Optimizing water treatment systems using artificial intelligence based tools

A. Pinto[1], A. Fernandes[2], H. Vicente[3] & J. Neves[4]
[1]*Águas do Centro Alentejo S.A., Portugal*
[2]*Municipal Laboratory of Water, Portugal*
[3]*Department of Chemistry and Chemistry Centre of Évora, The University of Évora, Portugal*
[4]*Department of Informatics, The University of Minho, Portugal*

Abstract

Predictive modelling is a process used in predictive analytics to create a statistical model of future behaviour. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. On the other hand, Artificial Intelligence (AI) concerns itself with intelligent behaviour, i.e. the things that make us seem intelligent. Following this process of thinking, in this work the main goal is the assessment of the impact of using AI based tools for the development of intelligent predictive models, in particular those that may be used to establish the conditions in which the levels of manganese and turbidity in water supply are high. Indeed, one of the main problems that the water treatment plant at Monte Novo (in Évora, Portugal) uncovers is the appearance of high levels of manganese and turbidity in treated water, which sometimes exceed the parametric values established in Portuguese Law, respectively 50 μg dm^{-3} and 4 NTU. In this study we tried to find answers to the above problem by building predictive models. The models we developed shall enable the prediction of manganese and turbidity levels in treated water, in order to ensure that the water supply does not affect public health in a negative way and obeys the current legislation. The software used in this study was the Clementine 11.1. The C5.0 Algorithm was also used as a means of introducing Decision Trees and the K-Means Algorithm was used to construct clustering models. The data in the database was collected from 2005 to 2006 and includes reservoir water quality data, treated water data and volumes of water stored in the reservoir.

Keywords: knowledge discovery from databases, data mining, decision trees, water quality, manganese, turbidity.

WIT Transactions on Ecology and the Environment, Vol 125, © 2009 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/WRM090171

1 Introduction

In attempting to apply Knowledge Discovery in DataBases (KDDB) to generate a predictive model from a water treatment system dataset, the usual steps nowadays are to pre-process the data to overcome the challenges of missing data, redundant observations, and records containing inaccurate data. However, new technological breakthroughs may provide new ways to create and store information. Indeed, in any organization there is great amount of information referring to different processes, which may be used to improve their behaviour, either by discovering innovative trends and specificities, or accelerating the course of efficient decision making. On the other hand, the conventional tools for data analysis have a great number of drawbacks, since they do not allow the detection of singularities inside such a data chaos. Undeniably, having in mind a response to a given number of difficulties (e.g. those resulting from the use of great amounts of data, multiple sources of data or several application domains) a new area of KDDB is being brought to life and its tools and techniques for problem solving have been since then enforced. The designation KDDB was formally adopted in 1989 and refers to a process that involves the identification and recognition of patterns in a Database, in an automatic process, i.e. obtaining relevant, unknown information, that can be useful in a decision making process, without a previous formulation of hypothesis [1, 2].

The interest in ecological mining has been growing in the last decades. Undeniably, several Data Mining techniques have been used to find patterns in water quality databases, such as Decision Trees (DTs), Artificial Neural Networks (ANNs) and Genetic Algorithms (GA) [3–8]. Although DTs have not been extensively used in ecological modelling they have the advantage of expressing regularities explicitly and thus being easy to inspect for ecological validity.

In this paper, an approach to establish the conditions in which the levels of manganese and turbidity in water supply are high by Data Mining models is exploited. Currently, the assessment of the treated water quality is done through analytical methods, after the conclusion of treatment process, which is a very restricted approach. Due to this constraint, the development of Data Mining based models in conjunction with the development of a Decision Support System is a better alternative for the quality management of a water treatment plant [9]. In reality, these models have shown to be very helpful to establish the conditions under which the high levels of manganese and turbidity in treated water occurs, and therefore anticipate the problems, and implement the contingency procedures to avoid them.

