Data mining techniques for water ecotoxicity classification for application on water resources management

Leonardo Bertholdo*

Faculty of Technology,
University of Campinas,
Limeira, São Paulo, Brazil
and
Center for Research and Development in Telecommunications,
Campinas, São Paulo, Brazil
E-mail: leo.btd@gmail.com
*Corresponding author

Celmar Guimarães da Silva,
Gisela de Aragão Umbuzeiro and
Luiz Camolesi Jr.

Faculty of Technology,
University of Campinas,
Limeira, São Paulo, Brazil
E-mail: celmar@ft.unicamp.br
E-mail: giselau@ft.unicamp.br
E-mail: camolesi@ft.unicamp.br

Abstract: Among the various forms of action that promote sustainability, technological innovation can be considered one of the most important. This paper applied data mining techniques to discover knowledge in the field of water quality monitoring data, providing useful and relevant support for decision-making in environmental management systems. At the current stage of research, a predictive modelling technique, known as rule-based classification, was used to find rules that can, based on the values of certain chemical parameters, predict the ecotoxicity level of a water sample. We used data from water analyses from main water bodies of São Paulo state in Brazil, from 2005 to 2010. We expect to get a reliable, fast and effective way to predict the ecotoxicity levels of water in rivers, lakes and reservoirs based on analyses of chemical parameters, or indicate the complementarity of these measurements for optimisation of monitoring networks and the consequent improvement natural resources management.

Keywords: water quality monitoring; water ecotoxicity; water bodies; chemical parameters; water resources management; sustainable development; Brazil; environmental management systems; knowledge discovery in databases; data mining; predictive modelling; rule-based classification; support for decision-making.

Reference to this paper should be made as follows: Bertholdo, L., da Silva, C.G., Umbuzeiro, G.A. and Camolesi Jr., L. (2014) 'Data mining techniques for water ecotoxicity classification for application on water resources management', *Int. J. Environment and Sustainable Development*, Vol. 13, No. 4, pp.408–424.

Biographical notes: Leonardo Bertholdo graduated in Information Technology and is currently Master in area of Information Engineering in the Faculty of Technology, at the University of Campinas (UNICAMP), Limeira – Brazil. He worked for six years as a Systems Analyst at the Center for Research and Development in Telecommunications, in Campinas. He has special interest in applying data mining techniques in environmental databases.

Celmar Guimarães da Silva completed his Doctoral degree in Computer Sciences in 2006, at the Institute of Computing, University of Campinas (UNICAMP), Limeira – Brazil. He is a faculty member at the Faculty of Technology, University of Campinas, since 2008. He also coordinates a computer technology course in this university since 2011. His research interest is related to the information visualisation area. He is a member of the Brazilian Computer Society.

Gisela de Aragão Umbuzeiro received her PhD in Molecular Biology at University of Campinas (UNICAMP), Limeira – Brazil. She is a Professor at the Faculty of Technology since 2009. She is a former President of SBMCTA – Brazilian Society of Environmental Mutagenesis, Carcinogenesis and Teratogenesis. She worked for the São Paulo Environmental Agency for 22 years. Her main interests are environmental and regulatory toxicology.

Luiz Camolesi Jr. is a Professor in the Faculty of Technology, University of Campinas (UNICAMP), Limeira – Brazil. He holds a PhD in Computational Physics from the Physical Institute at University of São Paulo (USP). Since the 1990s, he is a stakeholder of international scientific associations, conferences and workshops referee. His research interests include information engineering, database evolution, data mining and information quality.

This paper is a revised and expanded version of a paper entitled 'Data mining techniques for water ecotoxicity classification for application on water resources management' presented at VIII National Conference on Excellence in Management, Niterói, Brazil, June 2012.

1 Introduction

Water is essential for life existence and maintenance, as it constitutes the main component of living beings. In the human context, besides promoting survival, water also allows a great number of activities, such as public and industrial supply, agricultural irrigation, energy production and recreation activities. Nevertheless, demographic and industrial expansion in the last decades has affected several water bodies, such as rivers, lakes and reservoirs.

Freshwater is a limited natural resource, considering the high cost to obtain it from less conventional ways, such as seawater and groundwater. Therefore, the rational use of surface water and its quality control are totally relevant to preserve such a fundamental good (Alves et al., 2008).

Currently, the Environmental Agency of the State of São Paulo – CETESB is responsible for monitoring and survey on water body quality. Approximately 350 water sampling sites are monitored. Each sample is analysed according to physical, chemical and biological aspects, making up a dataset rich in information related to the environmental conditions of these water bodies.

The individual assessment of these data might not produce new relevant information, reason why it is fundamental the use of methodologies that allow synthesising the numbers surveyed into comprehensible, significant information as to enable the understanding of environmental sustainability of watersheds. At the moment, there are several 'index' to indicate the quality of water systems. The idea is to gather several parameters in a single number, such as the IVA (water quality index for protection of aquatic life), index used by CETESB, which considers for its calculation variables that specifically impact aquatic life, such as metals, dissolved oxygen, pH and toxicity (CETESB, 2011).

The toxicity of a water sample is also measured by ecotoxicological tests, which consist of determining the toxic effects in water organisms caused by one or more chemical agents. Two result types can be obtained. Acute toxicity is more drastic, caused by high concentration of chemical agents and in general occurs in a short period of expositions of the organisms. Chronic toxicity is more subtle, caused by low concentration of dissolved chemical agents. It is detected in longer periods of exposure or through adverse physiological responses in the reproduction and growth of living organisms (CETESB, 2011).

