Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 70 (2014) 864 – 871

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Computing and Control for the Water Industry, CCWI2013

Mining solution spaces for decision making in water distribution systems

J. Izquierdo[a]*, I. Montalvo[b], R. Pérez-García[a], E. Campbell[a]

[a]*Fluing-IMM, Universitat Politècnica de València, Cno. de Vera s/n, 46022, Valencia, Spain*
[b]*3S Consult GmbH, Albtalstraβe 13, 76137, Karlsruhe, Germany*

Abstract

Data mining solutions can be applied in combination with evolutionary algorithms to extract relevant information from solution spaces analyzed during optimization processes regarding water distribution system (WDS) design. Firstly, results from data mining can be introduced into the evolutionary algorithms to guide the search of solutions. Secondly, data mining techniques can be used not only to help explore the population of potential solutions but also to exploit data regarding the behavior of the WDS under different work conditions. As a result, applications can be developed for supporting decision making regarding WDS models in both offline and online contexts.

Keywords : Data Mining, Evolutionary Algorithms, Agent Swarm Optimization, Kohonen SOM, Bayesian Networks, Decision Making

1. Introduction

Almost any decision making process in engineering can be turned into an optimization problem. The flexibility introduced by evolutionary algorithms has allowed the use of virtually any objective function for evaluating solutions, even when these evaluations require running complex mathematical and/or procedural simulations of the systems under analysis. In this line of work, several applications have been developed within the field of urban hydraulics regarding design, calibration, energy saving, etcetera, Bei and Dandy (2012), Berardi et al. (2012), Wu and Behandish (2012), Shean and McBean (2010), Montalvo et al. (2010c).

Evolutionary algorithms have been extensively applied to solve optimization problems in the water industry, Liong and Atiquzzama (2004), Geem (2006), Izquierdo et al. (2008), Jin et al. (2008), Montalvo et al. (2010b). In

* Corresponding author. Tel.: +34 628028804.
E-mail address: jizquier@upv.es

doi:10.1016/j.proeng.2014.02.095

water distribution system (WDS) design they have brought a lot of flexibility for evaluating objectives in a way hardly achievable by other methods. Running mathematical simulations of water networks under different load conditions can be now part of a global decision making process where solutions evolve based on some “rules” defined in the evolutionary algorithm in use. In this context, rules are understood as the what, when and how things should be done to transform a population of solutions into better solutions during an evolutionary process.

Each algorithm has its own set of rules and most of them are based on the emulation of a paradigm. Genetic Algorithms (GA), Goldberg (1989), for example, are based on the theory of natural evolution. Particle Swarm Optimization (PSO), Kennedy and Eberhart (1995), is based on the way birds in a flock find its destination, which is more inspired on social evolution than on a genetic evolution. Ant Colony Optimization, Dorigo et al. (1996), is based on the foraging behavior of ants. Simulated annealing, Kirkpatrick et al. (1983), takes its name and inspiration from annealing in metallurgy. There are many other examples with different sources of inspiration.

To select what is the most appropriate algorithm for solving a specific problem is not easy task. Some algorithms can perform better than others in some problems and worse or very poorly in some others. This fact may indicate that, first, their rules apply better to certain problems than to others and, second, that, even if the population of solutions does evolve, the way it does is much more static and does not suitably evolve during the solution search process.

A simple example can be seen by analyzing what happens with the parameters used by evolutionary algorithms. These parameters certainly change how algorithms perform; however, frequently they remain static. Fine-tuning parameters to obtain better results from evolutionary algorithms is, in many cases, part of a hand-made meta-process where specialists, using their experience or recommendations from the literature, start changing parameters, testing algorithms performance and keeping the best parameter set of values.

An attempt to change parameters as part of the evolutionary process was done in Montalvo et al. (2010a). Also in Lessmann et al. (2011) a support vector machine was trained to generate PSO parameters while the solution space of a problem was explored. Those were just preliminary steps in changing the “behavior” of an optimization algorithm as part of the solution search process. But the real big step should not come by just better adjusting parameters to the problem being solved but by influencing more directly the way the search of solutions is performed.

2. Rule-based agents

WDS optimization using evolutionary techniques is not easy to achieve as the size and complexity of problems increase. Good results can be obtained from relative small networks – there are many examples in the literature – but good solutions are harder to obtain for larger problems and, in addition, those solutions can differ from what engineering good practice and sense could suggest. Now the question is: how the solution search process is done when the size of problems increase and the interrelations among variables and objectives get more complex?