The present study took place in Water Treatment Plant of Monte Novo which is part of the complex of Monte Novo Reservoir. This structure incorporates the company Águas do Centro Alentejo S.A. (the company), and is located 20 km southwest of the Portuguese city of Évora, which is classified by UNESCO as World Heritage.

The company operates in the sector of water supply, in the district of Évora, and the water treatment plant of Monte Novo is able to treat 1100 m^3 of raw water per hour, supplying 70 000 inhabitants currently.

The company mission is to produce an adequate and continuous supply of water that is chemically, bacteriologically and aesthetically pleasing, i.e. clear, colourless, low in organic content, non-corrosive, odourless, palatable, reasonably soft and safe, operating with ambition, innovation and competitiveness, fostering the respect for ethical and sustainability principles.

The paper is organized as follows: firstly, the data is presented and explained and the decision trees are introduced. The experiments are then performed and described, being the results analysed. Finally, closing conclusions are drawn.

2 Materials and methods

2.1 Sample collection and preservation

The collection of samples was performed two hours after the entry into operation of the water treatment plant to allow the renewal of water in the various units of treatment. Sampling points are located at the pipe of rising (raw water) and at the storage reservoir (treated water) (Figure 1). The sampling frequency was daily. The water samples were collected separately, directly to polyethylene bottles and, when necessary, preserved according SMEWW [10]. For alkalinity analysis the samples were collected in a polyethylene bottle of 200 cm^3 and refrigerated; for hardness analysis the samples were collected in a polyethylene bottle of 100 cm^3 and preserved with nitric acid 1:1, $pH \leq 2$; for turbidity analysis the samples were collected in a polyethylene bottle of 100 cm^3 and refrigerated. For iron, manganese and aluminium analysis the samples were collected in polyethylene bottles of 100 cm^3 and preserved with nitric acid 1:1, $pH \leq 2$.

2.2 Analysis

The determination of pH was executed by an electrometric method according to SMEWW 4500-H^+ B. In this method the basic principle is potentiometric measurement using a Crison GLP 22 pH meter equipped with a Crisolyt 50 14 electrode. The conductivity was evaluated according to SMEWW 2510 B using a WTW InoLab cond 720 conductivity meter. The turbidity was determined according to SMEWW 2130 B. Measurements were carried out on a Lovibond PC checkit turbidity meter. The dissolved oxygen was determined with a Crison OXI 45 oxymeter equipped with a DurOx 325 electrode according to SMEWW 4500-O B. The free chlorine and the combined chlorine were determined according to SMEWW 4500-Cl G using a Lovibond PC checkit chlorine meter. The iron was determined by molecular absorption spectrometry using the wavelength of 510 nm, according to SMEWW 3500-Fe D while the aluminium and manganese were determined using MercK photometric kits by molecular absorption spectrometry using the wavelength of 545 nm (aluminium) and 445 nm (manganese). The molecular absorption spectrometry measurements were carried out on a Helios Alpha UV-Vis spectrometer. Finally, the alkalinity was determined according to SMEWW 2320 B while the hardness was determined according to SMEWW 2340 C [10].