1.1 Objective

The general objective of research is to discover ecotoxicity classification patterns from monitoring data surveyed by CETESB between 2005 and 2010. Once discovered, these patterns might be used in toxicity prediction of future water samples, minimising the use of living organisms in ecotoxicological analyses in order to make the monitoring activity quicker. Furthermore, these patterns should be used to find the group of parameters/values adopted are poor to perform this prediction; and indicating the need of additional analyses or changes in the current patterns. This specific study aims at analysing the preliminary results related to the patterns identified by data mining techniques applied herein.

1.2 Methodology

The methodology applied in this research is based on the process known as knowledge discovery in databases (KDD), which is divided into five main stages: selection of rough data; selected data pre-processing; transformation; mining of transformed data; interpretation and evaluation of patterns found through mining. In the first two stages and the last stage, the research has the participation of a specialist on environmental sanitation with the aim to support data choice and preparation as well as the evaluation of results obtained.

In the data mining stage, the predictive modelling approach was applied. One of the central approaches of this discipline is to build a model to predict the value of a given attribute based on values of other attributes of the dataset. This modelling was carried out

through a rule-based classification technique, in which database records are classified from rules obtained through a learning mechanism.

In the scope of water quality data, each monitoring database record is represented by the analysis of a water sample collected in a given site of a water body, on a certain date, analysed according to several chemical parameters. In this context, the objective of the technique is to find rules which can define, based on the values of these parameters, the toxicity level of each water sample.

The choice of the mining technique was performed from a bibliographic research with the aim to survey methods already used in the environmental field with a strong concern with management of natural resources and sustainable development. The sequential covering technique was considered one of the most appropriate to our research because it allows extracting classification rules directly from data, opposed to other methods which extract rules indirectly, such as decision tree induction and neural networks.

In this study, the extracted classification rules are represented by the conditional '*If* < *values of chemical parameters* > *then* < *toxicity value* >', in which toxicity is the class to be attributed to test records. The latter are the water sample analyses that intend to predict the toxicity.

The performance of the classification rules was evaluated by two-part cross-validation method. The database is split into two subgroups with similar record quantity. First, one of the subgroups is used as training base from where the classification rules are extracted. Then the extracted rules are applied into other subgroup that plays the role of a test base.

Finally, the error rate of the rules applied into this test base is calculated. Roles are inverted, in a way that the training subgroup becomes the test subgroup and vice versa. The total error rate is then calculated through the average of the two procedures. Based on this rate, it is possible to infer the reliability of the rules produced and consequently to know how much they can be useful when applied in water quality management systems.

The preliminary results of this research are presented in this article, starting from Section 2, which show a brief history on water resources management and monitoring in Brazil, especially in the State of São Paulo. Section 3 describes the knowledge discovery process with highlight to its main stage: data mining. Section 4 presents the application of the rule-based classification technique into water monitoring data, as well as the preliminary results obtained. Section 5 discusses the final considerations related to this research.

2 Water resources management

Watershed management is becoming relevant in Brazil as environmental degradation available water resources rises (Jacobi and Barbi, 2007). The Brazilian Federal Constitution of 1988 gave fundamental importance to the society's participation in natural resources management, especially water management, becoming essential to lead all public policies in the sector. In the State of São Paulo, the State Constitution of 1989 had already incorporated new concepts to the water resources issue: decentralised, participative and integrated management; and multiple uses of water resources.

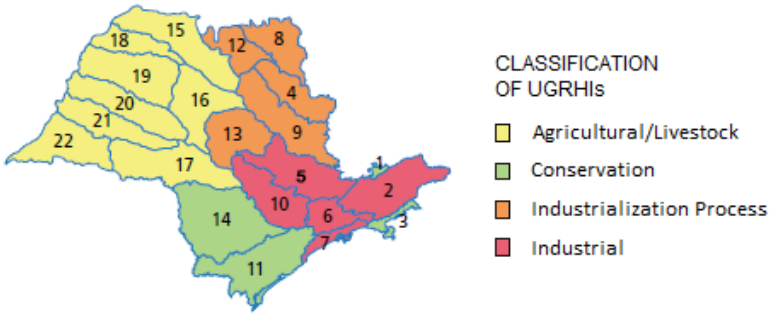
In 1991, the federal government sent to the National Congress the first law project regarding the National Policy on Water Resources. In the same year, the State of São Paulo instituted through the Law No. 7663 the State System of Water Resources. The

territory of São Paulo was divided into 22 water regions and the watershed management was established with the effective participation of the civil society in the decision-making process.

The state law reinforced dispositions of the Water Code of the Federal Constitution as it contemplated management instruments, such as the Watershed Plan, the charging for water use and the State Fund for Water Resources. The decision about the use of this funding is made by Watershed Committees – collegiate with deliberative power that gather representatives from municipalities, state departments and civil society organised for integrated, decentralised and participative water management (SOS Mata Atlântica Foundation, 2012).

In the State of São Paulo the Environmental Agency – CETESB is responsible for the control, inspection, monitoring and licensing of polluting activities, with the main concern to preserve and recover the water quality, air and soil (CETESB, 2012). Since 1974, CETESB has collected information on the fresh water quality in the State of São Paulo through a wide monitoring network distributed along 22 delimited Water Resources Management Units (UGRHIs). Each UGRHI has different sampling sites, from which water collected and analysed (CETESB, 2011). Figure 1 shows this division, classifying the UGRHIs into groups according to their respective vocations.

