TROPHIC DIATOM CLASSIFICATION USING NAÏVE BAYES

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

Knowledge discovery has been used in many different type of analysis and data types which lead to increased understanding of many natural processes and phenomena. This is why this process is important in the area of analysing environmental data. The topic and the goal of the paper is to used this process and the information contained in the measured data for given lake ecosystem and extract that information in an understandable form. This research aims to assess the relationships between the diatoms and the indicators of the environment using Naïve Bayes method learning technique. The diatoms are taken into account because they are ideal indicators of certain physical-chemical parameters and they can be classified into one of the trophic quality classes (TQCs). Before the algorithm processes the data, the input dataset is discretised. Then using the Naïve Bayes technique, several models for each TQC are obtained, presented and discussed. Then the obtained knowledge is verified with existing diatom ecological preference. Directions of future research and improvement for using this method for environmental data are given at the conclusion of the paper.

I. INTRODUCTION

The analysis of the environmental data can be conducted in many ways. Based on the needs, the data can be plotted, transformed on different axes, processed or build in form of models for regression or classification processes. In order to conduct the classification process, we need defined classes in which the diatoms will belong or in terms of ecology will indicate. Because in the literature there are known water classification systems based on several physico-chemical parameters and the diatoms are ideal indicators for some of these parameters, we can used them to classify the diatoms into one of these trophic quality classes. The TQCs are based on specific group of physico-chemical parameters responsible for certain processes like eutrophication is known as trophic state indexes or classes. Such parameters are concentration of total phosphorus, total nitrogen or secchi disk, which are vital in the living process of the organism in the lake ecosystem. [1]. Based on these facts, the classification process is straightforward if we consider the input data is divided into input data, which is desecrated and labelled for better interpretation of the data and the output part of the data - TQC The discretization of the input dataset is made for another reason. There is evidence that the method produces poor probability estimates [2, 3] on continues data with values near zero and using other techniques this can be eliminated.

Many research papers that are dealing with the discovering of the indicating properties of the diatoms, are using 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 [4]. While these techniques provide useful insights in the data, they are limited in terms of interpretability and the results are plotted on graph where the biological expert should understand and should give interpretation of the distances between groups and clusters of diatoms from them self or axes. That's why the direction of research to increase interpretability is moved toward more graphical and easy understanding inducing methods such as decision trees [5]. Using these methods and their variants not only improves the interpretability, but also we attempt to increase the classification accuracy of the models. Several attempts to model the diatom indicator relationship is made by [5, 6, 7, 8]. Beside the increased accuracy of the models, the obtained knowledge from these models were positively verified with the known literature for many diatoms. Furthermore, advancement of these methods is made by introducing new class of multi-target decision trees, in order to understand the dynamic nature of whole range of physico-chemical parameters of the ecosystem [8]. Even these methods were more precise and increased the interpretability; they were not robust of the dynamic nature of the ecological measurements. This important property must be somehow imported in the processes of classification.

In this direction, investigation other methods, such as the Naïve Bayes method could help in this ecological quest for better knowledge discovery algorithm. In the literature, there is a lot of research papers that point out that the Naïve Bayes and the decision trees (C4.5 [9]) that performs equally well as the C4.5 method [10, 11, 12] for many real dataset domains. This good performance of the method is sometimes surprising because many of the real world applications don't always satisfy the condition that for given class value all the attributes are independent. In this way, if we apply the Naïve Bayes method for diatom classification we assuming that the influence of the physico-chemical parameter on different diatoms is independent. This question should be further investigated, since in the lake ecosystem one diatom can be indicator of not only one, but several parameters and some of the organisms are competitive between them [1]. Further evidence on application of the Naïve Bayes technique for diatom classification comes from the Domingos and Pazzani [13] research, where they have shown that the Naïve Bayes method owes its good performance to the zero-one loss function. Another important property of the Naïve Bayes is that this algorithm have shown better results for classification rather then regression, according [14]. Nonetheless, from ecological point it is important to estimate the degree of membership for given environmental condition. To best our

knowledge this papers work is first for diatom classification based on trophic parameters of the lake ecosystem.

