First Principle Models Based Dataset Generation for Multi-Target Regression and Multi-Label Classification Evaluation

Ricardo Sousa[1] and João Gama[1,2]

[1] LIAAD/INESC TEC, Universidade do Porto, Portugal
rtsousa@inesctec.pt
[2] Faculdade de Economia, Universidade do Porto, Portugal
jgama@fep.up.pt

Abstract. Machine Learning and Data Mining research strongly depend on the quality and quantity of the real world datasets for the evaluation stages of the developing methods. In the context of the emerging Online Multi-Target Regression and Multi-Label Classification methodologies, datasets present new characteristics that require specific testing and represent new challenges. The first difficulty found in evaluation is the reduced amount of examples caused by data damage, privacy preservation or high cost of acquirement. Secondly, few data events of interest such as data changes are difficult to find in the datasets of specific domains, since these events naturally scarce.

For those reasons, this work suggests a method of producing synthetic datasets with desired properties(number of examples, data changes events, ...) for the evaluation of Multi-Target Regression and Multi-Label Classification methods. These datasets are produced using First Principle Models which give more realistic and representative properties such as real world meaning (physical, financial, . . .) for the outputs and inputs variables. This type of dataset generation can be used to produce infinite streams and to evaluate incremental methods such as online anomaly and change detection. This paper illustrates the use of synthetic data generation through two showcases of data changes evaluation.

Keywords: Synthetic data, data streams, Multi-Label Classification, Multi-Target Regression, First Principle Models

1 Introduction

In the areas of Machine Learning and Data Mining, datasets quality and quantity are crucial for evaluation stage of methods development [1]. Controlled evaluation environments with specified challenge problems are required to understand the behaviour of the methods [2]. Methods of Multi-Target Regression(MTR) and Multi-Label Classification (MLC) on online data streams are fair examples that imply these evaluation requirements. The importance of these methodologies has been growing due to reasonable modelling and predicting capabilities [3, 4]. Formally, let an unbounded data stream be represented by $\mathcal{D} = \{..., (\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_i, \mathbf{y}_i), ...\}$, where $\mathbf{x}_i = [x_{i,1} \cdots x_{i,j} \cdots x_{i,M}]$ is a M-dimensional vector of real values containing the data descriptive variables $x_{i,j}$ (input variables) of the i^{th} example (considering one example with the index of reference). For Multi-Target Regression, $\mathbf{y}_i = [y_{i,1} \cdots y_{i,j} \cdots y_{i,N}]$ denotes a real values vector of responses $y_{i,j}$ (output variables) of the i^{th} example. For Multi-Label Classification, $\mathbf{y}_i$ corresponds to a subset of nominal labels λ_k such that $\mathbf{y}_i \subseteq \{\lambda_1, ..., \lambda_k, ..., \lambda_L\}$, where L is the number of possible labels.

Typically, output set of labels $\mathbf{y}_i$ are transformed into a vector of outputs variables $[y_{i,1} \cdots y_{i,k} \cdots y_{i,L}]$, where $y_{i,k} \in \{0, 1\}$ are binary. If label λ_k is assigned to the i^{th} example then $y_{i,k} = 1$, otherwise $y_{i,k} = 0$. The outputs variables are redefined as $\mathbf{y}_i = [y_{i,1} \cdots y_{i,k} \cdots y_{i,L}]$. Finally, the objective of both MTR and MLC methods is to learn a function $f(\mathbf{x}_i) \rightarrow \mathbf{y}_i$ that maps the input values of $\mathbf{x}_i$ into the output values of $\mathbf{y}_i$.

In the evaluation of the methods, the number of examples are not sufficient in many cases by the reasons of sensitive data, data damage or high cost of acquisition [2]. Data changes scenarios are another prominent challenge for MTR and MLC methods [5]. Changes in the probability distributions, variables trends or variables rapid shifts of the inputs variables are events that have strong influence on the method's performance [5]. Similarly to the scarcity of examples, real world datasets that gather all desired data changes properties also lack [2]. Moreover, the data change events are not often annotated, since annotation is time consuming.

