European Journal of Operational Research 218 (2012) 97–105

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing and Logistics

Multicriteria variable selection for classification of production batches

Michel J. Anzanello [a,*], Susan L. Albin [b,1], Wanpracha A. Chaovalitwongse [b,2]

[a] Department of Industrial Engineering, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99, 5 andar, Porto Alegre, Brazil
[b] Department of Industrial and Systems Engineering, Rutgers University, 96 Frelinghuysen Road, CoRE Building, Room 201, Piscataway, NJ, USA

ARTICLE INFO

Article history:
Received 20 January 2011
Accepted 19 October 2011
Available online 26 October 2011

Keywords:
Multivariate statistics
Variable selection
Multiple criteria
Data mining
Batch manufacturing

ABSTRACT

In many industrial processes hundreds of noisy and correlated process variables are collected for monitoring and control purposes. The goal is often to correctly classify production batches into classes, such as good or failed, based on the process variables. We propose a method for selecting the best process variables for classification of process batches using multiple criteria including classification performance measures (i.e., sensitivity and specificity) and the measurement cost. The method applies Partial Least Squares (PLS) regression on the training set to derive an importance index for each variable. Then an iterative classification/elimination procedure using *k*-Nearest Neighbor is carried out. Finally, Pareto analysis is used to select the best set of variables and avoid excessive retention of variables. The method proposed here consistently selects process variables important for classification, regardless of the batches included in the training data. Further, we demonstrate the advantages of the proposed method using six industrial datasets.

1. Introduction

The enormous volume of data collected from industrial processes has challenged researchers to develop efficient methods to identify the most important process variables. In much of the existing work, the goal has been to find the most important variables for predicting outcomes, i.e., variables related to the product as in Gauchi and Chagnon (2001), Meiri and Zahavi (2006), Ozturk et al. (2006) and Olafsson et al. (2008). In contrast, the goal of this study is to select process variables that are most important for classification of production batches into two classes (e.g., good batches and failed batches).

The contribution here is a new method for identifying the most important process variables for classification. The proposed method classifies batches into two categories and selects the most important process variables based on multiple criteria which could include, for example, accuracy, specificity, sensitivity, and cost. In addition, the method is also applicable where there are many more variables than batches, a common situation in batch processing.

This work overcomes the limitations in Anzanello et al. (2009) where a method for identifying the most important process variables for classification is proposed based on one performance criterion only, namely, classification accuracy. Most important, simulation results presented here indicate that the previous method is inconsistent in that the number of process variables chosen to achieve maximum accuracy depends on the observations that are included in the training set. The inconsistency is an artifact of selecting the variables based on accuracy alone. For example, the previous method selects a set with 50 variables and accuracy 95% over a set with five variables and accuracy 94%. The method proposed here offers multicriteria variable selection, for example considering both maximizing accuracy and minimizing the number of variables, and consistently selects process variables important for classification. Further, we demonstrate the advantages of the method proposed here using six industrial datasets.

The most obvious single criterion for selecting variables in a classification task is classification accuracy – the fraction of batches that are correctly classified either as conforming or non-conforming. However there are situations where other criteria may be more critical. In pharmaceutical processing, for example, incorrectly classifying a non-conforming batch as a good batch may lead to serious consequences. In this case specificity, the fraction of non-conforming batches that are correctly classified may be one of the key criteria for selecting variables. In other words, a false negative is much more costly than a false positive. On the other hand, in situations where the major costs are associated with scrapping good batches sensitivity, the fraction of good batches that are correctly classified, may be one of the key criteria.

In many applications, one of the criteria for selecting process variables can be cost. The goal is to select a set of variables that minimizes the cost of measuring and collecting data.

* Corresponding author. Tel.: +55 51 3308 4423.
E-mail addresses: anzanello@producao.ufrgs.br (M.J. Anzanello), salbin@rci.rutgers.edu (S.L. Albin), wchaoval@rci.rutgers.edu (W.A. Chaovalitwongse).
[1] Tel.: +1 732 445 2238.
[2] Tel.: +1 732 445 5469.

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.10.015

Over the last decade, approaches for multicriteria variable selection using accuracy, sensitivity, and specificity have been proposed for several real life classification problems, especially in text recognition, business decisions and security systems; see Rose-Pehrsson et al. (2000), Doan and Horiguchi (2004), Piramuthu (2004), Pendaraki et al. (2005), Huang et al. (2006), Pasiouras et al. (2007) and Aragonés-Beltrán et al. (2008). For a comprehensive review on multicriteria classification and decision making, see Zopounidis and Doumpos (2002), Steuer and Na (2003) and Sueyoshi (2006). Recently, other data mining-based approaches for variable selection have been expanded to the biomedical area (Li and Li, 2008; Su and Yang, 2008), as well as in financial decisions (Gaganis et al., 2007; Kirkos et al., 2008).

The method proposed here works as follows. First, Partial Least Squares (PLS) regression is performed on the training set to compute a variable importance index for each of the real valued process variables. Then, using all process variables, the training set is classified using the *k*-Nearest Neighbor (KNN) based on the Euclidean distance of the real valued process variables; classification performance measures (accuracy, sensitivity, etc.) are then computed. Subsequently, the variable with the lowest importance index is eliminated. The training set is classified using the remaining process variables, and the classification performance measures are computed again. This variable elimination process is repeated until there is only one variable left, generating a number of candidate sets of selected process variables. To reduce the number of candidates we apply Pareto optimality analysis and identify the Pareto frontier sets. Among these, the variable set that is closest to the ideal solution (e.g., maximum sensitivity, maximum specificity, minimum cost) is then selected. The selected process variables are validated on the testing set.

