Diagnosis Failure Cause of complex Pharmaceutical System by Bayes Learning for Decision Support

Ngoc-Hoang Tran

Abstract*: This work proposes a real application of diagnosis protocol for complex pharmaceutical process drifts. Main challenge is to identify and classify failure causes of production process. The model which we have proposed is structured in the causal graph form, named "Hierarchical Naïve Bayes" (HNB) formalism. Our contribution is the presentation of a methodology that allows developing flexibility in particular complex pharmaceutical production context. A data extraction and processing prototype is performed in this paper from real pharmacy company to build Bayesian model. Diagnosis results are decision support elements that built based on HNB probabilities. Furthermore, this work can be applied in order to improve production quality in businesses competition.

Keywords*: Modeling and identification, Pharmaceutical production system, Data learning, Equipment diagnosis, Bayes Networks.

I. INTRODUCTION

Pharmaceutical production process is considered to be particularly complex, uncertainty and sensitive to production drifts. This process is characterized not only by complex manufacturing process and a highly uncertain environment [1], and more by direct intervention from process operation of operators, control recipe quality to decision - making. This make constant searching for maximum quantify risk accurately to improve product quality and reduce associated costs, fault isolation, detection and diagnosis have important one. According that, many diagnosis methods have been developed such as [6, 7] to localize more quickly and diagnosis accurately the failures causes of manufacturing process. These approaches model equipment as a single unit and collect data from sensors to identify equipment failures against product and process drifts.

However, sensors are not directly positioned on the product for technical reasons. Therefore, the manufacturing process has the risk of not observing perturbation that affects the product quality [2]. Many drifts are unavoidable in the production process. Indeed, recent increase of equipment breakdown in complex production process needs an improved methodology to ensure sustainable analyses capacities. The analysis on an experimental production data from a pharmaceutical Danapha's production process show that failure have significantly increased according its complex and important collected data.

With the objective is to support maintenance engineers for more accurate decisions about failures and correction actions on equipment/process, a proposed diagnosis methodology is presented in [3] are modeled as Bayesian network (BN) with unsupervised learning of structure using data collected from the variables (classified as symptoms) across production, process, equipment and maintenance databases. That approach contributes in increasing the effectiveness of diagnosis process failure but need proved in a real manufacturing process. Our work put in place between of development this diagnosis approach and instantiating in our real study case: Pharmaceutical Danapha's pharmaceutical production process. This paper is organized by follow structure: in next section, we present the element of diagnosis process, modeled by probabilistic Bayes Network (BN) approach. Then an application on pharmaceutical process and diagnosis results show how to our model work. Finally, conclusion close this paper.

II. DIAGNOSIS PROCESS

In this section, we present the diagnosis process consist of four steps described as Fig.1 that are designed based on the identification and classification variable process in [3]. This process in this paper is identified potential causes and propose of their characterized parameters on pharmaceutical process.

A. Instantiating characterized variables of pharmaceutical context

The first phase shows the analysis of the production system context where was characterized by high complexity and uncertainty in industrial manufacturing environment. Pharmaceutical system is even more complex by multiple processes exploitation on same production line with a large operations numbers to produce a finished product from raw material. Also, uncertain factor is presented by equipment drifts, human errors...that unscheduled, can impact the process control and product quality.

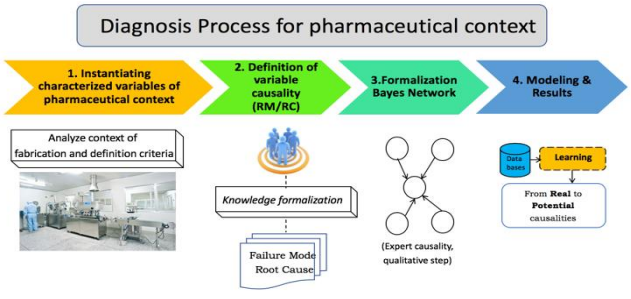

Fig. 1.Diagnosis process.