As an attempt to solve this problem, researchers produce synthetic datasets to create evaluation challenges with desired properties. This alternative allows to produce a significant amount of examples or even create a reasonable approximation of an infinite data stream. Moreover, few resources for storage and transmission are required. Despite high complexity, the produced datasets models do not reflect the real world conterpart. In fact, the latent models are based on abstract mathematical concepts.

However, datasets can be constructed through the employment of a First Principle Models (FPM) which are described by established laws without making assumptions (empirical or fitted parameters) [6]. FPM are used to create synthetic data for a wide range of areas such as Chemical Engineering (Industrial Chemical process) [7] and Mechanical Engineering(Mechanical Systems Diagnosis) [8]. In the area of Control Systems, Proportional-Integrate-Derivative (PID) systems modelling uses FPM extensively in several contexts of application [9]. For instance, FPM based software simulators(parametrized with inputs and outputs variables) that mimic those systems are created to reduce cost in the industrials trials [7]. The abundance of free software simulators and models of PID systems justified the focus on PID Systems. Thus, this work suggests a method to produce synthetic datasets that are reproducible for MTR and MLC evaluation. This method applies the FPM to produce more realistic and representative models.

Section 2 briefly reviews some existent methods of dataset generation. Section 3 describes the FPM method and the selected FPM model that is used to produce the synthetic datasets. Section 4 shows the production of the MTR and MLC synthetic datasets and their application through MTR and MLC methods evaluation showcases, under data changes scenarios. Finally, the results are presented and discussed in Section 5 and the main conclusions are reported in Section 6.

2 Related Work

In the literature, most of dataset generators produce Single-Label Classification (the output variable $\mathbf{y}_i$ is a label) datasets, since MLC is an emerging methodology [1]. Monedero et al [2], Frasch et al [10] and Narasimhamurthy et al [5] are representative examples of Single-Label Classification(SLC) dataset generators.

One possible strategy is to produce several Binary Classification (where $y_{i,k} \in \{0, 1\}$) outputs (one for each label) and combine them into MLC datasets. However, these datasets does not represent correlation between outputs.

Read et al propose a dataset generator that uses single-label generators and combine them according to configured label imbalance and probabilities of simultaneous label occurrence [11]. This dataset generator attempt to create more realistic datasets in terms of label imbalance and concept drifts through an empirical method.

Tomas et al propose a highly configurable dataset generators for MLC based on the creation of hyper-spheres [1]. However, the generator is not designed to produce data changes. In addiction, despite high flexibility, this dataset generator produces an abstract mathematical challenge.

Similarly, the majority of the existent datasets generator produce Single-Target Regression(also known as Multivariate Regression) datasets with simple but highly non-linear models. Friedman produced STR datasets from very simple and non-linear models to test the methods of Multivariate Adaptive Regresssion Splines [12] The same strategy used in MLC can be applied to produce MTR from a STR. For the best of our knowledge, no MTR dataset generators were found in literature. In fact, MLC methodologies received more attention by the researchers.

3 First Principle Model-Tennesse Eastman Process

In order to illustrate the application of FPM's to generate MTR and MLC datasets, the Tennessee Eastman Process (TEP) was chosen. TEP consists of an industrial process of continuous chemical production. The process is unstable, non-linear and controlled by PID system. Basically, PID systems are controlled mechanisms based on loop feedback that are used in process stabilization [9]. Figure 1 shows the model of a generic PID system. These systems involves essentially a Plant and a PID controller.

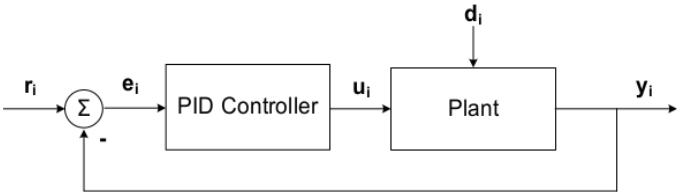

Fig. 1. Model of a generic PID system.

