Expert Systems with Applications 39 (2012) 2590–2596

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes

Doh-Soon Kwak [a], Kwang-Jae Kim [b,*]

[a] Samsung Electronics, San 61, Banwol-Dong, Hwasung, Gyeonggi 445-701, Republic of Korea
[b] Division of Mechanical and Industrial Engineering, Pohang University of Science and Technology, San 31, Hyoja-Dong, Nam-Gu, Pohang, Kyungbuk 790-784, Republic of Korea

ARTICLE INFO

Keywords:
Data mining approach
Missing values
Patient Rule Induction Method
Process optimization

ABSTRACT

Due to the rapid development of information technologies, abundant data have become readily available. Data mining techniques have been used for process optimization in many manufacturing processes in automotive, LCD, semiconductor, and steel production, among others. However, a large amount of missing values occurs in the data set due to several causes (*e.g.*, data discarded by gross measurement errors, measurement machine breakdown, routine maintenance, sampling inspection, and sensor failure), which frequently complicate the application of data mining to the data set. This study proposes a new procedure for optimizing processes called missing values-Patient Rule Induction Method (*m*-PRIM), which handles the missing-values problem systematically and yields considerable process improvement, even if a significant portion of the data set has missing values. A case study in a semiconductor manufacturing process is conducted to illustrate the proposed procedure.

1. Introduction

The use of data mining techniques in manufacturing industries has begun in the 1990s, gradually receiving attention from many manufacturing processes in automotive, LCD, semiconductor, and steel manufacturing for predictive maintenance, fault detection, diagnosis, and scheduling (Harding, Shahbaz, Srinivas, & Kusiak, 2006). Data mining techniques have also been used for process optimization in order to find optimum conditions for input variables that maximize (or minimize) output variables (Braha & Shmilovici, 2002; Kim & Ding, 2005).

Among many data mining techniques, the Patient Rule Induction Method (PRIM), originally proposed by Friedman and Fisher (1999), has been successfully applied for process optimization despite its recent emergence (Chong, Albin, & Jun, 2007; Chong & Jun, 2008; Kwak, Kim, & Lee, 2010; Lee & Kim, 2008). This method directly seeks a set of sub-regions for input variables, in which higher quality values are observed from the historical data.

An embedded assumption in existing PRIM works meant for process optimization is that missing values do not exist in the data sets, or the amount of missing ones is negligible. Although abundant data are readily available due to the rapid development of information technologies, missing values are a common occurrence in various industrial process data sets due to several causes (*e.g.*, data discarded by gross measurement errors, measurement machine breakdown, routine maintenance, sampling inspection, and sensor failure) (Arteaga & Ferrer, 2002; Muteki, Macgregor, & Ueda, 2005; Nelson, Taylor, & Macgregor, 1996). A large amount of missing values frequently complicates the application of data mining algorithms (including PRIM) to the data set, because most data mining algorithms have not been designed for them. Moreover, if missing values are not handled in principled ways, these can produce biased, distorted, and unreliable conclusions (Dasu & Johnson, 2003; Feelders, 1999). Thus, for the successful application of the existing PRIM works in process optimization, it is necessary to enhance existing works by systematically treating the missing-values problems.

The purpose of this paper is to develop a new PRIM-based method for optimizing processes, where a significant portion of the data set has missing values. This method will be referred to as the missing values-PRIM (*m*-PRIM). The remainder of the paper is organized as follows: PRIM is briefly reviewed in the next section; the proposed method is introduced, and the results of a case study are presented; finally, the conclusion and discussion are given.

2. Patient Rule Induction Method (PRIM)

The goal of PRIM is to discover a small box-shaped region, called a box, with a higher proportion of good observations compared to the entire region from a large data set $\{(y(r), \boldsymbol{x}(r)), r = 1, 2, \ldots, N\}$. In this set, $y(r)$ and $\boldsymbol{x}(r) = x_1(r), x_2(r), \ldots, x_p(r)$ are the output and p

* Corresponding author. Tel.: +82 54 279 2208; fax: +82 54 279 2870.
E-mail addresses: dskwak@samsung.com (D.-S. Kwak), kjk@postech.ac.kr (K.-J. Kim).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.114

input variables for the rth observation, respectively; and N is the number of observations in the entire region. Below is a brief description of PRIM as it pertains to the current research. Further details about PRIM can be found in Friedman and Fisher (1999) and Hastie, Tibshirani, and Friedman (2001).

2.1. Box and related statistics

A p-dimensional box B is defined as the intersection of sub-ranges of input variables such that:

$$B = s_{x_1} \times s_{x_2} \times \cdots \times s_{x_p}, \tag{1}$$

where $s_{x_j} = [x_j^l, x_j^u]$ is a sub-range of the input variable x_j ($j = 1, 2, \ldots, p$), and x_j^l and x_j^u denote the lower and upper bound of x_j in the box B, respectively.

