IMPROVING PIPELINE RISK MODELS BY USING DATA MINING TECHNIQUES

María Fernanda D'Atri [1], Darío Rodriguez [2], Ramón García-Martínez [2,3]

1. MetroGAS S.A. Argentina.
2. Área Ingeniería del Software. Licenciatura en Sistemas. Departamento Desarrollo Productivo y Tecnológico. Uiversidad Nacional de Lanús.
3. Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires.

Keywords: 1. pipeline risk management; 2. risk assessments models; 3. data mining; 4. authomatic risk detection

1 Introduction/Background

a. The scope

The consequences of the risks associated with accidents and emergencies that may occur in a gas pipeline are especially serious because, in general, they influence on population and environment. It is therefore of vital importance to develop an effective risk management, from assessment to mitigation, which require a reliable analysis and prediction model.

Risk analysis is a difficult task, mainly due to the nature of the data being handled. Potentially dangerous events are highly unlikely and, in turn, are due to many causes usually related, so the mere statistical analysis of historical data may not be effective.

This paper presents a different approach to traditional risk analysis: the basic idea is to introduce advanced techniques of exploiting information based on intelligent systems for optimizing risk models currently used.

The motivation behind the use of new techniques for risk analysis is given by the need to answer some questions that arise from the review of existing models

- If you have several risks at once, their impacts are additive?
- Can we integrate qualitative and quantitative methods to see the issue from a joint perspective?
- How important is the qualitative information about a process and how it fits in a quantitative analysis?
- How can you use historical data to predict future behavior?

b. Asset integrity

Risk management is the set of actions taken to control risk [1]. It implies the process for assessing this risk through execution of an action plan to control and reduce future risks. In general, this process is known as Asset Integrity Managing Process.

One of the basic rules to remember is to ensure that the process of managing risk is a continuous process that includes gathering information on a regular basis, assessment, analysis, prediction and mitigation of risks.

Therefore, risk assessment is the starting point of a program for monitoring the assets integrity. Different assessments can vary in scope and complexity and can use different methods or techniques. However, the ultimate goal of any assessment is to identify the most significant risks, to develop an effective plan and prioritized prevention, detection and mitigation of those risks.

c. Risk assessment models

Risk is governed by the probability of a risky event and the magnitude of loss, called in this context failure and consequence, respectively.

Risk assessment is the most critical and most difficult stage to implement because, in short, no one may be able to determine when or where a failure will occur. However, we can estimate which the most probable failure is, the places most affected are, as well as, the probability of occurrence and severity of the

consequences are. A risk assessment model should enable us to determine the value of risk in any sector of the gas pipeline, based on all the factors that influence in the failures and consequences.

All risk models, even the simplest ones, use statistical methods. That is, from a model that bases its decisions on the experience of experts on the subject to more rigorous mathematical models based on the failures history of the system.

Matrix model

One of the simplest risk assessments models is a decision matrix. This model classifies each pipeline risk according to the likelihood and the potential consequences of an event by a simple scale, such as high, medium or low. Each threat is assigned to a cell of the matrix based on its perceived likelihood and consequence. While this approach cannot consider all factors and their relationships, it does help to focus on the most dangerous threats. [2].

In general, in this model, the information to classify risks according to the cells of the matrix is based on knowledge and experience of experts.

This model is an evaluation of more qualitative than quantitative risk, however, is very useful as a first approach to classify the problem and separate it into two basic problems underlying the likelihood and consequence, that is, this representation can be useful for determine whether to attack the probability of failure or the consequence of a failure event.

Probabilistic model

The probabilistic model is the most rigorous and complex, known as PRA (Probabilistic Risk Assessment) or QRA (Quantitative Risk Assessment). This technique is most used in the nuclear, chemical and petrochemical, as it is an appropriate model for analyzing the frequency of occurrence of unlikely events [3].

This model uses the techniques of the events tree and / or failures tree. The mechanism is as follows: starting with an event that causes a glitch in the system, then the tree goes forward in search of all possible consequences and backward, in search of all possible causes or initiators of the events fails. All possible paths are quantified based on the probability of each branch of the tree and identifies the initiator of the event possible damage.