The Plant consists of a set equations that represent the behaviour of the controlled process. The process is driven by the manipulable variables $\mathbf{u}_i$ and observed by the measurement variables $\mathbf{y}_i$. Some processes present the disturbances variables $\mathbf{d}_i$ (binary variables)to simulate process impairments. The error $\mathbf{e}_i$ between the desired set points $\mathbf{r}_i$ and process measurements is computed with the purpose of being minimized. The PID controller consists of weighted sum of proportional(present values), integral(past values) and derivative (possible future values) terms which are calibrated in order to produce stable responses [9]. This component receives the error and computes new manipulable variables values that stabilizes the Plant process. Figure 2 represents the diagram of TEP Plant. Most process details were originally omitted for simplification and protection of intellectual property.

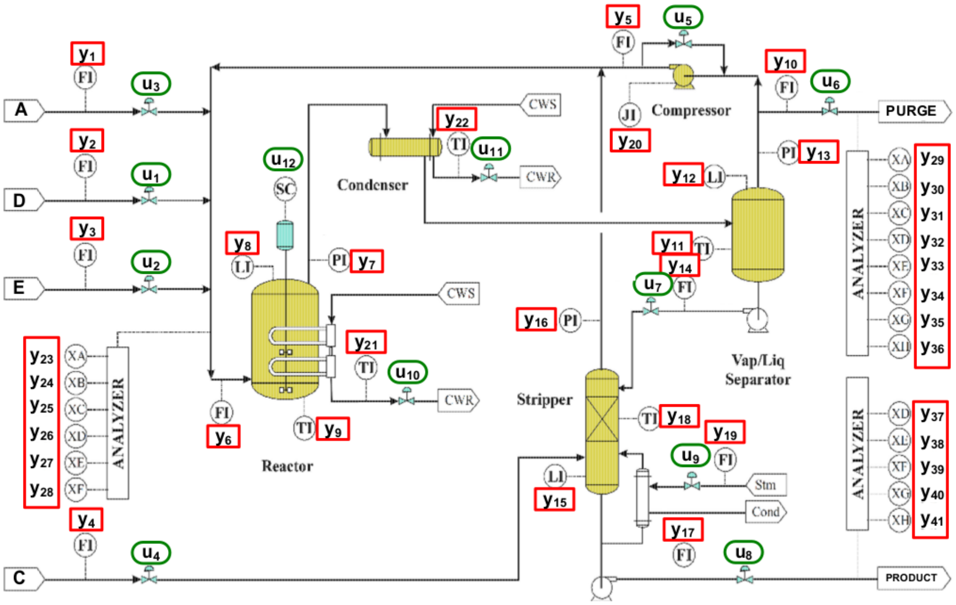

Fig. 2. Model of TEP Plant. The $u_j = u_{i,j}$ are the individual manipulable variables, $y_j = y_{i,j}$ are individual measurements variables.

Two products G and H (Product) are produced from four gaseous reactants (A, C, D, E). An inert product B is present but does not intervenes in the chemical reaction. A by-product F (Purge) results from the whole reaction. The chemical reactions are irreversible and exothermic. The process model comprises five inter-acting major units: a reactor, a condenser, a vapour-liquid separator, a stripper for the product stream and centrifugal compressor for the recycle stream. The model has 41 measurements and 12 manipulated variables and 12 set points. There is also 20 variables that simulate disturbances. The physical quantities that the variables represent are explained in detail in Bathelt et al [7]. Tables 6 to 5 gives the names of the variables and respective physical meaning (see in Appendix).