Traditionally, the solution search process has been totally ignorant of the specific problem being solved. The process has been the same no matter the size, the complexity and the problem domain. In these days of cloud computing, authors can be tempted to use more and more computing resources for solving “real world problems”. It is not a wrong idea but taking it literally and without carefully thinking on the improvement of the search process itself, this would lead researchers closer to brute force approaches to find the best objective than to clever ways of exploring and exploiting the solution space of such an NP-hard problem as the design of the network of a WDS by investing a reasonable amount of resources in the search process.

In the opinion of the authors, the search process should be as close as possible to the problem instance being solved. Algorithms adapting their behaviors to problems will have more chances to succeed. A method to achieve this can be done by combining the way evolutionary algorithms work with the introduction of rules based on the domain of the problem being solved. This idea can be supported with the use of rule-based agents combined with evolutionary techniques. The use of rule-based agents was one of the principles followed for developing Agent Swarm Optimization (ASO), Montalvo et al. (2010c), where various evolutionary techniques can be combined based in a common framework. As a matter of fact, ASO, more than just combining those techniques, is an attempt to include rules for taking the algorithms in use closer to the problem being solved.

In Montalvo et al. (2010c) and Montalvo (2011) rule-based agents were used for solving problems regarding WDS design. These works considered the idea of defining rules related to the problem domain and using those rules as part of the solution search process. Rules developed in these previous researches were exclusively focused on pipe sizing problems. Specifically, they tried to avoid pipe diameters increase from upstream to downstream and their application led to a reduced search space and the finding of good solutions with a better performance of PSO and GA.

Despite the ideas worked quite well, they had a couple of disadvantages. Firstly, they were "hard-coded": no changes could be enforced without changing the existing source code or adding more code to the software supporting the algorithms. Secondly, it is hard to discover new rules to be included in order to continue improving the solution search process.

Developing rule-based agents to be combined with evolutionary techniques in order to improve the performance of algorithms requires the active participation of specialists from the problem domain. It would be hard to develop good rules without a good understanding of the problems in the context of their domain. But even for people with a deep understanding of the problem domain it is hard to define rules that could be generalized and applied to work in combination with evolutionary techniques. It is much easier to analyze what should be better done to improve the search in a specific problem instance than to define a generalized way to do it. Even if a generalized way is found, it should be adjusted to be expressed in a programming language in order to be effectively used. The proposal of this paper is to include data mining techniques as a step for dynamically generating rules that could be used to improve the efficiency of solution search processes.

3. Techniques for mining solution spaces

During the execution of evolutionary algorithms, typically the amount of solutions evaluated represents quite a small percentage of the total solution space corresponding to the problem being solved. Nevertheless, the amount of solutions evaluated is still considerable, and most evolutionary techniques use just a small part of them at a time. Many of the solutions evaluated during the search process are "forgotten" after one generation and combined experience of several generations is typically not well exploited.

Data mining techniques can enable deeper insight into the many "good" solutions that have been just simply glimpsed and have been rapidly disregarded because they were dominated by better solutions during an ephemeral moment in the evolution process. Based on a database obtained by suitably recording certain of those disregarded solutions, data mining techniques can help better understand and describe how a system could react or behave after the introduction of changes.

This paper proposes to apply data mining techniques to the set of solutions evaluated after several generations of a single run of an evolutionary algorithm in order to extract rules intended initially to be used by the following generations. The kind of rules being tested during this research can be divided in two types:

- Rules trying to reduce the variables ranges, concentrating them in regions with higher probabilities of obtaining good solutions.
- Rules trying to define relations among variables by identifying those variables that should conveniently have values bigger/smaller than other existing variables.

Two basic methods have been explored for generating rules. One of them is based on Kohonen maps. The second one is based on the construction of a Bayesian network combined with association rules for better identifying variable dependences and regions with bigger probabilities of obtaining good solutions.

3.1. Kohonen maps and rule extraction

The map of Kohonen is known as an important paradigm of unsupervised neural network to analyze data, Kohonen (2001). The learning algorithm follows the pattern of competitive models, but the update rule produces an output layer in which the topology of the input patterns is preserved. This means that if two patterns are close in the input space (in the sense of some similarity measure, such as measures used in winner-take-all strategies) their corresponding active neurons are also topologically close in the output layer. A network that performs this function is called a map of characteristics. These maps not only group the input patterns in clusters but also visually describe the relationship between the clusters of the input space.

