Exploring an Ichthyoplankton Database from a Freshwater Reservoir in Legal Amazon

Michel de A. Silva, Daniela Queiroz Trevisan, David N. Prata, Elineide E. Marques, Marcelo Lisboa, and Monica Prata

Postgraduate Program in Computational Modeling Systems
Federal University of Tocantins (UFT)
Palmas – TO – Brasil
{michel,danielatrevisan,dnnprata,emarques}@uft.edu.br

Abstract. The purpose of this study is to use data mining techniques for the exploratory analysis of a database of ichthyoplankton samples from a freshwater reservoir in Legal Amazon. This database has already been analyzed using statistical techniques, but these did not find a relationship between biotic and abiotic factors. The application of the Apriori algorithm allows us to generate association rules that yield an understanding of the process of fish spawning in Tocantins River. In this case, we demonstrate the effective use of data mining for the discovery of patterns and processes in ecological systems, and suggest that statistical methods often used by ecologists can be coupled with data mining techniques to generate hypotheses.

Keywords: data mining, Apriori, database, ichthyoplankton.

1 Introduction

In recent decades, the evolution of hardware technologies has dramatically increased the amount of stored data. These data are kept in many structures, such as database systems, webpages, conventional files, spreadsheets, tablets, and smartphones.

Information technology, and specifically data mining, has provided information by applying several artificial intelligence techniques. The manipulation of databases has uncovered new approaches to the use of this information, broadening the discussion of the stored data and raising new study questions.

The increasing world population and technological advances in production processes have given rise to a world "guided by resources". The environmental impact of human exploitation is a topic that demands significant effort from the scientific community. Ecological data mining seeks answers to how we understand and use our natural resources more efficiently.

Ecologists have mainly used statistical methods to analyze the relationship between an observed response and a set of predictive variables in a dataset. This approach to

H. Motoda et al. (Eds.): ADMA 2013, Part II, LNAI 8347, pp. 384–395, 2013.

data analysis is more appropriate for hypothesis testing. When prior knowledge is minimal and the assumptions are not clearly developed, exploratory analyses are more appropriate than confirmatory analysis [1]. Data mining techniques are often more powerful, flexible, and efficient than statistical methods for conducting these exploratory studies (e.g., [2-4]).

In this paper, we describe the application of the Apriori algorithm [5] and association rules to search for patterns in an ichthyoplankton ecological database. The development of this process has disclosed information from a database domain that could not be detected through the use of statistical methods alone [6]. The application of this technique allows us to determine rules between certain proprieties, thus producing unexpected information that was previously hidden.

Our main goal is to answer the following question: "Is there any relationship between abiotic factors and the larval stage?"

In the first phase of modeling the problem, we pinpointed the topics for which ecologists are seeking answers, as well as relevant information for the application of each of the properties. At the second, preprocessing stage, we selected the data to be processed, integrating, reducing, and transforming them to add quality and ensure their feasibility. From this point, at the third stage, we started the experimental phase of data mining by applying the Apriori algorithm to the preprocessed data, using specified parameters for the degree of confidence and support. The fourth stage, known as post-processing, concerns the refinement and interpretation of the extracted rules, including an expert assessment to analyze the results.

2 Ichthyoplankton Database

In the context of conservation and natural resource management, databases are likely to become more extensive because of the enlargement of data collection, time series, and details of sampling networks. At present, the monitoring of occurrences that are taking place in natural systems, anthropogenic or not, is one of the major challenges to environmental management in tropical areas. Though desirable, the detection and description of these occurrences, identification of variables to indicate environmental quality, and construction and verification of models for future predictions in complex systems (such as aquatic ecosystems), are in the early stages of development in tropical areas. The relationship between high species diversity and environmental factors allows a wide range of scenarios and interactions that require different techniques to be identified.

In this scenario, information from monitoring fish fauna, both in the early stages of development (eggs and larvae) and in adulthood, associated with hydroelectric projects constitute a dataset that has not yet been explored.

In Tocantins River, as well as in other rivers in Brazil and across the world, fish fauna has been systematically modified by the construction of hydroelectric plants.

Seven projects operating in Tocantins River have dramatically changed the environment of fish and wildlife, and these may be further threatened by the construction of other plants already planned for this watercourse [7].

Nevertheless, monitoring the behavior of fish fauna, identifying key spawning sites, and studying the early stages of development are protection strategies that help to define the limits of conservation areas. These strategies can be incrementally improved and refined using the existing database, but they necessarily require the use of information technology.

