APPLICATION OF A KDD METODOLOGY TO A REAL INDUSTRIAL PROCESS

VICTORIA PACHÓN. ESCUELA POLITÉCNICA SUPERIOR, UNIVERSIDAD DE HUELVA. SPAIN. VPACHON@UHU.ES
JACINTO MATA. ESCUELA POLITÉCNICA SUPERIOR, UNIVERSIDAD DE HUELVA. SPAIN. MATA@UHU.ES
FRANCISCO ROCHE. ESCUELA POLITÉCNICA SUPERIOR, UNIVERSIDAD DE HUELVA. SPAIN. ROCHE@UHU.ES
JOSE C. RIQUELME. ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS. UNIVERSIDAD DE SEVILLA. SPAIN. RIQUELME@LSI.US.ES
JOSE MARÍA TEJERA. ATLANTIC COPPER SA.

ABSTRACT

By means of an acid plant it is possible to transform sulphuric dioxide (SO2) into sulphuric acid, using for this a set of chemical and physical processes. In the process of smelting copper mineral a large amount of SO2 is produced. This compound would be highly pollutant if it was emitted to the atmosphere. However, there are certain situations in the process of smelting copper mineral, in which SO2 escape to the atmosphere. This would be avoidable if we exactly knew under which circumstances this problem is produced. In this paper we present a practical application of KDD techniques to the chemical industry. With this technique we obtain rules that make possible the definition of proceedings that should help to optimise the functioning of the sulphuric acid production system. By means of the obtained results we show the viability of using automatic classifiers to improve a productive process, with an increase of the production and a decrease of the environmental pollution

KEYWORDS: Data Mining, KDD, Industrial Applications, Artificial Intelligence, Classification Rule, Data Preparation, Data Preprocessing, Automatic Classifier, Data Selection, Data Cleaning, Coding.

1. INTRODUCTION

In the process of smelting copper mineral a large amount of sulphuric dioxide (SO_2) is produced[1]. This compound would be highly pollutant if it was emitted to the atmosphere. By means of an acid plant it is possible to transform it into sulphuric acid, using for this a set of chemical and physical processes. In this way we dispose of a marketable product and, at the same time, the environment is protected. However, there are certain situations in which the gases escape to the atmosphere, creating pollutant situations. This would be avoidable if we exactly knew under which circumstances this problem is produced. In order to study the viability of the improvement of the control of the sulphuric acid production system, it is essential to know how the system works, not only from the point of view of its use but internally. Since at the present time we do not have a detailed model of the system, it would be necessary to give the first steps to obtain such model. Basically, an sulphuric acid production system consists in a gas flux system that contains sulphur in a higher or lesser quantity. By subjecting the gases to the appropriate processes, we can use this sulphur to produce sulphuric acid instead of releasing it to the atmosphere. In our case of study there are sensors that take measures of diferents properties or parameters of the industrial plant. This measures are stored in files, therefore we have a large database. The KDD (Knowledge Discovery in Databases) [1] process try to know why the system doesn't work propery. In this paper we present a practical application of KDD techniques to the chemical industry. By means of

[1] The research has been realised in collaboration wiht Atlantic Copper S.A.

the obtained results we show the viability of using automatic classifiers to improve a productive process, with an increase of the production and a decrease of the environmental pollution. In this research, we have specifically used classification |2| as Data Mining |3|. technique. With this technique we obtain rules that make possible the definition of proceedings that should help to optimise the functioning of the system. With the obtaining of the goals the company will have: an optimisation of the functioning of the sulphuric acid production system, a greater knowledge of the system that will be useful to help in a later taking of decisions and a reduction of the production costs at the same time that a greater protection of the environment. Using data from a real industrial process, we have developed a KDD proccess. We have obtained a set of classification rules that approach the system behaviour. We have two stages in our KDD process. In the first stage, we worked with two classes, and a high number of parameters. In the second stage, we worked with the most significative parameters of each subsystem, and 5 classes only. In this paper, we describe all the stage of the process and the most representative rules. The rules obtained from the KDD process have helped the human expert to know the range of values in which the system works properly. With this analysis we probe the viability of using authomatic classificators in a productive process, with an increase of the production and a decrease of the contamination.

