Naïve Bayes technique for diatoms classification with discretised input

Andreja Naumoski, Kosta Mitreski

University Ss. „Cyril and Methodius"Faculty of Electrical Engineering and Information Technologies, Skopje, Karpos 2 bb, Skopje, Macedonia
{andrejna,komit}@feit.ukim.edu.mk

Abstract. The challenge to discover knowledge from environmental data that has led to usage of methods and techniques such as data mining tools, can bridge the knowledge gap between the biological experts and organisms. This research aimed to assess relationships between the diatoms and the indicators of the environment with Naïve Bayes method. Diatoms are ideal indicators of certain physical-chemical parameters and they can be classified into one of the water quality classes (WQCs). The classification models are induced by using Naïve Bayes technique. The input dataset that is supplied for the naïve Bayes method is discretised. Based on the evaluation results, several models are presented and discussed. The obtain results from the models are verified with existing diatom ecological preference and for some diatoms new knowledge is added. To best of our knowledge, this is the first time the prosed method to be applied for diatom classification of any ecosystem.

Keywords: Naïve Bayes, diatom classification, indicators, water ecosystem

1 Introduction

The water quality classes define in the traditional way can be interpreted as a classification problem in the terms of data mining point of view. In this paper research, this property is used to discover the appropriate environment conditions for newly found diatom, which are an ideal bio-indicator of a certain physico-chemical parameter. Considering these facts, we deal with the typical classification problem, when we try to build a model that classifies the correct diatoms into one of the WQ classes.

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 [18]. 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, mainly decision trees. These methods, improves the interpretability and increases the prediction power of the models. First attempt to model diatom-indicator relationship for Lake Prespa, have been made by [4]. Several of the model produced, knowledge about the newly

M. Gusev (Editor): ICT Innovations 2010, Web Proceedings, ISSN 1857-7288

discovered diatom's relationships with the environment for the first time [4]. New class of multi-target decision trees later were used, in order to reveal the dynamic nature of the entire set of physical-chemical parameters of this lake ecosystem [15]. These methods were more precise and also have increased the interpretability. Nevertheless, these methods were not robust on data change. This is an important property, because the environmental condition inside of the lake changes over small periods of time.

Many empirical comparisons between naive Bayes and modern decision tree algorithms such as C4.5 (Quinlan 1993) showed that naive Bayes predicts equally well as C4.5 [3, 5, 7] for many real data domains. To best of our knowledge, this is the first usage of Naïve Bayes classifier for diatom classification. The good performance of naive Bayes is surprising because it makes an assumption that is almost always violated in realworld applications: given the class value, all attributes are independent. This is also true for the diatoms that are independent; one diatom can be indicator of one water quality class and one for another. Nevertheless, from ecological point of view it is very important to estimate the degree that diatom depends from the certain environmental conditions.

Domingos and Pazzani [8] present an explanation that naive Bayes owes its good performance to the zero-one loss function. In [9] the authors have shown that the performance of naive Bayes is much worse when it is used for regression (predicting a continuous value). Moreover, evidence has been found that naive Bayes produces poor probability estimates [11, 12]. That’s way the input dataset that we will use in the experiments shown in the paper are discrete and we donate a certain class for each diatom.

The rest of the paper is organized as follows: Section II provides the definitions for the Naïve Bayes classifier. In Section III we present the diatoms abundance water 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 the research directions are outlined.

2. Naïve Bayes method

Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. Typically, an 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. Let C represent the classification variable, and let c be the value of C.

A classifier is a function that assigns a class label to an example. From the probability perspective, according to Bayes Rule, the probability of an example E = (x_1, x_2,…, x_n) being class C is:

$$p(c \mid E) = \frac{p(E \mid c)p(c)}{p(E)} . \qquad \textbf{(1)}$$

Assume that all attributes are independent given the value of the class variable; that is,

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

The resulting classifier is then:

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

The function f_{NB} (E) is called a naive Bayesian classifier, or simply Naive Bayes (NB) (see eq.3). This is called conditional independence. In our paper it is obvious that the conditional independence assumption is true, meaning that each diatom is independent from one water quality class.