Figure 1 Classification of the 22 UGRHIs by vocations (see online version for colours)



Source: Adapted from CETESB (2011)

Each UGRHI has a number of sampling sites. In each site, a parameters group is analysed, which are related to physical, chemical, microbiological, hydrobiological and ecotoxicological aspects of the water. Annually, CETESB publishes on its web page the analyses performed in each sampling site in PDF files. Solely the basic network, which aims specifically at analysing the state water bodies, produces an annual data volume of 65,000 analyses (CETESB, 2011). Each analysis corresponds to a measuring of a parameter in a sampling site carried out on a specific date.

These analyses are compared to CONAMA Regulation No. 357/2005, an environmental regulated issued by the National Council for the Environment (CONAMA, 2005). This norm specifies classes of water offers environmental directives for its implementation and establishes conditions and patterns for the discharge of effluents (Umbuzeiro and Lorenzetti, 2010). This norm defines five classes for freshwater: special, 1, 2, 3 and 4. Each class presents a group of water conditions that if in compliance the water will serve for specific current and future uses. (Von Sperling, 2007).

Monitoring data are important tools for sustainable development management to warrant future generations needs (Brundtland, 1987). Monitoring data can be transformed in environmental indicators which can serve as 'sustainability indicators'. According to Maranhão (2007), the latter represent a deepening of environmental indicators as they integrate economic, social and environmental indicators, once sustainable development require an integrate view of the world.

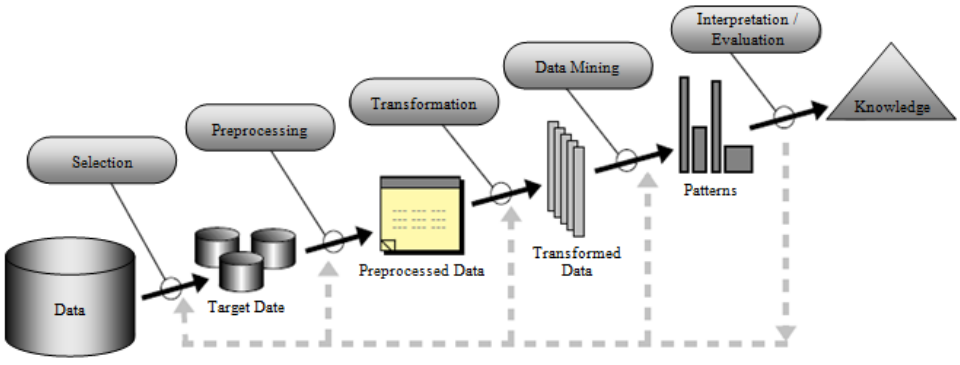
3 Knowledge discovery in databases

The capacity of an organisation to make decisions is frequently associated to the knowledge about its data domain. One of the problems experienced by information analysts is to transform data into relevant information for decision-making (Silva, 2007). The analyses carried out by CETESB create a valuable group of information about the quality of surface water. However, if they are analysed through conventional techniques, the discovery of inputs that might help the decision-making process becomes quite unlikely to happen.

In the last decades, processes that might help in the discovery of non-trivial information in great databases have been developed. They might offer a more representative, broader signification to the existing data in such databases. Among these processes, the KDD might be one of the most well-known and disseminated in the computing field.

According to Fayyad et al. (1996), KDD is a non-trivial process to identifying valid and new knowledge, potentially useful and comprehensible patterns in databases. This process is made up of stages since the selection of the dataset to be studied until the interpretation of the patterns e rules generated by approaches, such as data mining. Figure 2 presents five stages that comprise the KDD process.

Figure 2 Stages that compose the KDD process (see online version for colours)



Source: Fayyad et al. (1996)

In the selection stage, the dataset to be studied is chosen, comprising all variables that have the chance to be used during the process. In the pre-processing stage, adjusts are carried out in the selected dataset, such as: elimination of redundant data, recovery of incomplete data and treatment of discrepant data (outliers). The transformation stage

comprises the standardisation and centralisation of data selected and cleaned in the previous stages, in a way to reduce the processing time of the mining mechanisms.

The data mining stage is when algorithms are implemented, which are intelligent mechanisms responsible for the survey of implicit patterns and rules among the dataset. Finally, interpretation and evaluation verify the results obtained in the mining stage, with the aim to understand the meaning and relevance of the information discovered (Prass, 2004).

During most part of this process, it is essential the follow-up of a specialist in the domain, whose skills help decisively in the choice of the dataset to be studied, in the definition of the knowledge type to be discovered. Also, how such knowledge might contribute to support decisions (Duarte et al., 2011).

In data mining might the knowledge is obtained searching patterns and relationships among variables and their data. According to Berry and Linoff (2004), data mining consist of exploring and analysing great data quantity, with the aim to discover significant patterns and rules. In order to reach this objective, data mining uses techniques from different fields of knowledge, such as: statistics, databases, recognition of patterns, artificial intelligence, information visualisation, machine learning, among others. At the moment, this approach has been applied in the most diverse scenarios, such as: academy, finance, commerce, marketing, medicine, genetics, telecommunications and environment.