The rest of this paper is organized as follows. Section 2 presents a review of PLS mathematical fundamentals, KNN classification technique and Pareto optimality analysis. Section 3 describes the method for selecting the best subset of variables for classification. Section 4 presents the design of the simulation experiment used to compare the performance of the method proposed to the method in Anzanello et al. (2009). Section 5 shows that the method presented here is more consistent and selects fewer variables while leading to comparable accuracy. Section 6 applies the proposed method to six actual manufacturing data sets. Final conclusions are presented in Section 7.

2. Background

2.1. PLS regression

In this study, we focus on batch processing, common such as in chemical engineering applications. Batches are often characterized by many process variables and just one product variable. PLS regression is employed in our method because it has been widely applied to select variables for prediction of product variables; see Lindgren et al. (1994), Forina et al. (1999) and Sarabia et al. (2001). More recently it has been used in classification of production batches in Anzanello et al. (2009). PLS is known for performing well with correlated variables and high dimensional data frequently found in industrial applications; see Wold et al. (2001a), Kettaneh et al. (2005), Nelson et al. (2006) and Hoskuldsson (2001).

PLS constructs a small number of independent, linear combinations of the process variables. These new variables, called PLS components, account for much of the variance present in the original process variables and in the product variables. Typically only three or four PLS components can be used to represent dozens or even hundreds of process variables.

The key parameters that result from PLS regression are weights and loadings. These parameters can be calculated by means of the NIPALS algorithm; see Goutis (1997), Abdi (2003) and Geladi and Kowalski (1986). Further mathematical details of PLS can be obtained in Westerhuis et al. (1998), Wold et al. (2001a,b). The PLS regression can be performed using the PLS toolbox in statistical packages as Matlab® and R®.

The PLS regression can be formally defined as follows. Consider a matrix $\mathbf{X}$ consisting of N observations for each of J process variables and a matrix $\mathbf{Y}$ consisting of N observations for each of M product variables. (Note that in this study $M = 1$.) The process observation i is represented by the vector $\mathbf{x}_i$ $(x_{i1}, x_{i2}, \ldots, x_{iJ})$, while the product observation i is denoted by $\mathbf{y}_i$ $(y_{i1}, y_{i2}, \ldots, y_{iM})$, for $i = 1, \ldots, N$.

PLS constructs A independent PLS components that are linear combinations $t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \cdots + w_{Ja}x_{iJ} = \mathbf{w}_a'\mathbf{x}_i$ of the process variables, with $A \leqslant J$. The number of process components, A, is typically small, and can be defined by the cross-validation method in Hoskuldsson (1988). Vector $\mathbf{w}_a = (w_{1a}, w_{2a}, \ldots, w_{Ja})'$ represents the weights, which provide information about the way the variables combined themselves to generate $\mathbf{X}$ and $\mathbf{Y}$ as in Wold et al. (2001a).

Similarly, components are constructed for the product variables in $\mathbf{Y}$; i.e., $u_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + \cdots + c_{Ma}y_{iM} = \mathbf{c}_a'\mathbf{y}_i$, where $\mathbf{c}_a = (c_{1a}, c_{2a}, \ldots, c_{Ma})'$ is the product variable weights. The weight vectors $\mathbf{w}_a$ and $\mathbf{c}_a$ aim at maximizing the covariance of the PLS components $\mathbf{t}_a$ and $\mathbf{u}_a$. The weights are selected to yield components independent of one another; i.e., $\mathbf{t}_a$'s are orthogonal as in Xu and Albin (2002). Further, the loading vector, $\mathbf{p}_a = (p_{1a}, p_{2a}, \ldots, p_{Ja})'$, is obtained by the regression of the columns of $\mathbf{X}$ on $\mathbf{t}_a$, and provide relevant information about the process variables when associated to the process component $\mathbf{t}_a$.

2.2. k-Nearest Neighbor (KNN)

The *k*-Nearest Neighbor (KNN) technique is our choice of classification method as it is widely used, conceptually simple, and readily available in software packages. Anzanello et al. (2009) systematically compares KNN to Probabilistic Neural Networks, Support Vector Machines, and modifications of the KNN in the context of selecting the best variables for classification and KNN was identified there as the best choice. Some applications of KNN include gene classification in Golub et al. (1999), text recognition patterns in Weiss et al. (1999), and detection of abnormal brain activity in Chaovalitwongse et al. (2007).

The KNN algorithm can be formally defined as follows. Consider N observations in a J-dimensional training dataset, where J corresponds to the process variables. The objective is to classify a new observation in 0 or 1, denoting non-conforming or conforming, based only on process variables. The KNN algorithm measures the Euclidean distances between the new observation and the k nearest neighbors, i.e., existing observations. The class of each of the k neighbors is known, 0 or 1. A new observation is labeled as 0 if the majority of its k nearest neighbors belongs to 0. The number of neighbors, k, is selected by maximizing a classification performance measure in the training set where the class of each observation is known. Further details about KNN classification technique can be found in Ridgeway (2003).

2.3. Pareto optimal analysis

Pareto optimal analysis has been widely used for diverse applications including process design in Azapgic (1999), scheduling of manufacturing operations in Taboada and Coit (2008), and reliability optimization in power transmission in Taboada and Coit (2007). Given a set of choices (for example, possible subsets of variables)