Even when data collection is in progress, the application of different analytical techniques can help to evaluate sampling networks and spot the gaps that could be filled with the correct set of information.

Monitoring fish fauna in the area of influence of the Lajeado reservoir, on Tocantins River, produced one of the first systematized information bases for a research group in Legal Amazon, as data were being collected before, during, and after the construction of the reservoir. Changes in the distribution and abundance of fish larvae have been reported using traditional methods of analysis [6]. Accordingly, the application of data mining may enhance this research and identify new matters, methodologies, and results.

3 Data Collection

The studies were conducted at the Lajeado reservoir area (HPP Luís Eduardo Magalhães), shown in Figure 1. Construction of the dam was halted in 2002.

Fish larvae data were collected monthly from April 2010 to March 2012 at 12 points located around the reservoir (downstream and upstream) and the headwater and mouth of its major affluents (Santa Tereza, Santo Antônio, São Valério, Crixás, Areias, Matança, Água Suja, and Santa Luzia rivers) using the methodology suggested by [8]. At the same time, information concerning the air temperature, water conductivity, pH, and water transparency were recorded.

Water temperature data (°C) were obtained using a digital thermometer; the transparency of the water column (m) was measured using a Secchi disk of 0.30 m in diameter, while the hydrogen potential (pH) and electrical conductivity ($\mu S.cm^{-1}$) were obtained from portable digital potentiometers, and the concentration of dissolved oxygen was measured using a portable digital oximeter (YSI, $mg.L^{-1}$).

Samples were taken through horizontal hauls by a conic-cylindrical plankton net, with a 500 micron mesh and length of 1.5 m (conical part of 0.9 m and cylindrical part of 0.6 m), and a mechanical flowmeter coupled to the mouth of the network to obtain the volume of filtered water. Samples were taken at a depth of about 20 cm in the middle and near the left and right margins of the sites, which were more than 20 m wide.

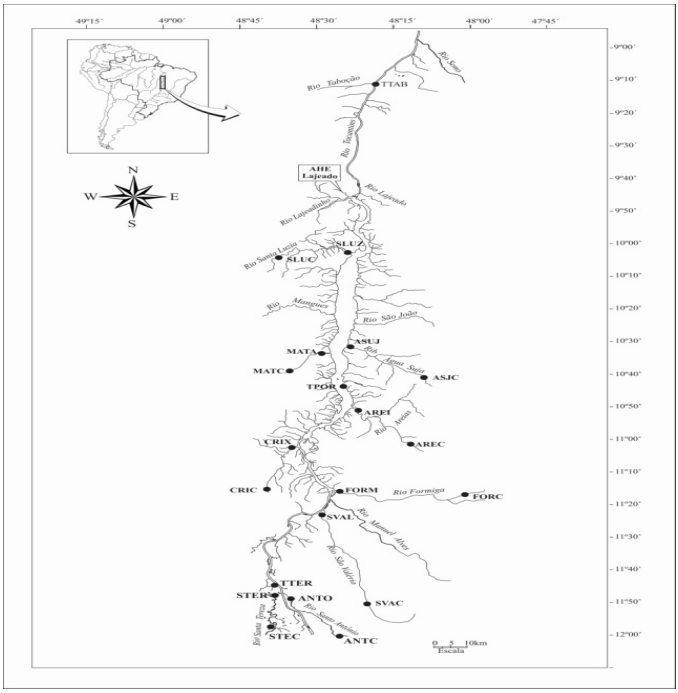


Fig. 1. Location of the sample points in the Lajeado reservoir and its area of influence

The samples were fixed in formalin and analyzed using a stereoscopic microscope. Larvae were separated according to the stage of development (larval yolk, pre-flexion, flexion, and post-flexion) considering criteria such as the presence of a yolk sac, mouth opening, and development of the tail, fin, and supporting elements, following the methodology in [8].

Larvae were identified in terms of their order and family level where possible, according to [8], which considers morphological characters such as the size and position of the eyes, shape and pigmentation of the body, position of the anal opening to the body, forming sequence and position of the fins, presence of wattles, and meristic data, such as the number of myomeres and fin rays. The obtained material was deposited in the Fish Collection of the Federal University of Tocantins.

The abundance of eggs and larvae was calculated per 10 cubic meters of filtered water, according to the method of [9], as modified by [8].