In the field of environmental management, KDD method has showed to be useful in promoting directives to transform rough data into information with strategic value. According to Silva (2007), the knowledge discovery in environmental monitoring databases using data mining techniques to evaluate water quality might be an important tool for the decision-making process. In terms of knowledge discovery process, the current state of this work regarding predictive modelling role is undergoing the stages of data mining and the interpretation and evaluation of results. The first results are obtained from the rule-based classification technique and analysed according to aspects of significance and relevance.

3.1 Related works

There are several works related to the use of data mining in the monitoring data classification of water resources. They aim at basically offering inputs that might help in the decision-making process and define future public policies on the sustainable management of these resources. Fernandes and Duarte (2009) presented a data warehousing system to store water quality data of a region in Portugal. Besides organising and standardising information in a database, the tool seeks to help knowledge discovery through the application of data mining techniques, such as classification and linear regression.

Magaia (2009) discussed the role of support systems to make decisions concerning the analysis of water quality. The author proposes the development of a system, which was applied in a wastewater treatment plant. The tool had the objective to collect and offer structures and ways for the multidimensional data exploration, as well as for the classification and generation of models through data mining mechanisms.

Seixas et al. (2008) investigated the correlation of space and time data that comprise the group of pollutants in the Rodrigo de Freitas Lagoon in Rio de Janeiro. The main objective was to obtain a methodology to classify water quality that can be applied in

other water bodies. The research included several knowledge discovery stages that were implemented to reach targets, as well as the use of data mining techniques to group and classify data.

Karimipour et al. (2005) investigated geospatial data mining for water quality management. A case study carried out in the region between Azerbaijan and Iran showed the correlation between pollution in big industrial centres and water quality indicators through geospatial data mining. It was clear the relationship between the quality and location of industrial pollution and the water quality indicators. Our work aims to find classification rules that allow to predict the water toxicity.

4 Discovery process of toxicity classification rules

The approach applied in this study was based on a KDD process which is considered a consolidated method in the technology field. This section presents the different stages of the work, since selection and preparation of rough water quality monitoring data, the mining stage of pre-processed data until analysis stage of the preliminary results obtained.

4.1 Data pre-processing

The application of data mining techniques has as a premise that the data to be researched should be pre-selected, uniform, normalised, centralised and with a satisfactory level of completion. Data preparation activities have the aim to optimise the significance and reliability of the results on the data mining stage.

Other important benefit of data pre-processing is the reduction of possible impacts in the mining performance with the consequent reduction of computing efforts for search for implicit and useful information in the dataset. In this study, the pre-processing stage comprised activities of data selection, transformation, centralisation, imputation and discretisation, which will be presented in the following sections.

4.1.1 Data selection

In the KDD process, data selection occurs before the pre-processing stage. However, some authors, such as Tan et al. (2009) consider selection part of this stage. This approach was also applied in this article, because all stages before data mining are strongly related and can be grouped in a single pre-processing stage.

In this work, the dataset analysed was selected based on general criteria, related to broader aspects of the data, and on specific criteria associated to more peculiar characteristics of the data. The general criteria and their respective application descriptions follow below:

- General criteria for data selection:
  - *Type of monitoring network* – The basic network sites were chosen, comprising almost 85% of the monitoring network sites of CETESB. This network type has the aim to uniquely evaluate the water in the rivers of the state of São Paulo, not comprising analyses of sediments and balneability of these rivers, nor analyses generated from automatic monitoring systems.

- *Time aspect* – Analyses carried out from 2005 to 2010 were chosen. Besides that CETESB publishes analysis data which took place from 2000, only the last six years available were used in order to restrict the research to the recent reality of water bodies.
- *Space aspect* – From the 22 UGRHIs that exist in the State of São Paulo, only four of them were considered, which are: Paraíba do Sul (2), Piracicaba/Capivari/Jundiaí (5), Alto Tietê (6) and Sorocaba/Médio Tietê (10). The purpose was to choose the most populous UGRHIs with approximately 70% of the inhabitants of the state and strongly industrialised, once the rivers of the regions with this profile usually are really impacted by industrial activity.

After the application of the general criteria, from the 317 sampling sites existing in an average of six years, 165 remained, located in the four UGRHIs selected and part of the basic network of CETESB.

The specific criteria for data selection took into account the completion issue, one of the basic premises so that the data miming stage was well-succeeded. Below are presented each of the specific criteria applied in the data selection as well as the order in which they were applied:

- Specific criteria for sampling sites selection:
  1 only sites of water bodies that have two or more sampling sites
  2 only sites that is present in all years
  3 only sites that have toxicity analysis, considering that this parameter is essential in this study
  4 only sites pertaining to Class 2.

In order to keep data uniformity, four sites were discarded, two pertaining to Class 0 (Special) and two pertaining to Class 3.

After the application of these criteria, from the 165 sampling sites selected based on general criteria, 144 remained, considering the sites with higher data richness and uniformity.

- Specific criteria for quality parameter selection:
  1 parameters which are present in at least 80% of the sampling sites
  2 parameters considered as offering more impact to aquatic life and human health and, consequently, with higher probability to generate relevant information.

The application of these specific parameters resulted in the selection of ten chemical parameters that supposedly could be directly or indirectly related, separately or grouped with toxic effects to the biota, besides toxicity. The parameters are: total cadmium, total lead, dissolved copper, total nickel, nitrate, nitrite, ammoniac nitrogen, dissolved oxygen, tensoactive substance and total zinc.

- Specific criteria for measuring group selection:
  - Only measuring groups of the collection sites and dates that contain the value measured in the toxicity field.