Revised Manuscript Received on May 21, 2020.
***** Correspondence Author
Ngoc-Hoang Tran*, The University of Danang – University of Technology and Education, 48 Cao Thang, Danang, Vietnam. Email: tnhoang@ute.udn.vn

B. Definition of variable causality (RM/RC)

In this phase, we analyze variable correlation data to identify a set of causal relationship of variable, consist of Failure Modes (FM) and Root Causes (RC). In this case, by using the Failure Modes, Effects and Critically Analysis (FMECA) approach which was supported by Danapha's company experts, this phase determines the considered priorities that using for the qualitative classification of failure modes.

C. Formalization Bayes Network model

A Bayes Network provide a simple representation graph including of nodes and arcs [8]. These nodes representation the system state or process condition: discrete or continuous, observable or unobservable variable. Respectively, the arcs of Bayes Networks represent the causal relations between their nodes. In order to model a probability distribution of Bayes Network, it is necessaire to be based on Bayes' theorem:

In general, given a set of node $X = \{X_1, X_2, \ldots, X_n\}$, we calculate the joint probability distribution $P(X)$ of these nodes by follow formula:

$$P(X) = \prod_{i=1}^{n} P(X_i / Parents(X_i)) \quad (1)$$

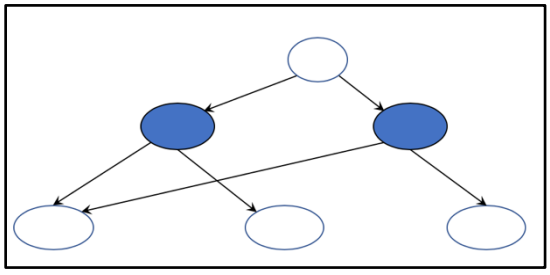

Fig. 2. Hierarchical Naïve Bayes models example.

A classical Bayes network, also called Naïve Bayes network or Bayes classifier is simplest structure which has two nodes levels: parent node (one single node) and children node (one or several nodes). This classifier is used common cause it's performance. In this network, its variables are all discrete. However, Naïve Bayes networks has many disadvantages on unsupervised process where it exists the unmeasured levels. In fact, they are developed in recent years such as Latent Hierarchical models [9], Hierarchical Naïve Bayes models [11]. These works provide an extension of classical Bayesian network with the latent nodes which their class (shown in blue in the Fig. 2) is unmeasured and unobserved.

To build a Bayesian network, two concepts of learning are normally used: parameter learning and structure learning. The first allows to estimate the conditional probability laws (parameters) given a structure of Bayes network by acquiring expert's knowledge. The second, it must determine an optimal graphical structure of the model by learning from observed production data. However, it should be noted that depend on learning data quality and it will grow in a super-exponential correspond this data set size. For Bayesian inference, there are many algorithms (Maximum Likelihood, Expectation – Maximization…) with multi libraries are large used (R-bayesm, BNT Matlab, BNJava,...) and quality software (Bayesfusion, BayesiaLab, AgenaRisk...). By learning data, the conditional probability table of each variable of Bayes network is computed based on the Bayes theory. These probabilities results support us to make decision of correction and maintenance process [5].

In this paper, the formalization of the Bayes Network (BN) model begin with the generation of all the variables of the model for the considered production system. In fact, the previous step can be considered as a definition step of the types of variables, this step will extract from the database the information necessary for the instantiating such as for example the list of sensors, involved on the equipment and thus generate the associated variables.

D. Modelling

Once the Causes and the Failure Mode is identified by experts in previous phase, we combine them to build a Bayesian graphical structure model. This structure can be approved by learning structure from a historical database dedicated to learning with appropriate algorithms. Result of this phase makes show us different probability distributions associate with each of the variables.

Finally, the testing and validation this model is consists in interrogating the model for predicting the failure modes or to diagnose their origins on a production data dedicated for testing.