The data were organized according to the dates and locations of sampling, as well as the sampling point (margin/medium; surface), and stored in two .xls files. One of these files contained abiotic data about the environment, and the other file stored biotic data about the eggs and larvae collected.

4 Methodology for Data Mining

The first stage of the work consisted of examining those data mining algorithms that can determine patterns among the properties of the data, thus identifying the unknown relationships between them.

Algorithms that are able to find such relationships are called algorithms of association rules, and extract frequent sets of attributes embedded in a larger set. These algorithms vary greatly in terms of their generation of subsets of the universe, and how the sets of chosen attributes are supported for the generation of association rules.

An association rule has the form $A \rightarrow B$, where the antecedent A and consequent B are sets of items or transactions. The rule can be read as: the attribute A often implies B [10]. To evaluate the generated rules, we used various measures of interest. The most often used [10] are the support and confidence. The authors [11] performed a survey of other metrics, and suggested strategies for selecting appropriate measures for certain areas and requirements. In this paper, we use the following measures:

- Support: P (AB). The support of a rule is defined as the fraction of items I that can be placed in sets A and B based on the given rule. If the support is not large enough, the rule is not worthy of consideration, or is simply deprecated and may be considered later.
- Confidence: P (A / B). This is a measure of the strength of a rule's support, and corresponds to statistical significance. It describes the probability of the rule finding B such that the transaction also contains A.
- Lift: P (B | A) / P (B) or P (AB) / P (A) P (B). Used to find dependencies, the lift indicates how much more common B becomes when A occurs. This value can vary between 0 and ∞.

Finding item sets with frequency greater than or equal to the user-specified minimum support is not trivial, as combinatorial explosion occurs when generating subsets of items. However, because frequent item sets are obtained, it is straightforward to generate association rules with confidence greater than or equal to the user-specified minimum value [12].

In this context, the Apriori algorithm is a seminal method of finding frequent item sets. Introduced in [5], and appointed by the IEEE International Conference on Data Mining [12] as the most promising generation algorithm for association rules, Apriori is one of the most popular approaches in data mining.

Using a depth-first search, the algorithm is able to generate sets of candidate items (recognized as the standard) with k elements from item sets of k-1 elements. The scan ends at the last element of the database, and infrequent patterns are discarded. In this paper, we implement the Apriori algorithm using the Waikato Environment for Knowledge Analysis (Weka) tool [13]. This version of Apriori iteratively reduces the minimum support until it finds the required number of rules with some minimum confidence parameter passed by the user. This ease of parameterization, and the fact

that it includes all evaluation metrics of association rules mentioned above, led to the adoption of Weka instead of a new implementation of the algorithm for rule extraction.

4.1 Preprocessing

The data are only considered to be of sufficient quality if they meet the requirements for their intended use. There are many factors that make up the quality of the data, including accuracy, completeness, consistency, timeliness, credibility, and interpretability [14].

To ensure these measures of quality, the following steps were applied as data preprocessing: (a) data integration, (b) data cleanup, (c) data reduction, (d) data transformation.

Data Integration. Ichthyoplankton samples collected at Lajeado reservoir and its area of influence were recorded in two different databases. To reduce redundancy and inconsistencies in the final dataset, we performed a secure integration that used the sample code as a key link between spreadsheets. The final data set was created in a new CSV (comma separated values) file. Redundant data were eliminated or grouped, depending on the value for the sample, avoiding inconsistencies in the set as a whole.

Data Cleaning. At this stage, various routines were performed to ensure data quality:

- Missing data were replaced by a global constant, indicated by "?" Thus, the algorithm could handle gaps without influencing the results.
- Some nonstandard values (outliers), such as values of "9999" for the water temperature, were removed. These data were considered missing, and were subsequently identified by the symbol "?"
- Inconsistent data, such as input "i" for the cloudiness attribute, which should receive only numeric values by default, were also reported as missing.

Reduction of Data. The final set of attributes from the original set was reduced by performing a dimension reduction in which irrelevant, weakly relevant, and redundant properties were detected and removed.

For this task we employed the CfsSubsetEval algorithm, which assesses the value of a subset of attributes by considering the predictive ability of each individual feature, along with the degree of redundancy between them. Subsets of features that are highly correlated with the class and have low intercorrelation are preferentially selected [15].

For this work, the combination of BestFirst (search method) and CfsSubsetEval (attribute evaluator) is as efficient as the best techniques (genetic algorithm, simulated annealing) for the selection of variables, but has the advantage of being faster than other approaches [16].