2. SULPHURIC ACID PRODUCTION SYSTEM

In order to study the viability of the improvement of the control of the sulphuric acid production system, it is essential to know how the system works, not only from the point of view of its use but internally. Since at the present time we do not have a detailed model of the system, it would be necessary to give the first steps to obtain such model. In Figure 1 we show a diagram of the aforementioned system. In the system there are producers and consumers of gases. The process consists in canalising in a suitable way the gases generated by the producers towards the consumers, in such a way that there is no production surplus at the same time that the necessities of the consumers are cared for. The produced gases are conveyed to the Blending Chamber (BC), from which they are distributed to the different consumers. The gases are conveyed from a point to another by blowers. Due to the complexity of the system and to the great number of parameters and situations that are continuously given, sometimes the system is not stable during a certain time (that is to say, there is pressure in the BC). Although at present time all the appropriate measures to avoid this have been taken, a KDD process would help to know which rank of values of the variables that operate in the system is the most suitable to maintain a stability in the BC and to discover, which are the reasons that provoke instability to the draught in the BC.

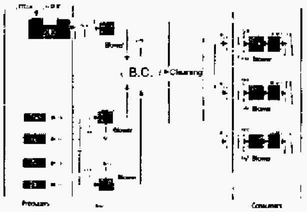

Figure 1. Sulphuric acid production system.

3. ACID FACTORY DATA

The results are obtained from the acid factory data. In our case of study there are sensors that take measures of different properties or parameters of the industrial plant. This measures are stored in files, therefore we have a large database. The KDD process tries to know why the system doesn't work properly.

Each database record consists of the measure datatime, and the numeric value of 52 diferent sensors. We have one different record each 2 seconds. We add a new column, called class, to indicate the system behaviour. We are working with a real system, so a change in one of the parameters in the instant of time t can be reflected in the class in the instant of time t+ t. t is a difficult value to know because it is the sum of the different phisical delays of the system. So, the value of the class in a record in a determinate instant of time could come about a change in the parameters of the previous instants of time. Changes will be recorded in previous database records. The results have been obtained without consideration of delay.

The tool used to get these results is C4.5 [4]. It is one of the most used in classification. We have chosen this tool because the resultant rules can be easily interpreted by the industrial chemistry experts.

4. RULE REPRESENTATION

A rule has two parts, the antecedent and the consequent. The antecedent is the group of conditions that are established on the parameter values and the consequent, in our case, is one of the classes assigned to the file. In Figure 2 it is shown an example of the decision rules obtained during the process.

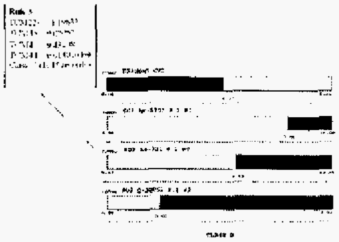

Figure 2. Example and Graphic representacion of decision rule.

The expression "Rule 3" indicates only one label assigned to the rule. The rule in Figure 2 express that if TCM22 <= 1.19833 and TCM35 > 9.02952 and TCM40 > 9.43208 and TCM44 > 0.0140164 then, the class is Class 0. If all the antecedent conditions are true, the system would not be working properly since a Class 0 is obtained as consequent. Moreover, parenthetically, it is shown the number of records from the training file that fulfil that rule. This value gives us a sign of the importance of the rule. If the rule is true in a high percentage of the records, this indicates that we are in front of a frequent situation for the system. If a record fulfils the antecedent of a rule, but not the consequent, that is, fulfils the parameter restrictions but is of a different class, it is considered as a learning mistake. To have a more intuitive interpretation of the rule for the industrial expert, a graph where the maximum and minimum values that the parameter can take (grey color) and the specified rule value (black color) are shown, has been included. When TCM22 takes low and medium values (below 1,19) and TCM35 does not take high values, when TCM40 takes medium and high values and TCM44 takes medium and high values, the system operation is not optimum (Class 0).

5. GETTING RESULTS WITH A KDD PROCESS

The KDD process is a cyclic process with six stages (Data selection, Cleaning, Enrichment, Coding, Data Mining and Reportint). The results obtained in each period change as the KDD process stages develop. In order to obtain enough accuracy in the rules extracted from the data, different preprocessing stage prototypes have been tested and, therefore, different Data Mining process results have been obtained.

