian membership functions (Eq. 4), which have mean and standard deviation and this two parameters are closely related to the property of the data. The equation is as follows:

$$f(x)=e^{\frac{-(x-\mu)^2}{2\sigma^2}}\ . \qquad (4)$$

, where μ is the mean value of the fuzzy membership function, and σ is the standard deviation. In order to achieve complete evenness of the fuzzy terms, we need to calculate the standard deviation, so that the intersection of the two Gaussian functions is 0.5.

$$\sigma=\sqrt{\frac{range^2}{8*\ln 2}}\ . \qquad (5)$$

By applying simple mathematics the standard deviation is given by the eq. 5. Using this data mining techniques then we learn pattern trees which can predict the outcome of the WQC from the data based on the particular diatom found in the tree.

IV. Dataset 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 [2]) with their abundance per sample, plus the three water quality classes for conductivity, pH and Saturated Oxygen. These measurements were made as a part of the TRABOREMA project [6]. The water quality classes are defined according the three physical-chemical parameters: Saturated Oxygen [16], Conductivity [17] and pH [16, 17] which are given in Table 3. Among the input parameters, 10 are numerical parameters and the rest 3 are nominal with number of possible classes from 4 up to 6.

In this work we induce simple and general pattern tree, which consists from 1 and 2 candidate trees, 0 and 3 for low level trees and depth = 5 (SPT5, PT5) and 10 (SPT10, PT10) for simple (SPT) and general pattern trees (PT), respectively. For similarity definition we use Jaccard and RMSE similarity and only AND and OR for fuzzy aggregation procedure. We set the number of evenly distributed Gaussian membership function to three (M=3). Then these results are compared with WEKA [19] crisp classifier group; C4.5, KNN, BayeNet, Bagging-C4.5, Boosting-C4.5 and MultiBoost-C4.5.

Table 2: Water quality classes for the physical-chemical parameters according [16, 17]

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

The configuration of the experiments is set up as fallows. 1) The entire dataset is used for training set, as a part of the training procedure of the algorithm and 2) Standard 10-fold cross validation is used for testing the prediction performance accuracy of the algorithm. Table 4 and Table 5 shows results of different experiments applied on the diatoms water quality classification dataset. And at the end of paper, we will compare the prediction accuracy of the pattern trees using 10-fold cross validation against standard crisp classifiers.

V. Pattern trees interpretation

In this section we present several important trees produced from the algorithm. Due extensive number of build tree and paper constrains we present several tree, one for each water quality class with highest similarity factor. Every tree can be transform into rule in several easy steps, which is done for each tree.

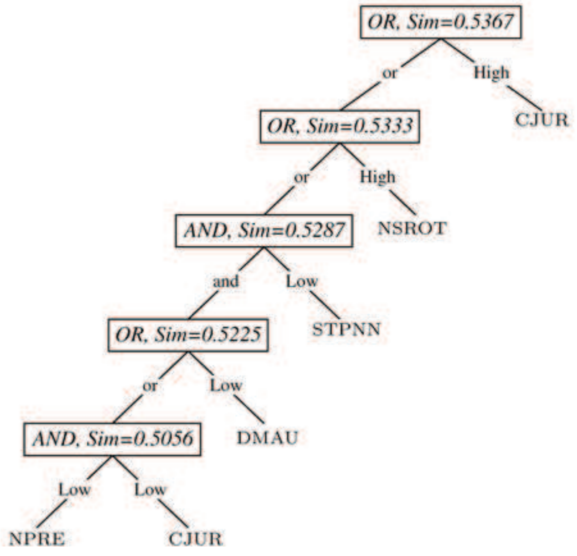

Figure 1: Pattern tree for Conductivity WQC – *fresh*

The pattern tree shown on Fig.1 clearly indicates that high abundance of the CJUR or NSROT indicates according the model tree that these diatoms can exist in *brackish* waters.

The diatoms that are referred with Low abundance on the right side of the tree, the model tree predict that these diatoms (NPRE, DMAU and STPNN) cannot exist in water with Conductivity WQC – fresh. Nevertheless, many of the diatoms predicted with this tree model, have moderate level of confidence factor.

This tree has highest similarity between the classes of 53.67 %. This tree is transformed into rule, as it has been shown below – **Rule1** for easy interpretation. The complete names of the Lake Prespa diatoms can be found in [2].