Indexed model

This model is the most used at present. In this approach, numerical values are assigned to each of the conditions and activities of the pipeline that can contribute to the risk, either increasing or reducing it. These are system variables and each variable is assigned a weight according to their relative influence on the risk assessment. The relative weights are based on statistics, if you have the necessary data or the view of experts, if not have them.

The main advantage of this technique is that you can include a much more comprehensive information on other models. Other advantages include the following: provides immediate answers, an analysis inexpensive and identifies opportunities for risk mitigation. On the other hand, one of the main criticisms of this model is the possible subjectivity of the relative weights given to variables.

Risk analysis model for a distribution company's high pressure system

We start from a risk assessment model for a gas pipeline standard featuring both models, indexed or probabilistic, following the guidelines of *ASME B31.8* standard [1]. It is a model of relative risk based on a qualitative analysis, which compares a section of pipe with the other. Risk is calculated on a section as a product of the probability by the consequence of failure.

In this approach, numerical values are assigned to each of the conditions and activities of the pipeline that can contribute to risk, either by increasing or reducing it, each one with a relative weight according to their influence on risk assessment. The advantages of this technique are that it includes a much more comprehensive information than other models, provides immediate answers and it requires an inexpensive analysis. However, one of the main criticisms is the possible subjectivity of the relative weights given to variables.

The threats and the consequences are the indexes of the model. The probability of failure is calculated as an algebraic sum of the threat and consequence of failure as the algebraic sum of the

consequences. The relative weights of each variable in the algebraic sums represent the relative importance of each in contributing to total risk.

$$R = P \; x \; C \qquad P = \sum_{i=1}^{5} a_i A_i \qquad C = \sum_{i=1}^{5} b_i B_i$$

R: Risk

P: Probability of failure

C: Consequence of failure

A_i: Threat

a_i: Relative weight of threat i

B_i: Consequence

b_i: Relative weight of consequence i

Each index in turn depends on a number of conditions, the variables of the model, that influence the threat or consequence. For example, the threat corrosion is dependent on soil type, the type of pipe coating, the efficiency of cathode protection, etc. The model is completed by assigning values to each of these variables depending on the properties of each segment of the pipe.

The Figure 1 shows the 5 Threats that contribute to the Probability of Failure and the 5 Effects contributing to the Consequence of Failure.

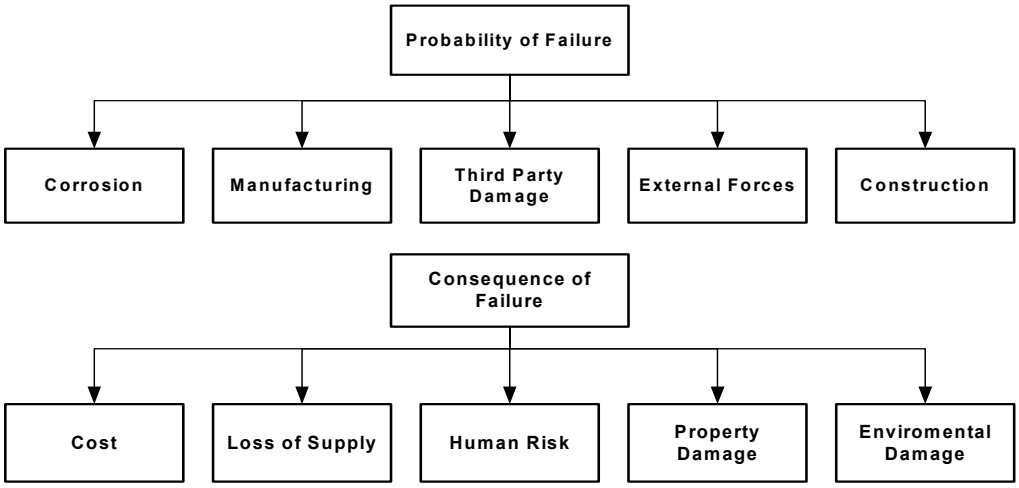

Fig. 1. Scheme of Threats and Consequences of the Model

This model relies heavily on the measurement of a number of important variables in each segment of pipe and the weightings of each variable and each of the indexes, but even with the influence they have on the outcome of risk calculation results, the weighting factors are based largely on the subjective knowledge of experts.