In order to estimate the probability that one diatom belongs into one water quality class we will use standardize normal distribution, express as:

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

The precision number is estimated by the Bayes classifier, together with the μ and σ for diatom and each WQC. 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 [11, 12]. The Naïve Bayes classifier algorithm was used from the WEKA machine learning toolkit [19]. The discrete class values are given below.

3. Data description and experimental setup

The datasets used in the experiments consist from 13 input parameters representing the TOP10 diatom species (diatom species that exist in Lake Prespa [2]) with their abundance per sample, plus the three WQC for conductivity, pH and Saturated Oxygen. Then one dataset is crated for each WQ class with the TOP10 diatoms.

Table 1. Water quality classes for the physical-chemical parameters [16, 17]

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

These measurements were made as a part of the TRABOREMA project [6]. The WQCs are defined according to the three physical-chemical parameters: Saturated Oxygen [16], Conductivity [17] and pH [16, 17] which are given in Table 1. Among the input parameters 10 are numerical parameters and the rest 3 are nominal with a number of possible classes from 3 to 6.

The experimental setup estimates the highest probability of diatom with water quality class. After the data is process by the algorithm, full classification model for each water quality class, then probability measured using normal distribution is estimated. The normal distribution takes as input value one discretised class term from the Table 2.

Table 2. Discretised input dataset into probability estimator

Diatoms	**DTerm 1 – DT1**	**DTerm 2 – DT2**	**DTerm 3 – DT3**	**DTerm 4 – DT4**	**DTerm 5 – DT5**
	Bad	Weak	Good	Very Good	Excellent
APED	0	3.25	6.5	9.75	13
CJUR	0	21.5	43	64.5	86
COCE	0	20.25	40.5	60.75	81
CPLA	0	10	20	30	40
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	7.75	15.5	23.25	31
STPNN	0	5.25	10.5	15.75	21

5. Experimental results

In this section, three models are given for each water quality class, to show the probability estimates from the Naïve Bayes classification. Later the results from the classification models with the known ecological reference of the diatoms are verified.

5.1 Interpretation of the classification models

All the induced classification models have a define range of discretised class terms, which later will be commented. Each diatom for certain class has a probability estimate, which is important measure of indicator properties of the diatom.

The results from the classification model for Conductivity water quality class are presented in Table 3. According to the classification model, the APED diatom is a bad indicator of brackish water with 99.73% of probability, while he is weak indicator of brackish fresh waters with probability of 14.46%. The model also identifies the APED diatom as good indicator of fresh brackish waters with probability of 5.35%, while the other estimates are very low. Similar conclusion can be made for all the TOP10 diatoms. We will summarize just a few of them. For example, COCE diatom is weak indicator of brackish waters (2.18%), on other hand he is good indicator of brackish fresh waters. According to the classification model the DMAU diatom is a weak indicator of brackish waters, while good indicator of brackish waters.

Table 3. Evaluation results from the classification model for Conductivity water quality class

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class	brackish	Brackish fresh	fresh brackish	fresh brackish	fresh brackish
APED	99.73%	14.46%	5.35%	0.60%	0.02%
Class	fresh brackish	brackish	brackish	brackish	Fresh brackish
CJUR	99.73%	2.18%	0.00%	0.00%	0.00%
Class	fresh brackish	brackish	Brackish fresh	Brackish fresh	Brackish fresh
COCE	99.73%	2.64%	2.13%	1.15%	0.29%
Class	fresh brackish	Brackish fresh	Brackish fresh	Fresh brackish	Fresh brackish
CPLA	99.73%	0.37%	0.00%	0.00%	0.00%
Class	Fresh brackish	brackish	Brackish fresh	Brackish fresh	Brackish fresh
CSCU	95.28%	6.09%	2.90%	0.44%	0.02%
Class	fresh brackish	Brackish fresh	brackish	brackish	brackish
DMAU	99.73%	14.19%	5.40%	0.67%	0.03%
Class	fresh brackish	Brackish fresh	Brackish fresh	Brackish fresh	Brackish fresh
NPRE	99.73%	14.50%	1.12%	0.01%	0.00%
Class	fresh brackish	brackish	brackish	brackish	brackish
NROT	99.73%	12.07%	0.89%	0.01%	0.00%
Class	fresh brackish	Brackish fresh	brackish	brackish	brackish
NSROT	99.73%	9.69%	0.60%	0.00%	0.00%
Class	fresh brackish	Brackish fresh	Brackish fresh	Brackish fresh	Brackish fresh
STPNN	99.73%	10.81%	1.92%	0.06%	0.00%

The STPNN diatom has weak indicator properties for brackish fresh waters, while the NROT diatom for brackish waters. 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 [15]. The classification model, classified the diatoms as bad indicators, because most of the data contained values of diatoms abundance near 0. We have assumed that low abundance of certain diatoms is bad indicator of given water quality class, but it was unknown for which class.

