CLEI ELECTRONIC JOURNAL, VOLUME 12, NUMBER 3, PAPER 3, DECEMBER 2009

Estimation of Distribution Algorithms: Applications to the Design of Process Sensor Networks

Mercedes Carnero
Universidad Nacional de Río Cuarto, Dpto. de Ciencias Básicas,
Río Cuarto, Argentina,
mcarnero@ing.unrc.edu.ar

José Luis Hernández
Universidad Nacional de Río Cuarto, Dpto. de Ciencias Básicas,
Río Cuarto, Argentina,
jlh@ing.unrc.edu.ar

and

Mabel Sánchez
Planta Piloto de Ingeniería Química (UNS- CONICET)
Camino La Carrindanga Km 7, Bahía Blanca 8000, Argentina
msanchez@plapiqui.edu.ar

Abstract

The optimal location of sensors involves the selection of type, number and location of sensors from a set of available instruments with certain values of cost, precision and reliability. The optimal design not only satisfies economic criteria, but also some requirements on the quality of key variable estimates. In this work, a methodology for solving the sensor network design problem based on Estimation of Distribution Algorithms is presented. These algorithms are included in the Evolutionary Computation paradigm and substitute the probability distribution estimation of a population composed by potential solutions and its subsequent sampling for the use of the classic crossover and mutation operators. The performance of the new strategy is evaluated and compared with that provided by other evolutionary techniques for the case of a steam metering network of a methanol synthesis plant.

Keywords: Sensor Network, Evolutionary Algorithms, Optimization.

1. INTRODUCTION

The commercial challenges that confront process industries today bring about the frequent application of strategies for on-line optimization, multivariate statistical process control, fault diagnosis, predictive and reliability centered maintenance, etc. The availability of process information is essential for the execution of all these activities.

The high development achieved by on-line monitoring and data storage systems allows for a huge volume of information of chemical plants. On the other hand, the quality of process knowledge, regarding its accuracy and precision, has been significantly enhanced since the application of Data Reconciliation procedures.

Given the important benefits attained using more accurate and precise process information when economic, safety and environmental aspects are evaluated, the attention is now directed towards the origin of process information, that is, the set of instruments installed on the plant.

Measurement planning is a complex multilevel task that involves the definition of the global objectives, the selection of measured variables and the specification of details, such as the sampling interval, the sampling technique, the type of operator interface, etc. The information that will be available from the process depends essentially on the selection performed in the second level, since the Degree of Estimability of a variable is a function of the plant topology and the set of measurements.

The selection of the subset of measured variables, i.e. the design of the sensor network, is performed during the formulation of the Process and Instrumentation Diagram. A common practice is to decide sensor locations based on previous experience with similar plants and using empirical rules. There are no comprehensive software packages that help the designer in that task. The development of strategies devoted to the optimal selection of sensors is of great interest, because it allows an optimal allocation of economic resources that assure the availability of the required information and acceptable safety levels.

The operation of a chemical plant can be represented by a mathematical model, i.e. a set of equations that relate the variables involved in the process. They can be divided into two sets: a) required variables or variables that should be estimated; b) non-required variables.

The design of a sensor network for monitoring purposes consists of determining if a process variable will be measured or not. If the first alternative is chosen, the number of installed sensors (hardware redundancy) and their features (cost, precision, failure rate) are set.

To optimally locate sensors that satisfy specific criteria, the Optimum Design of Sensor Networks is defined as an optimization problem formulated as follows

$$\begin{aligned} &Min/Max \quad f(\mathbf{q}) \\ &st. \\ &\qquad \mathbf{g}(\mathbf{q}) \leq \mathbf{g}^* \end{aligned} \tag{1}$$

where $\mathbf{q}$ is a vector of binary variables so that $q_i = 1$ if variable i is measured and $q_i = 0$ otherwise.

A wide variety of objective functions, $f(\mathbf{q})$, have been used: instrumentation cost, global error of variable estimates, system reliability, etc. As regards the set of process constraints, $\mathbf{g(q)}$, the design should guarantee the estimability of required variables, and also the conditions imposed on their precision, reliability and availability. Whichever the performance criteria and restrictions of the problem are, a combinatorial optimization problem arises that involves a huge number of binary variables.