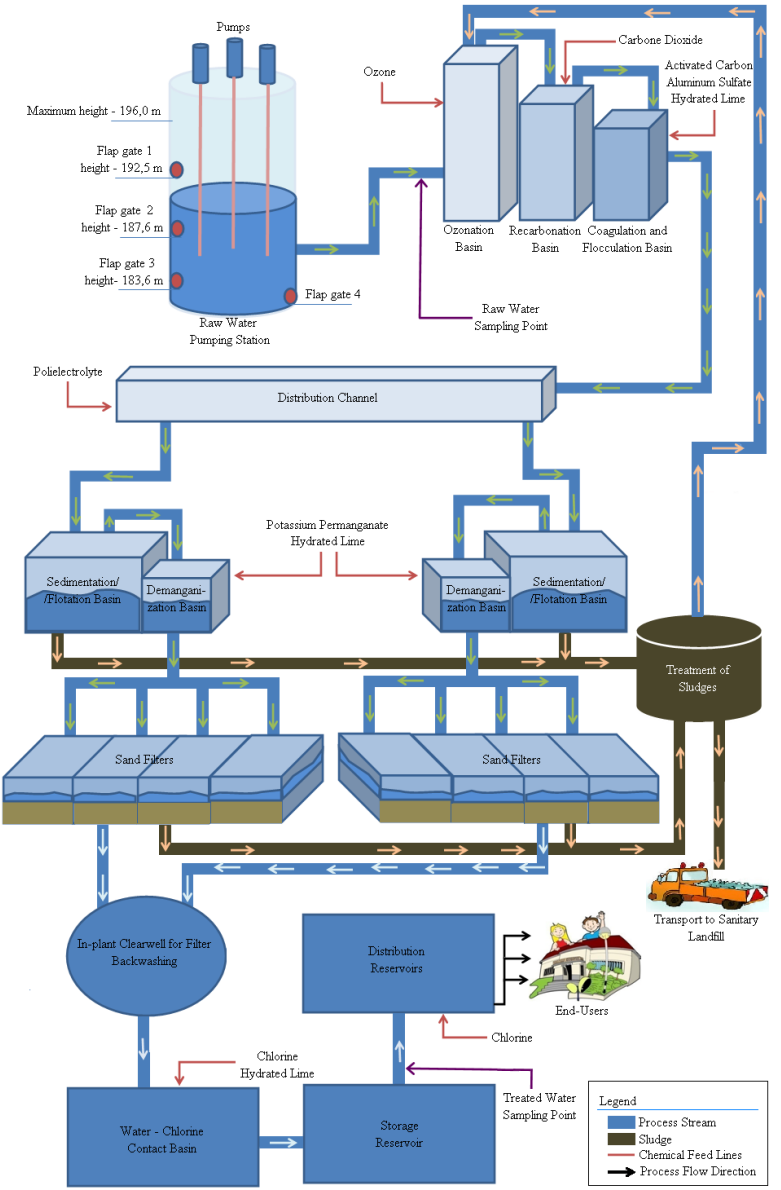

Figure 1: Monte Novo water treatment plant process diagram.

2.3 Database

The data used in this study was collected from April 2005 to December 2006, containing a total of 642 records with 35 fields. The main fields in the dataset were related with the reservoir, namely the volume stored and the water level, with the parameters of quality of raw water and with the parameters of quality of treated water.

The original dataset presented biased distributions: in 47.5% of observations the level of manganese in treated water was classified in Class NP, i.e. exceeded the parametric values established in The Portuguese Law and in 52.5% was classified in Class P, i.e. does not exceeded the parametric value. With regard to turbidity, the dataset presented biased distributions: in 48.2% of observations the turbidity in treated water was classified in Class NP, i.e. exceeded the parametric values while in 51.8% was classified in Class P, i.e. does not exceeded the parametric value.

Before attempting the DM modelling, the data was pre-processed. The original data set contained attributes with missing values. In particular, the chemical parameters presented some blank values. Since it was not possible to obtain the correct values, the blank registers were discarded. Furthermore, during April and May of 2006 an abnormal situation occurred and these registers were discarded too, remaining a total of 544 examples. In addition, in order to enhance the DTs learning procedure the data was normalized to the interval [0, 1] using the equation depicted below [11, 12]:

$$\overline{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where $\overline{X}$ denotes the variable normalized value, X denotes the variable value and Xmin and X_{max} denote, respectively, the minimum and the maximum values for the variable.

2.4 Decision trees

The prediction of the levels of manganese in treated water, according to the Portuguese Law criteria, was defined as a classification problem. The DT is one of the most efficient and popular DM classification methods. It adopts a branching structure of nodes and leaves, where the knowledge is hierarchically organized. Each node tests the value of a feature, while each leaf is assigned to a class (label).