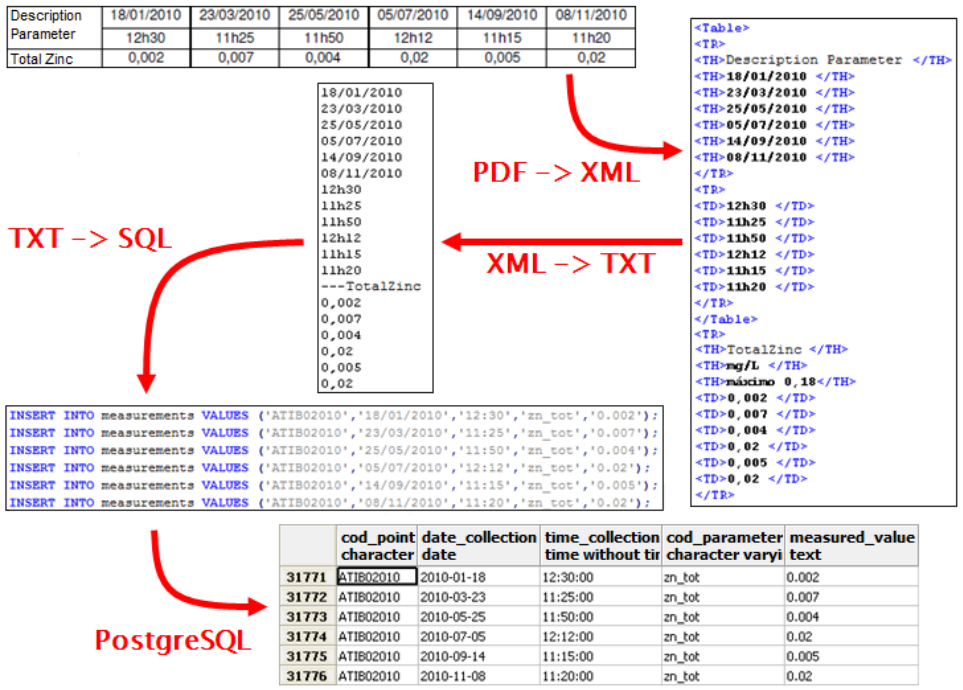
This last criterion eliminated approximately 30% of the measuring groups selected by then. It is important to explain that the term 'measuring group' refers to each group of 'measurements of *n* parameters + toxicity measuring', which is associated to a specific sampling sites and collection date.

4.1.2 Data transformation

After selected, the rough data were centralised in a database created through the database managing system called PostgreSQL. It was necessary to convert data, which were in a PDF file into a format adequate with the database structure. This activity was carried out in several stages and took most part of the pre-processing time, once the original archives has little differences among one to another, demanding treatment. Figure 3 shows the conversion process of the original data until its storage in the database.

First, the PFF archives were converted into the XML format (eXtensible Markup Language) with the support of the Adobe© Acrobat tool. After that, through two converters implemented in the Java programming language, the conversions from XML into text format (TXT) were carried out, and from this format into the SQL format (Structured Query Language). At last, The SQL commands produced were executed, allowing the insertion of data in the database preciously created in the PostgreSQL.

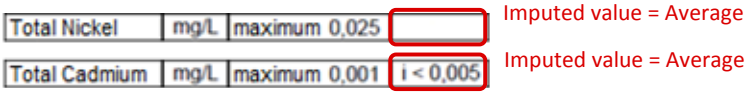
Figure 3 Conversion scheme of rough data (see online version for colours)



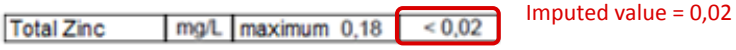
4.1.3 Imputation of missing data

The absence of values for some parameters or the inaccuracy of some might cause interferences in data mining and consequently generate distorted results. The most radical solution in such cases is the complete removal of the record, even if it has only one of the attributed with a missing value. To avoid the reduction of the quantity of valid groups, an imputation was applied. It consists in attributing values to the parameters based on one or more criteria.

In measurements with missing values or where it was not possible to detect if the value was under or above the standard establish by the CONAMA Regulation No. 357/2005 (CONAMA, 2005), the value was ignored and monthly average value of the parameter in six years was imputed (2005-2010). Examples:



In measuring below the CONAMA Pattern, although without an exact known value, the measured value was imputed. Example:



4.1.4 Data discretisation

Usually, the classification mechanisms require that continuous attributes be categorised through discrete values, process that is called discretisation. According to Tan et al. (2009), the best discretisation approach is the one that generates the best result for the data mining technique to be used. The conversion of a continuous attribute in discretion involves two tasks: to define how many categories must exist and how will be carried out the mapping of the continuous values to discrete values.

