17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Simultaneous fault diagnosis in chemical plants using Support Vector Machines

Ignacio Yélamos,[a1] Gerard Escudero, [b2] Moisès Graells, [a2] Luis Puigjaner[a1]

[a]Chemical Engineering Department-CEPIMA, [b]Sofware Department
[1]ETSEIB, Diagonal 647, Barcelona 08028, [2]EUETIB, Urgell 187, Barcelona 08036
[ignacio.yelamos, gerard.escudero, moises.graells, luis.puigjaner]@upc.edu
Universitat Politècnica de Catalunya (UPC) Spain

Abstract

One of the main limitations of the current plant supervisory control systems is the correct management of multiple simultaneous faults, which is crucial for supporting plant operators decision-making. In this work, Support Vector Machines (SVM) are used because of its proved efficiency dealing with multiclass problems in other technical areas. A Fault Diagnosis System has been developed implementing a multilabel approach using SVM and has been tested addressing a difficult diagnosis problem widely studied in the literature, the Tennessee Eastman process. Successful results have been obtained when diagnosing up to four simultaneous faults. These very first results are very promising since they have been achieved without any data processing or parameter tuning. Furthermore, they have been obtained just using training sets consisting of single faults, thus proving the achievement of a very powerful learning capacity.

Keywords: Support Vector Machines, Simultaneous Fault diagnosis

1. Introduction

In attention to the seriousness of accidents that may occur in chemical plants, incipient and reliable fault diagnosis is a significant requirement for preserving

public safety, as well as for enhancing the economy of the plant. Data-based diagnosis methods have faced such issues with different approaches, offering diverse solutions to the many difficulties arising in this area [1]. Yet, there still exist severe limitations on chemical plant fault diagnosis that have not been satisfactorily addressed. One of them is the management of simultaneous faults.
This problem may be addressed by creating new faults from the combination of single faults [2,3]. However, this methodology results unfeasible in actual industrial problems when dealing with large numbers of isolated faults and/or when facing combinatorial problems resulting from the presence of double, triple faults and so on. The creation of qualitative systems based on the cause effect events relationship has also been investigated [4,5] but such approaches usually generate too many spurious solutions and make the response unreliable for plant operators. A rigorous formulation of the complete diagnosis problem has been recently developed [6] focusing on non-simultaneous faults and the related monolabel classification approach and introducing the multilabel case that deals with the multiple faults diagnosis problem.
The Fault Diagnosis System (FDS) developed in this work is based on Support Vector Machines (SVM) [7], one of the most efficiently applied techniques developed by the Statistical Learning Theory. SVM allow properly dealing with the multiclass classification problem, using some binarization techniques from the Machine Learning field. They are based on the Structural Risk Minimization principle by the Statistical Learning Theory and have recently been gaining popularity in the Chemical Engineering field [8,9].

2. Problem formulation and results quantification

There are two ways of facing a classification problem such as the fault diagnosis is: the monolabel (mL) approach, which consists of classifying a set of patterns into an univocal class, and the multilabel (ML) approach, which allows assigning each input data to more than just one class.
The general fault diagnosis problem has been mostly addressed under the mL approach (i.e. [2,3]). This approach requires the generation of new artificial classes for each possible faults combination and results into the exponential drop of the classification performance (due to the new misclassification chances) and the exponential growth of the computational effort. The main advantages of the ML approach are the ability to successfully achieve the training of the diagnosis system by only using single classes, plus the ability to classify both, single and multiple class cases, the latter by decomposing the multiple class cases into the corresponding combination of single classes.
In this work, SVM under the ML approach have been used to take advantage of these features. The classification problem is next formulated and general indexes for evaluating solution performance are introduced. Consider a sample time t_s, a process measurements vector $X_s = \{x^1_s, x^2_s,\ldots, x^v_s\}$ obtained from plant data, and a given set of faults $f = 1, 2,\ldots, F$ that may be happening and

diagnosed. The faults happening and diagnosed for each t_s can be characterized by two different matrices (H and D respectively) containing samples in rows and possible faults occurring in columns. Both matrices are next shown:

$$H = \begin{pmatrix} h_{11} h_{12} \dots \dots h_{1F} \\ h_{21} h_{22} \dots \dots h_{2F} \\ \vdots \cdot \vdots \cdot \vdots \cdot \vdots \cdot \vdots \\ h_{S1} h_{S2} \dots \dots h_{SF} \end{pmatrix} \qquad D = \begin{pmatrix} d_{11} d_{12} \dots \dots d_{1F} \\ d_{21} d_{22} \dots \dots d_{2F} \\ \vdots \cdot \vdots \cdot \vdots \cdot \vdots \cdot \vdots \\ d_{S1} d_{S2} \dots \dots d_{SF} \end{pmatrix} \tag{1}$$