The organisation of the paper is at is follows: Section II provides the definitions of the used method. In Section III we present the description of the input dataset and the trophic quality datasets as well as the experimental setup. Section IV gives the experimental results and the verification of the model results and finally, Section V concludes the paper and presents direction for future research.

II. NAÏVE BAYES CLASSIFIER DEFINITION

The research work presented in this paper is represents classification problem that in the data mining terminology is defined as fallows. The goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. According to the standard definition of classification the example E is represented by a tuple of attribute values (x_1, x_2,.., x_n), where x_i is the value of attribute X_i. On another hand, the classification variable C is represented with his values c. This type of input dataset is processed y the algorithm that assigns a function and then this function appoint a class label to the given example. From the definition of the Naïve Bayes classifier, the probability of the given example $E = (x_1, x_2,.., x_n)$ being class C is:

$$p(c \mid E) = \frac{p(E \mid c)p(c)}{p(E)} \quad (1)$$

In this stage the algorithm assumes that all attributes are independent given the value of the class variable; that is,

$$p(E \mid c) = p(x_1, x_2, ..., x_n \mid c) = \prod_{i=1}^{n} p(x_i \mid c) \quad (2)$$

The resulting classifier is then:

$$f_{NB}(E) = p(C)\prod_{i=1}^{n} p(x_i \mid C) \quad (3)$$

The function $f_{NB}(E)$ is called a Naive Bayesian classifier, or simply Naive Bayes. In order to estimate the probability that one diatom belongs into one trophic quality class we will use standardize normal distribution or Gaussian distribution, express as:

$$F_X(x) = \Phi\left(\frac{(x-\mu) \pm (pr/2)}{\sigma}\right), where\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \quad (4)$$

The algorithm the both μ and σ variables for each diatom and each trophic class estimates the pr variable. The x value is inputted as discrete class terms, because of the ecological (uncertainty) nature of the diatom dataset, and the better performance reported by [2, 3]. The Naïve Bayes classifier algorithm was used as implemented in the WEKA machine learning toolkit [15]. The discrete class values are given below, together with the labels for better interpretability. The Diatoms Term 0 starts with 0 and they are labelled bad.

Table 1: TQC for the physico-chemical parameters

Diatoms	DTerm 2 – DT2	DTerm 3 – DT3	DTerm 4 – DT4	DTerm 5 – DT5
Label	Weak	Good	Very Good	Excellent
APED	3.25	6.5	9.75	13
CJUR	21.5	43	64.5	86
COCE	20.25	40.5	60.75	81
CPLA	10	20	30	40
CSCU	10.25	20.5	30.75	41
DMAU	3	6	9	12
NPRE	4.75	9.5	14.25	19
NROT	6	12	18	24
NSROT	7.75	15.5	23.25	31
STPNN	5.25	10.5	15.75	21

III. DATA DESCRIPTION AND EXPERIMENTAL SETUP

The experiments used datasets that consists from 12 input parameters that contains information about the abundance of the 10 most abundant diatoms (measured by the biological experts [16]) plus the two TQCs for secchi disk and concentration of total phosphorus. Several experiments are conducted; each of them takes as input the dataset for each TQC. The TQC are defined according two physical-chemical parameters: secchi disk [17] and the concentration of total phosphorus [17]. Their definition is given in Table 2. Among the input parameters, ten of them are numerical parameters and the rest of them are nominal with a number of possible classes from 3 to 6. These measurements were made as a part of the TRABOREMA project [18].

Table 2: TQC for the physical-chemical parameters

Physical-chemical parameters	Name of the TQC	Parameter range
Secchi Disk	*Oligotrophic*	SD >8m – 4m
	Mesotrophic	4m – 2m
	Eutrophic	2m - 0.5m
	Hypereutrophic	0.5m – 0.25m
Total Phosphorus	*Oligotrophic*	0-12 µg/L
	Mesotrophic	12-24 µg/L
	Eutrophic	24-96 µg/L
	Hypereutrophic	96-384+ µg/L

The experimental setup estimates the highest probability of diatom with TQC. After the data is process by the algorithm, full classification model for each TQC is obtained and then probability measured using normal distribution is estimated. The normal distribution takes as input value one discretised class term from the Table 1.