Table 1 Discretisation of continuous parameters

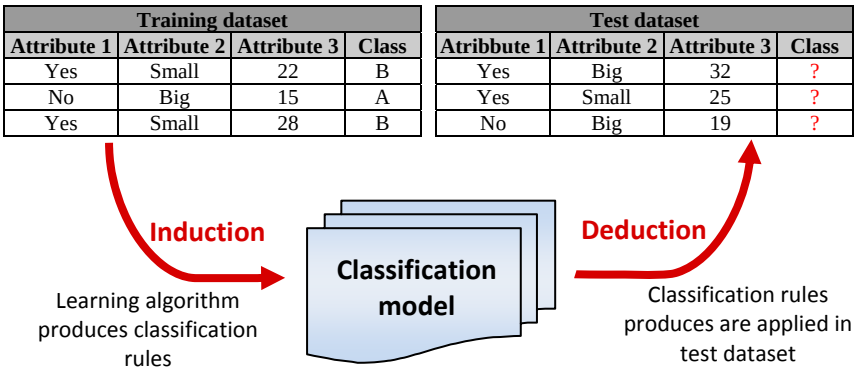
Continuous parameters	*Mnemonic*	*Description of discretised values*
Total cadmium, total lead, dissolved copper, total nickel, nitrate, nitrite, ammoniac nitrogen, dissolved oxygen, tensoactive substance, total zinc.	PC	CONAMA standard – within CONAMA standard
	AC	Above – above the CONAMA standard in up to three times.
	MA	Highly above – above the CONAMA standard in higher than three times.
Discrete parameter	*Mnemonic*	*Description of discrete values*
Toxicity	NT	Non-toxic – absence of physiological response of the micro crustacean *Ceriodaphnia dubia*.
	CR	Chronic – physiological response of the micro crustacean *Ceriodaphnia dubia*.
	AG	Acute – strong physiological response of the micro crustacean *Ceriodaphnia dubia*.

The discretisation of water quality monitoring data was performed empirically through data visual inspection, without the need of more refined supervised techniques. Table 1 shows how chemical parameters were discretised considering the CONAMA standards, the toxicity as a previously discretised parameter as of the response of a living organism as well as the mnemonics used to identify database values.

4.2 Rule-based water toxicity classification

The classification technique based on rules is a data mining approach that seeks to build a model from a group of records previously labelled, capable to classify records of other groups not labelled yet. Figure 4 illustrates the construction of a rule-based classification model. In the initial stage, a training dataset containing records with known classes is selected.

Figure 4 Construction of a rule-based classification model (see online version for colours)



This dataset is used as input to build a classification model, which is simply the group of classification rules found. In the next moment, this model is applied in a test dataset containing records with unknown classes. Finally, the model performance is evaluated based on the rate of errors produced in the classification of the records in the test base.

In the framework of this research, the algorithm learns a group of conditional rules from the training database, where each rule comprised by an antecedent, also called pre-condition, which contains the values of the chemical parameters already mentioned, and a consequent, related value of the predicted class, in the case this study the toxicity. After that, these rules learned through the algorithm are applied to the test base, in a way of attributing a toxicity value to each measuring group of this base. An example of rule produced would be:

$$\text{If } Nickel = AC \text{ and } Lead = AC \text{ and } Zinc = MA \text{ then } Toxicity = CR$$

To evaluate the quality of a classification rule, there are basic measures, such as covering and precision. The first one has the objective to determine the rate of registrations that fit in the antecedent of the rule and, therefore, discharge this rule. The second one defines the rate of registrations that fit both the antecedent and the consequent. Therefore, discharge this rule and also pertains to the class predicted by the rule. The calculations of these measures can be expressed as the following:

$$\text{Coverage} = \frac{\text{Records that satisfy the antecedent of rule}}{\text{Total number of records.}}$$

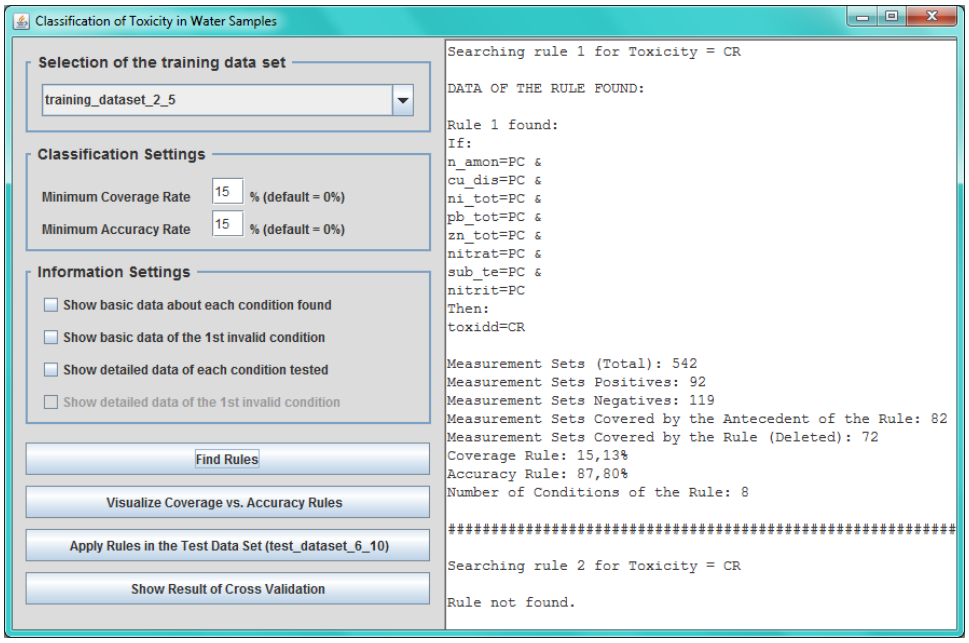
$$\text{Accuracy} = \frac{\text{Records that satisfy the antecedent and consequent of the rule}}{\text{Records that satisfy the antecedent of rule.}}$$

The classification model was generated by a sequential covering algorithm, which makes a search for the best rules to predict each class, in the case, toxicity values: non-toxic (NT), chronic (CR) and acute (AG). During the search for rules, all measuring groups with class equal to the one that is being researched are considered positive, and all other groups are considered negative. One rule is considered satisfactory if it covers the majority of the positive groups and a few of the negative groups.

