BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXII No. 1/2010	69 - 78	Seria Matematică - Informatică - Fizică

Pollution Level Analysis of a Wastewater Treatment Plant Emissary using Data Mining

Mădălina Cărbureanu

Petroleum-Gas University, Informatics Department, Ploieşti, Bd. Bucharest, 39
e-mail: mcarbureanu@upg-ploiesti.ro

Abstract

The wastewater treatment plant from Ploieşti, located in the South-Eastern part of the city, has only a mechanical step on its treatment wastewater line. The current plant emissary is Dâmbu, which joins Teleajen River. The pollution caused by an inefficient wastewater treatment can create severe damage to the population and the environment. In this paper an application of data mining in the wastewater treatment domain is presented, in order to determine and analyze the pollution level of a treatment plant emissary.

Keywords: *wastewater treatment plant, pollution level, data mining techniques, ID3*

About Data Mining Techniques

Data mining, also known as "Knowledge-Discovery in Databases" (KDD), has three generic roots, from which it borrowed techniques and terminology: statistics, artificial intelligence and database systems (DBS). According to [3], by data mining we understand the practice of patterns automated searching in huge data bases using statistic computational techniques, machine learning and patterns recognition, the extraction of potential useful information from data or data bases, the exploration and analysis of a great amount of data through automatic or semiautomatic means, for useful patterns discovery and also the process of information automatic discovery, the identification of hidden patterns and relations between data.

The literature identifies two main categories of data mining techniques, namely classical data mining techniques and next generation data mining techniques.

Classical data mining techniques involve techniques such as [9]:

- The nearest neighbor technique (K-NN), one of the easiest techniques to work with, because it operates similar with human reasoning;
- Clustering, a method through which similar instances are grouped in order to allow the user a better view of the database;
- Statistics, used for patterns discovering and for predictive models development.

The next generation data mining techniques are [9]:

- Decision and classification trees, which make the instances classification, by covering the tree from root nodes to leaf nodes;

- Artificial neural networks, which, unlike neural networks that are biological systems for patterns detection, prediction and learning, are programs that implement sophisticated patterns for detection and machine learning algorithms, in order to develop predictive models;
- Rule induction, whose goal is to generate rules and to supply useful information about the database;
- Genetic algorithms that are an identification technique to approximate the solutions, useful in optimization and searching problems.

In this paper, in order to determine and analyze the pollution level of a treatment plant emissary, we propose the use of the decision and classification trees technique, because of its advantages: the generated rules are easy to interpret, the continuous and nominal variables are easy to model, the technique is able to deal with rules-oriented domains and is able to identify the most relevant variable necessary for prediction.

From the data mining algorithms, such as CHIAD (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), Quest, C5.0, J48, ID3, we chose to use the last mentioned one, developed by J. Ross Quinlan, in 1979 [6].

The ID3 algorithm is an artificial intelligence inductive technique that generates decision and classification trees. These trees are similar to those used within the framework of the simple expert systems. In contrast with other statistical techniques, the decision trees are a plausible model for inductive representation of human knowledge [8]. According to [9] and [17], a series of improvements to ID3 algorithm were made through C4.5 and C5.0/See5 algorithms, regarding the working methods with numeric attributes, missing values, noisy data, rules generation etc.

Data Mining in the Wastewater Treatment Domain

The monitoring of the parameters in the treatment processes is an essential activity in wastewater treatment plants, being related with the environment monitoring. The goal is to analyze the treatment processes activity and efficiency and to ensure the conformity with the specific legislation [7].

Nowadays, lot of research is done to develop systems for monitoring and control the wastewater treatment plants, such as SCADA (Supervisory Control And Data Acquisition), HydroDat, Telemac, systems presented in [2].

From the systems mentioned above, within the framework of Telemac project (Telemonitoring and Advanced Telecontrol of Yield Wastewater Treatment Plants), started in 2001 and coordinated by ERCIM (European Research Consortium for Informatics and Mathematics), the work was focused on the data mining techniques usage [16]. This is an important approach, because data such as pH, temperature and other measurements, much more advanced (such as the volatile fat acids - VFA), are constantly cumulated. Data mining brings to light the perspective of learning from these data in order to improve the wastewater treatment plant management.

