Ciência Rural, Santa Maria, v.45, n.2, p.267-273, fev, 2015 http://dx.doi.org/10.1590/0103-8478cr20140147
ISSN 0103-8478

Comparative analysis of decision tree algorithms on quality of water contaminated with soil

Análise comparativa de algoritmos de árvore de decisão na classificação da qualidade da água contaminada por solo

Mara Andrea Dota[I*] Carlos Eduardo Cugnasca[I] Domingos Sávio Barbosa[II]

ABSTRACT

Agriculture, roads, animal farms and other land uses may modify the water quality from rivers, dams and other surface freshwaters. In the control of the ecological process and for environmental management, it is necessary to quickly and accurately identify surface water contamination (in areas such as rivers and dams) with contaminated runoff waters coming, for example, from cultivation and urban areas. This paper presents a comparative analysis of different classification algorithms applied to the data collected from a sample of soil-contaminated water aiming to identify if the water quality classification proposed in this research agrees with reality. The sample was part of a laboratory experiment, which began with a sample of treated water added with increasing fractions of soil. The results show that the proposed classification for water quality in this scenario is coherent, because different algorithms indicated a strong statistic relationship between the classes and their instances, that is, in the classes that qualify the water sample and the values which describe each class. The proposed water classification varies from excelling to very awful (12 classes).

Key words: *environmentalcontrol, runoff, wireless sensor networks, machine learning, data mining.*

RESUMO

Agricultura, estradas, fazendas de pecuária e outros usos da terra podem alterar a qualidade da água dos rios, barragens e outras águas doces superficiais. No monitoramentode processos ecológicos para a gestão ambiental, é necessário identificar com rapidez e precisão a contaminação de águas superficiais (em áreas como rios e represas) e subterrâneas, com o escoamento da água contaminada que,advinda, por exemplo, de áreas de cultivo e urbanas. Este artigo apresenta uma análise comparativa dos diferentes algoritmos de classificação aplicados a dados coletados a partir de uma amostra de água contaminada do solo, com o objetivo de criar um modelo de classificação para identificar a qualidade da água. A amostra foi parte de um experimento de laboratório, que partiu de uma amostra de água tratada, adicionando-se frações crescentes de solo. Os resultados mostram que a classificação proposta para a qualidade da água neste cenário é coerente, porque diferentes algoritmos indicaram uma forte relação estatística entre as classes e suas instâncias, ou seja, entre as classes que qualificam a amostra de água e os valores que descrevem cada classe. O modelo de classificação proposto utiliza 12 classes, que variam de excelente a muito péssima.

Palavras-chave: *monitoramento ambiental, enxurradas, rede de sensores sem fio, aprendizado de máquina, mineração de dados.*

INTRODUCTION

Siltation is the major problem deriving from soil runoff processes and causes several ecological damages. ZHOU & ZHANG (2012) described possible river damages caused by siltation and the costs involved to solve it. MARTÍN-VIDE et al. (2014) described a dramatic case of a Bolivianriver collapse caused by siltation and its influences in transnational rivers. The mixture of water and soil may cause other problems,such as clogging potable water pipelines. RASEKH & BRUMBELOW (2014) discuss the need to develop decision support models to help managers todecide about strategies to protect the water distribution systems.

The environmental control of the water quality near urban or agricultural areas can be

[I]Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica, Universidade de São Paulo (USP), Avenida Prof Luciano Gualberto, 380, Travessa 3, Butantã, 05508-010, São Paulo, SP, Brasil. E-mail: maraadota@gmail.com. *Corresponding author.

[II]Instituto de Ciências Agrárias e Tecnológicas, Universidade Federal de Mato Grosso (UFMT), Rondonópolis, MT, Brasil.

Received 02.01.14 Approved 05.29.14 Returned by the author 09.22.14
CR-2014.0147.R1

jeopardizedbylogisticalfactors.InBrazil,forexample, the need to travel long distances to check a watershed basin has limitedthe generation of information and decision-makingregarding environmental quality. For example, in the Rio Miranda basin (State of Mato Grosso do Sul) the watershed area corresponds to areasas large as countries such as The Netherlands. To evaluate 15 sample stations in the main river,more than 10 days of ground work and at least 30 days more of laboratory analyses arenecessary. Then, after at least 30 days other campaigns are performed.