being $h_{sf}, d_{sf} \in \{0,1\}, \forall i, j : 1 \le i \le S, 1 \le j \le F$

where $h_{sf} = 1$ and $d_{sf} = 1$ when fault *f* is happening or diagnosed respectively at sample time *s* and both are equal 0 otherwise.
The general ML classification problem seeks for matching both matrices when any distribution of binary values is allowed, including null rows for the normal case (no faults) and rows including several non-null values (simultaneous faults). Two different measurements for the matching degree are given in the machine learning literature, precision and recall:

$$\text{Prec}(f) = \Pr[h_{ij} = d_{ij} \,|\, d_{if} = 1 : 1 \le i \le S] \tag{2}$$

where precision for fault *f* has been defined as the conditioned probability of happening fault *f* conditioned to fault *f* has been diagnosed, and

$$\text{Rec}(f) = \Pr[h_{ij} = d_{ij} \,|\, h_{if} = 1 : 1 \le i \le S] \tag{3}$$

recall for fault *f*, which is defined as the conditioned probability of the FDS of predicting fault *f* conditioned to the sample is fault *f*. The F1 index is a measure that combines both, precision and recall, and is used along this work as the measure for the general performance of a FDS. It is evaluated as:

$$\text{F1} = (2 \times \text{Prec} \times \text{Rec}) / (\text{Prec} + \text{Rec}) \tag{4}$$

3. Methodology

The FDS is tested using the Tennessee Eastman (TE) process [10], which is a challenging and well-known benchmark allowing comparable results to the fault diagnosis community. The FDS is trained and tested with the original 20 faults and without any additional combination of faults. No preliminary filtering or

treatment is applied to the raw plant data. No tuning is applied to the SVM (default soft margin and lineal kernel). Hence, results presented do not show those optimal reachable, but outline the first crucial points from the multiple fault diagnosis analysis. Appropriate sizes for training and testing data sets were determined from a preliminary study on the F1 response (learning curve). Hence, 351 samples for each single fault were included in the training sets (351x20) whereas 1045 were used for building the testing sets (files are available at http://webon.euetib.upc.es/ciao/datos/escape2007.html).

4. Results and discussions

The performances of the mL and ML approaches are first tested and compared. Table 1 shows no significant differences on the values of the F1 index. It also shows that, both approaches, despite the ML advantage for managing simultaneous faults, provide a similar classification capability, even for those cases revealed more difficult (faults 3,5,9, etc.). From this preliminary study, faults 1,2,6,7,17 and 18 are selected for building up the new test data sets including simultaneous faults for validating the ML approach presented.

Table 1. Single-fault diagnosis. F1 index for the mono-label (mL) multi-label (ML) approaches.

	1	2	3	4	5	6	7	8	9	10	
mL	**98.1**	**98.0**	0.0	88.9	0.0	**100**	**99.8**	25.9	0.0	0.0	
ML	**96.1**	**98.0**	0.0	89.1	0.0	**100**	**99.5**	24.8	0.0	0.0	
	11	12	13	14	15	16	17	18	19	20	Average
mL	0.0	9.2	26.3	0.0	0.0	11.8	**97.4**	**89.9**	0.0	89.3	41.7
ML	0.0	0.0	22.1	0.0	0.0	11.5	**97.5**	**90.5**	0.0	84.9	40.7

For comparative purposes, the results for single-fault diagnosis (Table 1) are obtained without the different parameter tuning that allows improving separately the performance of each of these approaches. These worst case results show the parallel shortcomings. Although incomparable, for the manual ad-hoc parameter tuning, results up to an average F1 = 61% may be obtained for the mL case [6].

Combinations of two, three and up to four simultaneous faults were simulated and results when diagnosing them are shown in Table 2. The first three tested pairs are successfully diagnosed as both single faults are identified when happening simultaneously with very high performance (F1). Very good performance is also obtained when diagnosing three and four simultaneous faults, although it is not surprising to get lower F1 values. When faults 2, 7 and 17 are simulated at the same time, SVM is able to identify correctly the isolated faults that the system has been trained with. In the four simultaneous faults case, the F1 index just suffers a significant decrease for the 18^{th} fault, while faults 2,7

and 17 are being correctly diagnosed by the system. Finally, Table 3 shows the performance details for the case of four simultaneous faults, including precision, recall, and the individual class assignment. One of the most significant points to be highlighted from these detailed results is the very low misdiagnosis rate (b in Table 3) of the FDS, which reveals great possibilities for the future. Notice that precision, recall and F1 index are shown as percentages in Tables 1,2 and 3.