To evaluate the attributes, we compared values using the heuristic merit of each relationship, which is formalized as [17]:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\,\overline{r_{ff}}}} \quad (1)$$

The final formula of merit uses the Pearson correlation between a variable composite (sum or average) and a target variable (the class in question) [18].

The Weka CfsSubsetEval algorithm was executed with the initial set of data as input. From the 33 starting attributes, the base was reduced to 14. After being evaluated by experts in ichthyology, this number was further reduced to a total of 10 attributes considered essential for modeling the problem. The list of attributes and their respective scores of merit is shown in Table 1.

Table 1. Attributes selected by the CfsSubsetEval algorithm from Weka. The stage attribute refers to the larval stage of the fish.

Attribute Rated	Merit	Selected Attributes
Family	0.692	order, specie
Order	0.595	Smf, stage, family
waterConductivity	0.557	local, ph
Specie	0.543	local, waterTemperature, family
Ph	0.463	margin,dissolvedOxigen, waterConductivity, order
dissolvedOxigen	0.445	local, margin, airTemperature, ph
Depth	0.319	local, transparency
transparency	0.314	airTemperatura, depth, dissolvedOxigen, ph
Stage	0.268	order, family
Local	0.116	margin,dissolvedOxigen,ph, waterConductivity,família,specie
waterTemperature	0.096	local, airTemperature
airTemperature	0.094	waterTemperature,dissolvedOxigen, transparency
Margemmargin	0.088	local, dissolvedOxigen, ph, waterConductivity
Smf	0.01	local, margin, ph, waterConductivity, specie

Transformation of Data

- Decimal points input as "," were replaced by "." to ensure they were correctly interpreted by Weka.
- Date formats were standardized to "dd/mm/yyyy".
- Some numerical variables were discretized by mapping value ranges to in labels, as shown in Table 2. This enabled the Apriori algorithm, which requires nominal attributes, to be applied.

After completing this preprocessing, the data were gathered in a single file containing 9 attributes and 4913 instances (from the original 33 attributes and 5333 instances).

5 Results

In the experiments, we used different parameter values for the Apriori algorithm to find the best rules involving the attribute stage (stage of fish larvae) and the biotic and abiotic data.

Table 2. Discretized Attributes

Attribute	Values Gathered in Collection	Equivalent Values	Attribute	Values Gathered in Collection	Equivalent Values
Air Temperature	10 < airTemp <= 15	1	Depth	depth > 3 and <= 4	4
	15 < airTemp <= 20	2		depth > 4 and <= 5	5
	20 < airTemp <= 25	3		depth > 5 and <= 6	6
	25 < airTemp <= 30	4		depth > 6	7
	30 < airTemp <= 35	5	Water Conductivity	cond <= a 20	1
	35 < airTemp <= 40	6		cond > 20 and <= 40	2
Water Transparency	transp > 0 and <= 0.5	1		cond > 40 and <= 60	3
	transp > 0.5 and <= 1	2		cond > 60 and <= 80	4
	transp > 1 and <= 1.5	3		cond >80 and <= 100	5
	transp > 1.5 and <= 2	4		cond>100 and<= 120	6
	transp > 2 and <= 2.5	5		cond > 120	7
	transp > 2.5 and <= 3	6	Dissolved Oxigen	oxig > 0 and <= 1	1
	transp > 3	7		oxig > 1 and <= 2	2
Water Temperature	waTemp>10and<=15	1		oxig > 2 and <= 3	3
	waTemp>15and<=20	2		oxig > 3 and <= 4	4
	waTemp>20and<=25	3		oxig > 4 and <= 5	5
	waTemp>25and<=30	4		oxig > 5 and <= 6	6
	waTemp>30and<=35	5		oxig > 6 and <= 7	7
	waTemp>35and<=40	6		oxig > 7 and <= 8	8
PH	ph <= 4.0	1		oxig > 8 and <= 9	9
	ph > 4.0 and <= 6.0	2		oxig > 9 and <= 10	10
	ph > 6.0 and <= 7.0	3		oxig > 10 and <= 11	11
	ph > 7.0 and <= 9.0	4		oxig > 11 and <= 12	12
	ph > 9.0	5		oxig > 12	13
Depth	depth > 0 and <= 1	1			
	depth > 1 and <= 2	2			
	depth > 2 and <= 3	3			