2 Objectives of the paper

To develop a risk analysis model for a gas pipeline based on artificial intelligent systems tools, starting from failures history, to be able to predict the behavior of the pipeline and to eliminate the subjective aspects of relative weight of variables involved.

It is expected that the development of the novel techniques taken from the intelligent systems results in the improvement of risk assessments models in order to take wise decisions regarding risk mitigation.

Therefore, the idea is to increase the accuracy of the model trough the methods described below, and get the knowledge needed to develop a predictive model.

3 Development / Methods

It has been observed that the risk analysis models are essentially the same for all activities or processes, implementing them or the significance of its primary variables is what differentiates the activity to be monitored.

In the field of industrial processes, there is a trend towards the use of matrix methods and decision trees for risk assessment, based heavily on the experience of experts. These methods don't allow exploiting into the maximum potential the knowledge of failures history. Therefore, we propose to investigate the use of these techniques to perform risk analysis in a gas pipeline.

a. Data mining / Decision Trees

Data mining is the set of techniques and tools applied to the non-trivial process of extracting and presenting/displaying implicit knowledge, previously unknown, potentially useful and humanly comprehensible, from large data sets, with object to predict automated form tendencies and behaviors; and to describe automated form models previously unknown. The term intelligent data mining is the application of automatic learning methods to discover and enumerate present patterns in the data. For these, a great number of data analysis methods were developed, based on the statistic. In the time in which the amount of information stored in the databases was increased, these methods began to face problems of efficiency and scalability. This is where the concept of data mining appears. One of the differences between a traditional statistic based analysis of data and the data mining is that the first requires that the hypotheses are already constructed and validated against the data, whereas the second supposes that the patterns and the associated theses are automatically extracted from the data.

The tasks of the data mining can be classified in two categories: descriptive data mining and predictive data mining; one of the most common techniques of descriptive data mining are the decision trees (TDIDT), the production rules and self organized maps. On the other hand, an important aspect in the inductive learning is to obtain a model that represents the knowledge domain that is accessible for the user, it is particularly important to obtain the dependency data between the variables involved in the phenomenon; in the systems that need to predict the behavior of some unknown variables based on certain known variables, a representation of the knowledge that is able to capture this information on the dependencies between the variables is the Bayesian networks.

Carbonell, Michalski and Mitchell [4] identify three principal dimensions along which machine learning systems can be classified: the underlying learning strategies used, the representation of knowledge acquired by the system, and the application domain of the system. The product of learning is a piece of procedural knowledge. The members of TDIDT family [5] are sharply characterized by their representation of acquired knowledge as decision trees. This is a relatively simple knowledge formalism that lacks the expressive power of semantic networks or other first-order representations. As a consequence of this simplicity, the learning methodologies used in the TDIDT family are considerably less complex than those employed in systems that can express the results of their learning in a more powerful language. Nevertheless, it is still possible to generate knowledge in the form of decision trees that is capable of solving difficult problems of practical significance.

The underlying strategy is non-incremental learning from examples. The systems are presented with a set of cases relevant to a classification task and develop a decision tree from the top down, guided by frequency information in the examples but not by the particular order in which the examples are given. The example objects from which a classification rule is developed are known only through their values of a set of properties or attributes, and the decision trees in turn are expressed in terms of these attributes. The examples themselves can be assembled in two ways. They might come from an existing database that forms a history of observation.

The basis of the induction task [6] is a universe of objects that are described in terms of a collection of attributes. Each attribute measures some important feature of an object and will be limited here to taking a (usually small) set of discrete, mutually exclusive values. Each object in the universe belongs to one of a set of mutually exclusive classes. The induction task is to develop a classification rule that can determine the class of any object from its values of the attributes [7];[8]. The immediate question is whether or not the attributes provide sufficient information to do this. In particular, if the training set contains two objects that have identical values for each attribute and yet belong to different classes, it is clearly impossible to differentiate between these objects with reference only to the given attributes. In such a case attributes will be termed inadequate for the training set and hence for the induction task.