Different deterministic and stochastic strategies have been presented to solve this combinatorial problem. The design of a minimum acquisition-cost network of flowmeters, which assures the estimability of all mass process flowrates given the knowledge of mass balances, is solved using a deterministic greedy algorithm. It finds the optimal solution in polynomial time [4] [13]. Up to the present time there are no deterministic algorithms with the same feature for solving more complex design formulations. The existing deterministic algorithms are efficient to cope with small and medium size design problems [1] [9]. Consequently optimization methods based on metaheuristics arise as an alternative to tackle the design of large-scale plant sectors subject to complex objective functions and constraints [6] [11].

In this work, a metaheuristic technique based on Estimation of Distribution Algorithms (EDAs) is proposed to solve the optimal design of a minimum cost sensor network subject to estimability and precision constraints imposed on a set of key variable estimates. The EDAs comprise a set of techniques included in the Evolutionary Computation paradigm. They substitute the probability distribution estimation of a population composed of potential solutions, and its subsequent sampling for the use of classic crossover and mutation operators [12]. Given the stochastic nature of EDAs, they do not guarantee the convergence to the global optimum. Although their computational cost may be high, this problem can be overcome by incorporating specific problem knowledge by means of local search techniques [7].

2. OBJECTIVES

The design of a minimum acquisition-cost (*CT*) sensor network that satisfies estimability and precision constraints on the estimates of a set of key variables is formulated as follows

$$\begin{aligned} &Min \qquad CT = \sum_{i=1}^{n} c_i q_i \\ &s.t. \\ &\qquad \hat{\sigma}_j(\mathbf{q}) \leq \sigma_j^*(\mathbf{q}) \qquad \forall\ j \in S_J \\ &\qquad E_k(\mathbf{q}) \geq 1 \qquad\quad\ \ \forall\ k \in S_K \end{aligned} \tag{2}$$

where c_i is the cost of the available flowmeter for measuring the flow of stream i, E_k represents the Degree of Estimability of variable k [2], $\hat{\sigma}_j$ is the standard deviation of the j-th variable estimate after a data reconciliation procedure is applied, and n is the total number of measurable variables. Furthermore S_J and S_K are the set of key process variables with requirements in precision and estimability respectively, and S_J is a subset of S_K .

This work is devoted to analyze the application of EDAs for solving Problem (2) and to compare its performance with respect to the ad-hoc stochastic method based on Genetic Algorithms (GA) developed by Carnero et al. [5].

3. METHODOLOGY

3.1 Estimation of Distribution Algorithms

An EA is a probabilistic procedure that maintains a population of individuals $P(t)=\{\mathbf{q_1}(t), \mathbf{q_2}(t),...\}$ for iteration t. Each individual represents a potential solution to the problem. Each solution $\mathbf{q}_\mathbf{i}^t$ is evaluated to give some measure of its fitness. Then, a new population (iteration t+1) is formed by selecting the individuals that fit better. Some members of the new population undergo unary transformations $\mathbf{m_i}$ (mutation), which produce new individuals by a small change on a single individual ($\mathbf{m_i} : \mathbf{q} \rightarrow \mathbf{q}$), and higher order transformations $\mathbf{c_j}$ (crossover), which form new individuals by combining parts from several individuals ($\mathbf{c_j} : \mathbf{q} \times \mathbf{q} ...... \times \mathbf{q} \rightarrow \mathbf{q}$). The program is run a given number of generations or until a criterion is satisfied. The best individual is considered a near optimal solution.

The EDAs are heuristics that share some features of the EA but the potential solutions included in the population are assumed as realizations of multivariate random variables, whose joint probability distribution can be estimated and updated using different mechanisms. In this sense, a solution vector $\mathbf{q} =\{ q_1, q_2, ..., q_n\}$ can be considered as a sample of an n-dimensional vector $\mathbf{Q} = \{ Q_1, Q_2, ..., Q_n\}$ where Q_i is a binary variable.

Thus the joint probability distribution $f(Q_1,......,Q_n)$ is associated with $\mathbf{Q}$, and the marginal probability distribution $P(Q_i=q_i)= p_i$ is related with each unidimensional random variable Q_i.

Unlike EAs, whose specific operators use the information given by the population members to guide the search, the EDAs conduct the optimization procedure by means of the building and evolution of the solution-space probabilistic model. That is, the potential solutions are evaluated using the objective function, and the information obtained through a selection step of the best solutions is used to update the vector of probabilities, from which the next population is sampled.