Added to these difficulties is the needof reducing water analysis costs, especially in developing countries. Therefore, methods to speed and to reduce water analyses are necessary. In this context, the use of Wireless Sensor Networks (WSN) may help environmental monitoring,allowing to evaluate the changes thatoccurin real time. A WSN is a special type of *ad hoc* network with capacity for collecting and for processing information in an autonomous way, being these sensors distributed in a determined area (AKYILDIZ et al., 2002; TUBAISHAT et al., 2003; GAJBHIYE et al., 2008). They present great potential of use in agriculture and in environment monitoring, due to the possibility to cover a large area(hundreds or even thousands of sensors nodes), each one able tocollect data and to transmit them to a collecting node (named gateway), which directs the data to a host computer (GAJBHIYE&MAHAJAN, 2008). The proposal of using a WSNdid not replace the use of sample protocols, but it would generate information about what occurred in thecontrolling point of the system before and after the sample campaigns, largely complementing the information to environmental managers.

The need of a number of sensor nodes and short interval times for data collection generates a large volume of information to be analyzed. Moreover, the values collected should be analyzed in an integrated way, sincegathering all the valuesprovidesbetter accuracy than using only one value individually to represent the conditions of the environment as a whole (REN, 1995). For this, sensors can be related among themselves to provide information besides the data collected (SRIVASTA&BUCKMASTER, 2006).

In water quality monitoring, it is necessary to create a model that allows identifying this quality in real time. Our study proposes an approach using Sensor Fusion (SF) techniques for creating a water quality classification to be used in agriculture and in its environmentalissues.This classification will specifically indicate the contamination of river waters by soil spikes.Itintendsto use Artificial Intelligence (AI) techniques that are among the main sensor fusion used techniques, such as statisticand probabilistic,Dempster-Shaper Evidence Theory and others (KAFTANDJIAN, 2005) (SANTOS, 2007). The technique used in this study was abduction by Decision Tree, one of the most used according toWITTEN (2011).NAKAMURA (2007) observed "(…) *Even though it has not been formally used in WSNs, abduction has great potential for different applications such as fault diagnosis, event detection and explanation, and environmental phenomena assessment.* (…)".

Different Decision Treealgorithms were used (*Best-First Decision Tree Classifier*– BFTree, *Functional Trees* – FT, *Naïve Bayes Decision Tree*– NBTree, *Grafted* C4.5 *Decision Tree*– J48graft, C4.5 *Decision Tree*– J48, LADTree) to verify the coherence of the classification of the water quality proposed and to help to build a model that better represents it.

MATERIAL AND METHODS

Stock soil suspension (SSS)

The damaged scenario used in the classification of the water quality wassimulated ina controlled laboratory experiment. A soil suspension solution (SSS) wasmade using naturalsoil and prepared with theinitial concentration of 1.0gL^{-1}into distilled water, emulating the effect of a spike of soil in water pipelines (contamination of a water distribution system) and runoff in low-order stream. The SSS were maintained under constant stirrer conditionsatroom temperature (25ºC).

Sensors and water quality variables

A water multisensory probe (Hanna Instruments HI 9828) was immersed in a 1000ml glassware backer containing tap water. The experimental gadget used was of *batch* type, with constant stirrer, which simulates complete mixture condition of soil in the water column without volumetric losses typically observed in conditions of uniform contamination in water treated distribution pipelines and mixing zones of rivers with effluent contamination. After the parameters stabilization, the registrations began. Thefirst 30s of the experimentrepresent the initial condition. Eleven variables were collected: temperature, pH, pHmV, Redox Potencial (ORP), Dissolved Oxygen (DO ppm and saturation), conductivity (µ_S, µ_SCM, Mohm, Total Dissolved Solids (TDS) and Salinity.The parameters adopted as reference in this investigation are considered as basic descriptive variables of aquatic metabolism, being related to characteristics such as acidity, presence of dissolved salts and oxygenation (TUNDISI & TUNDISI, 2008). Thus, most of the changes in water quality may be evaluated by this basic set of variables.