IV. EXPERIMENTAL RESULTS

A. Interpretation of the classification models

The classification models obtained by the algorithm have a

definite range of the given discretised class terms, that play a vital role in the process of interpretation of the results. Because the output of the algorithm depicts the probability estimate the model interpreters the important measure of indicator properties of the diatom.

The results from the classification model for secchi disk TQC are presented in Table 3. All the diatoms are interpretable in the similar way, that why we will give several examples. For example, the classification model, found that the APED diatom is a weak indicator of *eutrophic* water with 20.82% of probability, while this diatom is weak indicator of *oligotrophic* waters with probability of 14.46%. Furthermore, the model also identifies the APED diatom as good indicator of *mesotrophic* waters with probability of 5.93%, while the other estimates are very low. Other diatoms have achieved similar probability. The COCE diatom is good indicator of *eutrophic* waters (2.45%), on other hand he is weak indicator of *oligotrophic* waters. According to the classification model the NPRE diatom is a bad indicator of *oligotrophic* waters, while good to excellent indicator of *eutrophic* waters. The STPNN diatom is a bad indicator of *oligotrophic* water, while good for *eutrophic* and etc. It is interesting to note that the low indicator properties is not a of impropriate method for classification, but more to the quality and quantity of the data. This was concluded for this diatom dataset in experiments with previous methods [5, 6, 7, 8]. Also important is to note that the classification model, classified some of the diatoms as bad indicators, because most of the data contained values of diatoms abundance near zero. In the processes of inducing the model we have assumed that low abundance of certain diatoms is bad indicator of given TQC, but it was unknown for which class. In this direction some of the results obtained from the model may or not fit in the known diatom literature.

Table 3: Evaluation results from the classification model for secchi disk TQC

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class	*oligotrophic*	*eutrophic*	*mesotrophic*	*mesotrophic*	*mesotrophic*
APED	14.49%	20.82%	5.93%	0.81%	0.03%
Class	*eutrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
CJUR	21.99%	1.15%	0.00%	0.00%	0.00%
Class	*mesotrophic*	*oligotrophic*	*eutrophic*	*eutrophic*	*eutrophic*
COCE	1.12%	2.24%	2.45%	1.03%	0.15%
Class	*eutrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
CPLA	59.54%	2.56%	0.00%	0.00%	0.00%
Class	*oligotrophic*	*eutrophic*	*mesotrophic*	*mesotrophic*	*mesotrophic*
CSCU	3.86%	8.46%	2.76%	0.38%	0.02%
Class	*eutrophic*	*mesotrphic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
DMAU	19.85%	14.13%	8.51%	4.25%	1.22%
Class	*oligotrophic*	*eutrophic*	*eutrophic*	*eutrophic*	*eutrophic*
NPRE	23.82%	13.01%	5.04%	0.65%	0.03%
Class	*eutrophic*	*oligotrophic*	*mesotrophic*	*mesotrophic*	*mesotrophic*
NROT	18.19%	14.92%	0.60%	0.00%	0.00%
Class	*oligotrophic*	*eutrophic*	*eutrophic*	*eutrophic*	*eutrophic*
NSROT	12.54%	10.41%	0.43%	0.00%	0.00%
Class	*oligotrophic*	*eutrophic*	*eutrophic*	*eutrophic*	*eutrophic*
STPNN	40.18%	12.74%	2.95%	0.12%	0.00%

Similarly we have obtained the results for the second physico-chemical parameter – Total Phosphorus. The evaluation results are presented in Table 4. From this model, it is easy to note that APED diatom is a good indicator of *oligotrophic* waters, and weak indicator of *mesotrophic* and bad indicator of *hypertrophic* waters. CJUR diatom is bad indicator of *hypertrophic* waters, but weak to excellent indicator of *oligotrophic* waters. The CSCU diatom is weak indicator of *mesotrophic* water and good to excellent indicator of *eutrophic* waters. Other diatoms like DMAU, NSROT and STPNN diatoms are good to excellent indicators of *oligotrophic* water, but with low probability according the model. As it can be easily noticed from the model, all the diatoms have around 99% probability to be bad indicators of certain water quality class – *hypertrophic* waters. Again as in the previous model the diatom abundance near zero have play a vital role in the process of classification.

