Classifier for Drinking Water Quality in Real Time

Jorge Camejo, Osvaldo Pacheco
Institute of Electronics and Telematics Engineering
University of Aveiro, Campus Universitário de Santiago,
3810-193. Aveiro, Portugal
jcamejo@ua.pt

Miguel Guevara
Institute of Mechanical Engineering
University of Porto. Rua Dr. Roberto Frias no 400.
4200-465. Porto. Portugal

Abstract—**Real time features are critical for automatic assessment of Drinking Water Quality (DWQ). This paper explores the use of real time features to feed machine learning classifiers for DWQ. Two different representative datasets were used from: a) The Provincial Water Quality Monitoring Network from Ontario, Canada and b) National Hydrologic Information System from Central Region of Portugal. The procedure followed in this study was: (1) automatically computing a Water Quality Index to classify the datasets elements in five classes (excellent, good, medium, bad and very bad) using the Kumar method; (2) selecting best performed real time features on results of classified datasets; and (3) exploring machine learning algorithms (e.g. Decision Trees, Artificial Neural Networks and k-Nearest Neighbor) for producing DWQ classifiers. In this work, we perform the classification of two classes (good and medium) out of the five possible categories, due to the absence of vectors in the datasets.**

Keywords- Drinking Water Quality, Hydroinformatics, Data Mining, Machine Learning

I. Introduction

The general quality of water in lakes, rivers and coastal areas is periodically assessed by scientific and environmental institutions. This classification is based on laboratory analysis of water quality samples collected from stations at select locations and data on water samples are stored in computer systems. From these recorded parameters of water, researchers have been able to take important decisions to safeguard the environment in general and their own consumption.

There are several ways to study and evaluate water quality. One of the most used is the Water Quality Index (WQI), not only for its ability to generate understandable classifications, but also for its potential to facilitate behaviour studies over time. In this work a WQI [1] is used for reference classification to feed machine learning algorithms.

The Water Quality Index (WQI) makes available a single number that expresses the overall water quality in a region based on a reduced number of parameters. The idea of this index is to turn complex water quality data into information that is reasonable and quantifiable by the general public. Usually associated with this numerical rating, a qualitative categorization is attributed. This qualitative classification makes the water quality evaluation process even simpler and may be used in the learning process by artificial intelligence algorithms. Some of the advantages of WQI are as follows: (i) Easy to disseminate through the non-specialized people; (ii) more importance than the individual parameter values; (iii) an average of various variables into a unique number combining different measurement units into one unity is represented. The main WQI disadvantage is that some information regarding individual variables and their interactions may be lost.

Many researchers and environmental institutions have presented WQI methodologies or have modified the previous methods, [2], [3], [4], [5] , [1], [6], [7], [8], [9] [10], [11-21], [22], [23].

For the WQI methods the numbers of the parameters to be used are dependent on the water to be classified, i.e. drinking water, environmental protection or estuaries and coastal waters. In general, 6 to 22 parameters are reported to be necessary for the WQI. However, only a few of these parameters can be measured in real time.

The main problems of automatic assessment of drinking water quality are the difficulty of real time measurement of some physical, chemical, and microbiological parameters and the human and financial resources spent to obtain their values. However a real time monitoring would provide continuous data so that daily, seasonal, and event-driven fluctuations are not missed. It makes it possible to immediately recognize changes in water-quality conditions. The main objective of the present study is to propose a drinking water quality classification system based only real time parameters, specificically, we proposed to use four parameters that can be measured in real time but, at the same time, capable of guaranteeing good predictive accuracy for decision making regarding the water quality behaviour. The selected parameters were: pH, Dissolved Oxygen, Nitrates and Temperature.

II. Materials and Methods

In this study two datasets, located in different regions, were collected from 2000 to 2011 as shown in Table I. One from Provincial Water Quality Monitoring Network (PWQMN), Ontario, Canada (DS1), [24], and one dataset from central regions in Portugal (DS2), [25] , [26].

The original datasets do not have a standardized format to allow the manipulation of instances (features vectors). They also do not contain a water quality classification and no option to evaluate the behaviour of quality control is included.

TABLE I. Dates of sampling of each dataset

Central Portugal	Ontario Canada
Sep 2000 - March 2011	Jan 2002 - Dec 2009

Foundation for Science and Technology (FCT)

In order to make an automatic classification as an initial reference for this study a pre-processing based on Extraction, Transformation and Loading (ETL) [27] was necessary. According to this procedure three main steps were applied: (i) the selection of only the required parameters for initial qualification; (ii) the transformation of parameter values in order to unify the different units and dimensions; (iii) the elimination of the vectors with missing values. All datasets were stored in a SQL Sever Database and the ETL process was performed; (iv) a new selection of parameters that: a) can be measured in real time b) in order to create a base dataset of this work. The selected parameters were: (Dissolved Oxygen, pH, Nitrates and Temperature).

