Expert Systems With Applications 63 (2016) 35–48

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Data-based fault detection in chemical processes: Managing records with operator intervention and uncertain labels

Mahdieh Askarian[a,c], Raúl Benítez[b], Moisès Graells[c,*], Reza Zarghami[a]

[a] School of Chemical Engineering. College of Engineering, University of Tehran. PO Box 11155-4563, Tehran, Iran
[b] Automatic Control Department. Universitat Politècnica de Catalunya. EUETIB, Comte d'Urgell 187, 08028-Barcelona, Spain
[c] Chemical Engineering Department. Universitat Politècnica de Catalunya. EUETIB, Comte d'Urgell 187, 08028-Barcelona, Spain

ARTICLE INFO

Article history:
Received 27 January 2016
Revised 17 June 2016
Accepted 18 June 2016
Available online 23 June 2016

Keywords:
Fault detection
Mislabeling
Label noise
Underlying states
Operational intelligence
Interactive learning

ABSTRACT

Developing data-driven fault detection systems for chemical plants requires managing uncertain data labels and dynamic attributes due to operator-process interactions. Mislabeled data is a known problem in computer science that has received scarce attention from the process systems community. This work introduces and examines the effects of operator actions in records and labels, and the consequences in the development of detection models. Using a state space model, this work proposes an iterative relabeling scheme for retraining classifiers that continuously refines dynamic attributes and labels. Three case studies are presented: a reactor as a motivating example, flooding in a simulated de-Butanizer column, as a complex case, and foaming in an absorber as an industrial challenge. For the first case, detection accuracy is shown to increase by 14% while operating costs are reduced by 20%. Moreover, regarding the de-Butanizer column, the performance of the proposed strategy is shown to be 10% higher than the filtering strategy. Promising results are finally reported in regard of efficient strategies to deal with the presented problem.

1. Introduction

Monitoring of chemical processes, as many other activities, is required for determining the need for corrective actions and subsequent efficient operation. Indeed, abnormal situation management (ASM) is an essential task for loss prevention and safe operation of chemical plants. To achieve this aim, multiple protection layers (Fig. 1) are applied in industrial plants as per the international standard IEC61511 (2003), each one consisting of equipment and/or administrative controls coordinated with other protection layers (Isermann, 1994). Most automatic protection layers are triggered by actuators, while sensors' readings indicate the violation of limiting thresholds. Despite all the progress in automatic risk reduction systems (hardware and software), the operator supervision and the corrective action in ASM is still indispensable. Indeed, the operator and the automatic controls, together, constitute operational intelligence (Rajaram & Jaikumar, 2000).

The fault detection (FD) system is a core component of ASM that has attracted a lot of attention recently. Moreover, it is expected to be explicitly included in the standard IEC61511 in the near future. In fact, a fault consists of an unpermitted deviation of at least one property or parameter of a system from its acceptable, usual or standard condition (Isermann & Balle, 1997). FD methods are categorized in three main groups: quantitative model-based methods, qualitative model-based methods and data-driven methods (Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003). Qualitative model-based FD methods are not often deployed for complex chemical process, because the corresponding analytical description is rarely available. In addition, quantitative model-based FD methods, so-called inference methods, are developed based on explicit structural knowledge and causalities (Korbicz, Koscielny, Kowalczuk, & Cholewa, 2012). These methods, which rely on experts' knowledge in a specific domain, are often costly and time-intensive to obtain. Thus, FD is commonly addressed by process history based methods, since for operating plants a large amount of historical process data is available (MacGregor & Cinar, 2012; Qin, 2012).

Data-driven FD systems early developed based on multivariable statistical analysis e.g. principal component analysis (PCA) (Lu, Yao, Gao, & Wang, 2005), partial least squares (PLS) (Chiang, Braatz, & Russell, 2001). Recently, FD has been considered as a classification problem as well, and Machine Learning provides various tools for classification, which are categorized below (Isermann, 2006):

* Corresponding author.
E-mail addresses: mahdiehaskarian@ut.ac.ir (M. Askarian), raul.benitez@upc.edu (R. Benítez), moises.graells@upc.edu (M. Graells), rzarghami@ut.ac.ir (R. Zarghami).