According to [11], a number of possibilities are brought into discussion, such as the models and useful rules development in predicting dangerous situations, using the data evolution supplied by sensors, the damaged sensors detection through their inconsistent data and the partial replacement of the expensive sensors with combined data supplied by regular sensors.

Although an important part of the activity consists in data mining, the goal is usually knowledge discovery. Data mining is not a matter of rule induction algorithms or neural networks running. A huge amount of preliminary work is necessary, and Telemac does not make an exception from this rule [11]. The process, on the whole, includes steps such as data selection, data

cleaning, transformation of data lines into a convenient form, data enrichment or reduction, data processing, data mining itself, reports, statistics etc. The software used in Telemac project is Clementine.

In this project, the rule induction techniques are used to estimate the values of data supplied by sensors and to determine if the developed models are reliable in time.

Telemac project currently uses data mining as a component of its distance control centre (Telecontrol Center), responsible for the monitoring of a number of plants. It runs periodically for database updating.

One of the special characteristics of the project is the idea of distance control centre, where several monitors or stations function as an expertise center, bringing together the experts knowledge. Data mining has been acknowledged as a technology that may have a meaningful contribution in this domain.

The Telemac system architecture is presented in figure 1.

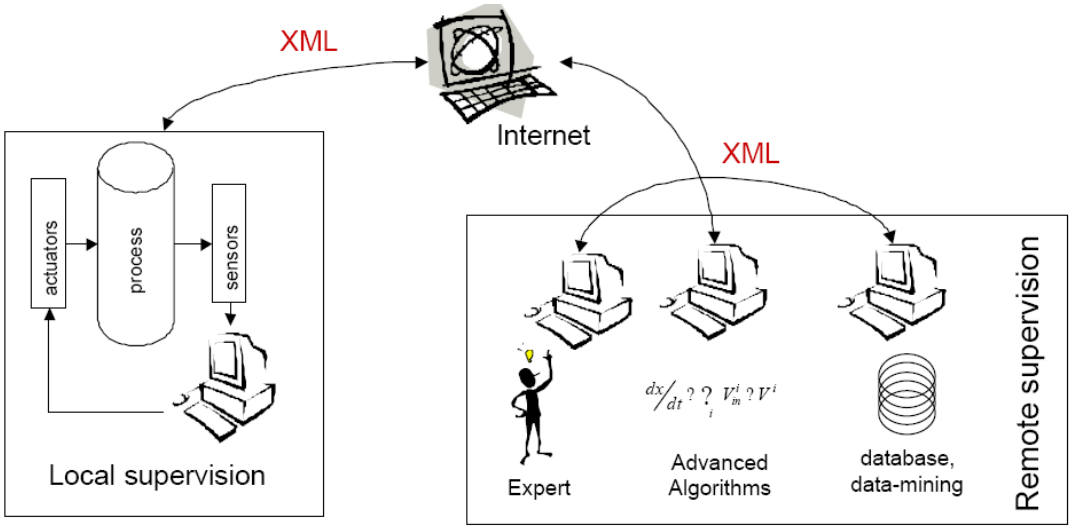

Fig. 1. Telemac architecture [12]

The challenges for data mining (DM) in the wastewater treatment domain are the following:

- DM must characterize the current state of the plant;
- The usage of DM must reduce the number of sensors necessary to determine the plant state, issue very important because some sensors are quite expensive;
- DM must achieve failure detection, diagnosis, estimation of the sensor failure moment etc.;
- DM must supply illustrative, eloquent, demonstrative techniques for helping the human expert to interpret data mining results (visual data mining);
- DM must integrate its validated results into Telemac system.

Although some of these challenges have been solved, there are still some aspects that require a quick solving, such as: some wastewater treatment plants have a small number of sensors, consequently fewer data than necessary are available, the problem of plant's evolution in time, when the system receives results obtained by data mining, the problem of how well the system learns from revised models when new data are available emerges.

The wastewater treatment plant process analysis is a difficult task because of its complex, mostly dynamic behavior [13]. Decision and regression trees, as a data mining technique, are an useful instrument for analyzing the plant's activity, if a set of input data are available.