The C5.0 Algorithm [13] was used to induce the DTs, under the SPSS Clementine System. In fact, the use of DTs enables the automatic extraction of production rules that can be incorporated in a Decision Support System, using, for example, either the Structured Query Language (SQL) commands or the Predictive Model Markup Language (PMML), which is an XML-based language, developed by the Data Mining Group (DMG), which provides a way for applications to define statistical and data mining models and to share models

between PMML compliant applications (The Data Mining Group (DMG) is an independent, vendor led group which develops data mining standards, such as the Predictive Model Markup Language (PMML). DMG's XML related standards can be found at Source Forge).

2.5 Clustering models

The study of the problem of turbidity in treated water, based on parametric values established in the Portuguese Law, was defined as a segmentation problem. The K-Means Clustering Method is one of the most efficient and popular DM segmentation algorithms. Clustering models focus on identifying groups of similar records and labelling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. These models are often referred to as unsupervised learning models, since there is no external standard by which to judge the model's performance. There are no "right" or "wrong" answers for these models. Their value is determined by their ability to capture interesting groupings in the data and provide useful descriptions of those groupings.

The basic idea in K-Means Clustering Method is to try to discover k clusters, such that the records within each cluster are similar to each other and distinct from records in other clusters. K-Means is an iterative algorithm, based on measuring distances between records and between clusters. The method defines a fixed number of clusters, iteratively assigns records to clusters, and adjusts the cluster centres until further refinement can no longer improve the model (or the number of iterations exceeds a specified limit). The K-Means Algorithm was used to induce cluster analysis, under the SPSS Clementine System [14].

3 Results

3.1 Framework

Attending to the fact that the manganese levels are only known after the conclusion of the whole treatment process, it was decided to develop models to forecast the manganese class in treated water (i.e. P, NP), using the KDDB methodology for problem solving already presented on the top of the database referred to above, using the C.5 algorithm. Figure 3 shows the DT obtained and Table 1 gives a picture of the corresponding decision rules.

3.2 Tests

The classification model was developed using the C5.0 algorithm. To ensure statistical significance of the attained results, 10 (ten) runs were applied in all tests, being the accuracy estimates achieved using the Holdout method [15]. In each simulation, the available data is randomly divided into two mutually exclusive partitions: the training set, with 2/3 of the available data and used during the modelling phase, and the test set, with the remaining 1/3 examples, being used after training, in order to compute the accuracy values.

A common tool for classification analysis is the coincidence matrix [16], a matrix of size L x L, where L denotes the number of possible classes. This matrix is created by matching the predicted (test result) and actual (manganese real class) values. L was set to 2 (two) in the present case.

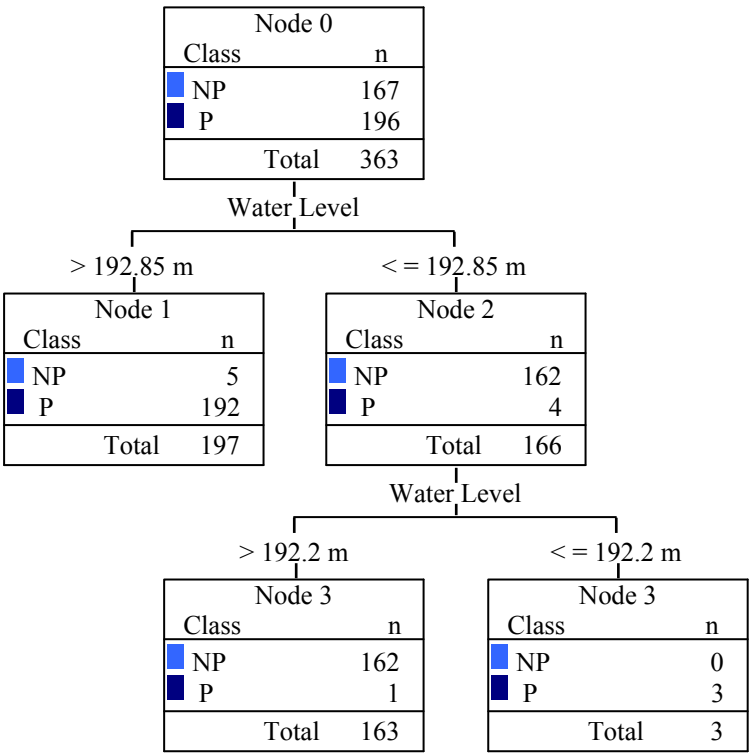

Figure 2: Decision tree model to predict the manganese levels in treated water.