4 Methods

This section gives the description of procedures for the generation of MTR and MLC datasets. Two evaluation showcases of MTR regressor and MLC classifier which predict over the generated datasets are also described. The generation of the datasets was focused on data changes robustness tests. Data changes in streams can be analysed in two aspects: nature and rate. The nature reflects the variables statistics that changed such as mean and variance [13]. The rate of change is an important aspect that influences the performance of most MTR and MLC methods. Abrupt changes (concepts drifts) are identified when the change occur from one example to the next (inexistence of transition phase). Gradual changes (concepts shifts) present a transition phase where the changing statistic is continuously varying [13]. This work is focused on the abrupt changes (concept drifts) of the mean statistic which is one of the most studied topics in data streams [14].

In this work, a simulator that implements the model described in Section 3 was used. The files of this simulator can be found at http://depts. washington.edu/control/LARRY/TE and was originally developed by Ricker [15]. This simulator was partially developed in C and in Matlab(Simulink). TEP simulator is composed essentially by a plant function developed in C language and posteriorly built into a Matlab mex file. The PID controller (a set of small PID controllers) function was developed in Simulink.

To produce data changes events, the set points variables are manipulated. Some set points such as Production Set Point allows to produce gradual changes (concept shifts) and abrupt changes (concept drifts) events, since the variables are related to measurements variables convergence to stable values. The rate can be calibrated using a rate limiter in the simulink environment. Figure 3 shows the parameters of data change event. The curve reflects the variation of a statistical parameter from a value **a** to a value **b**. The parameter **t** is the example index where the change starts and **W** is the number of examples in the transition phase. The change in the set point value create similar changes in the Plant parameters. The changes can be whether a descending and ascending of a parameter.

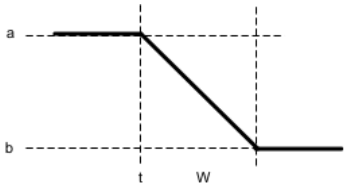

Fig. 3. Model of a data change reflecting a statistical parameter variation.

The generated datasets were produce with 100000 examples each, for both MTR and MLC evaluation. In this experiments, the changes were created with the periodicity of 5000, 10000 and 20000 examples. The W values were 0 in order to simulate a concept drift. Two special cases were also produced. One dataset does not present any drift(base line) and the other presents a constant decreasing change. Regarding the MTR dataset production, the dataset joins measurement variables $\mathbf{y}_i$ and manipulable variables $\mathbf{u}_i$ with purpose of predicting $\mathbf{u}_i$ from $\mathbf{y}_i$. Each data example is defined as $\mathbf{e}_i = (\mathbf{y}_i, \mathbf{u}_i)$. As performance measures, the error was used for the error evolution and RMSE was used for global evaluation and comparison between algorithms. The error curve was smooth with a median sliding window due to the spiky form. The window length is 1000 examples without overlapping. For MLC dataset generation, the disturbances variables $\mathbf{d}_i$ and measurement variables $\mathbf{y}_i$ are joined. The purpose is to predict the disturbance variables $\mathbf{d}_i$ from $\mathbf{y}_i$. The $\mathbf{d}_i$ disturbance variables are already in the form used in MLC problem transformation [11]. Each data example is defined as $\mathbf{e}_i = (\mathbf{y}_i, \mathbf{d}_i)$. A Poisson process was used to choose the instant of a disturbance occurrence with a user defined duration. The duration of the disturbances was 5000 examples. In this evaluation, the inputs variables were smoothed with a low-pass filter (sliding windows of 1000 examples) in other to have stability. The F-Measure, Precision, Recall, Accuracy and Exact Match were used for the classification scenario.

The Multi-Target regressor MT-AMRules and the Multi-Label classifier ML-AMRules were used to exemplify the application of synthetic datasets in MTR and MLC methods evaluation, respectively. Both MTR regressor and MLC classifier were tested using a prequential mode.

5 Results

In this section, two simple showcases with MTR and MLC are demonstrated. The types of involved data changes events are also visualized. A comparison between a scenario without data changes and several scenarios where several types of data changes occur is performed. Figure 4 depicts 4 data changes in the mean statistic with different rates using the Reactor Coolant variable. The porpose is to show the flexibility and the diversity of the adapted software to create several types of data changes.