Table 4: Evaluation results from the classification model for Total Phosphorus TQC

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class	*hypertrophic*	*mesotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
APED	99.73%	14.25%	7.01%	0.93%	0.03%
Class	*hypertrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
CJUR	99.73%	4.10%	0.01%	0.00%	0.00%
Class	*hypertrophic*	*eutrophic*	*oligotrophic*	*oligotrophic*	*mesotrophic*
COCE	99.73%	2.43%	2.15%	0.69%	0.11%
Class	*hypertrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
CPLA	99.73%	7.25%	0.53%	0.00%	0.00%
Class	*hypertrophic*	*mesotrophic*	*eutrophic*	*eutrophic*	*eutrophic*
CSCU	99.73%	6.84%	2.78%	0.68%	0.07%
Class	*hypertrophic*	*oligotrophic*	*mesotrophic*	*mesotrophic*	*mesotrophic*
DMAU	99.73%	15.51%	6.15%	1.19%	0.09%
Class	*hypertrophic*	*eutrophic*	*eutrophic*	*eutrophic*	*eutrophic*
NPRE	99.73%	12.46%	1.68%	0.04%	0.00%
Class	*hypertrophic*	*oligotrophic*	*mesotrophic*	*mesotrophic*	*mesotrophic*
NROT	99.73%	13.13%	0.89%	0.01%	0.00%
Class	*hypertrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
NSROT	99.73%	9.81%	1.90%	0.05%	0.00%
Class	*hypertrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*	*oligotrophic*
STPNN	99.73%	10.94%	1.96%	0.07%	0.00%

B. Verification of the results from the models

The ecological references for the 10 most abundant diatoms are taken from the diatom ecology publication by [19], used in several recently published papers [5, 6, 7, 8], and database (European Diatom Database - http://craticula.ncl.ac.uk/Eddi/jsp/index.jsp). Regarding the ecological reference of the 10 most diatoms in Lake Prespa, the CJUR and NPRE diatoms are newly described organisms with no record for their ecological preferences in the literature. Therefore, some of the results from the classification models are the first known ecological reference for certain TQC.

If we compare what is known for the APED diatom in the relevant literature [19], according to the models based on the secchi disk class revealed that this diatom is *mesotrophic* and for TP is *oligotrophic*, which indicates that further studies are required for made stronger conclusion. If we compare the CSCU with the known literature [19] and the models, we can agree that this diatom is *eutrophic* indicator for total phosphorus classes. On another hand, the COCE diatom according to the secchi disk is *eutrophic*, which is relevant with the known literature [19], while for the TP classes classify this diatom in different classes. According to the model, the STPNN diatom based on the TP classification is *hypereutrophic*, while for the secchi disk is *oligotrophic*. The algorithm correctly classified the *hypereutrophic* indication, since this diatom is known by this property in the known literature.

The other ecological references for the rest of the diatoms are new and they have to be further investigated, before any solid conclusion is made.

V. CONCLUSION

The method that is described in this paper has obtained models that are presented in a form of tables and rules derived from them. Then these models are verified with the known ecological knowledge. The purpose of the method is to present the probability of given diatom to be member of certain class. In this way, we strongly believe that classification of the indicating properties of the diatoms can be improve with this method not just for Lake Prespa, but from any lake ecosystem, since the geographical location plays no role in the bio-indicator properties of certain diatom [20].

The experiments on diatom datasets show that the Naïve Bayes method can be a good tool for diatom classification. For each of the defined TQC, the method has found a relationship between the diatoms and given class with certain probability. The relationship is between the labelled term, which we associate with a certain class in defined range and the given diatom. This mainly is depended on the quality and the quantity of the data. Another fact that mainly influences the outcome of the algorithm and his models are the changing ecosystem conditions, which adds a degree of uncertainty in the process of diatom classification. In this direction is the use of the Naïve Bayes classifier, because estimates the probability of a diatom in a certain TQC and reduces the uncertainty that is accompanied with the environmental data.

As we mention before, important factor of the models is the clarity, compared with previously used statistical methods [4], where the results from the models were interpreted

graphically. These models can be broken into rules and inputted in other knowledge discovery algorithms for further analysis. The estimated probability that the models produce can be used for advancing the algorithm and combining other techniques to increase the accuracy of the models. The experiments that are conducted showed that machine learning tools could extract valuable knowledge in a relatively comprehensible form, even when the application area is so complex for humans and the data is far from being perfect.