As mentioned above, a classification rule will be expressed as a decision tree [9]; [10]; [11]. Leaves of a decision tree are class names, other nodes represent attribute-based tests with a branch for each possible outcome. In order to classify an object, it starts at the root of the tree, evaluate the test, and take the

branch appropriate to the outcome. The process continues until a leaf is encountered, at which time the object is asserted to belong to the class named by the leaf. Only a subset of the attributes may be encountered on a particular path from the root of the decision tree to a leaf; in this case, only the outlook attribute is tested before determining the class. If the attributes are adequate, it is always possible to construct a decision tree that correctly classifies each object in the training set, and usually there are many such correct decision trees. The essence of induction is to move beyond the training set, to construct a decision tree that correctly classifies not only objects from the training set but other (unseen) objects as well.

In order to do this, the decision tree must capture some meaningful relationship between an object's class and its values of the attributes. Given a choice between two decision trees, each of which is correct over the training set, it seems sensible to prefer the simpler one on the grounds that it is more likely to capture structure inherent in the problem. The simpler tree would therefore be expected to classify correctly more objects outside the training set.

b. Data Mining Process

The problem in which knowledge discovery was focused was identifying knowledge pieces (rules) associated with the risk involved in an incident of any kind produced in the gas network.

The scheme of knowledge discovery process is presented in Figure 2 [12]. A data query is applied to Pipeline Example Incidents Data base and a view with the identified class attribute and related ones is built.

This resulting database is used in the TDIDT based knowledge discovery process to obtain rules that characterized each class associated with the different values of the identified class attribute.

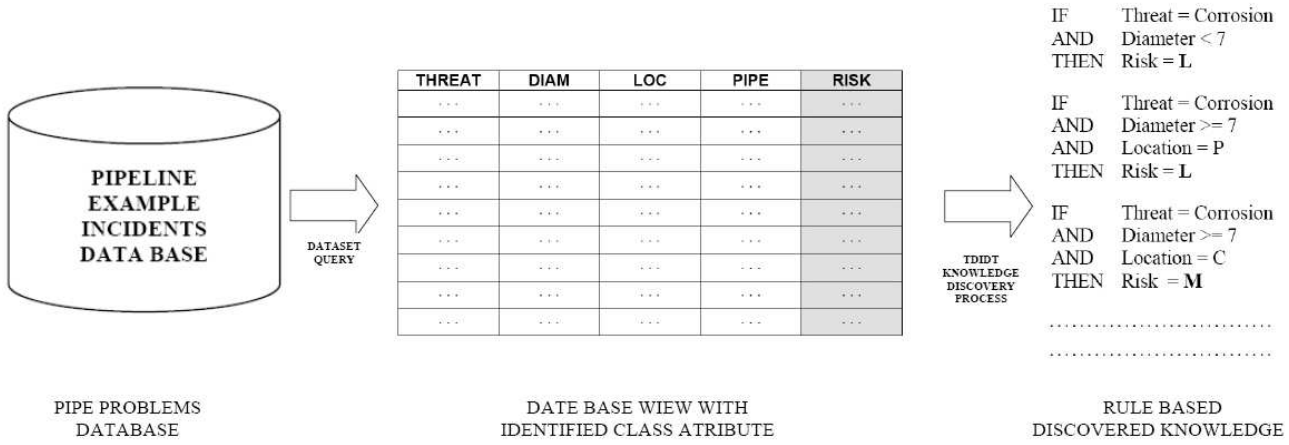

Fig. 2. Scheme of knowledge discovery process.

4 Results

To a first approximation to the problem was to start with a simplified model, in which only considered the most significant threats to the network MetroGas:

- Third party damage (THPD)
- Corrosion (CORR)

The process undertaken to discover rules of behavior from the data set available, outlined in Figure 2, can be summarized in the following phases:

1. Identification of sources of information: The source of information used is the historical record of repairs performed on the pipeline. Variables surveyed at the time of repair are:

- Date of repair (DATE)
- Address (ADRESS)
- Location (LOC)
- Type of pipe (PIPE)
- Pipe diameter (DIAM)
- Pressure (PRESSURE)

Furthermore, the incident is classified as the threat is appropriate: THPD or CORR and depending on the Risk of it: High (H), Medium (M) or Low (L).