III. APPLICATION TO PHARMACEUTICAL PROCESS

A. Description

In our framework of this paper, we interest in pharmaceutical manufacturing of dosage forms and its production process. Fig. 3 illustrates a part of an operation poste for manufacturing of pharmaceutical dosage-form products in DANAPHA company. From stock, Pharmaceutical material blends are be milled by a milling machine, direct after that to dry granulation or high mixed speed to obtain the desired physical properties, before their formulation as a finished drug product. In the granulation, the active ingredients and excipients are wetted with aqueous or solvent solutions to produce course granules with enlarged particle sizes. The granules are dried, mixed with lubricants (e.g., magnesium stearate), disintegrates or binders, then compressed into tablets at the end of this process [12]. Many different faults that are different nature are detected in this process. They are for example malfunction component, production drift (variation of certain variables values) or actuators breakdowns. In this paper, we propose to work on only two faults that can be allowed presenting (FM_1 and FM_2) as table I.

B. Diagnosis model

Corresponding to two fault FM1 and FM2, a set of variables which consist of process parameters and latent causes is proposed as table I. These variables are inherently based on knowledge of Danapha's experts that pilot process directly. This table is also confirmed by an inference mechanism from historical collected data. In general, we propose three distinct categories of variables:

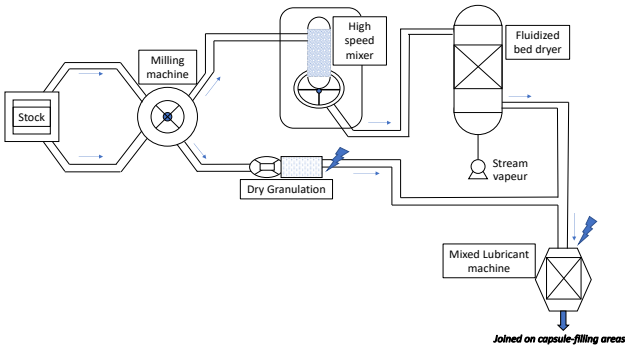

Fig. 3.Danapha's pharmaceutical manufacturing of dosage- form product.

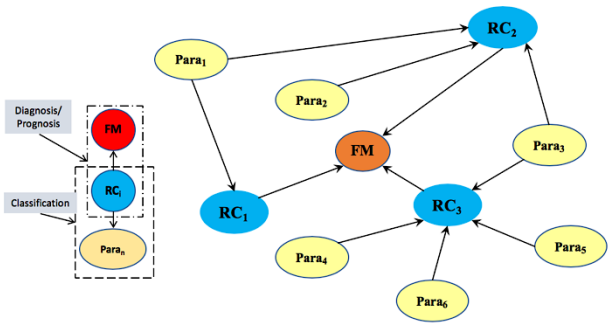

Fig. 4.Diagnosis model by Hierarchical naïve Bayes network.

• Primary class: These Failure Modes (FM_i) that we assume to represent the happen states in manufacturing process. These variables take two possible values (detected, not detected): two failure variables FM_1 and FM_2 correspond to express the "Wet granulation" and "Mixed lubricant gear deflection".

• Second class: The Root Causes (RC_j) are quantitative variables defined by expert opinion. They correspond to three elements causes of process. As above variables, these RC_j take binary mode (observed or unobserved).

Table- I: Process variables of diagnosis model

Node	Description
FM_1	Wet Granulation
FM_2	Mixed Lubricant gear deflection
RC_1	High Mixed Volume
RC_2	Low Milling Quality
RC_3	Air Heater Pressure
$Para_1$	Material Volume
$Para_2$	Milling's rotor speed
$Para_3$	Dry Granulation's roller speed
$Para_4$	Drying pressure
$Para_5$	Drying temperature
$Para_6$	Drying volume

• Third class: The Parameter descriptions $Para_n$ (n=1→50): they are determined by the real Danapha's process. There are totally 50 parameters that identified in this process by Danapha's operators. In this paper, we propose table I in which present only six collected parameter which are considered. In order to model, these variables have either a binary mode (true/ false) which is resulting from discretization process.