The Developed Application

Because the wastewater treatment plant from Ploieşti has only a mechanical step, whose structure has been presented in [1], we decided to apply, from the decision tree data mining techniques, the ID3 algorithm in order to analyze the pollution level of Dâmbu brook, the treatment plant's emissary. Taking into consideration the fact that the proposed application doesn't use training data with missing attribute values or noisy data, we consider that there can be applied the ID3 algorithm.

For developing the application, we chose to use Weka (Waikato Environment Knowledge Analysis) software. Weka is "a collection of machine learning algorithms for data mining, in Java language" [14]. The algorithms can be directly applied on data sets or can be called from the programmer's code.

The ID3 algorithm is implemented in Weka, in the package weka.classifiers.trees.Id3. The input data are organized in an .arff file, whose format is recognized by Weka. The characteristics of this file are the three sections indicated through key words: *@relation*, *@attribute*, *@data*, which corresponds to the name of the data base, the name and type of the fields and the name of the recordings, respectivelly. After running an algorithm, Weka supplies a set of statistic results, such as: the confusion matrix, the precision, F measure, TP Rate (True Positive Rate), FP Rate (False Positive Rate), ROC area (Receiver Operating Characteristic), kappa statistic etc.

The main quality indicators measured in the wastewater mechanical treatment step are CBO5 (biochemical consumption of oxygen) and MTS (floating materials). The values for these two indicators are presented in table 1 and are to be found in the wastewater treatment plant technical documentation, representing values sampled at the plant's input and output [10].

Table 1. The quality indicators values

Indicator	Sampling point	2009							
		February	**March**	**April**	**May**	**June**	**July**	**August**	**Average values**
MTS(mg/l)	Input	229.46	239.58	241.96	235.79	237.1	236.85	236.92	236.8
	Output	143.12	143.98	131.58	119.84	119.93	131.26	131.55	131.6
CBO5(mg/l)	Input	88.3	91.47	91.38	89.23	90.72	90.79	92.85	90.68
	Output	42.51	43.31	38.85	37.89	38.33	37.74	39.79	39.77

The efficiency of the plant's mechanical treatment step represents the reduction percent of a part from a certain substance (in our application, MTS and CBO5), so that, after the discharging of the treated wastewater into the emissary, the last one must fulfill the quality conditions imposed by the legislation [4]. The mechanical treatment step's efficiency values in removing MTS and CBO5 are presented in table 2, and are supplied by the plant technical documentation [10].

Table 2. The efficiency values

Indicator	Sampling point	2009							
		February	**March**	**April**	**May**	**June**	**July**	**August**	**Average values**
MTS	Efficiency (%)	37.63	39.90	45.62	49.18	49.42	44.58	44.47	44.40
CBO5	Efficiency (%)	51.86	52.65	57.48	57.54	57.75	58.43	57.15	56.12

In order to determine the pollution level of Dâmbu brook with effluents from the plant, namely with MTS, from the indicators presented in tables 1 and 2, we chose to use MTS at the plant's input, named MTSI and the mechanical step efficiency in MTS removal, named EFFMTS. Also, we introduced a new indicator, named GRPOLMTS, which represents the effluent charge with MTS at the plant's output, measured in mg/l.

To obtain the decision rules, the ID3 algorithm divides the training set into a number of different sets. It searches the most relevant attribute and uses a measure to restrict the searching area, named entropy. It is a greedy algorithm which develops a decision tree from up to bottom, at each nod selecting that attribute that best classifies the local training examples. The best attribute is that having maximum information gain, by the formula:

$$\text{Information\_gain}(M, A) = \text{Entropy}(M) - \sum_{v \in Values(A)} \frac{|M_v|}{|M|} \text{Entropy}(M_v) \quad (1)$$

Here, *Values (A)* represents the set of all possible values for the attribute A, and M_v is the M set for which attribute A has v value [5].

For a better classification and data manipulation, the ID3 algorithm requires that for each variable to be established three intervals, presented in table 3 and table 4. To do that, we used the average values from tables 1 and 2. For instance, the average for MTSI is 236.8 mg/l, and the values higher than this value belong to *big* interval. For *medium* and *small* intervals, we chose a value of 228 mg/l, representing the smallest value for MTSI registered in 2009, according to the technical documentation [10]. For the other indicator, the average for CBO5 at the plant's input is 90.68 mg/l, and the higher values belong to *big* interval. For *medium* and *small* intervals, we chose a value of 87 mg/l, representing the smallest value for CBO5I in 2009 [10]. For establishing the EFFMTS and EFFCBO5 intervals, we used the same reasoning.