In this work, the methodology originally proposed by Baluja [3], who introduced the concept of competitive learning (typical in artificial neural networks), was used to guide the search. It is assumed that random variables are independent, thus the product of their marginal distributions constitutes the joint distribution of all variables. This is updated to take into account the structure of best current solutions. Although a simplified model of variable relationships is considered, the approach has shown good results for solving complex combinatorial problems such as channel assignments for mobile systems and task planning problems [8][15].

The pseudocode of a basic EDA is as follows

```
Initiate the probability vectors p
while (stopping criteria = .FALSE.)
    Generate N individuals by simulation according to p
    Evaluate the fitness function F for each member of the population
    Select the best solution
    Upgrade p using the best solution and the learning rate LR
    Mutate p using a probability of mutation PMUTA and a quantity of MS mutation
endwhile
```

3.2 Sensor Network Design Methodology Based on EDAs

To compare the results of the proposed methodology with those provided by the evolutionary strategy developed by Carnero et al. [5], some features of this procedure are maintained in the new technique. It has the following distinctive characteristics:

1. The initial population satisfies the estimability condition of key variables
2. The marginal probability of the variables for the first iteration is estimated using the initial population. Each unidimensional random variable Q_i (i =1:n) follows a Bernoulli Distribution. The maximum likelihood estimate of the expected value of Q_i is its sample mean. Therefore the vector of sample means for each instance is evaluated as follows

$$p_i = \frac{1}{N}\sum_{1}^{N} q_i \quad (i = 1 : n) \tag{3}$$

3. Fitness function F, proposed by Deb [10] is applied. It takes into account constraint violations as follows

$$F = \begin{cases} \sum_{i=1}^{n} c_i q_i & \text{if } \mathbf{q} \text{ is feasible} \\ CT_{\max} + Q(\mathbf{q}) & \text{if } \mathbf{q} \text{ is unfeasible} \end{cases} \tag{4}$$

where

$$Q(\mathbf{q}) = \begin{cases} (CT_{\max} - \sum_{i=1}^{n} c_i q_i)\left(\dfrac{ncu}{nr}\right) & \text{if } \mathbf{q} \text{ does not satisfy } S_K \text{ restrictions} \\ \sum_{i=1}^{n} c_i q_i \left(\dfrac{1}{R}\sum_{r=1}^{R}\dfrac{\sigma_r - \sigma_r^*}{\sigma_r}\right) & \text{if } \mathbf{q} \text{ satisfies restrictions on } S_K \text{ but not on } S_J \end{cases} \quad (5)$$

$CT_{\max}$ is the cost of measuring all variables, R and *ncu* stand for the number of variables in S_J and S_K whose constraints are unsatisfied respectively and, *nr* is the number of variables in S_K.

4. A local search procedure is run after population update. A portion of the population that satisfies the estimability of required variables and has the best values of the fitness function is selected. The neighborhood of each solution is inspected to find a new individual that has an F value lower than the current solution, by elimination of one measurement or by interchange of one measurement by an unmeasured variable. The estimability of required variables is maintained. If a better solution is found, the current one is replaced by the new one. The local search is accomplished using the formulation derived for the classification of process variables of linear systems [16].

5. The probability vector is updated, position by position, using the learning rate *LR* as follows

$$p_i^u = p_i^c(1-LR) + s_i\, LR \quad (i = 1:n) \quad (6)$$

where p_i^u is the updated probability, p_i^c stands for the current probability and s_i represents the *i*-th element of the best solution. In this way, the algorithm incorporates the knowledge of the best current solutions, which are determined through the selection procedure. The set of selected solutions may contain only one element, i.e. the best current solution, as an extreme case.

The value of *LR* is essential to the convergence of the algorithm. High values of *LR* introduce a bias towards specific solution structures avoiding the exploration of different regions of the search space; consequently they originate problems of premature convergence.

6. To introduce diversity in the search, each element of the probability vector is involved in a mutation procedure with *PMUTA* probability as follows

$$p_i^u = p_i^c\ (1-MS) + rand(0,1)\,MS \quad (i = 1:n) \quad (7)$$

where *MS* is the mutation amount.