5.1. KDD first stage

In the Data Selection stage, study of the correlation between columns was realized. One column was selected from those that were highly correlated (correlation higher than 0.8), eliminating the others since we can consider that they provide the same information to the KDD process.

In the Cleaning stage, as the data come from a real industrial system, we have to check some problems that could appear, and in that case we have to solve them:

- It is possible that some parts of the system are not working in the moment of the data acquisition (for example, because of a failure), so errors in the sample taking can be produced. These errors are shown as out of range values for that parameter in the whole data or as a constant value. Each possible error was detected and eliminated from the respective parameter study.
- During some specified periods of time, because of unknown reasons, the data acquisition can be incorrect (of all the parameters or only some of them). We have to look for each row of the file corresponding to incorrect data acquisition and eliminate from the study the respective period of time information. After the data selection and cleaning stages, 36 columns were selected from the starting 57.

In the Coding stage, a new column named label or class with a Class 1 or Class 0 value was added. The value of this column is set by the value of the parameter named TCM12 (Real value of pressuere in BC). What we have to study is when the draught exists in the BC. Talking to the expert, it was established that there is pressure in BC when the TCM12 value is in the interval [-2.7, -2.3]. In this case, we can assume that the system is working in an optimum way. Then, we have a Class 1 record. On the other hand, it is said that the system is not working in an optimum way when the TCM12 value is not between the interval [-2.7, -2.3]. In this case, we have a Class 0 record. The assignment of these classes is necessary to create the training file which the Data Mining algorithm works with, since it is a supervised learning process.

So, before preprocessing we have a database with 43201 records and 57 columns. Before preprocessing we obtain a database with 43201 records and 36 columns. This database has 9804 records of Class 1 and 33397 records of Class 0. When the data preparation step is finished, the knowledge extracted from the Data Mining algorithm is represented as decision rules.

5.2. Results and conclusions of the first stage.

In this first stage, only rules whith apparition frequency is 1% or higher have been taken into account. The file has 43201 records, so, only rules that are true in more than 430 records will be shown. In further studies, rules with a lower apparition frequency can be taken into account, if neccesary. The number of classification errors compared with the number of instances gives us an idea of how "reliable" a rule is.

Having analized the rules, it is found that the parameter TCM37 (Factory 3 blower speed) is the most determining. High values or very high values of TCM37 imply pressure in BC.This is the way this fact might be explained: TCM37 refers to the blower speed which extracts the gas from BC to the acid factory. Besides, this blower is the most powerful one among the three blowers used to extract the gas from the BC. When the blower reaches a very high speed for a long time, it is not possible to extract more gas because it is working at its maximum. That produces an accumulation of gas in the BC which creates pressure, and therefore, Class 0 records.

The extracted rules involve parameters from different subsystems in its predecessor, which makes its explanation difficult to an expert. Then, it was necessary a higher decrease in the number of parameters if we wanted that the ones which took part were the most significant ones. This idea would make the rules easier to understand. To reduce the system as much as possible, it was thought to work with the TCM to obtain a productive system equivalent to the 3 productive systems and another one with the consumers. This way the system would only consist of a productive system, the blending chamber and a consuming system. The Class 1 rules that can be deduced from the data, gather a very low records pecentage, and therefore, not very much significant. In order to know if the system is working properly, more Class 1 data are needed, or at least, a significant number of them compared with the Class 0 ones. Besides, the study could be improved if the TCM12 group of values could be separated in more classes, not only Class 0 and 1. We could have 3, 4 or 5 m ore classes to express the different states of the system.

From the results obtained in the First Stage, we can conclude that, although these results were satisfactory, more useful rules could be found if the number of parameters was reduced and the number of classes was increased. With these changes, a second stage of the KDD process could be developed. In this stage, the preprocessing step could be done again to prepare the data for the new circumstances.