Table 2. Multiple-fault diagnosis. F1 index for multi-label (ML) approach.

	1	2	6	7	17	18
2 Simultaneous faults (1,2)	98.7	97.8	-	-	-	-
2 Simultaneous faults (6,7)	-	-	98.8	99.9	-	-
2 Simultaneous faults (17,18)	-	-	-	-	95.4	79.1
3 Simultaneous faults (2,7,17)	-	94.9	-	99.5	95.5	-
4 Simultaneous faults (2,7,17, 18)	-	94.5	-	71.3	93.5	14.6

Table 3. Details for the classification of a case consisting of four simultaneous faults.

	1	2	3	4	5	6	7	8	9	10
a	0	**898**	0	0	0	0	**556**	0	0	0
b	0	**0**	0	0	10	21	**0**	1	206	100
c	0	**105**	0	0	0	0	**447**	0	0	0
d	1045	**42**	1045	1045	1035	1024	**42**	1044	839	945
Prec	-	**100**	-	-	0.0	0.0	**100**	0.0	0.0	0.0
Rec	-	**89.5**	-	-	-	-	**55.4**	-	-	-
F1	-	**94.5**	-	-	-	-	**71.3**	-	-	-
	11	12	13	14	15	16	17	18	19	20
a	0	0	0	0	0	0	**881**	**79**	0	0
b	0	415	1	56	15	83	**0**	**0**	0	0
c	0	0	0	0	0	0	**122**	**924**	0	0
d	1045	630	1044	989	1030	962	**42**	**24**	1045	1045
Prec	-	0.0	0.0	0.0	0.0	0.0	**100**	**100**	-	-
Rec	-	-	-	-	-	-	**87.8**	**7.9**	-	-
F1	-	-	-	-	-	-	**93.5**	**14.6**	-	-

(a) Samples happened and diagnosed (b) Samples diagnosed but not happened (c) Samples happened and not diagnosed (d) Samples not happened and not diagnosed.

5. Conclusions

Fault diagnosis in chemical plants has been usually addressed as a monolabel (mL) classification problem. This work adopts the multilabel (ML) approach for which a general formulation is provided. It has been implemented using SVM and a FDS has been developed for addressing the actual problem of diagnosing faults occurring simultaneously. Thus, the technique presented does not need creating new faults for modeling a group of simultaneous faults, which results in allowing training with only the original faults, avoiding the artificial increase of the computational burden, and keeping the classification performance.
The system has been validated using a well-established benchmark and well-established performance indexes from the machine-learning field. Examples have been prepared for addressing up to four simultaneous faults and successful results have been obtained. Moreover, these very first results are especially promising since they have been obtained without attending some key aspects: data processing, parameter tuning, and training with simultaneous faults.

Acknowledgements

Support from Spain “Ministerio de Educación y Ciencia” (FPI program) and the E.U. (MRTN-14CT-2004-512233 and RFCS-PR-03013) is fully appreciated.

References

1. Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S.N. (2003). A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. Comput. Chem. Engng., 27, 293.
2. Raich A., & Çinar A. (1996). Statistical process monitoring and disturbance diagnosis in multivariable continuous processes. Aiche J., 42, 995.
3. Raich, A., & Çinar, A. (1997). Diagnosis of process disturbances by statistical distance and angle measures. Comput. Chem. Engng.., 21, 6, 661.
4. Vedam, H., & Venkatasubramanian, V. (1998). PCA-SDG based process monitoring and fault diagnosis. Control Engineering Practice, 7, 903-917.
5. Maurya, M.R., Rengaswamy, R., & Venkatasubramaian, V. (2004). Application of signed digraphs based analysis for fault diagnosis of chemical flowsheets. Engng. Applications of Artificial Intelligence, 17, 501.
6. Yélamos, I., Escudero, G., Graells, M., Puigjaner, L. Performance assesment of novel fault diagnosis system based on support vector machines. Comput. Chem. Engng. (submitted).
7. Boser, B., Guyon, I., & Vapnik, V. (1992). Atraining Algorithm for Optimal Margin Classifiers. In Proceedings of the Workshop on Computational Learning Theory, COLT.
8. Kulkarni, A., Jayaraman, V.K., Kulkarni, B.D. (2005). Knowledge incorporated support vector machines to detect faults in TE Process. Comput. Chem. Engng., 29, 2128.
9. Chiang, L.H., Kotancheck, M.E., & Kordon, A.K. (2003). Fault diagnosis based on Fisher discriminant analysis and support vector machines. Comput. Chem. Engng., 28, 1389.
10. Downs, J., & E. Vogel, E. (1992). A plant-wide industrial process control problem. Comput. Chem. Engng., 17, 3, 245.