Table 3. The indicators intervals

Indicator	Small	Medium	Big
MTSI	<228	[228; 236.8]	>236.8
EFFMTS	<35(reduced_eff)	[35;44.40](medium_eff)	>44.40(raised_eff)
GRPOLMTS	<43	[43;60]	>60

We must note that: to establish the intervals for GRPOLMTS indicator, we used the charging limit value with MTS for treated wastewater at the evacuation into the natural emissary, value that is 60.0 mg/dm^3 [15].

It was necessary to build two distinct databases, one for MTS and one for CBO5, because these indicators are separately measured at the mechanical step input and output [1].

Using the training data from tables 1 and 2 and the intervals from table 3, we obtained a database for MTS implemented in Weka, as depicted in figure 2.

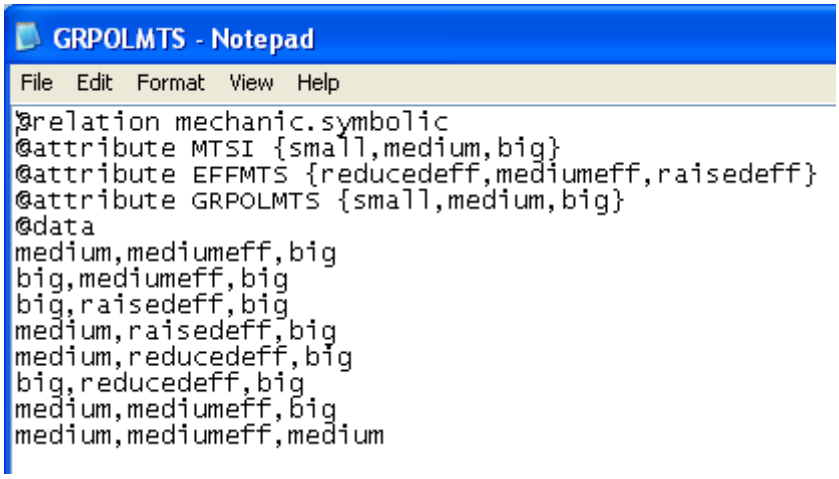

Fig. 2. GRPOLMTS.arff database

Applying the ID3 algorithm on this database, we obtained a decision tree in which the root node is EFFMTS and the leaf nodes represent the values obtained for GRPOLMTS. The decision tree is presented in figure 3.

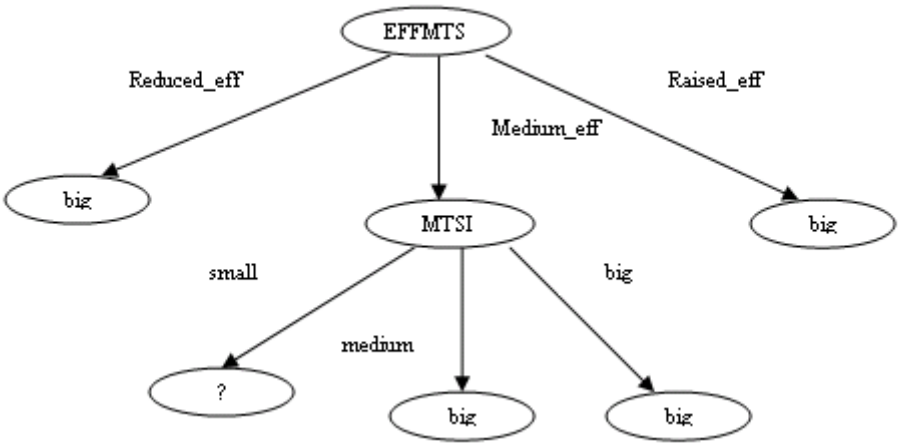

Fig. 3. The decision tree

Interpreting the decision tree obtained using ID3 algorithm, we extracted a set of decision rules, as follows:

- If EFFMTS is Reduced_eff, then GRPOLMTS is big;
- If EFFMTS is Medium_eff and MTSI is small, then "It cannot be said anything about the pollution level with MTS;
- If EFFMTS is Medium_eff and MTSI is medium, then GRPOLMTS is big;
- If EFFMTS is Medium_eff and MTSI is big, then GRPOLMTS is big;
- If EFFMTS is Raised_eff, then GRPOLMTS is big.