A graph structure model with these three variables classes is proposed based on Bayes 's rules. That classify the diagnosis failure causes with two hierarchical classes RC_j and $Para_n$. In which the RC_j is represented by knowing the parameter $Para_n$ nodes who is considered also as the causes. This model offers at the final the probability distributions associated with each of variable in Fig. 4. In follow section, our result would be presented in next section.

C. Diagnosis Results

Our contribution is at first Bayes learning simulation data learning by collecting multi production data in case study. A matrix (11 x 10.000) corresponding to 11 variables (as we described) and 10.000 real collected samples from Danapha's company for learning the probabilities are created by BNT Matlab © library [10]. The calculation of probabilities is calculated by MLE (Maximum Likelihood Estimation) algorithm. The principal idea of this algorithm is estimating the probability of a variable based on its occurrence in the considered dataset.

In pharmaceutical manufacturing, the production process is normally controlled and monitored in real time by Fault Detection and Classification (FDC), Statistical Process Control (SPC). This manufacturing process is located in the coordination level in CIM pyramid. Therefore, planning & supervision systems such as the Enterprise Resource Planning system (ERP), Computerized Maintenance Management System (CMMS), Manufacturing Execution System (MES) [11], SCADA system (Supervisory Control and Data Acquisition) are integrated to manage from production demands to collect the production information of control process. Therefore, our production dataset for learning consist of production time, machine's name, machine ID, function, recipe, human factor. Moreover, the dataset of maintenance consists of Time, Machine's name, Machine state evaluation. The dataset is collected in six months from the CMMS, the metrology detects the product quality where final products are detected as good or bad product.Also, the metrology data composes of time, product type, Lot number and Product quality in real time. Our learning data is built by collecting from FDC, SPC, RMS, metrology data, production data as shown in Fig. 5.

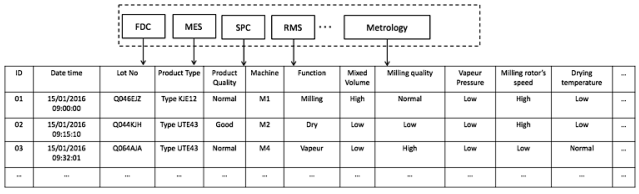

Fig. 5.Collected production data for Bayes learning

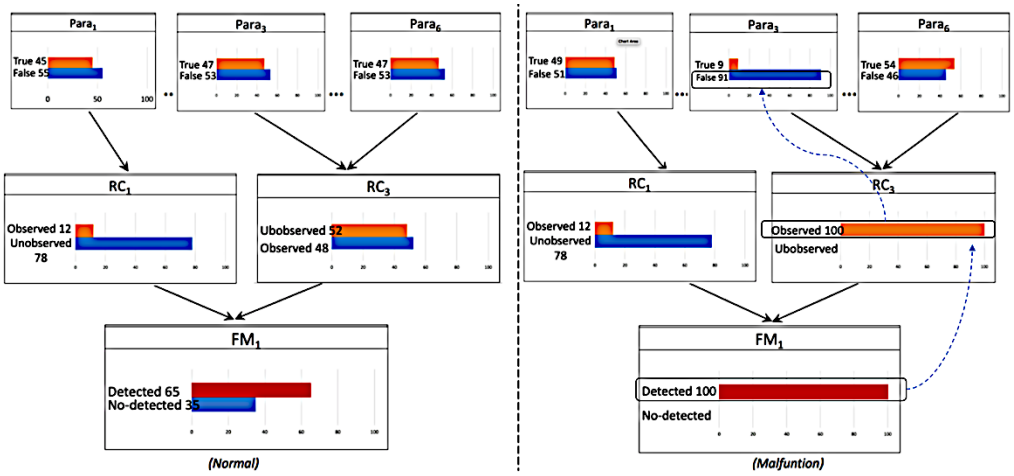

Fig. 6. Diagnosis results by probabilites of variables

Then, Fig. 6 present results illustrative a scenario inference whereas have only the probability distributions of a part of variables on table I after learning from database. The probabilities of these variables from two case without and appear of failure mode FM_1 in system. By comparing it's probabilities in two case, operator can make a correct decision. $P(RC_i|FM)$ and $P(Para_n|RC_i)$ of each variable show us how's our model work since the observation a failure mode. In fact, probabilistic inference is obviously based on learning results. The result show in Fig. 6 that the proposed method performs good detection capability by showing the root cause RC_3 and $Para_3$ who cause the high speed of roller of drying.