A Kohonen map is a two dimensional array of fully connected neurons with the input vector organized in a square or a hexagon (Fig. 1). Hexagonal arrangement is advised because at the end of the learning process it provides better visualization of the structure of the input space.

The topology preserving property is obtained by a learning rule involving the winning neuron and its neighbors in the update process. So, close neurons learn to activate when presented with similar patterns. During training, the network allocates a position to the neurons on the map based on the effect of the dominant feature of the input pattern. For this reason Kohonen maps are called self-organizing maps (SOM).

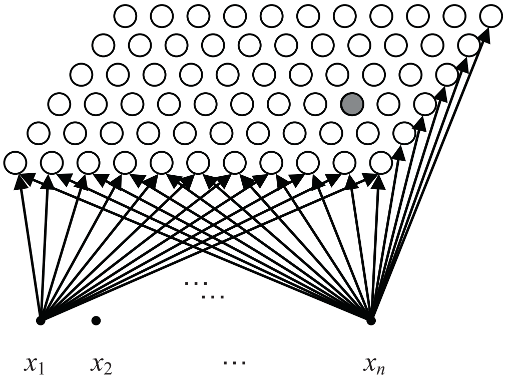

Fig. 1. Two-dimensional Kohonen SOM showing the winner neuron (solid)

If the input space is highly dimensional, Kohonen maps can be interpreted as linear projectors onto a two-dimensional array of neurons that takes into account the probability density of the data and preserves the topology of the original input pattern. Preserving the topology of the original input pattern is a great advantage for visualizing results. Nevertheless, despite their power of visualization, SOMs cannot provide a full explanation of their structure and composition without further detailed analysis. One method towards filling this gap is the unified distance matrix or *U*-matrix technique of Ultsch et al. (1993) and Malone et al. (2006). The *U*-matrix technique calculates the weighted sum of all euclidean distances between the weight vectors for all output neurons. The resulting values can be used to interpret the clusters created by the SOM, Malone et al. (2006).

A SOM should be trained first. In the case of this research, the training data are the solutions already evaluated by an evolutionary algorithm including also the values of the objectives involved in the optimization problem being solved. Next, boundaries from the components/*U*-matrix are identified. Once the process of selecting boundaries is complete, important input features/variables are chosen to be used for constructing the rules. This process is explained in detail in Malone et al. (2006).

Extracted rules are in the format of symbolic, propositional rules in conjunctive normal form. Each rule consists of an antecedent that describes the cluster characteristics and a consequent pertaining to a cluster. In the problems studied in this research the goal is to find those rules associated with the existence of good solutions. In this context, good solutions could be considered as those solutions with at least one objective with a value no more than 5% different compared to the best known value for the objective in question.

Rules generated this way are not evaluated using expert knowledge directly. Their evaluation is only based on data and structure of patterns found. Despite it can be helpful in many cases, it is much better to include expert knowledge in the evaluation of rules whenever that expert knowledge is available. The implementation of this idea moved our research to explore the use of Bayesian networks and association rules.

3.2. *Bayesian networks and association rules*

Bayesian networks, Pearl (1988), Heckerman (1995), are a type of probabilistic graphical models that are characterized by modeling causal relationships. These probabilistic graphical models arise as a result of the union between graph theory and probability theory, since when building a probabilistic mathematical model, it is essential to take into account two important components regarding the information that is available: qualitative and quantitative information of the problem, Susi (2007). The qualitative information associated introduces dependencies between the variables of the model. Based on graph theory, this information can be summarized by a graph, where nodes represent the variables of the problem and the edges of the graph the dependence and causal relations between them, so that the lack of edges induces independence relationships. Furthermore, when building a mathematical model, probabilistic information is also available concerning the probability distribution of the variables of the problem, also known as quantitative information of the problem. The theory of probability is fundamental to obtain relations between the probability distributions of the variables of the problem. These distributions can be estimated from a set of data or, sometimes, from information provided by experts in the field on the problem at hand.

With the qualitative and quantitative information of the problem a probabilistic graphical model associated with it is defined. This is the link between graph theory and probability theory, given by a pair (G, P) where G is the graph representing the qualitative information of the problem, and P is the set of conditioned distributions providing the joint probability distribution of the problem.