For the second part of the application, in order to determine the pollution level of Dâmbu brook with CBO5, from the indicators presented in tables 1 and 2, we used CBO5 at the plant's input, named CBO5I and the mechanical step efficiency in CBO5 treatment, named EFFCBO5. We also introduced a new indicator, named GRPOLCBO5, that represents the CBO5 concentration level in the effluent at the plant's output, measured in mg/l.

We must specify that, in order to establish the intervals for GRPOLCBO5 indicator, we related to the charge limit value with CBO5 for treated wastewater at the evacuation into the natural emissary, that is 25 mg/dm^3 [15].

Using the same reasoning above-mentioned, we obtained the intervals presented in table 4.

Table 4. The indicators intervals

Indicator	Small	Medium	Big
CBO5I	<87	[87;90.68]	>90.68
EFFCBO5	<20(reduced_eff)	[20;56.12](medium_eff)	>56.12(raised_eff)
GRPOLCBO5	<17.5	[17.5;25]	>25

Using the training data from tables 1 and 2 and the intervals from table 4, we obtained a database for CBO5, implemented in Weka, in which the inputs are represented by CBO5I and EFFCBO5 and the target is GRPOLCBO5, as we can observe in figure 4.

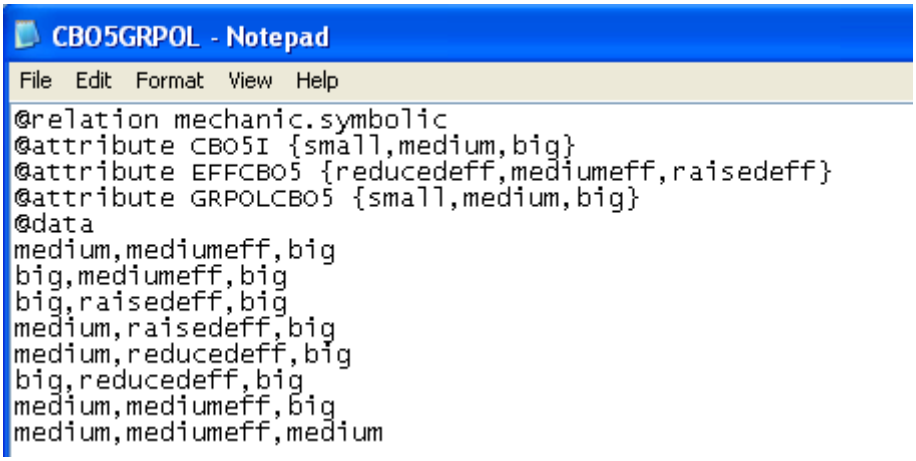

Fig. 4. GRPOLCBO5.arff database

Applying the ID3 algorithm on the database presented in figure 4, we obtained the decision tree from figure 5.

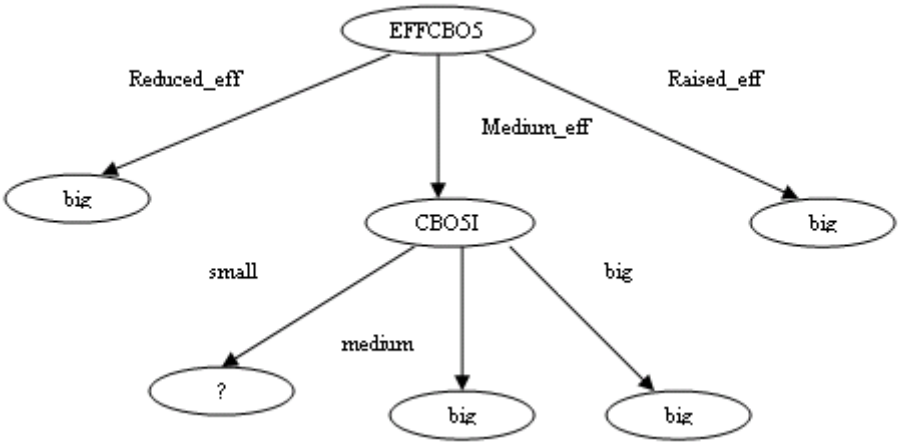

Fig. 5. The decision tree

Interpreting the decision tree obtained using ID3 algorithm, we extracted a set of decision rules, as follows:

- If EFFCBO5 is Reduced_eff, then GRPOLCBO5 is big;
- If EFFCBO5 is Medium_eff and CBO5I is small, then "It cannot be said anything about the pollution level with CBO5";
- If EFFCBO5 is Medium_eff and CBO5I is medium, then GRPOLCBO5 is big;
- If EFFCBO5 is Medium_eff and CBO5I is big, then GRPOLCBO5 is big;
- If EFFCBO5 is Raised_eff, then GRPOLCBO5 is big.

The two sets of rules obtained are very useful in analyzing the pollution level of Dâmbu brook with MTS and CBO5, because, by knowing the values for MTS and CBO5 at the plant's input and the values for the mechanical step efficiency in removing MTS and CBO5 and using the decision rules, we can establish the pollution level with MTS and CBO5 for the plant emissary.

Therefore, for the indicators values presented in tables 1 and 2, using the decision rules, we obtained the results presented in tables 5 and 6.

Table 5. GRPOLMTS level

Month	MTSI(mg/l)	EFFMTS (%)	GRPOLMTS-charge level with MTS(mg/l)
February	229.46(medium)	37.63(medium_eff)	>60 (big)
March	239.58(big)	39.90(medium_eff)	>60 (big)
April	241.96(big)	45.62(raised_eff)	>60 (big)
May	235.79(medium)	49.18(raised_eff)	>60 (big)
June	237.1(big)	49.42(raised_eff)	>60 (big)
July	236.85(big)	44.58 (raised_eff)	>60 (big)
August	236.92(big)	44.47(raised_eff)	>60 (big)

Table 6. GRPOLCBO5 level

Month	CBO5I(mg/l)	EFFCBO5 (%)	GRPOLCBO5-concentration level with CBO5(mg/l)
February	88.3(medium)	51.86(medium_eff)	>25 (big)
March	91.47(big)	52.65(medium_eff)	>25 (big)
April	91.38(big)	57.48(raised_eff)	>25 (big)
May	89.23(medium)	57.54(raised_eff)	>25 (big)
June	90.72(big)	56.84(raised_eff)	>25 (big)
July	90.79(big)	58.43(raised_eff)	>25 (big)
August	92.85(medium)	56.40(raised_eff)	>25 (big)

From the results obtained in table 5 and table 6 using the decision rules, we can easily observe that in February-August 2009, the plant emissary - Dâmbu brook - has registered a raised level of pollution with effluents from the wastewater treatment plant, respectively with MTS and CBO5.

Furthermore, by using the obtained decision rules, there can be achieved the monthly level pollution prediction for the plant's emissary.

Also, we can observe that, no matter the mechanical step's efficiency level in removing MTS and CBO5 from wastewaters, the pollution level is high, related to the limit values assessed by NTPA 001/2002, fact that highlights the necessity of the mechanical step's modernization and the building of a biological step [15].

From the data presented in table 2, one may notice that the maximum mechanical treatment efficiency level in removing MTS is 49.42% and in removing CBO5 - 58.43%. Taking into consideration the specific normative, the condition for defining the efficiency of a wastewater

treatment plant (with mechanical and biological step) in MTS and CBO5 removing is CBO5>75% and MTS>50%, fact that justifies the raised values obtained for Dâmbu pollution level with CBO5 and MTS.

Conclusions

Using a data mining technique, such as decision trees, we have developed an application in order to determine and analyze the pollution level of Dâmbu brook, the emissary for the wastewater treatment plant from Ploieşti. Data mining has various applications in the wastewater treatment plant domain, due to the fact that it supplies useful instruments for analyzing the current state of the plant and its evolution in time.

What we can conclude from this article is the fact that, no matter how efficient the mechanical step is in removing MTS and CBO5 from wastewater, the Dâmbu brook pollution level is high, compared to the admissible limit values established through NTPA 001/2002.