After carrying out this process, the number of the vectors decreased from 26120 to 4102 distributed as shown in Fig. 1.

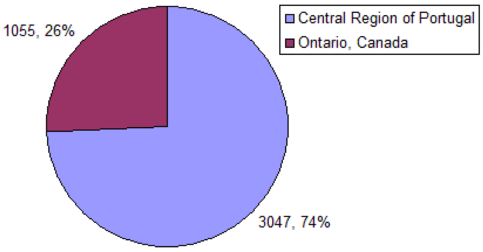

Figure 1. Vectors Distribution by dataset.

A. *Water Quality Index Methods*

The next step was to apply a water quality index method to associate a class to each instance on the dataset, to use as the classification criterion. In general, these types of qualification are developed in three key steps: (i) selection of the variables out of a large number of possible variables; (ii) developing the sub-indices function and weightages; and (iii) aggregation of the candidate variables selected [1]. As referred above there are different methods to classify the water quality.

Brown et. al developed a water quality index [28] comparable in structure to Horton's index, [3] but with much greater rigour in selecting parameters, developing a common scale, and assigning weights for which elaborate Delphi technique was performed. This effort was supported by the National Sanitation Foundation (NSF) from the United States. Brown et al. assembled a panel of 142 experts from various professions throughout the United States with expertise in various aspects of water quality management. The members of the panel were mailed three questionnaires [1].

Many researchers and science institutions in various countries have applied this method to water classification [14], [29] and other researchers have applied these methods to check other methods [30], [31], [32], [23], [12]. The reasons of method selection are [1]: i) brought out the shortcomings and improvements in the formulation of the NSF-WQI, ii) that is able to classify each water sample, on contrary of other methods as the Canadian Environmental Quality Guidelines [6] that classify the water for a period of time, usually for 6 months. Moreover, the Kumar [1] method is based on Brown, [28] and National Sanitation Foundation of the United States (NSF) and requires only 9 parameters (Table II), it is proper for drinking water classification and it has been the most used in the last years.

TABLE II. TABLE BASED ON [1]. WEIGHTAGES CALCULATED IN DIFFERENT METHODOLOGIES. WA-[28], WB-NATIONAL SANITATION FOUNDATION, WC-PROPOSED BY [1].

No.	Parameters	W A	W B	W C
1.	Dissolved Oxygen	0.15	0.17	0.13
2.	Fecal Coliform	0.14	0.16	0.13
3.	pH	0.12	0.11	0.12
4.	Biochemical Oxygen Demand	0.12	0.11	0.11
5.	Nitrates	0.10	0.10	0.10
6.	Phosphates	0.10	0.10	0.11
7.	Temperature	0.10	0.10	0.11
8.	Turbidity	0.08	0.08	0.10
9.	Total Solids	0.09	0.07	0.09

The mathematical formula to compute the WQI based on this method is given by:

$$WQI = \prod \theta\iota * \omega\iota \qquad (1)$$

Where θ_ι is the quality of ith parameter, a number between 0 and 100, resulting from the "mean curve of quality variation" as a function of concentration; and the $\omega\iota$ consists in the weight of ith parameter, a number between 0 and 1 given as a function of its importance for the global water quality (see Table II). The final WQI is a numeric value between 0 and 100. Also the qualitative classification is associated with this index, as it is shown in Table III.

TABLE III. QUALIFICATION EQUIVALENT OF NUMERIC VALUES OF WQI

Range of WQI	Qualification
90 - 100	Excellent
70 - 90	Good
50 - 70	Medium
25 - 50	Bad
0 - 25	Very Bad

The Fig.2, shows the boxplot after the calculation of the WQI for the datasets of this study. The significant global water quality differences between Ontario, Canada and Central Region of Portugal water indexes are observed. Here can be observed that in this formed dataset are represented only two water classes: "MEDIUM" and "GOOD". Another critical point is the distribution of classes where these datasets: 95.8 % of the samples in Portugal regions belong to "MEDIUM" class, while 89.6 % of Ontario samples fit into "GOOD" class. The existence of such unbalanced levels of qualification in both regions (only 4.2 % of "GOOD" class in Portugal regions and 10.4 % of "MEDIUM" class in Ontario region) limited the application of the Machine Learning algorithms, namely, the truthfulness and generalisation. Therefore, another pre-processing step was necessary.