Association analysis, introduced in Agrawal et al. (1993) is one of the descriptive models used in data mining. It is aimed to examine items that are seen together frequently in data set and to reveal patterns that help decision making. These patterns are presented as “association rules” or “frequent itemsets” in association analysis. A problem encountered with association rules is that a great number of patterns are generated even for small data sets. To figure out this problem, patterns obtained by association analysis should be evaluated according to their interestingness levels and the patterns which are found interesting according to the evaluation should be eliminated, Ersel and Günay (2012). These evaluations can be categorized into “objective interestingness measures” and “subjective interestingness measures”. While objective measures are based on data and structure pattern, subjective measures are also based on expert knowledge in addition to data and structure of the pattern, Tan et al. (2006). Subjective interestingness measures are generally specified through belief networks and here is precisely where Bayesian networks can be used because of their capacity for representing beliefs.

Bayesian network can be pre-elaborated by experts on the domain of the problem being solved. It makes it possible to use the preliminary Bayesian network created by experts to generate subjective interestingness measures for evaluating association rules. Existing belief networks originated with specialist criteria can be extended or improved by using the information resulting from the association analysis. New solutions explored by evolutionary algorithms will add more data that can be also used in improving both the structure and parameters of Bayesian networks. Several methods to achieve this can be found in Margaritis (2003). The interchange between Bayesian network and results from association rules brings dynamics to the process of knowledge discovery by updating iteratively both generated rules and beliefs.

4. Exploiting information

Both changing parameters dynamically based on the “search experience” of agents and the use of generated rules based on previously evaluated solutions can be really useful when solving optimization problems using evolutionary techniques. Dynamically adjusting parameters help automatize the process of fine-tuning the performance of algorithms. Fig. 2 provides a good example of what happens while evolving PSO parameters during a PSO execution. It shows the evolution of *c*1 and *c*2 – individual and social learning parameters – for the particles that eventually got the leadership in 100 runs of the algorithm.

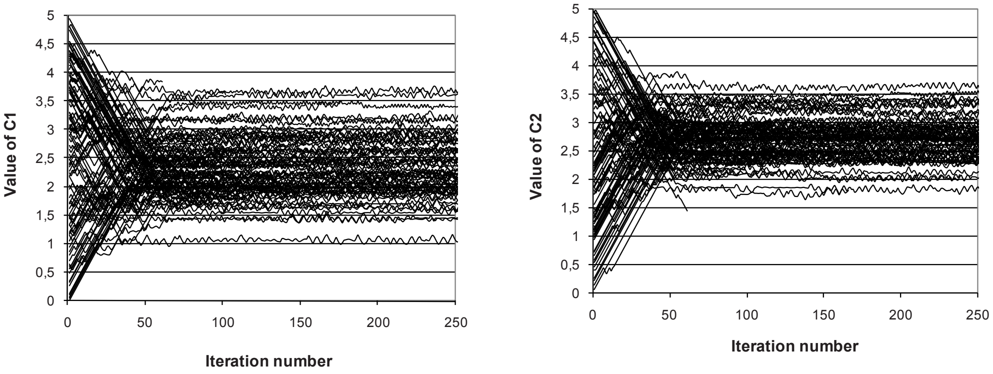

Fig. 2. Evolution of learning rates parameters for the eventual leaders of one hundred runs of PSO, Montalvo et al. (2010a).

It can be seen a clear tendency to approach certain values within a range between 2 and 3 for both parameters. As a consequence, we claim that if there is not a clear idea of what parameter value to use then the process of evolving parameters during a run can help identify the regions where the best parameter values could be enforced.

Also, setting rules makes it possible to reduce considerably the search space of problems. This reduction, together with the way the search was done after introducing rules, brought very good results in several examples, Montalvo et al. (2010c), Montalvo (2011). In a well-known simple benchmark problem, the two loop network described in Alperovits and Shamir (1977), the reduction of its search space amounted to more than 97% after using the rule of downstream-decreasing diameters. Despite that reduction, there is still a big set of solutions to explore but the probability of finding good solutions in the reduced search space is much bigger.