Soil-water spike simulations and data acquisition

In this experiment, the solution described before followed a pattern of dosage of 1.0ml for each 240s of interval of the 1000ml glassware beaker. Continuous agitation was preformatted at around 30rpm. During the successive solutions additions, the multisensory probe registered the values continuously.After 10 successive spikes (a 40-minute total) the assay ends. The steps for the assay are: First stage: 1.0ml of the solution for every 240s (10 successive applications);Second stage: 10.0ml of the solution for every 240s (10 successive applications). The estimated SSS nominal concentration pertime (time interval) can be seen in table 1.

Data processing and analysis

The data collected during the experiment were submitted to different classification algorithms for verifying whether the classification proposed is coherent; in other words, if the data of each class have strong statistic relation among themselves. Furthermore, the classification algorithms were used to build a model to identify water quality given any input. Nine of the eleven variables collected were considered in the experiment: pH, pHmV, ORP, OD, ODppm, µScm, µScmA, MOhmcm, SDTppm. Figure 1 shows some statistics values to each variable.

Table 1 - Nominal Stock soil suspensionconcentrations over time.

Application sequence (n)	Time interval (s)	Concentration (gL^{-1})
0		0.0000
1	0	1.0 . 10^{-6}
2	240	0.0020
3	480	0.0030
4	241	0.0040
5	481	0.0050
6	242	0.0060
7	482	0.0070
8	243	0.0079
9	483	0.0089
10	244	0.0100
11	484	0.1089
12	245	0.1176
13	485	0.1262
14	246	0.1346
15	486	0.1429
16	247	0.1509
17	487	0.1589
18	248	0.1667
19	488	0.1743
20	249	0.1818
21	489	0.2703

The decision tree algorithms used were: *Best-First Decision Tree Classifier*– BFTree, *Functional Trees* – FT, *Naïve Bayes Decision Tree*– NBTree, *Grafted* C4.5 *Decision Tree*– J48graft, C4.5 *Decision Tree*– J48 andLADTree. These algorithms are part of Data Mining and Machine Learningcollectionimplemented in the WEKA software - WEKA 3.6.10 (HALL, 2009). The WEKA software (Waikato Environment for Knowledge Analysis) began to be written in 1993, using Java, in Waikato University, New Zealand and is nowadays licensed under a General Public License (GPL).

The data collected in the laboratory experiment, produced a sample of 5100 readings. Two experiments were carried out with those data:

Experiment 01: data were divided into two groups: one training set (3400 data) and one test set (1700 data). These data were prepared with routines in C programming language, formatted according to a WEKA input file (.ARFF extension). The division into two files was performed randomly, removing data lines from the original file (with 5100 data) to build the test data file. It is necessary to make this division to avoid the overfitting problem,which happens when a lot of data are considered tobuild the model (using decision tree) and this model becomes perfect for this unique data set (ABERNETHY, 2010). In fact, the intention is to create a model to predict the output of other data setsdifferent from the ones used for building it;the training data file is used to build a model and the test data file is used to verify if its accuracy does not reduce with different data sets, guaranteeing that the model will predict the output to unknown values accurately.

Experiment 02: Using the whole data set and applying the k-fold-cross-validation (k = 10). It consists in dividing the training data set into K-test times (usually, K=10). The model is trained with nine training sets and one test set is applied to validate it. This is done 10 times (K=10).

In this investigation, it was adopt by definition of water quality or alteration in water quality, when any change happens to the initial condition of the water sample under test that can be detected observing the collected data from readings in the experiment using the probe. Quality classes were defined in this study as being as follows: the full data set (5100 values) was divided into 12 classes of equal size with 425 readings each. And the first class is the best water quality and the twelfth is the worst, because, according to the experiment conduction, it is