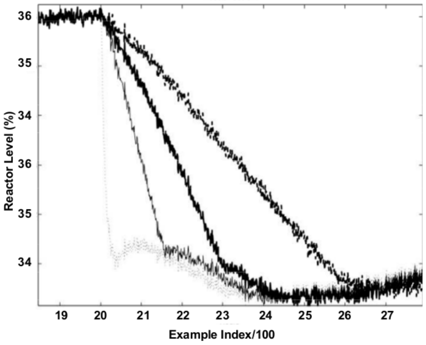

Fig. 4. Examples of data changes with different rate of variation observed in Reactor Coolant variable.

These gradual and abrupt changes are created using rate delimiters in Simulink implementation. The thinner curve represents an abrupt change (concept drift) which is the event that is intended to be evaluated in MTR and MLC showcases.

In the following paragraphs, the MTR and MLC evaluation showcases are presented. Figure 5 shows the evalution of the error for several MTR datasets.

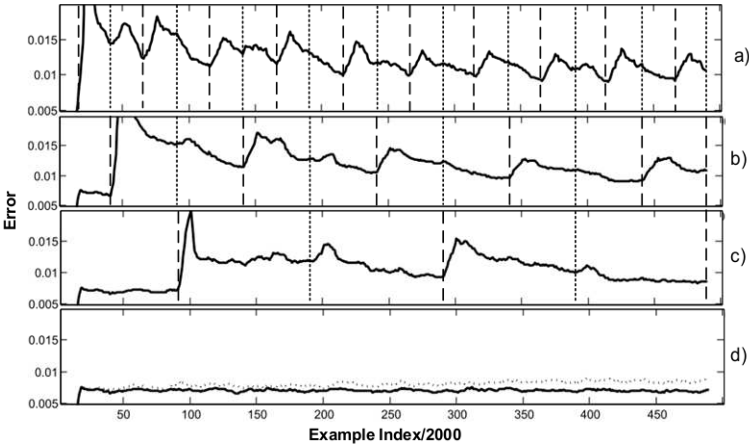

Fig. 5. Error evolution that show the effect of data changes.

The plots a) to c) represent the smoothed error curve for 5000 (MTR_5k dataset), 10000 (MTR_10k dataset) and 20000 (MTR_20k dataset) examples of concept drift periodicity, respectively. The plot d) represent in solid line the scenario where the dataset presents no data change (MTR_NoChange dataset) and the scenario where the data change is constant (MTR_Const dataset) in dotted line. The error curves show the effect of the concept drifts. Figures 5 shows descending (dashed) and ascending (solid) concept drift events. The descending variation cause error increasing with more impact than ascending variation. The impact of ascending variation can be observed in the plot c). Table 1 show the results of MTR regression using several datasets with different challenges.

Table 1. RMSE of the MTR evaluation.

Dataset	MTR_5k	MTR_10k	MTR_20k	MTR_NoChange	MTR_Const
RMSE	0.158	0.151	0.146	0.101	0.119

Table 1 shows that the more often data changes, the higher is the RMSE. This fact is espected, since the drifts lead to the relearning of the model. Interestingly, Figure 5 d) shows that the constant and gradual change lead to a gradual increasing of the error compared to the drift scenario where no drifts occurs.

Table 2 show the results of MLC regression using several algorithms. Performance measures of the MLC evaluation for datasets of 5000 (MLC_5k dataset), 10000 (MLC_10k dataset) and 20000 (MLC_20k dataset) example periodicity. It also presents the performance measures for scenarios of no data change (MLC_NoChange dataset) occurred and constant change (MLC_Const dataset).

