Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-8 (2015) 900–905

Development of Ant Colony Optimization (ACO) Algorithms Based on Statistical Analysis and Hypothesis Testing for Variable Selection

C. M. Pessoa, C. Ranzan, L. F. Trierweiler, and J. O. Trierweiler

Group of Intensification, Modeling, Simulation, Control, and Optimization of Processes (GIMSCOP)
Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre, RS, Brazil (Tel: +55(51) 33084167,
e-mail: {carolpes, cassiano, luciane, jorge}@ enq.ufrgs.br)

Abstract: Obtaining reliable models from experimental data is a point of deep interest in all areas of research. Since the quality of the model depends on the number of selected variables, it is important to develop methods that identify the best ones. This work proposes a method of variable selection based on the Ant Colony Optimization (ACO) algorithm. Using data from a *Saccharomyces cerevisiae* fermentation, several criteria for trail update and model comparison were implemented and the obtained models were compared. The use of the length of the confidence interval produced the best results, finding the optimal model more frequently.

Keywords: Modelling, Variable Selection, Spectral Data, Ant Colony Optimization, Hypothesis Testing, Confidence Interval

1. INTRODUCTION

In order to achieve production increment, one of the alternatives is to invest in optimization techniques. This often means investment in new or better control methods, which is linked to two important tasks: obtaining a good model for the variable of interest and monitoring the process (Yamuna & Ramachandra, 1999).

Modelling consists on finding a causal relationship between variables. Regression analysis, combined with statistical techniques to quantify the confidence of the model, appears as the main tool used for this purpose, (Sykes, 1993). Among the different forms of regression, linear is certainly the most widely used. In this context, techniques like Partial Least Squares (PLS), Principal Component Analysis (PCA) and Principal Component Regression (PCR) appear as the primary regression methods, useful in the quantitative analysis of data (Geladi *et al.*, 2004).

The quality of the model depends not only on which variables are used in the regression, but on how many. A small subset of predictor variables is often preferable against using all available data, because it reduces costs and time spent in the measurements, tends to present a more simple physical interpretation, and, in the case of multiple linear regression (MLR), reduces the uncertainty of prediction, since this uncertainty increases with the ratio between the number of explanatory variables and the number of samples used in the calibration (Brown *et al.*, 2009). It is important to notice that, in general, each variable has different difficulty levels and cost involved in measuring them, and this aspect must be taken into account during variable selection.

Process monitoring, however, is highly dependent on the type and quality of sensors used. In recent years, optical sensors have become increasingly important in biotechnological applications. Optical sensors can be interfaced through glass window in reactors. Therefore, it is an in-situ, non-invasive method that gives real-time measurements (Hantelmann *et al.*, 2006, Scheper *et al.*, 1999). Several types of spectroscopy are possible through this technique, fact that makes models capable of dealing with spectral data so attractive. In this context, fluorescence sensors have being investigated for the determination of biomass and viable cells, bioreactor characterization, metabolic studies (transition aerobic/ anaerobic) and especially the monitoring of bioprocesses (Solle *et al.*, 2003, Hitzmann *et al.*, 1998). The development of a method capable of working with spectral data in order to identify spectral regions related to response variable can enable the development of optical sensors tailored to specific process, which would improve its control, making it more efficient and economic.

In the case of variable selection, a fairly common approach is the combination of suitable criteria that evaluate the quality of a subset of predictors combined with an algorithm that optimizes these criteria (Brown *et al.*, 2009). This approach is used in this work, applying the Ant Colony Optimization (ACO) algorithm as optimization method. Due to the advantages of spectral data, this work addresses the use of ACO for selecting components from spectroscopic analyses. When applied for different kind of data, the algorithm must take into account the difficulty level and cost involved in each variable measurement when selecting them.

Ant Colony Optimization algorithm is based on the hypotetical collective behavior of ants when searching for food sources. During this search, the ants secrete pheromones to mark their path, but they evaporate over time. In nature, ants that travel the shortest path return to the nest more quickly, so that the path traveled by these individuals has a

10.1016/j.ifacol.2015.09.084

higher concentration of pheromone. This trail acts as a decoy for other ants and, in time, all individuals of the colony tend to go trough this optimal (shortest) way (Allegrini & Olivieri 2011).