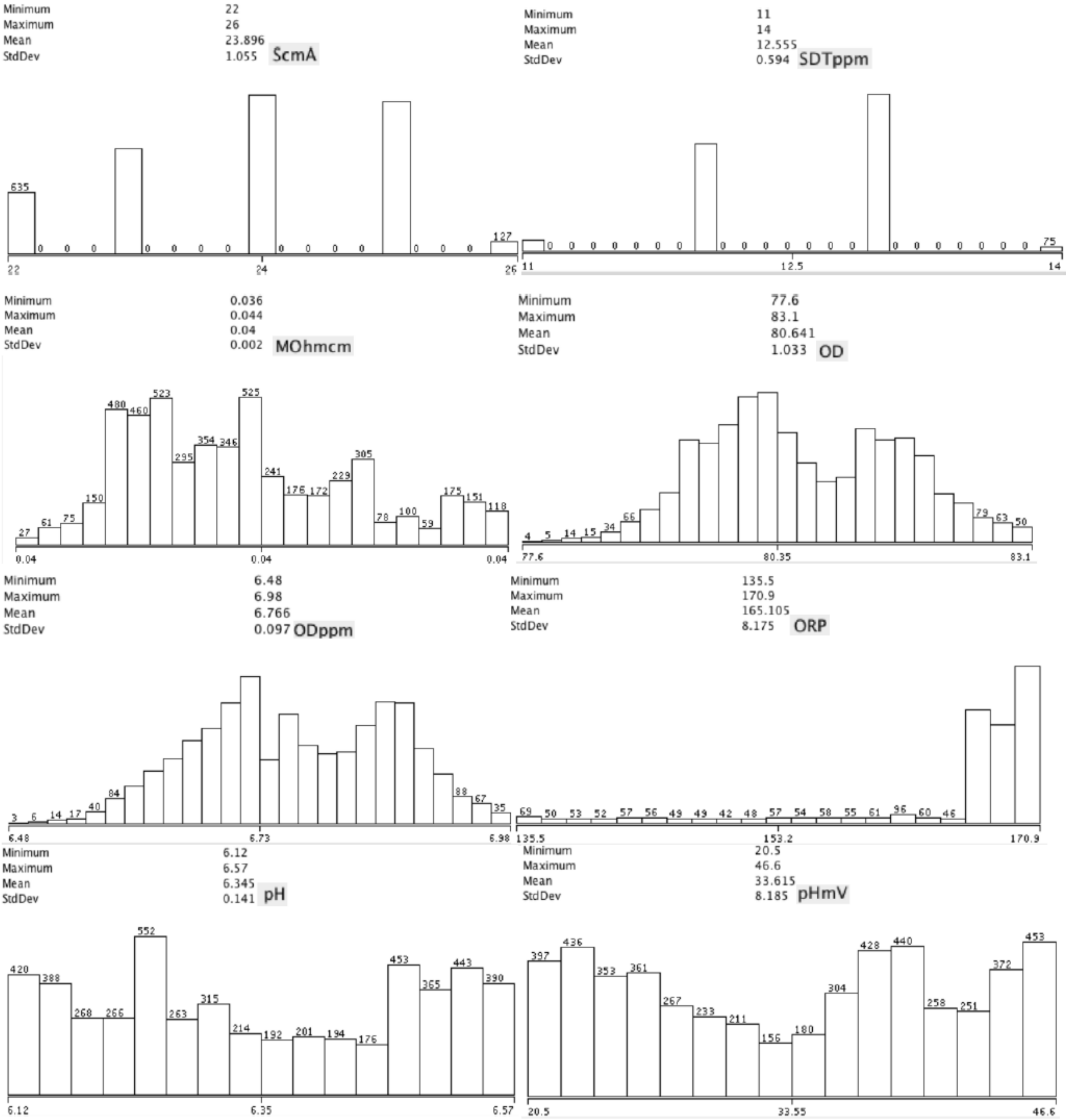

Figure 1 - Some statistics values to each variable are present here, as minimum and maximum values, mean and Standard Deviation. DO = dissolved oxygen; TDS = Total dissolved Solids; pH = Hydrogen Potential; ORP = Oxidation-Reduction Potential; Scm = electric conductance; MOhm = electric resistance.

known that the first data class has no contamination and the last class is the most contaminated of the experiment.Twelve classes were used because preliminary studies demonstrate that this was the most appropriate (CONAMA, 2005) and they were named as excelling (first class), excellent, nearly excellent, very good, good, not so good, a little bad, bad, very bad, nearly awful, awful, very awful (twelfth class).

To evaluate the classification coherence, the algorithms selected were executed using an experimental procedure as suggested by WEKA (using a training sample and a test sample). The training sample is used to build the model. The test sample verifies if the model is correct.