Our research paper, showed that studies like ours that combines the ecological knowledge for the processes in the lake ecosystem together with the information technologies, are necessary to provide understanding of the physical, chemical and the biological processes and their relationship to aquatic biota. Verification of the obtained models with the known ecological information in the literature has successfully classified certain known diatoms. The other knowledge obtained from the models must be further investigated before any strong conclusion is made. Encourage by the results from the models, further research regarding the Bayes method should be focused on different probabilistic functions instead of Gaussian faction to better describe the large number of low abundance data or near zero data. Other methods could be also combined and used for diatom classification.

Acknowledgement: This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] S. R. Carpenter, D. L. Christensen, J. J. Cole, "Biological control of eutrophication in lakes," *Environmental Science and Technology* vol. 29, pp. 784–786, 1995.

[2] B. P. Bennett, "Assessing the calibration of Naïve Bayes' posterior estimates, " in *Technical Report* No. CMUCS00-155, 2000.

[3] S. Monti, G.F. Cooper, "A Bayesian network classifier that combines a finite mixture model and a Naïve Bayes model," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1996 Morgan Kaufmann, pp. 447-456.

[4] E. F. Stoermer and J. P Smol, *The diatoms: Applications for the Environmental and Earth Sciences, 2nd Edition*. Cambridge, UK: Cambridge University Press, ISBN-13: 978-0521509961, pp. 42-46, 2010.

[5] A. Naumoski, D. Kocev, N. Atanasova, K. Mitreski, S. Krtić, S. Džeroski, "Predicting chemical parameters of water quality form diatoms abundance in lake Prespa and its tributaries," in *Proceddings of the 4th International ICSC Symposium on Information Technologies in Environmental Engineering - ITEE 2009*. Springer Berlin Heidelberg press, Thessaloniki, Greece, 2009, pp. 264-277.

[6] A. Naumoski, K. Mitreski, "Pattern tree spatial models for ecological classification," in *Proceedings of the 9th International Conference for Informatics and Information Technology (CIIT 2012)*. Bitola, Macedonia, 2012.

[7] A. Naumoski, S. Krstic, K Mitreski, "Novel Inverse Sigmoid Fuzzy Approach for Water Quality Diatom Classification," in *Proceedings of ICT Innovations 2011, AISC 150, L*. Kocarev (Eds.), Springer-Verlag Berlin Heidelberg, 2012, pp. 207-217.

[8] 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.

[9] J. Quinlan, J, "C4.5: Programs for Machine Learning," Morgan Kaufmann: San Mateo, CA, 1993.

[10] P. Langley, W. Iba, K. Thomas, "An analysis of Bayesian classifiers," in *Proceedings of the Tenth National Conference of Artificial Intelligence.* AAAI Press, 1992, pp. 223-228.

[11] I. Kononenko, "Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition," In *Wielinga, B., ed., Current Trends in Knowledge Acquisition.* IOS Press, 1990.

[12] M. J. Pazzani, "Search for dependencies in Bayesian classifiers," in Fisher, D., and Lenz, H. J., ed. *Learning from Data: Artificial Intelligence and* Statistics. Springer Verlag, 1996.

[13] P. Domingos, M. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," *Machine Learning*, vol. 29, pp. 103-130, 1997.

[14] J. Friedman, "On bias, variance, 0/1-loss, and the curse of dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, 1996.

[15] WEKA 3.6.2 - Machine Learning Toolkit - http://www.cs.waikato.ac.nz/ml/weka/, accessed on 13 January 2013.

[16] S. Krstič, "Description of sampling sites, "FP6-project TRABOREMA: Deliverable 2.2, 2005.

[17] R. E. Carlson, J. Simpson, "A Coordinator's Guide to Volunteer Lake Monitoring Methods," *North American Lake Management Society* vol. 96, 1996.

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

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

[20]. C. Gold, A. Feurtet-Mazel, M. Coste, A. Boudou, "Field transfer of periphytic diatom communities to assess shortterm structural effects of metals (Cd Zn) in rivers," *Water Research*, vol. 36, 2002, pp. 3654-3664.