Dorigo and Gambardela (1997) developed the first version of ACO seeking solution for the Traveling Salesman problem, a problem of combinatorial optimization search in the space of permutations (Ranzan, 2014). Currently, several studies have been published regarding the application of the ACO method for screening variables, among which can be mentioned the work of Ranzan (2014), Allegrini and Olivieri (2011), Hemmateenejad *et al.* (2011), Mullen *et al.* (2009) and Socha *et al.* (2008).

Ranzan *et al.* (2014) applied the Sum of Squares Errors (SSE) as a criterion for updating the pheromone trail and to compare models. The goal was to predict the content of protein in different brands of flour based on NIR spectral data. The results showed the use of ACO as a filtering tool made possible the selection of important spectral regions, increasing the coefficient of determination of generated models by 60% compared to other methods which used the full spectrum, such as PCA and PCR.

Other optimization algorithms have also been used in variable selection, and the two stochastic optimization algorithms most known and applied in the field of chemometrics are simulated annealing and genetic algorithm (Cerny, 1985, Kirkpatrick *et al*., 1983). Moreover, other methods, such as tabu search, artificial colonies of bees, particles swarm and harmonic search can also be used for this application (Ghasemi *et al*., 2012, Mello & Pinto, 2008).

2. STATISTIC METRICS

Establishing appropriate criteria to evaluate the generated models is also crucial in order to obtain the optimal result. Among the parameters useful in this evaluation, the most used are the root mean square error of calibration (RMSEC) and prediction (RMSEP), which examines the fit of the model to the set of calibration and testing data evaluating the reproducibility of the data, and the coefficient of determination R^2, which is a measure of the proportion of variability explained by the fitted model.

This coefficient is used quite frequently due to its simplicity, but there are some disadvantages in its interpretation, such as the increase of its value by the addition of terms in the model. The adjusted coefficient of determination (R_a^2) is a variation that can be used to solve these problems, since it takes into account the number of degrees of freedom associated with the sum of squared error (SSE) and the sum of total squares (SST) (Walpole *et al.*, 2012).

Also, the use of hypothesis tests is very useful when analyzing models. The t-Student test (or t-test), for example, allows to test hypotheses about the coefficients and build their confidence intervals (Wilcox, 2012). Basically, the hypotheses being tested are:

$$H_0: \beta_j = 0 \qquad H_1: \beta_j \neq 0 \qquad (1)$$

where β_j is a given model parameter j and j=0,1,…,k.

The rejection or not of the hypothesis H_0, called null hypothesis, depends on the level of significance chosen, on the parameter estimator and its variance and on the standard deviation of errors. If the null hypothesis is not rejected, the variable associated with β_j explains an insignificant amount of change in y in the presence of the others regressors, and therefore can be removed from the model.

Even considering the estimators as unbiased, they are unlikely to estimate the parameters β_j accurately. Thus, it is preferable to determine an interval where it is possible to assume, with a given confidence, that it contains the true value of parameter β_j, called the confidence interval. The smaller the confidence interval, there is less uncertainty in the model parameter (Walpole *et al.*, 2012).

The use of such tests has the advantage of evaluate the contribution of each parameter separately, rather the adequacy of the model as a whole. Optimization methods can, therefore, use this information in combination with statistical models to identify and select the most relevant variables.

Another way to assess the contribution of each predictor is making use of F-tests to compare subsets of variables against the full model. The higher the F value, the worse the submodel is when compared to the full model. This aspect will be better discussed in the next section.

3. ACO MODIFICATIONS

The version of ACO implemented in this study is a modification of the one used by Ranzan *et al.*, (2014), which is based on pheromone trail evolution during spectral group scanning. Initially, all spectral components are marked with the same pheromone concentration. The ACO routine selects random spectral components to compose a group that is evaluated using the objective function for process variable prediction. Based on objective function error, the pheromone concentration, associated with each spectral component at the evaluated spectral group, is updated. For the subsequent spectral group selection, the random selection chooses spectral components associating the same random trigger and a cumulative density of pheromone for the full range of spectral elements. This association brings into evidence significant elements inside the spectral range, and, after few iterative runs, a pheromone profile is established, and the regions with high pheromone density highlight the significant excitation/emission pairs for process variable prediction.

A schematic summary of steps within ACO implementation used in this work can be seen in Fig. **1**. See Ranzan *et al.* (2014) for more detail about this algorithm