RESULTS AND DISCUSSION

Table 2 presents some results from experiment 01. The algorithms are placed in decreasing order of the Correctly Classified Instances

Table 2 - Kappa is a value that represents the relationship between instances of the same class, ranging from 0 to 1, with 1 being a strong relationship. Columns 2, 3 and 4 refer to experiment 01. Columns 5, 6 e 7 refer to experiment 02.

Algorithm	CCI (%)	ICI (%)	Kappa	CCI (%)	ICI (%)	Kappa
BF Tree (training)	99.5000	0.5000	0.9945	98.902	1.098	0.988
BF Tree (test)	99.0588	0.9412	0.9897			
FT (training)	99.4706	0.5294	0.9942	98.6863	1.3137	0.9857
FT (test)	99.0000	1.0000	0.9891			
NB Tree (training)	99.2647	0.7353	0.9920	98.6863	1.3137	0.9857
NB Tree (test)	98.6471	1.3529	0.9852			
J48graft (training)	99.2647	0.7353	0.9920	98.8039	1.1961	0.987
J48graft (test)	98.7647	1.2353	0.9865			
J48 (training)	99.2647	0.7353	0.9920	98.8824	1.1176	0.9878
J48 (test)	98.8824	1.1176	0.9878			
LAD Tree (training)	94.7059	5.2941	0.9422	93.7255	6.2745	0.9316
LAD Tree (test)	94.0588	5.9412	0.9352			

(CCI), considering the training group. One instance in this case is a group of readings of a determined time of the 9 variables considered. The value of the CCI indicates how many instances were classified in an appropriate way to their class, and the values of table 2 demonstrate that,from the construction of the model to its verification, the CCI number was high. The last two algorithms of table 2 were the ones with lowest CCI.

In table 2, it is possible to note that the algorithms BFTree, FT, NBTree, J48graft, J48 obtained a high percentage of CCI. It is also possible to verify this in the graph of figure 2 (left), which presents the percentualvalues of CCI and ICI (Incorrectly Classified Instances).

The “kappa” value is an indexanalogous to the correlation coefficient. Kappa is equal to zero in the absence of any relation and approximated to one in a strong statistic relation between the class and the attributes of the instances (ABERNETHY, 2010). It is possible to note by the kappa value shown in table 2 to the same algorithms that show high CCI degree also present kappa value very close to one, showing that the relation among the different instances inside each class are correctly classified.

Another interesting result is relatedto CCI and kappa presented by the training and test samples. The CCI between the two samples submitted to the same algorithms show little difference, indicating that the model accuratelypredictsthe outputfor values different from the ones used in its creation.

Results from experiment 02 also indicate that many instances were classified according to their class because a high percentage of CCI was obtained (see table 2). In this case, these results were obtained from an average of 10 runs. High percentages of CCI show a high index of instances classified correctly. Analyzing the CCI indexes, it may be noted that BFTree, J48graft and J48 algorithms were the best-rated classes (Figure 2– right). The results of experiment 02 are more accurate than those of experiment 01 because they are the result of an average of 10 runs. Kappa also shows a strong relationship between instances of the same class, close to 1 (Table 2).

These results show that the classification proposed is coherent, since the different algorithms proved a strong statistic relation between the classes and their instances; in other words, among the classes that qualify the water sample from Excelling to Very Awful and the values that describe the sample. The algorithms that better rated the sample were BFTree, J48graft and J48.

CONCLUSION

The importance of this study lies in the need to quickly and accurately identify the contamination of superficial water bodies with runoff contaminated by soils suspension, for example, from cultivation areas and urban areas into the low-order streams and water distribution pipelines. The process used to identify these changes was by ground sample and the best methods to control contamination in real time involve high costs. The proposal is to use data from the sensors (WSN) to determine this contamination in real time, aiming to aid environmental managers’ decisions.