The evaluation results for the pH water quality class are presented in Table 4. From the model, it is easy to note that APED diatom is a good indicator of *alkaliphilous* waters, and weak indicator of *alkalibiontic* and bad indicator of *circumneutral* waters. NPRE diatom is bad indicator of *acidophilous* waters, but good to excellent indicator of *acidophilous* waters. NROT, NSROT and STPNN diatoms are good to excellent indicators of indifferent waters, but with low probability according the model. All the diatoms more or less have around 15% probability to be good indicators of certain water quality class.

Table 4. Evaluation results from the classification model for pH water quality class

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class APED	circumneutral 19.19%	alkalibiontic 14.16%	alkaliphilous 8.00%	alkaliphilous 2.51%	alkaliphilous 0.34%
Class CJUR	acidophilous 30.31%	alkalibiontic 3.24%	alkalibiontic 0.00%	alkalibiontic 0.00%	acidophilous 0.00%
Class COCE	acidophilous 1.55%	Indifferent 2.51%	Indifferent 2.38%	alkaliphilous 0.80%	alkaliphilous 0.17%
Class CPLA	Indifferent 95.46%	alkalibiontic 6.11%	alkalibiontic 0.14%	alkalibiontic 0.00%	alkalibiontic 0.00%
Class CSCU	Indifferent 99.73%	circumneutral 7.06%	alkalibiontic 3.82%	alkalibiontic 1.01%	alkalibiontic 0.09%
Class DMAU	circumneutral 14.08%	Indifferent 19.27%	acidophilous 7.75%	acidophilous 3.96%	acidophilous 1.24%
Class NPRE	acidophilous 99.73%	alkaliphilous 12.20%	alkaliphilous 3.60%	alkaliphilous 0.33%	alkaliphilous 0.01%
Class NROT	acidobiontic 99.73%	Indifferent 13.96%	Indifferent 2.70%	Indifferent 0.08%	Indifferent 0.00%
Class NSROT	acidobiontic 99.73%	Indifferent 10.00%	Indifferent 2.69%	Indifferent 0.13%	Indifferent 0.00%
Class STPNN	acidobiontic 99.73%	Indifferent 8.87%	Indifferent 0.79%	Indifferent 0.01%	Indifferent 0.00%

Concerning the last water quality class – Saturated Oxygen, the results from the classification model (see Table 5) shows that the TOP10 diatoms have all bad indicator properties for the *polysaprobous* WQC class.

Table 5. Evaluation results from the classification model for Saturated Oxygen water quality class

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class APED	Poly saprobous 99.73%	Oligo saprobous 14.89%	β-meso saprobous 5.07%	β-meso saprobous 0.42%	β-meso saprobous 0.01%
Class CJUR	Poly saprobous 99.73%	α-meso saprobous 6.54%	α-meso saprobous 0.16%	α-meso saprobous 0.00%	α-meso saprobous 0.00%
Class COCE	Poly saprobous 99.73%	β-meso saprobous 2.79%	α-meso saprobous 2.15%	α-meso saprobous 0.80%	α-meso saprobous 0.11%
Class CPLA	Poly saprobous 99.73%	Poly saprobous 10.26%	β-meso saprobous 1.54%	β-meso saprobous 0.07%	β-meso saprobous 0.00%
Class	Poly saprobous	α-meso saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous

CSCU	99.73%	6.82%	2.80%	0.39%	0.02%
Class	Poly saprobous	β-meso saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous
DMAU	99.73%	15.07%	6.33%	1.33%	0.11%
Class	Poly saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous
NPRE	99.73%	13.25%	1.68%	0.03%	0.00%
Class	Poly saprobous	Oligo saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous
NROT	99.73%	12.29%	0.98%	0.01%	0.00%
Class	Poly saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous
NSROT	99.73%	9.70%	0.74%	0.00%	0.00%
Class	Poly saprobous	α-meso saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous
STPNN	99.73%	9.78%	0.49%	0.00%	0.00%