In the examples presented in Montalvo et al. (2010c, 2013) an extensive use of rule-based agents has been done. Without including rules, the evolutionary algorithms used had many difficulties for finding good solutions, and most of the solutions found were not suitable for practical engineering purposes. Introducing rules changes the situation completely but the dependencies of the rules defined by the expert were crucial. Leaving the extraction of rules to be applied to data mining techniques improves results compared to the use of evolutionary algorithms without any rule included. That was the case of applying SOM considering just the available data and patterns found for the evaluation of extracted rules. Much better results were obtained using a combination of association analysis and Bayesian networks touched by the hand of specialists. In this second case, results were similar than those obtained in Montalvo et al. (2010c, 2013) despite this time part of the rules used were extracted dynamically from explored solutions during run time. Fig. 3 can give an idea of the size and complexity of one of the networks used as benchmarking for pipe sizing in Montalvo et al. (2013). It is the result of applying ASO incorporating rules extracted from a Bayesian network combined with association analysis.

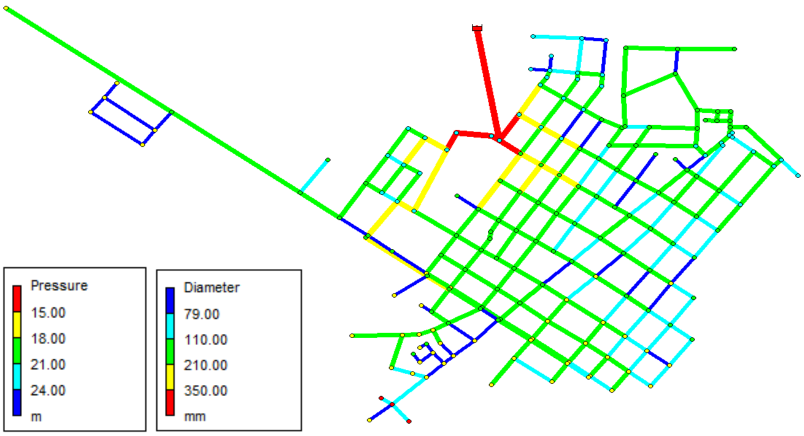

Fig. 3. San José´s network solution after sizing all pipes

5. Conclusions

Despite big efforts have been made on the development of optimization strategies, the problem-dependent (hydraulic engineering in our case) component of the problems should not be underestimated. In fact, the best strategy is to be able to introduce as much information as possible from the problem itself into the optimization solution to be used. This idea can only be achieved by having both deep understanding of the problem to be solved and a flexible framework to combine that understanding with good optimization techniques. In this contribution we have argued that the use of data mining techniques can provide the flexible framework sought, due to its recognized ability to explore and exploit complex and massive databases in many fields.

The available computation capacity of these days is a great opportunity to improve evolutionary algorithms with data analysis methods. This way better understanding of solution search spaces will be gained when solving optimization problems. Additionally, it will prepare the basis for solving real time problems where conditions could change rapidly, and answer should be given without the possibility of performing too long calculations. The combination of evolutionary algorithms and data mining techniques is surely a good approach for solving coming challenges.

Acknowledgements

This work has been supported by project IDAWAS, DPI2009-11591, of the Dirección General de Investigación of the Ministerio de Ciencia e Innovación of Spain, and ACOMP/2011/188 of the Conselleria d’Educació of the Generalitat Valenciana.

References

Alperovits, E., Shamir, U., 1977. Design of optimal water distribution systems. Water Resources Research 13(6): 885-900.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining associations rules between sets of items in large databases, in Proc. ACM SIGMOD international conference on Management of data, 1993, pp. 207-216.

Bei, W., Dandy, G.C., 2012. Retraining of metamodels for the optimization of water distribution systems, in Proc. Water Distribution System Analysis Conference, Adelaide, Australia, 2012, pp. 36-47.

Berardi, L., Laucelli, D., Giustolisi, O., 2012. A decision support tool for operational optimization in WDNETXL, in Proc. Water Distribution System Analysis Conference, Adelaide, Australia, 2012, pp. 48-65.

Dorigo, M., Maniezzo, V., Colorni, A., 1996. The ant system: optimization by a colony of cooperating ants, IEEE Transactions on Systems, Man and Cybernetics—PartB, 26(1), 1–13.

Ersel, D., Günay, S., 2012, Bayesian networks and association analysis in knowledge discovery process. Istatistikciler Dergisi 5, pp 51-64.

Geem, Z. W., 2006. Optimal cost design of water distribution networks using harmony search. Engineering Optimization 38(3): 259-280.