4. RESULTS

The procedure previously described was applied to solve the instrumentation design problem of the steam metering network (SMN) for a methanol production plant. The process consists of 11 units interconnected by 28 streams and is represented in Fig. 1. Also Table 1 shows the true value of mass flowrates, the standard deviation of measurement errors, and the cost of available flowmeters. They were obtained from Sen et al. [17].

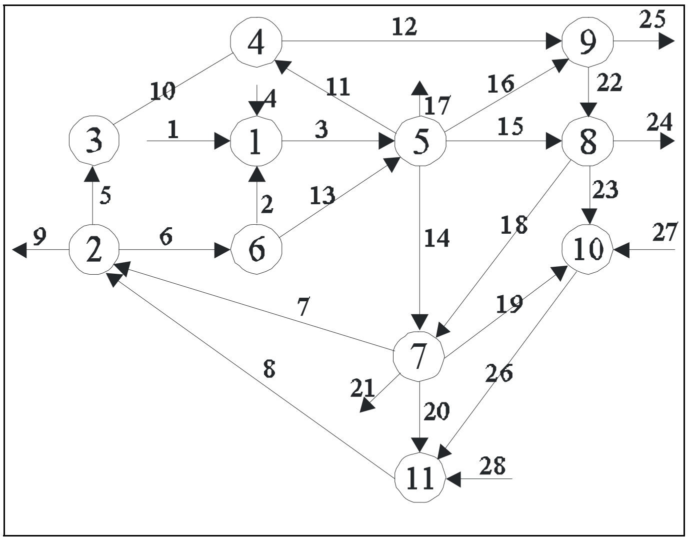

Figure 1: Steam Metering Network Flowsheet

Table 1: Steam Metering Network Data

Stream	$C^m{}_i$	σ_i	c_i
1	0.86	0.0215	3.7
2	1.	0.025	4.5
3	111.82	2.8	132.2
4	109.95	2.749	129.2
5	53.27	1.332	65.3
6	112.27	2.807	132.4
7	2.32	0.058	5.0
8	164.05	4.101	193.9
9	0.86	0.0215	2.06
10	52.41	1.31	62.8
11	14.86	0.3715	20.2
12	67.27	1.682	80.0
13	111.27	2.782	130.4
14	91.86	2.296	109.8
15	60.	1.5	71.6
16	23.64	0.591	29.7
17	32.73	0.8182	39.5
18	16.23	0.4057	20.4
19	7.95	0.1987	11.1
20	10.5	0.2625	13.6
21	87.27	2.182	102.9
22	5.45	0.1362	8.1
23	2.59	0.0648	6.3
24	46.64	1.166	55.5
25	85.45	2.136	101.0
26	81.32	2.033	93.7
27	70.77	1.769	84.7
28	72.23	1.806	85.4

Three design cases were analyzed that correspond to different sets of key variables and precision constraints. They are shown in Table 2. It was assumed that there was no restriction for the location of sensors on any stream, and therefore the search space was made up of 2^{28} solutions.

Table 3 contains the optimization results which are the same provided by the previously developed EA.

Different *LR* values were tested and the best performance of the algorithm for all case studies was obtained by using *LS*=0.1. For the probability vector upgrade, the best solution was considered as the only member of the selection set. Furthermore the parameters associated with the probability vector mutation were fixed at *PMUTA* = 0.02 and *MS*=0.05 values, taking into account the recommendations found in the literature.

Table 2: Constraint Bounds for each Case Study

Case	Constraint Bounds
Case 1	S_K : streams 1 2 6
	$\sigma_2^* = 0.025 \quad \sigma_6^* = 1.7851$
Case 2	S_K : streams 4 8 17 21 23 25
	$\sigma_4^* = 2.199 \quad \sigma_8^* = 3.281$
	$\sigma_{21}^* = 1.754 \quad \sigma_{25}^* = 1.709$
Case 3	S_K : streams 4 5 7 8 12 16 18 20 27 28
	$\sigma_4^* = 2.199 \quad \sigma_5^* = 1.0654$
	$\sigma_8^* = 3.281 \quad \sigma_{12}^* = 1.3454$
	$\sigma_{27}^* = 1.4154 \quad \sigma_{28}^* = 1.4446$

Table 3: Optimal Solutions

Case	Measured Variables	Cost
1	1 2 6 7 9 10 13 20 26 28	533.56
2	1 4 6 7 10 11 14-24	752.26
3	1 2 4 5-7 9-11 13 15-24 26-28	1178.06

5. CONCLUSIONS

In this work, a strategy based on EDAs is presented for the design of process sensor networks. Its performance is compared with that provided by an existing heuristic inspired in classic EA. The same results are obtained for three case studies using both techniques.