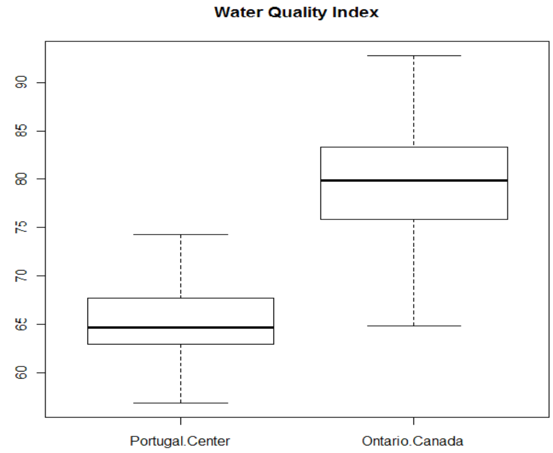

Figure 2. Comparison between WQI in Ontario, Canada and Portugal regions.

In order to equilibrate the "GOOD" and "MEDIUM" instances a new balanced dataset (DS3) was created. The DS3 presented biased distributions of 858 instances per class with a total of 1716 instances.

B. *Decision Framework.*

The classifier proposed in this study consists of three principal steps (see Figure. 3) as follows:

- From the reviewed literature, the Kumar's method was chosen, programmed and applied to our dataset.
- Among the parameters used for WQI calculation only some of those that can be measured in real time were selected. The parameters are: Dissolved Oxygen, pH, Nitrates and Temperature.
- The mixtures of two Cross Validation methods were applied. One of them was programmed in order to evaluate each iteration and the standard deviation and means were computed so as to compare the different machine learning algorithms. . The second cross validation method was applied in each iteration of the first cross validation method in order to guarantee the generalisation of these experiences as follows: About 80% of these data were applied to train the model, 10% were used for validation and about 10 % were used for a final test of models produced by machine learning algorithm.
- Machine learning algorithms, such as Partial Decision Tree, Feed-Forward Artificial Neural Net and k-Nearest Neighbor, were applied to the selected parameters in order to obtain a water quality classifier in real time.

The results of these processes will be shown and discussed in the next section.

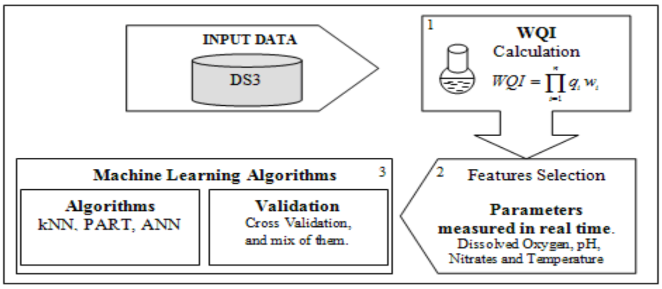

Figure 3. Decision Framework Steps.

III. Results and Discussions

Three different machine learning methods were applied to DS3. The results were validated using Cross Validation method and presented in Table 5 as representatively of one of the algorithms used. Ten folds cross validations was programmed to evaluate all iterations, which means that the average and standard deviations were evaluated. By each iteration about 80 % of these data were applied to train a model, 10 % for validation and 10 % to test for the three models produced.

Similar tendencies are observed comparing the three algorithms by iterations. The use and application of these algorithms in the field of water resource are thus well justified. However, kNN algorithm gave slightly better results. The Bregman Divergences and Mahalanobis distance were used to support nearest neighbor method. In the case of PART algorithm, the confidence was 95 % and the third method was an Artificial Neural Network backpropagation with three hidden layers with 3 and 1 neurons respectively.

In addition to the feasibility of the classification mentioned above, there are other important factors in the application of this framework (table IV).

TABLE IV. Model Assessments

Comparison Points	WQI [1]	Real Time Classifier proposed
Cost of the results analysis	Costly	Reasonable
Time to obtain analytical results	Delay	Real time
Number of parameters	9	4
Precision of WQI	Microbiological parameter is present	No microbiological parameter is present

Due to the absence of vectors with three of the five classes of the Kumar method, the proposed work is capable of guaranteeing good predictive accuracy for classifying water in two classes: medium and good. The high correlation between the microbiological and some chemical and physical parameters is used in order to overcome the difficulty of the microbiological parameters measured in real time.

This work will be set up in the region of Aveiro based on a water quality monitoring system called SIMOQUA [30] which has been developed at the University of Aveiro. SIMOQUA monitors, develops, and calibrates the sensors in real time for some chemical and physical drinking water parameters for three water sources from the Aveiro region: Silval, Vales das Maias and JK10.