5.3. KDD Second Stage

In this stage, the study is a consequence of the possible improvements that were found after the first stage. The differences between them are:

- In the first stage, the value of the class depended on a calculus on the TCM12 value. However, Class 0 was detected corresponding with a correct behaviour of the system was set by the expert to solve some situations. To solve this problem, it is neccesary to think of another way to calculate the class. The new classes are obtained calculating the error between the draught set point value in the BC (measured by TCM42) and the real draught value in the BC measured by TCM12. We obtain: Pressure, High pressure, Draught and High Draught.
- To reduce the number of parameters, we must take into account the blowers and valves that take the gas to BC and the blowers and valves that extract gas from BC. Table 1 shows the range of values corresponding to each class and the amount of instances of each class. We have selected only those rules that are true in more than 1% of the instances of its class. The fourth column expresses the value that indicates if we can consider that a rule is true in a significative number of cases.

Range of values (TCM42-TCM12)	Class	Number of instances of each class	1%
(-0,5..0,5)	OK	29843	298
(0,5..1,5)	Pressure	7393	739
(1,5...+inf)	High pressure	1295	13
(-1,5..-0,5)	Draught	3825	38
(-inf...-1,5)	High Draught	845	8

Table 1. Number of instances of each class.

5.4. Results and conclusions of the second stage

Developing the second stage, we can get rules that describe the optimal behaviour (class OK). However, these rules are easier to interprete by the human expert. We obtain the same conclusion than First Stage about TCM37. High or very high values of TCM37 blower imply pressure in BC.

In addition, high or very high values of other blowers (TCM30 and TCM46) imply pressure in BC. Medium or low values of these blowers imply Class OK. Figure 3 shows the some of the most representative rules obtained in the second stage. We don't have a lot of instances of classes: High pressure, Draught and High Draught , so the rules with these classes are not categorical enough.

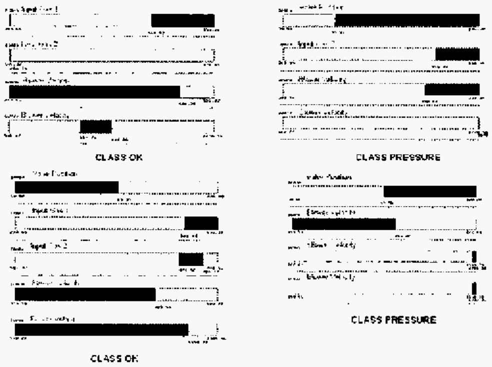

Figure 3 . Graphic representation of Second Stage Rules

6. CONCLUSIONS

Using data from a real industrial process, we have developed a KDD proccess. We have obtained a set of classification rules that approach the system behaviour. We have two stages in our KDD process. In the first stage, we worked with two classes, and a high number of parameters. In the second stage, we worked with the most significative parameters of each subsystem, and 5 classes only. In this paper, we describe all the stages of the process and the most representative rules. The rules obtained from the KDD process have helped the human expert to know the range of values in which the system works properly. With this analysis we probe the viability of using authomatic classifiers in a productive process, with an increase of the production and a decrease of the contamination.Future works will study the errors due to the system delays and develop a method of obtaining rules with delay.

7. REFERENCES

[1]U.M. Fayyad, G. Piatetsky-Shapiro y P. Smyth. *"From Data Mining to Knowledge Discovery in Databases"*, Ai Magazine. 1996. Pages 37-54.

[2]M.S. Chen, J. Han y P.S. Yu. "Data Mining: An Overview from Database Perspective". IEEE Transactions on knowledge and Data Engineering, 1996. Pages 866-883.

[3]Perner, P., 2001. "Tutorial part on data mining". In: Perner, P., Ahlemeyer-Stubbe, A. (Eds.), Proceedings of the 1st Industrial Conference on Data Mining, ICDM 2001, IBal Report, Leipzig. 2001

[4]J.R. Quinlan. *C4.5*," Programs for Machine Learning". Morgan Kaufmann Publishers. San Mateo, California, 1993

[5]Santen, A., Koot, G.L.M. y Zullo, L.C., "Statistical data analysis of a chemical plant". Computers and Chemical Engineering 21, S1123–S1129,1997.

[6]Alex Berson, Stephen Smith y Kurt Thearling, "Building Data Mining Applications for CRM". McGraw-Hill. New York, 2000.