The inefficiency of the current mechanical step from Ploieşti wastewater treatment plant, determined by the reduced technological performance and ageing, requires the step's modernization with a higher technology and the building of a biological step. Thus, there can be reduced the pollution level of the plant emissary. At this moment these two objectives are priorities for the local administration.

As future work, the decision rules obtained in this article will be refined using improved data mining algorithms and it will be integrated in a system with fuzzy logic and data mining for analyzing the pollution level of Ploieşti wastewater treatment plant emissary.

The analysis of the pollution level of a plant emissary is absolutely necessary, taking into account the fact that nowadays it is experimented the possibility of using treated wastewater as a source of drinking water, fact that imposes a rigorous analysis of its quality.

References

1. Cărbureanu, M. - *The Efficiency Level Analysis for the Wastewater Mechanical Treatment Process using Data Mining and Fuzzy Logic*, Petroleum-Gas University of Ploieşti Bulletin Mathematics-Informatics-Physics Series, Vol. LXI, No. 2/2009, pp. 59-66
2. Cărbureanu, M. - *Sistem bazat pe logică fuzzy şi data mining pentru analiza calităţii apelor uzate epurate,* MSe thesis, Petroleum-Gas University of Ploieşti, 2010
3. Gorunesu, F. - *Data Mining. Concepte, modele şi tehnici*, Editura Albastră, Cluj-Napoca, 2006
4. Ianculescu, O., Ionescu, Ghe., Racoviţeanu, R. - *Epurarea apelor uzate*, Editura Matrix Rom, Bucureşti, 2001
5. Oprea, M. - *Sisteme bazate pe cunoştinţe. Ghid teoretic şi practic*, Editura MatrixRom, Bucureşti, 2002
6. Quinlan, J.R.- *Machine Learning, Induction of decision trees,* vol. 1, Springer, Netherlands, 1986
7. Robescu, D. et al. - *Controlul automat al proceselor de epurare a apelor uzate,* Editura Tehnică, Bucureşti, 2008
8. Schrodt, P.A.- *Predicting Interstate Conflict Outcomes Using a Bootstrapped ID3 Algorithm,* Political Analysis, Oxford Journals, 1990
9. Witten, I.H., Frank, E.- *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition),* Elsevier, San Francisco, 2005
10. ***- *The technical document for the objective - Wastewater collecting and treatment system modernization from Ploieşti city*, RASP Ploieşti, 2009
11. ***- *Data Mining Applied in Anaerobic Wastewater Treatment,* http://www.ercim.org/publication /Ercim_News/enw56/lambert.html, accessed 27 March 2010

12. * * * - *Experience with data mining for the anaerobic wastewater treatment process,* http://epublicns03. esc.rl.ac.uk/bitstream/1581/TELEMAC_Water_Science_v4.pdf, accessed 20 March 2010
13. * * * - *Modeling of wastewater treatment plant with regression trees,* http://library.witpress.com/pages/PaperInfo.asp?PaperID=1291, accessed 20 March 2010
14. * * * - *Weka Software*, http://www.cs.waikato.ac.nz/ml/weka/, accessed 20 March 2010
15. * * * - *The NTPA-001/2002 normative for establishing the loading limits with pollutants for industrial and city wastewater at the evacuation into the natural receivers, published in the Romania Official*, part I, no. 187, 20 March , 2002
16. * * * - *Telemac* , http://www.ercim.eu/telemac/, accessed 13 June 2010
17. * * * - *See5/C5.0* , http://www.rulequest.com/see5-comparison.html , accessed 13 June 2010

Analiza gradului de poluare a emisarului unei staţii de epurare a apelor uzate folosind data mining

Rezumat

Staţia de epurare a apelor uzate din Ploieşti, situată in zona de sud-est a municipiului, dispune numai de treaptă mecanică pe linia de epurare a apei uzate. Emisarul actual al staţiei este pârâul Dâmbu, care se varsă în râul Teleajen. Poluarea cauzată de tratarea ineficientă a apelor uzate poate crea prejudicii importante pentru populaţia din zonă şi mediul înconjurător. În cadrul acestei lucrări este prezentată o aplicaţie a tehnicii de data mining în domeniul epurării apelor uzate, pentru determinarea şi analizarea gradului de poluare a emisarului unei staţii de epurare.