The APED diatom and the CPLA diatom, according the classification model are weak indicator of *oligosaprobous* water class, but good to excellent indicators for *β-mesosaprobous* with very low probability. The CSCU, NPRE, NROT, NSROT and STPNN diatoms according to the models are weak to very good indicators of *oligosaprobous* waters. The rest of the diatoms are less or more weak to good indicators of *α-mesosaprobous* waters. Once more, the classification model has low values for the probability estimates, except for the first class.

5.2 Verification of the results from the models

Ecological references for the TOP10 diatom are taken from the latest diatom ecology publications [14], used in several recently published papers [1, 2, 4, 15], and database (European Diatom Database - http://craticula.ncl.ac.uk/Eddi/jsp/index.jsp). Concerning ecological reference of the TOP10 dominant diatoms in Lake Prespa, CJUR and NPRE are newly described taxa (diatoms) 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 WQC classes.

In the relevant literature APED diatom is known to be *alkaliphilous*, *fresh-brackish*, nitrogen-autotrophic (tolerates elevated concentrations of organically bound nitrogen), high oxygen saturation (>75%), *β-mesosaprobic* and *eutrophic* (because of Organic N tolerance) diatom indicator [14]. According to the classification models the APED diatoms is found to be an *alkaliphilous* and *fresh-brackish* indicator. Regarding the Saturated Oxygen WQ classes, APED is a weak indicator of *oligosaprobous,* but good indicator of *β-mesosaprobic* environment.

Concerning CSCU diatom indicator affinity, the model for pH WQC has revealed that this diatom is *alkalibiontic*. According to the models for conductivity WQ class the CSCU diatom is *brackish* to *brackish fresh* diatom, while the Saturated Oxygen WQC model shows the weak affinity of this diatom to α-*mesosaprobous*, but with good indicator properties for *oligosaprobous* waters. In the relevant literature the CSCU is known as *alkalibiontic*, *freshwater* to *brakish* water taxon, being *oligosaprobic* indicators with eutrophic references [14].

The COCE diatom is known as *meso-eutro* taxon [14], while concerning the pH properties of this diatom, there is no known ecological reference. According to the classification models, the COCE diatom is relative good indicator for *brackish fresh* waters, *indifferent* to *alkaliphilous,* and for the saturated oxygen demand he is a weak indicator of *β-mesosaprobous* but relatively good for *α-mesosaprobous* environments. Further experiments to investigate the trophic indicator affinity of this diatom should be made by using trophic state index classes.

The STPNN diatom, in the literature is known as *hyper-eutrophic* (*oligo-eutrophic*; indifferent) taxon frequently found on moist habitats, while the classification models have been found to be *alkalibiontic* taxon. According to the classification models, this diatom is a weak to good indicator of *brackish* waters, while *indifferent* taxon for pH WQC. The model for the saturated oxygen showed that this diatom is weak indicator of *α-mesosaprobous* waters, but relatively good for *oligosaprobous*.

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. Nevertheless, many of the known ecological references are verified with the classification method, thus proving the reliability of the proposed method for diatom classification.

6. Conclusion

The proposed method has verified the known diatom ecological knowledge and for some of them, added a new knowledge. Classifying the diatoms from measure data can be greatly improved with the proposed method, not just from Lake Prespa, but from any lake ecosystem, since the geographical location plays no role in the bio-indicator properties of certain diatom [13].

The experiments on diatom WQC datasets show that the Naïve Bayes method can be a good tool for diatom classification. For each of the defined WQ classes, the method has found a relationship between the diatoms and the indicator with certain probability. The input data in the proposed method is divided into classes, with labeled a term, associate with a define range. With this process, the classification accuracy of the proposed method is higher, based on the research work done previously on other datasets. Also, another fact is the changing ecosystem conditions, which adds a degree of uncertainty in the process of diatom classification. That's way, we use the Naïve Bayes classifier, because estimates the probability of a diatom in a certain WQC class and reduces the uncertainty which is accompanied with the environmental data.