Table 2. Performance measures of the MLC evaluation

Dataset	MLC_5k	MLC_10k	MLC_20k	MLC_NoChange	MLC_Const
Accuracy	0.80	0.79	0.80	0.83	0.79
Exact Match	0.62	0.61	0.62	0.66	0.60
Precision	0.65	0.65	0.64	0.62	0.65
Recall	0.68	0.66	0.68	0.63	0.67
F-Measure	0.65	0.64	0.64	0.61	0.65

Table 2 shows that the drifts and the gradual change produce little effect on the performance measures.

6 Conclusion

This work presented a framework for data set generation for MTR and MLC evaluation. A realistic and representative datasets for MTR regression and MLC Classification were obtained for method evaluation. However, the main limitations are the limited number of inputs and outputs. As future work, the main goal is to implement a MTR and MLC data streamer that produces data trough a configurable FPM in a standalone and portable application.

7 Acknowledgements

This work was partly supported by the European Commission through MAESTRA (ICT-2013-612944) and the Project TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE -01-0145-FEDER-000020 is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

1. Jimena Torres Tomás, Newton Spolaôr, Everton Alvares Cherman, and Maria Carolina Monard. A framework to generate synthetic multi-label datasets. *Electronic Notes in Theoretical Computer Science*, 302:155 – 176, 2014. Proceedings of the {XXXIX} Latin American Computing Conference (CLEI 2013).
2. Javier Sánchez-Monedero, Pedro Antonio Gutiérrez, María Pérez-Ortiz, and César Hervás-Martínez. An n-spheres based synthetic data generator for supervised classification. In *Advances in Computational Intelligence - 12th International Work-Conference on Artificial Neural Networks, IWANN 2013, Puerto de la Cruz, Tenerife, Spain, June 12-14, 2013, Proceedings, Part I*, pages 613–621, 2013.
3. Changsheng Li, Weishan Dong, Qingshan Liu, and Xin Zhang. MORES: online incremental multiple-output regression for data streams. *CoRR*, abs/1412.5732, 2014.
4. Jesse Read, Albert Bifet, Geoff Holmes, and Bernhard Pfahringer. Streaming multi-label classification. In *Proceedings of the Second Workshop on Applications of Pattern Analysis, WAPA 2011, Castro Urdiales, Spain, October 19-21, 2011*, pages 19–25, 2011.
5. Anand Narasimhamurthy and Ludmila I. Kuncheva. A framework for generating data to simulate changing environments. In *Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, AIAP'07, pages 384–389, Anaheim, CA, USA, 2007. ACTA Press.
6. Manuel Rodríguez and David Pérez. First principles model based control. In Luis Puigjaner and Antonio Espuña, editors, *European Symposium on Computer-Aided Process Engineering-15, 38th European Symposium of the Working Party on Computer Aided Process Engineering*, volume 20 of *Computer Aided Chemical Engineering*, pages 1285 – 1290. Elsevier, 2005.
7. Andreas Bathelt, N. Lawrence Ricker, and Mohieddine Jelali. Revision of the tennessee eastman process model. *IFAC-PapersOnLine*, 48(8):309 – 314, 2015.
8. Jonathan Cagan and Alice Agogino. Innovative design of mechanical structures from first principles. *IA-EDAM*, 1(3):169 – 189, 1987.
9. K. J. AAström and T. Hägglund. *PID Controllers: Theory, Design, and Tuning.* Instrument Society of America, Research Triangle Park, NC, 2 edition, 1995.
10. Janick V. Frasch, Aleksander Lodwich, Faisal Shafait, and Thomas M. Breuel. A bayes-true data generator for evaluation of supervised and unsupervised learning methods. *Pattern Recogn. Lett.*, 32(11):1523–1531, August 2011.
11. Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Generating synthetic multi-label data streams. In *ECML/PKKD 2009 Workshop on Learning from Multi-label Data (MLD'09)*, 2009.

12. Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.
13. João Gama. *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC, 1st edition, 2010.
14. Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Richard Kirkby, and Ricard Gavaldà. New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 139–148, New York, NY, USA, 2009. ACM.
15. N. Lawrence Ricker. Decentralized control of the tennessee eastman challenge process. *Journal of Process Control*, 6(4):205 – 221, 1996.