http://dx.doi.org/10.1016/j.eswa.2016.06.040
0957-4174/

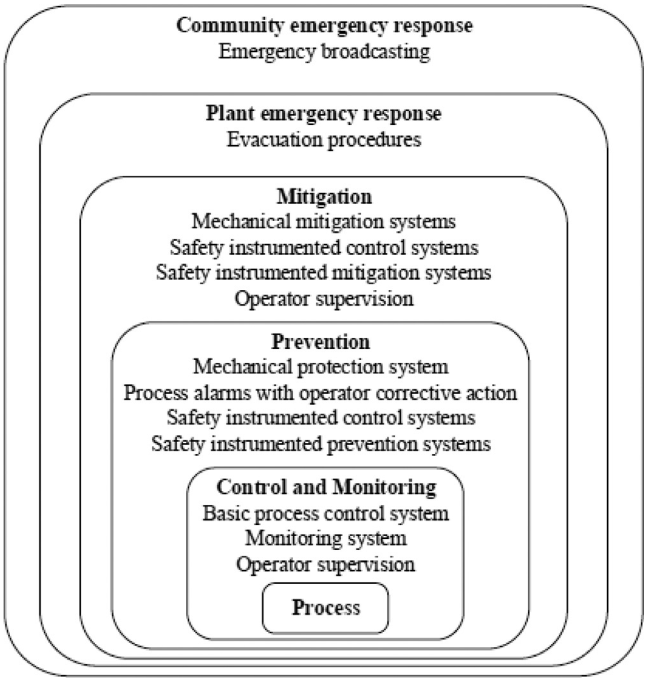

Fig. 1. Protection layers in process industries (IEC61511, 2003).

Nomenclature

a	number of true positive samples
a'	number of faulty samples by the operator and the classifier
b	number of false positive samples
b'	number of samples diagnosed as faulty in contrast with operator opinion
c	number of false negative samples
c'	number of samples labeled faulty by operator but not diagnosed
C_A	concentration of A in the outlet stream
C_{Af}	concentration of A in the feed stream
C_B	concentration of B in the outlet stream
d	number of true negative samples
d'	number of samples which both the operator and the classifier have agreement on no fault happened.
$\boldsymbol{D}_M$	Mahalanobis distance
D_p	dataset at each phase
F	fault
F_t	identification matrices of observation equation
G_t	identification matrices of state equation
$I_{interaction}$	interaction index
m	number of attributes
p^{th}	counter of phase
R	number of runs
r^{th}	counter of runs
T	temperature of the reactor
t	time
T	time horizon
T_f	temperature of the inlet stream
T_j	temperature of the jacket
v_t	Gaussian random vectors of observation equation
V_t	variance matrices of observation equation
w_t	Gaussian random vectors of state equation
W_t	variance matrices of state equation
Y_t	attributes

Greek symbol

$\boldsymbol{\theta}_t$	state variables
Σ	covariance matrix
μ	mean
σ	standard deviation
π	probability

Acronyms

ASM	abnormal situation management
ARL	acceptable risk level
CSTR	continuous stirred tank reactor
DLM	dynamic linear model
FD	fault detection
FDA	Fisher discriminant analysis
FPR	false positive rate
FNR	false negative rate
GNB	Gaussian naïve Bayes
HMM	hidden Markov model
KPI	key performance indicator
MLE	maximum likelihood estimation
PCA	principal component analysis
PID	proportional, integral and derivative
PLS	partial least squares
SVM	support vector machines
TPR	true positive rate

- **Geometric classifier**. e.g. k-nearest neighbourhood (kNN) (Pandya, Upadhyay, & Harsha, 2013);
- **Probabilistic classifier**. e.g. Gaussian naïve Bayes (GNB) (Askarian, et al., 2015; Sáez-Atienzar, et al., 2015) and the hidden Markov model (HMM) (Li, Fang, & Huang, 2015);
- **Approximation classifier**. e.g. polynomial support vector machines (SVM) (Danenas & Garsva, 2015; Namdari & JazayeriRad, 2014);
- **Soft computing techniques**. e.g. fuzzy classifier (Serdio, Lughofer, Pichler, Buchegger, & Efendic, 2014) and artificial neural networks (Duda, Hart, & Stork, 2001).

The main advantage of FD using classification methods is the ability at dealing with unstructured information and implicit knowledge. However, each method poses some limitations that are discussed in detail by Isermann (2006). The major weak point of classifiers is vulnerability to mislabeling, which is the issue explored in this work.

In Machine Learning, the standard approach consists in training a classifier from a labeled dataset to predict the class of new samples accordingly. Usually, labels are considered given and the labeling process is assumed to be reliable (Bootkrajang & Kabán, 2012). However, in industrial practice and process plants, assigning labels to training data may need attention and careful examination. Indeed, true labels corresponding to the state of system are usually unavailable. Mislabeling may occur for several reasons including expert errors, lack of information or data labeling by non-experts (de França & Coelho, 2015). Label uncertainty is an important issue in classification, because most classifiers are built on the hypothesis of a perfectly labeled training set. Some Machine Learning literature exist regarding effects of uncertain labels, which shows that mislabeling may detrimentally affect the classification performance and the reliability of the learned models (Brodley & Friedl, 1999; Frénay, de Lannoy, & Verleysen, 2011).

Numerous methods have been proposed to deal with label noise. Filter approaches aim at identifying and removing any mislabeled instances (Brodley & Friedl, 1999; Zhang, Li, Yang, & Yong, 2014). A residual-based fault detection is developed which solely