Table 1: Decision rules corresponding to DT Model of Figure 2.

Rule for NP Class	Rules for P Class	
Rule 1	Rule 1	Rule 2
If water level ≤ 192.85 and water level > 192.2	If water level ≤ 192.2	If water level > 192.85
Then → NP	Then → P	Then → P
n = 163	n = 3	n = 197
confidence = 0.994	confidence = 1.0	confidence = 0.976

3.3 Discussion

Table 2 presents the coincidence matrixes for this model. The values denote the average of the 10 (ten) runs. The results reveal that the model gives correct answers in 98.3% of cases for the training set and 96,7% for the test set. The algorithm for DT induction only used the field Water Level of stored water in reservoir to predict the concentration of manganese in treated water. When the level of stored water is under the height of 192.85 m the problems with manganese levels in treated water arise. Combining this information with the height of the flap gate 1 (Figure 2), the model shows that problems occur when the flap gate 1 is closed and flap gate 2 is open. When this change happens the water treatment plant must be prepared to treat raw water with free manganese. In these circumstances, it is absolutely necessary to add hydrated lime in demanganization basins to promote the conditions to remove the manganese by chemical precipitation.

Table 2: The coincidence matrix for the model.

	Training Set		Test Set	
Class	NP	P	NP	P
NP	162	5	58	3
P	1	195	3	117

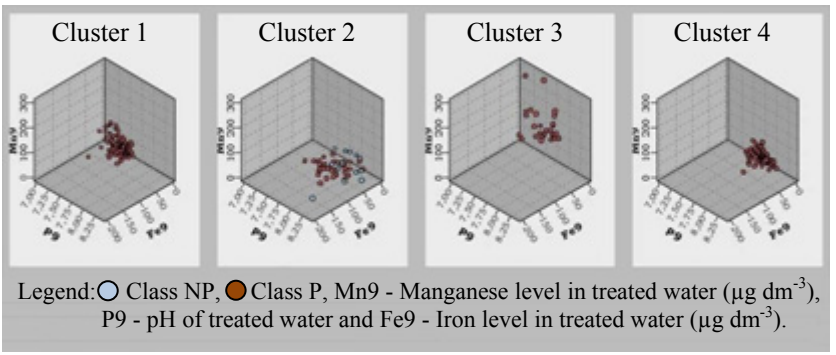

Figure 3: Clusters obtained in the study of turbidity levels in treated water.

In the study of the problem of turbidity in treated water there were identified 4 (four) clusters (Figure 3). Cluster 1 joins 198 examples, cluster 2 includes 102 cases, cluster 3 is formed by 98 examples and cluster 4 includes 146 examples. The turbidity levels for all examples in clusters 1, 3 and 4 are included in class P, i.e. do not exceed the parametric value established in Portuguese Law. Cluster 2 is formed by all the examples of class NP and by some others of class P. This cluster is characterized by low levels of manganese, below 50 μg dm^{-3} and pH upper than 7.25. These results underline the importance of the final adjustment of treated water pH. On the one hand the pH of treated water must be high

enough to preserve the pipelines and the infrastructures and, on the other hand, must be low enough to avoid the appearance of turbidity after the conclusion of the water treatment process.

4 Conclusion and future work

The use of DM techniques can solve complex problems in water supply industry, such as the real-time diagnosis of manganese and turbidity levels in treated water. The use of these techniques allows the reduction of costs and real-time intervention, which will preserve and increase the final product quality.