4.3 Preliminary results

As to configure the classification processing and to visualise the results produced by this processing stage, it was implemented a tool in Java programming language that its main interface is presented in Figure 5. This interface can be divided into two parts: the control panel, on the left hand side, which related to classification configurations and visualisation as well as to the command buttons; and the processing area, on the right hand side, where the processing results can be visualised.

Figure 5 Tool for search of water ecotoxicity classification rules (see online version for colours)



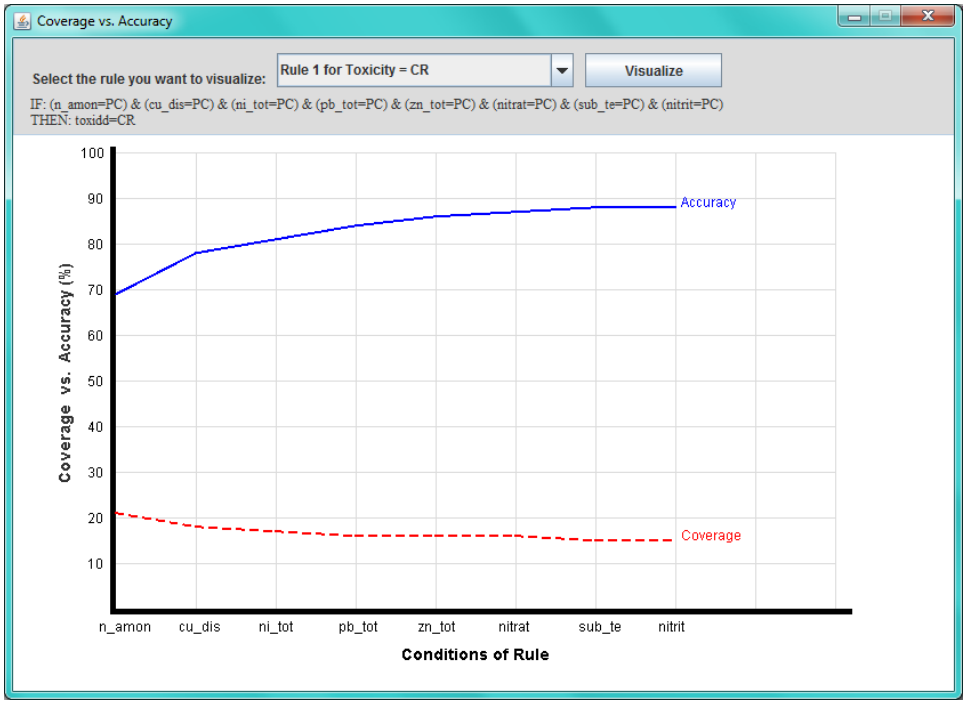
Note: In_amon, cu_dis, ni_tot, pb_tot, zn_tot, nitrat and sub_te refer to, respectively, the parameters: ammonium nitrogen, dissolved copper, total nickel, total lead, total zinc, nitrate and tensoactive substance.

Before starting the classification, it is necessary, to select the training database, which will serve as input for the algorithm learning. Then, it is possible to configure the minimum covering and precision rates that must be considered in the search for rules; if they are not configured, a search for all possible rules is made, regardless of their covering and precision rates. It is also possible to define some visualisation options that allow configuring up to four detailing levels of the processing information. At last, the button search rules start the process of search for water toxicity classification rules.

After generating the rules, the behaviour of the covering and precision rates can be visualised during the formation of each rule found. Through a linear graphic, this functionality allows to evaluate in a quick, efficient way the performance of each rule produced, then helping the decision-making process of which rules must be considered or discarded for water toxicity classification.

Figure 6 shows how covering and precision rates tend to follow opposed directions as the rule is being increased with new conditions (or parameters). This phenomenon indicates that, in general, the higher the precision of a rule, the lower its covering will be, and vice versa.

Figure 6 Visualisation of covering and precision of the rules produced in this study (see online version for colours)



Once the rules are generated, they are applied in the test base. At that point, the quantity of measuring groups incorrectly classified are calculated and presented, as well as the precision rate of the rules applied. It should be noted that this verification is possible because the classes of the measuring groups are known both in the training base and in the test base, essential characteristic for the application of the two-part cross-validation method. The next step consists of exchanging the roles of the two bases used and

repeating the same procedure, in a way that the training base becomes the test base and vice versa. Finally, the performance of the rules produced by the two interactions can be evaluated through the button visualise cross-validation result.

The preliminary results indicate that the maximum precision rate reached by the classification rules produced was of approximately 77%. This means that for each 100 measuring groups classified by the rules produced, in 23 of them toxicity (non-toxic, chronic or acute) was incorrectly classified.

Besides the considerable error rate, the results were not what we expected because the classification rule obtained were associated, in their majority, to the 'non-toxic' toxicity value, while it was expected to obtain rules that predicted toxic values, such as 'chronic'" or 'acute'. Even that the tool has produced some rules for 'chronic' toxicity, it was not possible to generate rules for 'acute' toxicity, as all possible rules tested always produced more mistakes than hits.

5 Conclusions

This article presented the use of specific data mining techniques for knowledge discovery in the domain of water quality monitoring. During this research, we noted the level of relevance of the subject discussed for sustainability management. There is a great volume of research related to the application of computing science in the environmental area, especially in water resources management, fact that shows the great concern of the scientific community with the future of our watersheds.