With the results, the proposed classification is expected to motivate the creation of a water quality classificationfor this context obtained by the models generated by the Decision Tree algorithms; this index is expected to represent the real conditions of the contamination of the water by soil, or at least to serve as a reference to emit alerts concerning

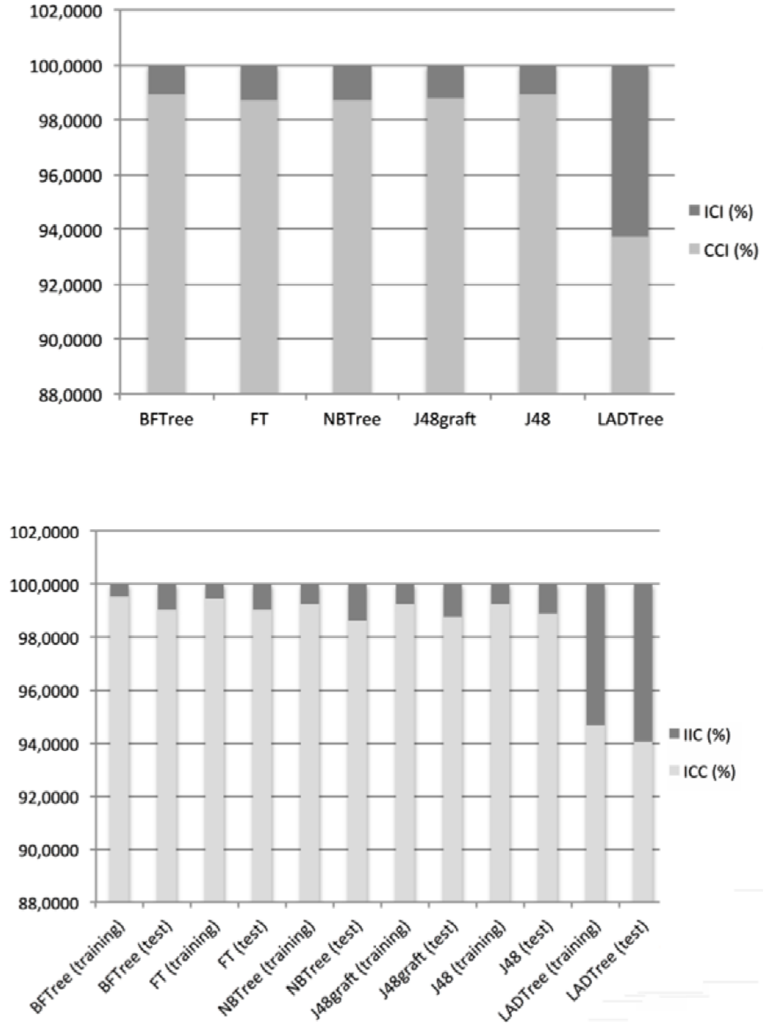

Figure 2 - Graphic on the left shows results of CCI and ICI values for both samples submitted to each algorithm proportionally exposed (experiment 01). Note that the two algorithms on the right obtained the lowest CCI values.Graphic on the left shows results of the CCI and the ICI values from experiment 02, exposed in a proportional way. It is possible to note that BFTree, J48graft and J48 obtained the highest CCI values.

quality alterations, allowing the investigation *in situ* to confirm or not the emitted alert. The next step is to investigate, among the algorithms that obtained the highest ICC, which one best contributes to the construction of a water model classification. As a future study, the authors intend to use the WSN for the environmental monitoring collecting data in real time.

ACKNOWLEDGMENTS

The authors thank the Foundation for Research Support of the State of Mato Grosso (FAPEMAT) for supporting this research (project nº 278718/2010 and nº 313438/2011), and the Coordination of Improvement of Higher Education Personnel (CAPES) via the inter-institutional Ph.D. program with the Engineering School at the University of São Paulo and the Federal University of Mato Grosso (UFMT). Thanks also to Professor Dr. Pablo A. Dalbem de Castro (IFSP, São Carlos/SP) for the collaboration related to Artificial Intelligence area.

REFERENCES

ABERNETHY, M. **Data mining with WEKA, Part 2**: classification and clustering. Series: Data Mining with WEKA, Part2. DeveloperWorks, IBM Technical Library. 2010. Available from: <http://www.ibm.com/developerworks/library/os-weka2/os-weka2-pdf.pdf>. Accessed: Dec. 10, 2012.

AKYILDIZ, I.F et al. Wireless sensor networks: a survey. **Journal Computer Networks**: The International Journal of Computer and Telecommunications Networking, v.38, n.4, p.393-422, 2002.Available from: <http://dl.acm.org/citation.

cfm?id=585777>. Accessed: Dec. 10, 2012. doi: 10.1016/S1389-1286(01)00302-4.