Appendix: Tables of TEP variables

Table 3. Manipulable Variables

Number	Variable Name
1	D feed flow (stream 2)
2	E feed flow (stream 3)
3	A feed flow (stream 1)
4	A and C feed flow (stream 4)
5	Compressor recycle valve
6	Purpc valve (stream 9)
7	Separator pat liquid flow (stream 10)
8	Stripper liquid product flow (stream 11)
9	Stripper steam valve
10	Reactor cooling water flow
11	cooling water flow
12	Agitator speed

Table 4. Setpoints

Variable Number	Variable Name
1	Production Set point
2	Strip Level Set point
3	Separator Set point
4	Reactor Level Set point
5	Reactor Pression Set point
6	Mole G Set point
7	A Set point
8	C Set point
9	Recycled Valve Position
10	Steam Valve Position
11	Stripper steam valve
12	Agitator Setting

Table 5. Disturbances variables

Variable Name	Variable Number	Type
1	A/C feed ratio, B composition constant (stream 4)	Step
2	B composition A/C ratio constant (stream 41)	Step
3	D feed temperature(stream i)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser c4mting water inlet temperature	Step
6	A feed loss (stream I)	Random
7	C header pressure losereduced availability (stream 4)	Random
8	A, B, C feed composition (stream 4)	Random
9	D feed temperature (stream 2)	Random
10	C feed temperature (stream 4)	Random
11	Reactor cooling water inlet temperature	Drift
12	Condenser cooling water inlet temperature	Stiction
13	Reaction kinetics	Stiction
14	Slow drift	Random
15	Reactor cooling water valve Sticking	Random
16	Condenser cooling water valve Sticking	IDV(16)
17	Unknown	Random
18	Unknown	Random
19	Unknown	Random
20	Unknown	Random

Table 6. Variables of measurements.

Number	Variable Name	Unit
1	A feed (stream 1)	kscmh
2	D feed (stream 2)	kg/h
3	E feed (stream 3)	kg/h
4	A and C feed (stream 4)	kscmh
5	Recycle flow (stream 8)	kscmh
6	feed rate (stream 6)	kscmh
7	Reactor pressure	kPa gauge
8	Reactor level	%
9	Reactor temperature	°C
10	Purge rate (stream 9)	kscmh
11	Separator temperature	°C
12	Product separator level	%
13	Separator pressure	kPa gauge
14	Separator underflow (stream 10)	m3 /h
15	Stripper level	%
16	Stripper pressure	kPa gauge
17	Stripper underflow (stream 11)	m3 /h
18	Stripper temperature	°C
19	Stripper steam dew	kg/h
20	Compressor work	kW
21	Reactor cooling water outlet temperature	°C
22	Separator cooling water outlet temperature	°C
23	of A in Reactor feed (stream 6)	mol %
24	Concentration of B in Reactor feed (stream 6)	mol %
25	Concentration of C in Reactor feed (stream 6)	mol %
26	Concentration of D in Reactor feed (stream 6)	mol %
27	Concentration of E in Reactor feed (stream 6)	mol %
28	Concentration of F in Reactor feed (stream 6)	mol %
29	Concentration of A in Purge (stream 9)	mol %
30	Concentration of B in Purge (stream 9)	mol %
31	Concentration of C in Purge (stream 9)	mol %
32	Concentration of D in Purge (stream 9)	mol %
33	Concentration of E in Purge (stream 9)	mol %
34	Concentration of F in Purge (stream 9)	mol %
35	Concentration of G in Purge (stream 9)	mol %
36	Concentration of H in Purge (stream 9)	mol %
37	Concentration of D in stripper underflow (stream 11)	mol %
38	Concentration of E in stripper underflow (stream 11)	mol %
39	Concentration of F in stripper underflow (stream 11)	mol %
40	Concentration of G in stripper underflow (stream 11)	mol %
41	Concentration of H in stripper underflow (stream 11)	mol %