The data selected for the study comprised a significant sample of the water quality monitoring data produced in the State of São Paulo. However, the selected dataset had to be drastically reduced in relation to the original set. One of the reasons was the great quantity of incomplete measurements, once the essential parameters for this research did not present a measured value. As a consequence, the data reduction strategy was adopted to preserve the quality of the dataset, as the mining result is directly related to this factor.

The knowledge discovery is intrinsically an exploratory and interactive process, demanding several adjustments and new interactions and experiments in the search for patterns among data. For this reason, is will be necessary to re-evaluate particularly the group of selected parameters as well as data imputation and discretisation, once they can significantly influence mining responses.

Even though the techniques used did not produce relevant water toxicity classification rules, the initial results of this research show the potential that data mining has in the support of implicit information extraction in water quality monitoring data. The inability to create rules for acute toxicity, for example, shows that the chemical parameters used in the concentrations measured might not influence a measuring at a site in which it reaches this toxicity level. Information like this might represent valuable subsidies for decision-making regarding water resources management and environmental sustainability.

References

Alves, E.C., Silva, C.F., Cossich, E.S., Tavares, C.R.G., Souza Filho, E.E. and Carniel, A. (2008) 'Evaluation of water quality of the river basin Pirapó – Maringa, Parana State, through physical, chemical and microbiological parameters', *Acta Scientiarum – Technology*, Vol. 30, No. 1, pp.39–48, Maringá, Brazil (in Portuguese).

Berry, M.J.A. and Linoff, G.S. (2004) *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 672p, Wiley Publishing, Inc., Indianapolis.

Brundtland, G.H. (1987) *Our Common Future – Report on the World Commission on Environment and Development*, United Nations Environmental Programme, New York.

CETESB (2011) *Report of Surface Water Quality State of São Paulo*, CETESB, São Paulo [online] http://www.cetesb.sp.gov.br/agua/aguas-superficiais/35-publicacoes-/-relatorios (in Portuguese).

CETESB (2012) *Institutional – CETESB – Environmental Agency of the State of São Paulo – History* (in Portuguese).

CONAMA (2005) National Council on the Environment. Resolution No. 357, 17 March 2005, CONAMA Brasília [online] http://www.mma.gov.br/port/conama/res/res05/res35705.pdf (in Portuguese).

Duarte, A.A.A., Bertholdo, L., Umbuzeiro, G.A., Camolesi Junior, L. and Silva, C.G. (2011) 'Processing and data visualization for knowledge discovery in systems monitoring water quality', in *III Workshop on Applied Computing for Management of the Environment and Natural Resources*, Natal, pp.1409–1418 (in Portuguese).

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery: an overview', in *Advances in Knowledge Discovery and Data Mining*, Vol. 17, No. 3, pp.37–54, AAAI Press/The MIT Press, England.

Fernandes, J. and Duarte, A.S. (2009) 'A data warehousing system for area water quality', 16p, University of Minho, Portugal [online] http://www3.di.uminho.pt/~prh/uce15-0809/g16.pdf (in Portuguese).

Jacobi, P.R. and Barbi, F. (2007) 'Democracy and participation in water resources management in Brazil', Katálysis, Florianópolis, Brazil, Vol. 10, No. 2, pp.237–244 (in Portuguese).

Karimipour, F., Delavar, M.R. and Kinaie, M. (2005) 'Water quality management using GIS data mining', *Journal of Environmental Informatics*, Vol. 5, No. 2, pp.61–72, Canada.

Magaia, L.P.T. (2009) *The Role of Decision Support Systems for the Analysis of Water Quality* Master thesis, University of Minho, Portugal (in Portuguese).

Maranhão, N. (2007) *Indicator System for Planning and Management of Water Resources Watershed*, 422p, PhD thesis, Federal University of Rio de Janeiro, Rio de Janeiro (in Portuguese).

Prass, F.S. (2004) 'KDD: Process of knowledge discovery in databases', *Interest Group on Software Engineering*, Florianópolis, Brazil, Vol. 1, No. 1, pp.10–14 (in Portuguese).

Seixas, J.A., Lima, B.S.L.P. and Ebecken, N.F.F. (2008) 'Mining spatial and temporal data to classify water quality: a case study', in *Data Mining IX: Data Mining, Protection, Detection and Other Security Technologies*, Chapter 9, pp.83–91, Wit Press.

Silva, I.A.F. (2007) *Knowledge Discovery in Environmental Monitoring database for Assessment of Water Quality*, 134p, Master Thesis, Federal University of Mato Grosso, Cuiabá, Brazil (in Portuguese).

SOS Mata Atlântica Foundation (2012) 'A public policy for water' [online] http://www.rededasaguas.org.br/politicas-publicas/ (in Portuguese).

Tan, P., Steinbach, M. and Kumar, V. (2009) *Introduction to Data Mining*, 900p, Publisher Ciência Moderna, Rio de Janeiro.

Umbuzeiro, G.A. and Lorenzetti, M.L. (2009*) Fundamentals of Water Quality Management* Resolution CONAMA 357/2005, Unicamp/CPEA, Limeira, Brazil (in Portuguese).

Von Sperling, M. (2007) *Studies and Modeling of Water Quality in Rivers*, Department of Sanitary and Environmental Engineering – Federal University of Minas Gerais Belo Horizonte, Brazil, Vol. 7, 588p (in Portuguese).