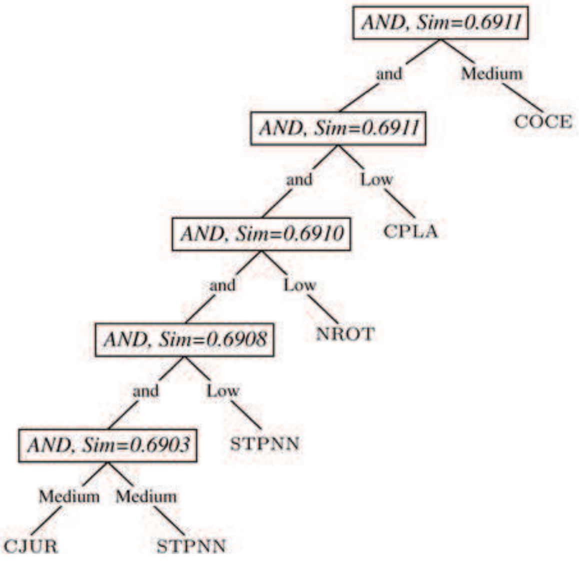

Figure 2: Pattern tree for pH WQC – *alkaliphilous*

Rule1: If (CJUR is Low and NPRE is Low) or DMAU is Low and STPNN is Low or NSROT is High or CJUR is High then the class is *fresh* (with confidence of 0.5367).

Rule2: If (CJUR is Medium and STPNN is Medium) and STPNN is Low and NROT is Low and CPLA is Low and COCE is Medium then the class is *Class4* (with confidence of 0.6911).

The tree shown on the Fig. 2, predicts the existing diatoms using pH WQC, and then is transform into **- Rule2.** This tree has high level of 69.11 % of similarity between the fuzzy terms and compared with the previous one has higher similarity.

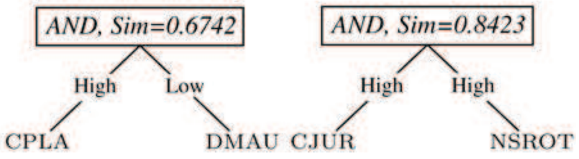

Figure 3: Pattern tree for SatO WQC – is *α-meso / polysaprobous* and *polysaprobous*

The medium abundance of COCE and CJUR diatoms indicate according the tree model that this diatoms can be found in waters were the pH WQC is *alkaliphilous* according the model tree.

Rule3-1: If (CPLA is High and DMAU is Low) then the class is *α-meso / polysaprobous* (with confidence of 0.6742).

Rule3-2: If (CJUR is High and NSROT is High) then the class is *polysaprobous* (with confidence of 0.8423).

According the tree for the Saturated Oxygen WQC (see Fig. 3), high abundance of CPLA and indicates that this diatom can exist in waters were the Saturated Oxygen level is *α-meso / polysaprobous*. The model tree predicts that the high/low abundance for the CJUR and NSROT diatoms exist in *polysaprobous* waters.

Table 3: 10 fold cross validation classification accuracy of classical crisp classifiers and four variants of PT (in %)

DataSet	C4.5	kNN	Bayse Net	Bagging C4.5	Boosted C4.5	MultiBoost C4.5	SPT5	SPT10	PT5	PT10
Conductivity 10-cross xVal-J	65.60	66.51	68.81	63.30○	63.76	69.72	69.50	69.50	70.45●	70.45●
Conductivity 10-cross xVal-R	65.60	66.51	68.81	63.30○	63.76	69.72●	68.16	68.64	69.07	68.14
Saturate Ox. 10-cross xVal-J	54.73	47.26○	58.71●	53.23	56.22	55.72	56.00	56.00	57.50●	57.50●
Saturate Ox. 10-cross xVal-R	54.73	47.26○	58.71●	53.23	56.22	55.72	54.50	54.50	53.00	55.00
pH 10-cross xVal-J	55.50	46.33○	61.47●	56.42	49.54	57.40	56.71	57.16	57.16	57.62●
pH 10-cross xVal-R	55.50	46.33○	61.47●	56.42	49.54	57.40	57.62●	57.16	56.73	56.28

●, ○ statistically significant improvement or degradation

A. Experimental performance evaluation

Most of the classic decision trees – classification algorithms, produce very strict interpretability of acquired knowledge from the measurements. Also, these algorithms are not very robust on data change, which is not the case with the pattern trees. In order to improve the classification accuracy and interpretability of the results, we use PT tree which are also resistant to overfitting.

Table 4: Classification accuracy per WQC, for M=3 using Gaussian distribution

Conductivity WQC –Prediction Accuracy (in %)				
	SPT5	SPT10	PT5	PT10
Train-J	70.18%	70.18%	72.02%	72.02%
xVal-J	69.50%	69.50%	70.45%	70.45%
Train-R	71.10%	71.10%	73.39%	73.39%
pH WQC – Prediction Accuracy (in %)				
	SPT5	SPT10	PT5	PT10
Train-J	57.80%	58.26%	59.17%	59.17%
xVal-J	56.71%	57.16%	57.16%	57.62%
Train-R	61.01%	59.63%	60.55%	59.63%
Saturated Oxygen - Prediction Accuracy (in %)				
	SPT5	SPT10	PT5	PT10
Train-J	58.71%	58.71%	58.71%	58.71%
xVal-J	56.00%	56.00%	57.50%	57.50%
Train-R	59.70%	59.70%	59.70%	59.70%

Using the pattern trees, we have induced several tree which according the performance analysis have highest confidence (similarity) factor. Beside the confidence factor the classification accuracy of each tree is obtain and presented in Table 4 and evaluation with other algorithms in Table 5. Train-J and xVal-J are acronyms for train and test procedure evaluated with Jaccard similarity measure and Train-R and xVAl-R are represented for RMSE similarity measure.