In this work, a classification model and segmentation model were tested, using DTs and clustering analysis. The results of the models may be very useful, once they reveal when the problems arise, therefore allowing the establishing of contingence in order to find the best solutions.

Therefore, and according to the objectives of Águas do Centro Alentejo, S.A., not only in establishing a compromise of integrating Health, Safety and Environment on the strategy and activities of the company, but also close by its continuous improvement on its performance, giving a decisive contribute to achieve a sustainable development and excellence in water supply business, it is necessary to act on a pro-active way. Consequently, the use of predicting methods that have inherent processes of decision making, and that come from the scientific area of AI, may contribute to solve or minimize the problems referred to above. The proposed time-line that encompasses further work opens the room for the development of automatic tools for decision support, which are expected to enhance the water treatment plant response.

Acknowledgements

The authors would like to thank The University of Minho for providing the software facilities for implementing the present solution and to Águas do Centro Alentejo, S.A., in making available the data collected, which was fundamental to conduct this work.

References

[1] Fayyad, U., Piatetshy-Shapiro, G., Smith, P. & Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining*, MIT Press: Massachusetts, 1996.

[2] Thuraisingham, B., *Data Mining Technologies, Techniques, Tools and Trends*, CRC Press LLC: USA, 1999.

[3] Dzeroski, S., Environmental Sciences. *Handbook of Data Mining and Knowledge Discovery*, ed. W. Klösgen & J. M. Zytkow, Oxford University Press: Oxford, U.K., pp. 817-830, 2002.

[4] Vicente, H., *Especification and Prototyping of Management and Control Systems of Water Quality in Reservoirs*, PhD Thesis, University of Évora: Évora, Portugal, 2004.

[5] Santos, M. F., Cortez, P., Quintela, H., Neves, J., Vicente, H. & Arteiro, J., Ecological Mining - A Case Study on Dam Water Quality. *Data Mining VI - Data Mining, Text Mining and their Business Applications*, ed. A. Zanasi, C. A. Brebbia & N. F. F. Ebecken, WIT Press: Southampton, UK, pp. 523-531, 2005.
[6] Chau, K.-W., A review on integration of artificial intelligence into water quality modelling. *Marine Pollution Bulletin*, **52**, pp. 726-733, 2006.
[7] Kuo, J.-T., Hsieh, M.-H., Lung, W.-S. & She, N., Using artificial neural network for reservoir eutrophication prediction. *Ecological Modelling*, **200**, pp. 171-177, 2007.
[8] Fernandes, A., Vicente, H. & Neves, J., Solving Challenging Problems in the Oil Industry Using Artificial Intelligence Based Tools. *Proc. of the 7th Industrial Simulation Conference, ISC'2009*, Loughborough, 2009.
[9] Turban, E., Aronson, J. E. & Liang, T.-P., *Decision Support Systems and Intelligent Systems*, Prentice Hall: New Jersey, USA, 2004.
[10] Eaton, A., Clesceri, L., Rice, E. & Greenberg, A., (eds). *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association: USA, 2005.
[11] Pyle, D., *Data Preparation for Data Mining*, Morgan Kaufmann: San Francisco, USA, 1999.
[12] Han, J. & Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kauffmann Publishers: San Francisco, U.S.A., 2006.
[13] Quinlan, J. R., Bagging, boosting and C4.5. *Proc. of the AAAI'96 National Conference on Artificial Intelligence*, AAAI Press, pp. 725-730, 1996.
[14] Bradley, P. S. & Fayyad, U. M., Refining Initial Points for K-Means Clustering. *Proc. of the 15th International Conference on Machine Learning (ICML98)*, eds. J. Shavlik, Morgan Kaufmann: San Francisco, pp. 91- 99, 1998.
[15] Souza, J., Matwin, S. & Japkowicz, N., Evaluating Data Mining Models: A Pattern Language. *Proc. of the 9th Conference on Pattern Language of Programs*, 2002.
[16] Kohavi, R. & Provost, F., Glossary of Terms. *Machine Learning*, **30**, pp. 271-274, 1998.