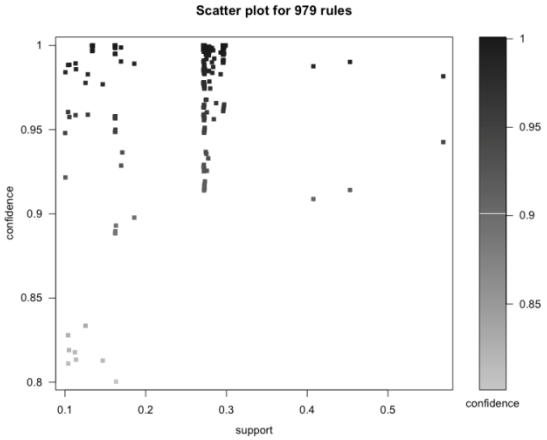
The coverage, or support, of an association rule is taken as the number of instances that are correctly predicted by the rule. Its accuracy, or confidence, is the number of instances that the rule correctly predicts, expressed as a percentage of all instances to which it applies [19].

The minimum and maximum support values were set to 0.1 and 1.0, allowing the rules to be generated freely. Parameter values for the confidence and interest varied

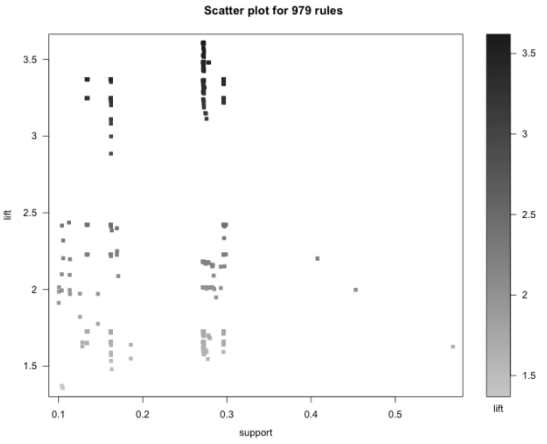
from 0.8 to 1.0 and from 1.5 to 3.5, respectively. In total, 979 association rules were generated, as shown in Graph 1.

A large number of rules were generated with confidence higher than 0.9 and support between 0.2 and 0.3, as shown in Graph 1.

For the metric of interest, the number of rules with lift values between 3 and 3.5 is concentrated between support values of 0.2 and 0.3. Slightly lower lift values, between 1.5 and 2.5, have support in the wider range of 0.1 to 0.3, as can be seen in Graph 2.

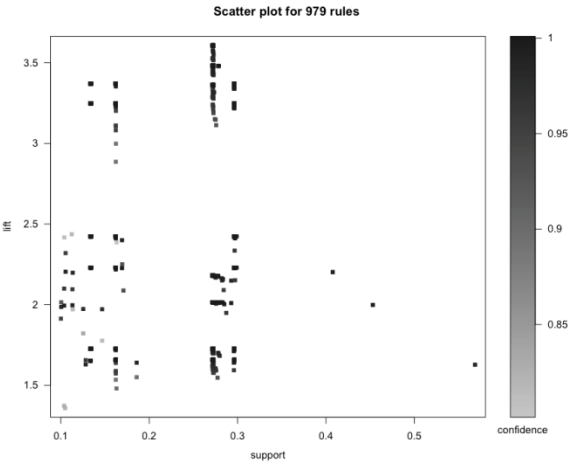


Graph 1. Relation between confidence and support for the 979 generated association rules and their bounds



Graph 2. Number of rules and the relation between lift and support parameters

An analysis of the results showed that rules with a higher confidence value have a support measure of between 0.2 and 0.3, as shown in Graph 3.



Graph 3. Number of rules and the relation between lift, confidence, and support

After analyzing the results, the data mining was refined by considering which of the rules found thus far lack semantic significance, according to experts on fish populations.

The top 10 rules, shown in Chart 1, can thus be interpreted as follows. The value before the symbol "==>" indicates the support of the rule, that is, the number of items covered by a premise(s). The value that appears after the consequent attribute is the number of items for which the consequent of the rule is valid. The confidence value of the rule is given in parentheses. Thus, we can read Rule 1 as follows: "if transparency=6 and pH=5 then depth=7".