More important is the interpretation of the classification models, compared with the classical statistical methods such as: PCA, CCA, DCA and other methods, used previously, the proposed method is more directly interpretable. The obtained models have openly stated prediction and probability in terms of finding correct diatom-indicator relationship. The experiments showed that machine learning tools can extract valuable knowledge in a relatively comprehensible form, even when the application area is so extremely complex for humans and the data are far from being perfect.

We believe that studies like ours that combines the ecological together with information technologies, especially in the area of eco informatics, are necessary to provide understanding of the physical, chemical and biological processes and their relationship to aquatic biota for predicting a certain effect. Verification of the obtained models showed that the proposed method, have successfully classified certain known diatoms, and added new ecological knowledge for the unknown diatoms for certain WQCs.

Further research needs to be focused on developing classification models base on the Naïve Bayes method for trophic state index classes. Other methods for classification could be suitable for diatoms classification that needs to be explored.

References

1. Krstič, S.: Description of sampling sites. FP6-project TRABOREMA: Deliverable 2.2. (2005).
2. Levkov. Z., Krstič. S., Metzeltin. D., Nakov. T.: Diatoms of Lakes Prespa and Ohrid (Macedonia). Iconographia Diatomologica, vol. 16, pp. 603 (2006)
3. Langley, P., Iba, W., and Thomas, K.: An analysis of Bayesian classifiers. In Proceedings of the Tenth National Conference of Artificial Intelligence. AAAI Press, pp. 223—228 (1992)
4. Naumoski, A., Kocev, D., Atanasova, N., Mitreski, K., Krtić, S., Džeroski, S.: Predicting chemical parameters of water quality form diatoms abundance in lake Prespa and its tributaries. The 4th International ICSC Symposium on Information Technologies in Environmental Engineering - ITEE 2009. Springer Berlin Heidelberg press, Thessaloniki, Greece pp. 264--277. (2009)
5. Kononenko, I.: Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition. In Wielinga, B., ed., Current Trends in Knowledge Acquisition. IOS Press, (1990)
6. TRABOREMA Project WP3.: EC FP6-INCO project no. INCO-CT-2004-509177 (2005-2007)
7. Pazzani, M. J.: Search for dependencies in Bayesian classifiers. In Fisher, D., and Lenz, H. J., eds., Learning from Data: Artificial Intelligence and Statistics V. Springer Verlag, (1996)
8. Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Machine Learning 29, pp. 103--130 (1997)
9. Friedman, J.: On bias, variance, 0/1-loss, and the curse of dimensionality. Data Mining and Knowledge Discovery 1, (1996)
10. Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA, (1993)
11. Bennett, P. N.: Assessing the calibration of Naïve Bayes' posterior estimates. In Technical Report No. CMUCS00-155 (2000)
12. Monti, S., Cooper, G. F.: 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. Morgan Kaufmann. Pp. 447--456 (1996)
13. Gold, C., Feurtet-Mazel, A., Coste, M., Boudou, A.: Field transfer of periphytic diatom communities to assess shortterm structural effects of metals (Cd Zn) in rivers. Water Research Vol. 36, pp. 3654--3664 (2002)

14. Van Dam, H., Martens, A., Sinkeldam, J.: A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology Vol. 28, Issue. 1, pp. 117—133 (1994)
15. Kocev. D., Naumoski. A., Mitreski. K., Krstić. S., Džeroski. S.: Learning habitat models for the diatom community in Lake Prespa. Journal of Ecological Modelling, vol. 221, No. 2, pp. 330--337 (2009)
16. Krammer. K., Lange-Bertalot. H.: Die Ssswasserflora von Mitteleuropa 2: Bacillariophyceae. 1 Teil. pp. 876, Stuttgart: Gustav Fischer-Verlag (1986)
17. Van Der Werff. A., Huls. H.: Diatomeanflora van Nederland. Abcoude - De Hoef (1957, 1974)
18. Stroemer. E. F., Smol. J. P.: The diatoms: Applications for the Environmental and Earth Sciences, Cambridge University Press, Cambridge (2004)
19. WEKA 3.6.2 - Machine Learning Toolkit - http://www.cs.waikato.ac.nz/ml/weka/