The experimental results confirm these findings, by comparing the PT used for the diatoms dataset with other algorithms, whose results are presented in Table 4.

We have tested the calssification algorithm performance, by building simple and general pattern trees, techniques describe in detail by authors in [1]. The results are given in Table 3.

VI. Conclusion

In this paper we present a work which takes the diatoms property as bio-indicator and with classification algorithms: pattern trees we have extracted knowledge that predicts the quantity of diatoms that belongs to certain water quality class. The extracted knowledge is with satisfied classificatory accuracy and similarity between the classes. According the performance table given in the paper, the highest accuracy is achieved using PT5 and PT10 for conductivity and SatO, while pH WQC highest accuracy was achieved with Byes Net (train procedure).

Regarding the similarity factor, the SatO WQC - *α-mesosaprobous* has highest value. Nevertheless, other produced rules do not downgrade the importance of the proposed method. Low quality of data is one of the main reasons for low classification accuracy and the main reason to use pattern trees. Overall conclusion about the experiment performance is that the conductivity WQC has the best classification accuracy using evenly Gaussian membership function. In fact many of the pattern trees, such as the tree presented with Fig. 3 clearly indicate that SatO WQC can be indicated with high abundance of CPLA. Some of them involve several diatoms, which completely consistent with the ecological understanding of the complex ecosystem interaction. Nevertheless, other pattern trees indicate that they can be used for extracting knowledge from diatoms data with certain confidence factor.

Many of the produced rules can be validate with the knowledge of the biological expert, but in many cases new diatoms are discovered and their ecological preference is unknown. Presented classification algorithm in this paper, leads to improvement of the process of faster classification of newly discover diatoms.

In this paper we classify only the three WQC, but nevertheless each important physical-chemical parameter can be represented with quality class. In this direction, the presented algorithm can be used to predict the influence of each physical-chemical parameter, not just for WQC presented in this paper.

Advantages of this proposed approach, is the fuzzy value (low, medium, high) in certain range assign to each measurement, indentifying the bio-indicator (diatom) for certain WQC, not dealing with exact values.

In future work, we plan to investigate more diatoms and improve the classification accuracy by implementing more fuzzy membership functions and similarity definitions.

REFERENCES

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

[2] Z. Levkov, S. Krstič, D. Metzeltin, and T. Nakov, “Diatoms of Lakes Prespa and Ohrid (Macedonia),” Iconographia Diatomologica, Vol. 16, pp. 603, 2006.

[3] L. T. Kóczy, T. Vámos, and G. Biró, “Fuzzy signatures,” EUROFUSE-SIC, pp. 210 – 217, 1999.

[4] B. S. U. Mendis, T. D. Gedeon, and L. T. Kóczy, “Investigation of aggregation in fuzzy signatures,” 3rd International Conference on Computational Intelligence, Robotics and Autonomous Systems, Singapore, Vol. CD-ROM, 2005.

[5] 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, Springer Berlin/Heidelberg, 2005.

[6] “TRABOREMA Project” WP3, EC FP6-INCO project no. INCO-CT-2004-509177, 2005-2007

[7] B. Schweizer, and A. Sklar, “Associative functions and abstract semigroups,” Publ. Math. Debrecen, vol. 10, pp. 69 – 81, 1963.

[8] C. Olaru, and L. Wehenkel, “A complete fuzzy decision tree technique,” Fuzzy Sets and Systems, Vol. 138, pp. 221–254, 2003.

[9] Y. Yuan, and M. J. Shaw, “Induction of fuzzy decision trees,” Fuzzy Sets and Systems, Vol. 69, no. 2, pp. 125 – 139, 1995.

[10] J. R. Quinlan, “Decision trees and decision making,” IEEE Transactions on Systems, Man, and Cybernetics, Vol. 20, no. 2, pp.339 – 346, 1990.

[11] C. Z. Janikow, “Fuzzy decision trees: issues and methods,” IEEE Transactions on Systems, Man, and Cybernetics, Vol. 28, No. 1, pp.1–14, 1998.

[12] 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.

[13] 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.

[14] A. Suárez, and J. F. Lutsko, “Globally optimal fuzzy decision trees for classification and regression,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, no.12, pp. 1297–1311, Dec 1999.

[15] Y. Yuan, and M. J. Shaw, “Induction of fuzzy decision trees,” Fuzzy Sets and Systems, Vol. 69, no. 2, pp. 125 – 139, 1995.

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

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

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

[19] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann, 2005.