The proposed work will allow the study of the aquifer systems for human consumption in real-time, will highlight the historical trends of water quality in the studied regions and will provide preventive measures to protect the drinking water sources, advancing any anomaly/deviation occurred between the time periods of usual checks. Moreover, one of the most important advantages of this system will be to help the government entities as a second opinion in the water management.

TABLE V. K-NEAREST NEIGHBOR ALGORITHM, BREGMAN DIVERGENCES AND MAHALANOBIS DISTANCE WERE USED

		kNN algorithm on mixed datasets								**Iteration**
		Testing		**Accuracy**		**Good**		**Medium**		
		Good	**Medium**							
TRUE INSTANCES CLASSIFIED 85 GOOD AND 85 MEDIUM	**Good**	83	2	98.82		97.65		100		1
	Medium	0	85							
	Good	81	4	97.65		95.29		100		2
	Medium	0	85							
	Good	84	1	96.47		98.82		94.12		3
	Medium	5	80							
	Good	85	0	100		100		100		4
	Medium	0	85							
	Good	84	1	99.41		98.82		100		5
	Medium	0	85							
	Good	85	0	94.71		100		89.41		6
	Medium	9	76							
	Good	82	3	98.24		96.47		100		7
	Medium	0	85							
	Good	82	3	98.24		96.47		100		8
	Medium	0	85							
	Good	82	3	98.24		96.47		100		9
	Medium	0	85							
	Good	90	3	98.39		96.77		100		10
	Medium	0	93							
Training per Iteration		**Good**	**Medium**	**Means**	**Sdv**	**Means**	**Sdv**	**Means**	**Sdv**	
		773	773	98.02	1.50	97.68	1.64	98.35	3.65	

IV. CONCLUSIONS

In this work, the Artificial Intelligence techniques in combination with Delphi method have been applied to the classification of the drinking water quality. The Delphi technique used by the NSF was improved using the Kumar method and PART, ANN and KNN algorithms have been used for the water quality classification in real time. A number of 1716 vectors from data collected in water stations of Portugal Regions and Ontario, Canada have been used for the system validation together with specialists from AdRA. Indeed, in the future work, it is intended to: 1) include in this real time classification, like the most important task, the "no proper" class in order to warn the most critical regions and safeguard the human health. To achieve this, due to the absence of vector with critical quality in the present work, the data acquisition in water sources with less desirable qualities is mandatory (Other countries with major problems in natural water sources). 2) Study a way to have indicators in real-time with direct impact on microbiological parameters in order to be more unfailing a real-time classification.

ACKNOWLEDGMENT

The authors acknowledge the Institute of Electronics and Telematics Engineering of the University of Aveiro (IEETA/UA) and Águas DA Região de Aveiro (AdrA) for the collaboration in this work. Jorge Camejo is thankful to the Foundation for Science and Technology (FCT) for the fellowship, to European Social Fund (FSE) and to the Operational Human Potential Program (POPH). Prof. Guevara acknowledges POPH - QREN-Tipologia 4.2– Promotion of scientific employment funded by the ESF and MCTES, Portugal.

REFERENCES

[1] D. Kumar, Alappat, B., "NSF-Water Quality Index : Does it represent the experts´ opinion?," Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management, 2009.

[2] T. Horton and E. S. Chase, "A Study of the Application of the Score. System to the Sanitary Quality of Public Water Supplies in New York State," American Journal of Public Health, vol. 7, pp. 380-390, Apr 1917.

[3] R. K. Horton, "An index number system for rating water quality," J. Water Pollut. Control Fed., pp. 300-306, 1965.

[4] J. T. Brown and W. L. Duncan, "Legal Aspects of a Federal Water Quality Surveillance System," Michigan Law Review, vol. 68, pp. 1131-1166, 1970.

[5] INAG, "The Portuguese Water Management Service," ed, 2000.

[6] CCME, "Canadian Water Quality Guideline. Canadian Council of Ministers of the Enviroment," W. Q. I. T. s. o. t. C. W. Q. Guidelines, Ed., ed, 2001.

[7] C. J. Rickwood and G. M. Carr, "Development and sensitivity analysis of a global drinking water quality index," Environmental Monitoring and Assessment, vol. 156, pp. 73-90, Sep 2009.

[8] T. S. Seilheimer, et al., "Comparative study of ecological indices for assessing human-induced disturbance in coastal wetlands of the Laurentian Great Lakes," Ecological Indicators, vol. 9, pp. 81-91, 2009.