CONAMA (CONSELHO NACIONAL DO MEIO AMBIENTE). 2005. **Resolução Conama** n.357. Available from: <www.mma.conama.gov.br/conama>. Accessed: May 25, 2014.

GAJBHIYE, P.; MAHAJAN, A. A survey of architecture and node deployment in Wireless Sensor Network. In: **Applications of Digital Information and Web Technologies**, 2008. ICADIWT 2008. First International Conference on the. p.426-430. 4-6 Ago. 2008. Available from: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4664386&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4664386>. Accessed: Dec. 10, 2012.doi: 10.1109/ICADIWT.2008.4664386.

HALL, D.L.; LLINAS, J. An introduction to multidata sensor fusion. **Proceedings of IEEE**, v.85, n.1, p.6-23,1997.

KAFTANDJIAN, V. et al. The combined use of the evidence theory and fuzzy logic for improving multimodal NDT systems. In: Instrumentation and Measurement. **IEEE Transactions**, v.54, n.5, p.1968-1977, 2005. Available from: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1514651&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D1514651>. Accessed: Dec. 10, 2012.doi: 10.1109/TIM.2005.854255.

MARTÍN-VIDE, J.P. et al.Collapse of the Pilcomayo River. **Geomorphology (Journal), Discontinuities in Fluvial Systems**, v.205, p.155-163,2014.ISSN 0169-555X. Available from: <http://www.sciencedirect.com/science/journal/0169555X/205>. Accessed: Dec. 10, 2012. doi: 10.1016/j.geomorph.2012.12.007.

NAKAMURA, E.F. et al. Information fusion for wireless sensor networks: methods, models, and classifications. **Journal ACM Computing Surveys (CSUR) Surveys**, v.39, n.3, p.1-55, 2007. Available from: <http://dl.acm.org/citation.cfm?id=1267073>. Accessed: Dec. 10, 2012.doi: 10.1145/1267070.1267073.

RASEKH,A.; BRUMBELOW, K. Drinking water distribution systems contamination management to reduce public health impacts and system service interruptions. **Journal Environmental Modelling & Software archive**, v.51, p.12-25, 2014. Available from: <http://dl.acm.org/citation.cfm?id=2561292>. Accessed: Dec. 10, 2012. doi: 10.1016/j.envsoft.2013.09.019.

REN, C.L.; MICHAEL, G.K. **Multisensor integration and fusion for intelligent machines and systems**. Norwood, NJ: Ablex, 1995. ISBN:0-89391-863-6.

SANTOS, T.G. et al. Fusão de dados em ensaios não destrutivos utilizando decisão fuzzy para a avaliação de soldas obtidas pelo processo de fricção linear. **Revista Soldagem Insp**, v.12, n.3, p.124-132, 2007.

SRIVASTAVA, A.K. et al. Precision agriculture.In: McCANN, P. (Ed). **Engineering principles of agricultural machines**. Michigan: ASABE, 2006. C.6, p.123-138.

TUBAISHAT, M.; MADRIA, S.Sensor networks: an overview. In: **Potentials, IEEE**, v.22, n.2, p.20-23, 2003. Available from: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1197877&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F45%2F26953%2F01197877>. Accessed: Dec. 10, 2012.doi: 10.1109/MP.2003.1197877.

TUNDISI, J.G.; MATSUMURA-TUNDISI, T. **Limnologia**. São Paulo: Oficina de Textos, 2008. p. Total de p.?790.

WITTEN, I.H. et al.The WEKA data mining software: an update. In: **Newsletter ACM SIGKDD Explorations Newsletter**, v.11, n.1, p.10-18, 2009. Available from: <http://dl.acm.org/citation.cfm?id=1656278>. Accessed: Dec. 10, 2012. doi: 10.1145/1656274.1656278.

WITTEN, I.H. et al. **Data mining**: practical machine learning tools andtechniques.3.ed.SanFrancisco:MorganKaufmann.2011.p.629.

ZHOU, J; ZHANG, M.Coarse sediment and lower Yellow River siltation. **Journal of Hydro-environment Research**, v.6, n.4, p.267-273,2012.ISSN 1570-6443. Available from: <http://www.sciencedirect.com/science/article/pii/S1570644312000512>. Acessado em: 10 de dez. de 2013. doi: http://dx.doi.org/10.1016/j.jher.2012.05.005.