Computers and Chemical Engineering 117 (2018) 32–41

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Harvest time prediction for batch processes

Max Spooner[a,*], David Kold[b], Murat Kulahci[a,c]

[a] DTU Compute, Technical University of Denmark, Asmussens Alle 322, Kgs. Lyngby 2800, Denmark
[b] Chr. Hansen A/S, Hvidovre, Denmark
[c] Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

ARTICLE INFO

Article history:
Received 14 November 2017
Revised 1 May 2018
Accepted 20 May 2018
Available online 24 May 2018

Keywords:
Batch process
Prediction
Dynamic time warping
Partial least squares
Lasso regression

ABSTRACT

Batch processes usually exhibit variation in the time at which individual batches are stopped (referred to as the harvest time). Harvest time is based on the occurrence of some criterion and there may be great uncertainty as to when this criterion will be satisfied. This uncertainty increases the difficulty of scheduling downstream operations and results in fewer completed batches per day. A real case study is presented of a bacteria fermentation process. We consider the problem of predicting the harvest time of a batch in advance to reduce variation and improving batch quality. Lasso regression is used to obtain an interpretable model for predicting the harvest time at an early stage in the batch. A novel method for updating the harvest time predictions as a batch progresses is presented, based on information obtained from online alignment using dynamic time warping.

1. Introduction

Batch processes are characterised by a beginning, when the raw materials are loaded into a reactor vessel, a finite period of transformation or growth, and an end when the finished product is harvested from the reactor. The time at which to harvest the batch is often defined based on some features in the process which from experience ensure the desired product specifications. There is often batch to batch variation in the time at which the harvest criterion occurs, and so batches have different durations. This is especially the case in bio-based industrial processes where the harvest time is dependent on the activity of living organisms. In order to ensure the batch is harvested at the optimum point in time, it must be monitored closely by a technician who must react quickly when the harvest criterion is reached. In this work we consider the problem of predicting the harvest time at an early stage in the process. Obtaining good harvest time predictions is of value for two reasons. Firstly, such predictions provide a guide for the technicians on when to focus on the process and when it is safe to work on other tasks. Secondly, the predictions facilitate scheduling of downstream processes such as packaging.

The general problem of predicting some response variable based on data measured during a batch process has been investigated extensively. The predominant approach, Multi-way Partial Least Squares (MPLS), was pioneered by Nomikos and MacGregor (1995) where the method was used to make predictions of 5 quality variables using batch process data. This approach has been applied and adapted by several authors including Ündey et al. (2003), Wang (2011) and Gunther et al. (2009). Besides MPLS, a wide range of machine learning methods have been applied to the problem of predicting end-of-batch quality using online process data including neural networks (Lennox et al., 2001), support vector regression (Desai et al., 2006) and lasso regession (Yan et al., 2014).

There is limited research on batch process prediction where the response variable is harvest time of the batch rather than end-of-batch quality. In Marjanovic et al. (2006) MPLS is used to predict the end time of a batch based on only the first two hours of process data. Others apply more ad-hoc methods to detect the optimal fermentation time of a batch process (Latrille et al., 1993).

In this paper we present a real case study of predicting the harvest time of a batch bacteria fermentation process for the bioscience company Chr. Hansen A/S using a novel statistical approach, and apply the methods to datasets from two different bacteria fermentations. A need for more accurate harvest time prediction was identified by the company. The process is characterised by two phases which influenced our approach for harvest time prediction. In brief, the proposed method consists of predicting the harvest time at the end of the first phase using a lasso regression model, and then updating the predictions from this model online during the second process phase using the time-information provided by online dynamic time warping (DTW). An overview of the steps involved in the method is shown in Fig. 1. Each step is explained in detail in Section 2

* Corresponding author.
E-mail address: mpsp@dtu.dk (M. Spooner).

https://doi.org/10.1016/j.compchemeng.2018.05.019
0098-1354/

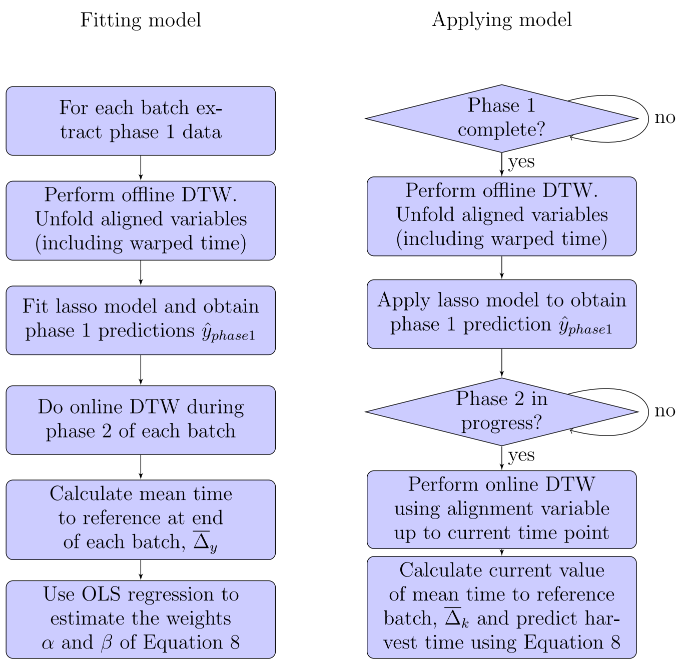

Fig. 1. Summary of model fitting and application.