[9] A. Said, et al., "An innovative index for evaluating water quality in streams," Environmental Management, vol. 34, pp. 406-414, Sep 2004.

[10] WEP, "Lower Great Miami Watershed Enhancement Program (WEP)." 1996.

[11] A. R. Karbassi, et al., "Development of Water Quality Index (WQI) for Gorganrood River," International Journal of Environmental Research, vol. 5, pp. 1041-1046, Fal 2011.

[12] M. R. Nikoo, et al., "A probabilistic water quality index for river water quality assessment: a case study," Environmental Monitoring and Assessment, vol. 181, pp. 465-478, Oct 2011.

[13] M. Vasanthavigar, et al., "Application of water quality index for groundwater quality assessment: Thirumanimuttar sub-basin, Tamilnadu, India," Environmental Monitoring and Assessment, vol. 171, pp. 595-609, Dec 2010.

[14] M. K. Chaturvedi and J. K. Bassin, "Assessing the water quality index of water treatment plant and bore wells, in Delhi, India," Environmental Monitoring and Assessment, vol. 163, pp. 449-453, Apr 2010.

[15] S. M. Yidana and A. Yidana, "Assessing water quality using water quality index and multivariate analysis," Environmental Earth Sciences, vol. 59, pp. 1461-1473, Feb 2010.

[16] A. M. Jinturkar, et al., "Determination of water quality index by fuzzy logic approach: a case of ground water in an Indian town," Water Science and Technology, vol. 61, pp. 1987-1994, 2010.

[17] P. N. Rajankar, et al., "Groundwater quality and water quality index at Bhandara District," Environmental Monitoring and Assessment, vol. 179, pp. 619-625, Aug 2011.

[18] P. Y. Li, et al., "Groundwater Quality Assessment Based on Improved Water Quality Index in Pengyang County, Ningxia, Northwest China," E-Journal of Chemistry, vol. 7, pp. S209-S216, Dec 2010.

[19] S. Ramesh, et al., "An innovative approach of Drinking Water Quality Index-A case study from Southern Tamil Nadu, India," Ecological Indicators, vol. 10, pp. 857-868, Jul 2010.

[20] H. Boyacioglu, "Utilization of the water quality index method as a classification tool," Environmental Monitoring and Assessment, vol. 167, pp. 115-124, Aug 2010.

[21] A. K. Yadav, et al., "Water Quality Index Assessment of Groundwater in Todaraisingh Tehsil of Rajasthan State, India-A Greener Approach," E-Journal of Chemistry, vol. 7, pp. S428-S432, Dec 2010.

[22] F. B. Semiromi, et al., "Evolution of a new surface water quality index for Karoon catchment in Iran," Water Science and Technology, vol. 64, pp. 2483-2491, 2011.

[23] F. Soroush, et al., "A Fuzzy Industrial Water Quality Index: Case Study of Zayandehrud River System," Iranian Journal of Science and Technology Transaction B-Engineering, vol. 35, pp. 131-136, Feb 2011.

[24] M. o. t. Environment, "Provincial Water Quality Monitoring Network (PWQMN) dataset ", ed. Ontario, 2010.

[25] ARH.Centro, "Administração da Região Hidrográfica do Centro I.P," ed. Coimbra, 2011.

[26] AdRA, "Águas da Região de Aveiro (AdRA) former Serviços Municipalizados de Aveiro (SMA)," 2011.

[27] F. V. a. S. L. Adzic. J, "Chapter IV: Extraction, Transformation, and Loading Process in Data WareHouse and OLAP. Concepts, Architectures and Solutions ", ed, 2007.

[28] R. M. e. a. Brown, "A water quality index - Do we dare?," Water Sewage Works 11, pp. 339-343, 1970.

[29] M. N. Varnosfaderany, et al., "Water quality assessment in an arid region using a water quality index," Water Science and Technology, vol. 60, pp. 2319-2327, 2009.

[30] J. P. Bhatt and M. K. Pandit, "A macro-invertebrate based new biotic index to monitor river water quality," Current Science, vol. 99, pp. 196-203, Jul 25 2010.

[31] A. G. Perez-Castillo and A. Rodriguez, "Physicochemical water quality index, a management tool for tropical-flooding lagoons," Revista De Biologia Tropical, vol. 56, pp. 1905-1918, Dec 2008.

[32] F. B. Semiromi, et al., "Water quality index development using fuzzy logic: A case study of the Karoon River of Iran," African Journal of Biotechnology, vol. 10, pp. 10125-10133, Sep 5 2011.