1. transparency=6 ph=5 1458 ==> depth=7 1458 conf:(1)
2. dissolvedOxigen=12 transparency=6 1456 ==> depth=7 1456 conf:(1)
3. dissolvedOxigen=12 ph=5 1456 ==> depth=7 1456 conf:(1)
4. dissolvedOxigen=12 ph=5 1456 ==> transparency=6 1456 conf:(1)
5. dissolvedOxigen=12 transparency=6 1456 ==> ph=5 1456 conf:(1)
6. dissolvedOxigen=12 transparency=6 ph=5 1456 ==> depth=7 1456 conf:(1)
7. depth=7 dissolvedOxigen=12 ph=5 1456 ==> transparency=6 1456 conf:(1)
8. depth=7 dissolvedOxigen=12 transparency=6 1456 ==> ph=5 1456 conf:(1)
9. dissolvedOxigen=12 ph=5 1456 ==> depth=7 transparency=6 1456 conf:(1)
10. dissolvedOxigen=12 transparency=6 1456 ==> depth=7 ph=5 1456 conf:(1)

Chart 1. Ten best generated rules.

We can observe that the attribute stage has not been found among the top 10 rules. The attribute stage is included in rules 17 and 22 (Chart 2), which have very high confidence values of 0.99 and 0.96, respectively.

17. transparency=6 stage=pre 524 ==> depth=7 518 conf:(0.99)
22. depth=7 stage=pre 541 ==> transparency=6 518 conf:(0.96)

Chart 1. Association rules with stage attribute

These rules, presented in Chart 2, answer the initial objective of the research, which was to determine the existence of any relationship between abiotic and biotic factors (in this case, the larval stage). The two rules were validated by experts in fish fauna as being important for understanding the process of spawning fish.

6 Final Considerations

The association rules found in this work are consistent with the reality of fish fauna found at the sampling sites, and were semantically validated by ichthyology experts. The data used in this study had previously been analyzed using statistical methods, but no relationships between biotic and abiotic factors were determined. The application of data mining techniques identified new association rules, providing new insights into Amazon fish fauna.

During this research, we found many errors in the input data and/or an inability of existing software to perform the necessary tasks. For the development of future work, and for the application of data mining techniques to other ichthyofauna databases, we are developing specific software to collect data from samples, as well as information visualization routines and decision support systems.

References

1. Hochachka, W.M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., Kelling, S.: Datamining discovery of pattern and process in ecological systems. Journal of Wildlife Management 71(7), 2427–2437 (2007)
2. Breiman, L.: Bagging predictors. Journal Machine Learning 24, 123–140 (1996)
3. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Journal Machine Learning 36, 105–139 (1999)
4. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York (2009)
5. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), pp. 487–499 (1994)
6. Agostinho, A.A., Marques, E.E., Agostinho, C.S., Almeida, D.A., Oliveira, R.J., Melo, J.R.B.: Fish ladder of Lajeado Dam: migrations on oneway routes? Neotropical Ichthyology 5(2), 121–130 (2007)
7. Empresa de pesquisa energética – EPE: Plano Decenal de Expansão de Energia 2021. MME/EPE, Brasília (2012)
8. Nakatani, K., Agostinho, A.A., Baumgartner, G., Bialetzki, A., Sanches, P.V., Makrakis, M.C., Pavanelli, C.S.: Ovos e larvas de peixes de água doce: desenvolvimento e manual de identificação. EDUEM. Maringá, 378 p. (2001)
9. Tanaka, S.: Stock assessment by means of ichthyoplankton surveys. FAO Fisheries Technical Paper, vol. 122, pp. 33–51 (1973)
10. Agrawal, R., Imielinski, T., Swami, A.: Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering, Special Issue on Learning and Discovery in Knowledge-Based Databases 5, 914–925 (1993)

11. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3), Article 9, 9–es (2006)
12. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowledge and Information Systems 14(1), 1–37 (2007)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
14. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco (2011)
15. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. Ph.D thesis, Waikato University, Hamilton, NZ (1998)
16. Tetko, I.V., Solov'ev, V.P., Antonov, A.V., Yao, X., Doucet, J.P., Fan, B., Hoonakker, F., Fourches, D., Jost, P., Lachiche, N., Varnek, A.: Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure–Property Relationship Studies of Metal Complexation with Ionophores. Journal of Chemical Information and Modeling 46, 808–819 (2006)
17. Ghiselli, E.E.: Theory of psychological measurement. McGraw-Hill (1964)
18. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: ICML 2000 Proceedings of the Seventeenth International Conference on Machine Learning, pp. 359–366 (2000)
19. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco (2011)