Given a box B and the data set $\{(y(r), \boldsymbol{x}(r)), r = 1, 2, \ldots, N\}$, there are two statistics indicating the properties of the box. The first one is the support (β_B), which denotes the proportion of the observations contained in B given by:

$$\beta_B = n_B/N, \tag{2}$$

where n_B is the number of observations inside B, calculated by:

$$n_B = \sum_{r=1}^{N} 1(\boldsymbol{x}(r) \in B). \tag{3}$$

Here, the function $1(\cdot)$ takes one when the argument is true, and zero otherwise. The support clearly ranges from zero to one. The second statistic is the box objective (Obj_B), which is the mean of the output variable in B given by:

$$Obj_B = (1/n_B) \sum_{\boldsymbol{x}(r) \in B} y(r). \tag{4}$$

If the output variable has the most desirable value (*i.e.*, target), the box objective in (4) is expressed as:

$$Obj_B = (-)(1/n_B) \sum_{\boldsymbol{x}(r) \in B} (y(r) - t\,\mathrm{arg}\,et)^2. \tag{5}$$

2.2. Algorithm

A prepared data set of interests is randomly split into a learning set and a test set. Then, PRIM starts with box B_0, which includes all observations in the learning set. From B_0, PRIM creates $2 \times p$ candidate boxes, $\{C_{1-}, C_{1+}, C_{2-}, C_{2+}, \ldots, C_{p-}, C_{p+}\}$. The C_{j-} and C_{j+} ($j = 1, 2, \ldots, p$) are obtained by peeling $100\alpha\%$ of the observations inside the box from the left and right side of the jth input variable x_j. Here, α is the peeling parameter which determines the number of observations peeled off at each iteration and is typically set to a small value (between 0.05 and 0.1). Then, PRIM chooses the one with the largest box objective among the candidate boxes and lets this box be B_1. Boxes $B_1, B_2, \ldots, B_k$ are iteratively generated until the support becomes less than the predetermined stopping parameter β (*e.g.*, 0.05). By peeling off a small number of observations in each iteration, a long sequence of boxes is thereby created.

To avoid over-fitting, the box objectives in the generated boxes $B_1, B_2, \ldots, B_k$ are recalculated using the test set. The one with the largest box objective from the test set is selected as the optimal box (*i.e.*, the optimum condition on the input variables) and is given by:

$$\mathrm{x}^* = \left(x_1^*, x_2^*, \ldots, x_p^*\right) = \left(\left[x_1^L, x_1^U\right], \left[x_2^L, x_2^U\right], \ldots, \left[x_p^L, x_p^U\right]\right), \tag{6}$$

where x_j^L and x_j^U denote the lower and upper bound of x_j ($j = 1, 2, \ldots, p$) in the optimal box, respectively. The advantage of the PRIM algorithm compared with other rule discovery algorithms such as CART (Breiman, Friedman, Olshen, & Stone, 1984) and C4.5 (Quinlan, 1994, 1995) is its patient strategy, which has a small value of the peeling parameter α, allowing for the possibility of creating many peeling steps. Thus, each peeling becomes less important in determining the final box, and unfortunate peelings that remove good observations can be mitigated in subsequent steps.

3. The proposed method: *m*-PRIM

In this section, m-PRIM, which considers the missing-values problem and its role in optimizing processes is presented. The left side of Fig. 1 shows a brief procedure to optimize processes based on PRIM, where the amount of missing values is negligible. The overall procedure of m-PRIM, which has three additional steps after "Step 0: Prepare the data set," is presented in the dotted box in Fig. 1.

The basic idea of m-PRIM is to convert an incomplete data set, which has missing values in the data set, into k-imputed complete data sets, generate k optimal boxes from each of k-imputed completed data sets, and aggregate the k optimal boxes.

In this study, an incomplete data set is converted into k-imputed complete data sets by multiple imputation (MI). Originally proposed by Rubin (1987), MI has been used as an alternative to traditional methods such as case deletion, mean imputation, hot-deck, regression approach and single imputation using Expectation Maximization (EM), among others, in a wide variety of missing-values problems (Schafer & Graham, 2002).

MI is a simulation-based approach, where each of the missing value is replaced with $k > 1$ plausible values from their predictive distribution. Further, MI is implemented based on the assumption that missing values would be missing at random (MAR). In this study, NORM is used for multiple imputations. As developed by Schafer, 1999a, 1999b, NORM performs multiple imputations under a multivariate normal model. This model makes no distinctions between input or output variables, although it treats all as a multivariate variable. In NORM, proper multiple imputations are created through data augmentation (Tanner & Wong, 1987), where EM-estimated is used as starting values for the parameters. The rule of thumb suggested by Schafer is being used to guarantee the convergence of the data augmentation. Additionally, in terms of number of imputations (k), the use of more than five to ten imputations (Schafer, 1999a, 1999b) tends to have little or no practical benefit. The details of MI can be found in Rubin (1996), Schafer (1997) and Schafer and Olsen (1998).

In the following study, D^{inc} and D^c denote the incomplete data set and the complete data set, respectively; $D^{(j)c}$ ($j = 1, 2, \ldots, k$) is the imputed complete data set, where k is the number of imputations. Additionally, $\boldsymbol{x}^{a*}$ is the representative optimal box aggregated from the k optimal boxes $\boldsymbol{x}^{(j)*}$ ($j = 1, 2, \ldots, k$) that have been generated from each of the imputed complete data sets $D^{(j)c}$ ($j = 1, 2, \ldots, k$). Each step of the proposed method is described below.

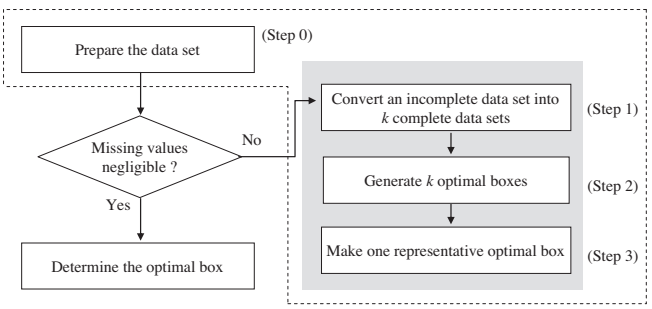

Fig. 1. Overall m-PRIM procedure.