The proposed method provides a valuable early warning of the expected harvest time to enable scheduling of downstream tasks. As a benchmark, we compare the phase 1 lasso model to a MPLS model, which is the more usual approach to dealing with the high dimensionality of the unfolded phase 1 data. We advocate the use of lasso regression for two main reasons. Firstly, lasso regression results in a far simpler model than MPLS, as it automatically selects only a subset of variables for which to give non-zero model coefficients. The lasso model is therefore easier to interpret than the PLS model, which is a big advantage for implementation in an industrial setting. Secondly, the test error, estimated using nested cross validation (Cawley and Talbot, 2010), is found to be smaller for the lasso model than for the MPLS model. Therefore, in this case study, the lasso model is expected to provide more accurate harvest time predictions than the MPLS model when used on new batches.

In summary, the novel contributions of this work consist firstly, in the presentation of real data from an existing batch process and a real problem to be solved: that of predicting the harvest time. Secondly, we demonstrate the advantages of lasso regression as a variable selection method, in contrast to the more predominant latent structure method MPLS that retains all variables in the model regardless of their relevance. Thirdly, we present a novel method that updates the harvest time predictions using dynamic time warping. This method exploits directly the ability of DTW to indicate time information regarding a batch, whereas previous applications for batch processes has mostly used dynamic time warping as a preprocessing step.

2. Methods

2.1. Process

The fermentation process used by Chr. Hansen to produce bacteria cultures consists of the following steps

1. A small volume of concentrated bacteria cells is added to a fermenter vessel which has been pre-filled with growth medium
2. The bacteria cells grow and multiply, thereby producing acid which lowers the pH inside the fermenter
3. When pH reaches a predefined set point, the pH level is automatically controlled at the set point level by adding a base to the fermenter
4. Based on expert judgement the batch is stopped and the contents of the fermenter are transferred to downstream processing

Data for two different bacteria fermentations was obtained which we refer to as product 1 and product 2. For both products, 6 variables are measured during the fermentation: pH, Base Flow (rate of base addition), Base Quantity (total amount of base added), Temperature, Level and Pressure. However, our interest was limited to the first three of these variables because they are most closely linked to the biological process. The temperature variable was not used because it was tightly controlled and maintained at a constant level. The product 1 and product 2 data consisted of 44 and 43 batches respectively and were all taken from normal operating conditions. The data is shown in Fig. 2 where the two phases of the process can be distinctly seen. In the first phase, pH is not controlled and is decreasing until it reaches the set point. In the second phase, base is added in order to maintain the pH at the set point. The observed changes in pH and Base Flow are closely linked to the state of the process and reflect rates of bacteria growth and metabolism. There is variation between batches in the time taken to reach different stages, as well as in the magnitudes of the variables, due to differences in raw materials and inoculation material.

The criterion to start the harvest is product specific and is based on a combination of process parameters. The harvest process should be initiated manually when the criterion is met, but may be delayed as it relies on human judgement and taking additional factors into account, such as whether equipment downstream is available.

Due to the above mentioned variations and equipment limitations the harvest is done inconsistently as reflected in Fig. 2. Therefore, the aim of this work was to predict in advance the time at which the harvest criterion will be met. This will reduce the need for human evaluation thereby enabling less process variation and better planning in regards to utilization of the downstream equipment.

Batches which do not attain the harvest criterion cannot be used for model building, as for such batches the response variable, correct harvest time, is missing. Therefore, due to the batch to batch variation, the harvest criterion for each product was redefined to features attained by all batches so that it would be possible to assess the performance. The harvest criterion for product 1 was defined as the moment when Base Flow falls to 95% of its maximum value. For product 2 the harvest criterion was defined to be when Base Quantity reaches 0.72 (in scaled units). These criteria are very similar to those used in practice by the company, but have the advantage of being attained by all the batches. Of course, an alternative approach would be to use the harvest criteria actually in use, and use some missing data imputation method to fill in the correct harvest time values for batches which were harvested too soon. However, then assessment of model performance would depend on how the missing data are imputed. Therefore, we use the former approach so that there is no doubt that the harvest time to be predicted corresponds to the same criterion in all batches.

2.2. Phase 1 models

The clear division of the process into two phases suggested a goal of predicting the harvest time at the end of the first phase. A prediction at this point is early enough to be useful for scheduling purposes, whilst enough data has been accumulated to make predictions realistic. In Ündey and Çinar (2002) approaches to statistical monitoring of multiphase and multistage batch pro-