Goldberg, D. E., 1989. Genetic algorithms in search, optimization and machine learning, Addison-Wesley, Reading, Ma.

Heckerman, D., 1995. A Tutorial on Learning With Bayesian Networks". Technical Report, Msr TR-95-06, Microsoft Research, Redmond, WA.

Izquierdo, J., Montalvo, I., Pérez, R., Tavera, M., 2008. Optimization in water systems: a PSO approach. Business and Industry Symposium (BIS), Ottawa, Canadá.

Jin, X., Zhang, J., Gao, J. L., Wu, W. Y., 2008. Multi-objective optimization of water supply network rehabilitation with non-dominated sorting Genetic Algorithm-II." Journal of Zhejiang University SCIENCE A 9(3): 391-400.

Kennedy, J., Eberhart, R. C., 1995. Particle swarm optimization, in Proceedings of the IEEE International Conference on Neural Networks, Piscataway, NJ, 1942-1948.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., 1983. Optimization by Simulated Annealing. Science 220 (4598): 671–680.

Kohonen, T., 2001. Self-Organizing Maps. Springer-Verlag, Berlin, Heidelberg.

Lessmann, S., Caserta, M., Montalvo, I. 2011. Tuning metaheuristics: A data mining based approach for particle swarm optimization, Expert Systems with Applications: An international Journal, 38(10), 12826-12838.

Liong, S. Y., Atiquzzama, M., 2004. Optimal design of water distribution network using shuffled complex evolution. Journal of The Institutios of Engineers, Singapore 144(1): 93-107.

Malone, J., McGarry, K., Wermter, S., Bowerman, C., 2006, Data mining using rule extraction from Kohonen self-organising maps. Neural Computing & Applications, 15(1), pp 9-17.

Margaritis, D., 2003, Learning bayesian network structure from data. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.

Montalvo, I., 2011. Diseño óptimo de sistemas de distribución de agua mediante Agent Swarm Optimization. PhD doctroal dissertation. Universitat Politècnica de València, Valencia, Spain.

Montalvo, I., Izquierdo, J., Pérez-García, R., Herrera, M., 2010a. Improved performace of PSO with self-adaptive parameters for computing the optimal design of water supply systems. Engineering Applications of Artificial Intelligence 23(5): 727-735.

Montalvo, I., Izquierdo, J., Schwarze, S., Pérez-García, R., 2010b. "Multi-objective particle swarm optimization applied to water distribution systems design: An approach with human interaction." Mathematical and Computer Modelling 52: 1219-1227.

Montalvo, I., Martínez Rodriguez, J. B., Izquierdo, J., Pérez-García, R., 2010c. Water Distribution System Design using Agent Swarm Optimization. Proc., 12th Water Distribution Systems Analysis Symp, Tucson, Arizona: K. Lansey, C. Choi, A. Ostfeld, and I. Pepper, 2010.

Montalvo, I., Izquierdo, J., Herrera, M., Pérez-García, R., 2013, Water supply system computer-aided design by Agent Swarm Optimization. Computer-Aided Civil and Infrastructure Engineering, under second review.

Pearl, J., 1998. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Representation and Reasoning Series (2nd printing ed.). San Francisco, California: Morgan Kaufmann.

Shen, H., McBean, E., 2010. Hydraulic calibration for a small water distribution network. Proc., 12th Water Distribution Systems Analysis Symp, Tucson, Arizona: K. Lansey, C. Choi, A. Ostfeld, and I. Pepper, 2010.

Susi, R., 2007. Análisis de sensibilidad en las Redes Bayesianas gaussianas. Doctoral dissertation, Universidad Complutense de Madrid, Spain.

Tan, P., Steinbach, M., Kumar, V., 2006, Introduction to data mining, Addison-Wesley, Boston, 769p.

Ultsch, A., Mantyk, R., and Halmans, G., 1993, Connectionist knowledge acquisition tool CONKAT, in: J. Hand, ed., Artificial Intelligence Frontiers in Statistics AI and Statistics, Vol. III (Chapman and Hall, London, UK, 1993) 256-263.

Wu, Z. Y., Behandish, M., 2012. Real-time pump scheduling using genetic algorithm and artificial neural network based on graphics processing unit, in Proc. Water Distribution System Analysis Conference, Adelaide, Australia, 2012, pp. 1088-1099.