The diversity of an EA is provided by the crossover operator, which allows the exploration of different regions of the search space, maintaining the constructive blocks of the most promising solutions. The difficulty associated with the design of this operator increases with the complexity of the combinatorial problem. In contrast, EDAs use global information about the population to estimate its probability density function and employ it to generate new solutions. This has the advantage of avoiding the rupture of building blocks.

Future works involve the implementation of parallel EDAs and the development of sophisticated models for the joint probability distribution that can manage more complex designs.

References

[1] Bagajewicz, M. and E. Cabrera (2002) New MILP Formulation for Instrumentation Network Design and Upgrade, AIChE Journal 48, 2271

[2] Bagajewicz, M.; Sánchez, M. (1999) Cost Optimal Design and Upgrade of Non-Redundant and Redundant Linear Sensor Networks. AIChE J., 45, 1927-1938.

[3] Baluja, S. Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning (Technical Report CMU-CS-94-163); Carnegie Mellon University; Pittsburgh, PA, USA, 1994

[4] Carnero, M.; Hernández, J.; Sánchez, M.; Bandoni, A. (2001) An Evolutionary Approach for the Design of Non-Redundant Sensor Networks. Industrial and Engineering Chemistry Research, 40, 5578-5584.

[5] Carnero, M; Hernández, J; Sanchez, M; Bandoni, A. (2005). On the Solution of the Instrumentation Selection problem, Industrial & Engineering Chemistry Research. ISSN:0888-5885 Vol 44, N° 2. pp 358-367.

[6] Chao-An, L., C. Chuei-Tin, K. Chin-Leng and C. Chen-Liang, (2003). Optimal Sensor Placement and Maintenance for Mass-Flow Networks, Industrial and Engineering Chemistry Research, 42, 4366.

[7] Chaves, J. Domínguez, D, Vega, M. Gomez, J., Sánchez, J. (2008). Parallelizing PBIL for Solving a Real-World Frequency Assignment Problem in GSM networks. 16 Euromicro Conference on Parallel, Distributed an Network-Based Processing. IEEE Computer Society. Pp 391-398.

[8] Chaves, J. Domínguez, D, Vega, M. Gomez, J., Sánchez, J. (2008). SS vs PBIL to Solve a Real-World Frequency Assignment Problem in GSM networks. EvoWorkshops 2008. LNCS 4974. Springer-Verlag Berlin Heidelberg. Pp. 21-30.

[9] Chmielewski, D., Palmer, T.; Manousiouthakis, V. (2002) On the Theory of Optimal Sensor Placement. AIChE J., 48, 1001–1012.

[10] Deb, K. (2000) An Efficient Constraint Handling Method for Genetic Algorithms. Comp. Meth. Appl. Mech. Eng., 186, 311-338.

[11] Gerkens, C. and G. Heyen, (2004). Use of Parallel Computers in Rational Design of Redundant Sensor Networks. Escape 14 Congress, Lisboa, Portugal.

[12] Larrañaga, P, Lozano, J., Mühlenbein, H. (2003). Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. N° 19. pp 149-168.

[13] Madron, F. (1992) Process Plant Performance. Measurement and Data Processing for Optimisation and Retrofits. Ellis Horwood Ltd., Chichester.

[14] Narasimhan, S.; Jordache, C. (2000) Data Reconciliation and Gross Error Detection. Gulf Publishing Company, Houston.

[15] Pang. H, Hu, K. Hing, Z. (2006). Adaptive PBIL Algorithm and its Application to SolveAcheduling Problems. Proceedings of then 2006 Conference on Computer Aided Control System Design. Munich. Pp 784-789.

[16] Romagnoli, J.; Sánchez, M. (1999) Data Processing and Reconciliation for Chemical Process Operations. Academic Press, San Diego.

[17] Sen, S.; Narasimhan, S.; Deb, K. Sensor network design of linear processes using genetics algorithms. Comput. Chem. Eng. 1998, 22, 385-390.