A real problem drift of Air Heater Pressure of manufacturing process is happened in reality on April 2016 in this Danapha's atelier. Base on learning and inference results, a similar inference is founded as comparing with happens reality. This show how our model works for supporting to make a correct decision.

However, it must be note that the structure of classifier is not easily established even either by learning from data or expert's opinion if it's existant too many variables representing production process. Therefore, it should be necessary to propose weights primarily for each variable depend on their differences properties in order to make optimal distribution. These indicators can be proposed by operator's experience or by learning from historical production data in some case.

IV. CONCLUSION

This paper deal with diagnosis problem of industrial pharmaceutical production. Our methodology is presented with detailed steps from definition of characteristics of pharmaceutical context to particular Bayes approaches for modelling diagnosis process. In simulation result, this method is effectiveness for diagnosis failure cause on complex data in Danapha's case study.

ACKNOWLEDGMENT

The author wishes to thank Danapha Pharmaceutical Company. This work was supported in part by their description documents of manufacturing process, learning production data and real scenarios.

This research is funded by University of Technology and Education – The University of Danang under project number T2019-06-131.

REFERENCES

1. Zio, Enrico. (2013). "System Reliability and Risk Analysis." *The Monte Carlo Simulation Method for System Reliability and Risk Analysis. Springer*, London, 7-17.
2. Bouaziz. M.-F, Zamaï. E, Duvivier. F. (2013). "Towards Bayesian Network Methodology for Predicting the equipment Health Factor of Complex Semiconductor Systems". *International Journal of Production Research*, Volume 51, Issue 15, 4597-4617.
3. Tran, N. H., Bouaziz, M. F., & Zamaï, E. (2014). "Identification and classification protocol for complex systems". In *2nd European Conference of the Prognostics and Health Management Society, PHME 2014* (pp. 58-65).
4. Pearl J., (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan – Kaufmann, San Diego.
5. Pearl, Judea. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.
6. M. Sampath, S. Lafortune and D. Teneketzis, (1998). Active diagnosis of discrete-event systems, Auto- matic Control, IEEE Transactions on, 43(7), pp 908–929.
7. E. Deschamps and E. Zamai, (2007). Diagnosis for control system reconfiguration, In IFAC Manage- ment and Control of Production and Logistics, volume 4, no.1, pp. 377–382.
8. Jensen F.V., (1996). *Introduction to Bayesian networks*, UCL Press, London.
9. Bishop, C. M. and Tipping M. E., (1998). "A hierarchical latent variable model for data visualization". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(3), 281–293.
10. Murphy K.,. The Bayes Net Toolbox for Matlab. (2011). *Computing Science and Statistics: Proceedings of Inference*, vol. 33.
11. Tran, N. H., Henry Sébastien, and Eric Zamaï. (2016). "Generic and configurable diagnosis function based on production data stored in Manufacturing Execution System." *Third European Conference of the Prognostics and Health Management Society 2016*. Vol. 7. No. 057.
12. Gibson, Mark, ed. *Pharmaceutical preformulation and formulation: a practical guide from candidate drug selection to commercial dosage form*. CRC Press, 2016.

AUTHORS PROFILE

Mr Ngoc-Hoang Tran was born in Danang, Vietnam in 1986. He completed his Master's degree on engineering of complex system in 2013 and PhD degree in automation and civil production from Grenoble-INP, France in 2018. He is currently working on Mechatronics department-Faculty of Mechanical engineering, University of Technology and Education – The University of Danang. His main research interests include equipment diagnosis, IoT supervision and AI recognition with Bayes network technology.