2. Integration of records of information: Since you use a single source of information records are integrated.

3. Identification of the class attribute: The attribute class to consider is Risk (H, M, L). We took this as a class attribute or objective function as it realizes the gravity associated with each incident.

	Variable	Values	Description
V1	Location	C	City
		P	Province
V2	Tipe	R	Mayor pipe
		S	Service pipe
V3	Diameter	0,5	Minimum [Inch]
		24	Maximum [Inch]
V4	Pressure	P22	22 [Bar]
		P10	10 [Bar]
V5	Threat	CORR	Corrosion
		THPD	Third party damage
	Risk	H	High
		M	Medium
		L	Low

Table 1. Classification of variables

4. Induction Algorithm Implementation: The result of applying this algorithm generates a decision tree associated with rules that are listed in the Table 2:

Rules	THREAT	DIAM	LOC	PIPE	RISK
1	CORR	< 7			L
2		> 7	P		L
3			C		M
4	THPD	< 5		R	H
5				S	M
6		> 5			H

Table 2. Rules of behavior

The interpretation of each of the rules is as follows:

Rules 1-2-3: If the incident was caused due to corrosion of the pipe, the severity of it is low, unless the pipe diameter is greater than 7 inches and is located in Capital; and in any case, depends on the type of pipe.

Rules 4-5-6: If the incident is due to third party damages, the risk of it does not depend on the location and is high, except in the case where the diameter is less than 5 inches and the type of pipe is S.

Another conclusion to be drawn from these results is that the system pressure is not relevant in determining the severity of the incident. This does not imply that the system pressure does not affect the value of risk calculated according to the quantitative model presented above, indicates that there is not one relevant variable of the problem to determine a new rule of behavior.

These simple rules of behavior can be used to determine mitigation actions. For example, from Rule 1 can be defined that is more important take mitigation actions of corrosion in pipes of larger diameter.

We have begun working with another database, which includes all the variables that feed the risk analysis model, to advance the objective of achieving the qualitative feedback and quantitative model. Rules of behavior we hope to get richer than those presented, to have more variables and greater number of cases.

5 Summary / Conclusions / Perspectives

First, we introduced the risk analysis model used in risk management of a distribution company's high-pressure system, pointing out its strengths and weaknesses. Then, some techniques based on intelligent systems as tools for improving the current model.

The results shown should be taken only by way of example, since the database wasn't enough records or sufficient amount of input variables to be able to obtain conclusive results. However, the results are promising in that it shows a possible way, based on the failures history of the system, determining rules of behavior of the pipeline.

It is desirable, on the one hand, to have a feedback between the model and rules of behavior to estimate more accurately the relative weights associated with each variable that underpin the model. Secondly, identify the relevant variables in risk management when defining a single procedure for collecting information.

The next steps will determine the validity of the proposed method with a more powerful database, both in number of incidents and in many characteristics of the pipe at the time of the incident revealed.

Furthermore, we will investigate the exploitation of the information collected with other intelligent tools such as Bayesian networks, which would identify if there is some degree of interdependence between the variables in the model from the construction of so-called interdependence weight tree and predictive learning.

It is expected that the development of the novel techniques taken from the intelligent systems, as data mining, results in the improvement of risk assessments models and could be successfully used in the risk management program of the gas pipeline under study.

On the other hand, we expect the model to predict the potential risk for failure and to anticipate or mitigate the consequences, as far as possible. The mechanism of risk prediction using these techniques could be used in the prediction of all types of industrial risks, for which the probabilistic analysis is not always effective.

Finally, referring to the questions posed at the beginning, we believe the integration between quantitative and qualitative methods is the way forward to improve risk analysis models. We also note that the use of historical data with advanced techniques of data mining is an interesting approach to discover rules of behavior of the system, from which to predict future failures and to identify risk mitigation actions.

References

[1] ASME B31.8S.2001, *Managing System Integrity of Gas Pipelines*, The American Society of Mechanical Engineers. http://www.asme.org.

[2] Muhlbauer K., *Pipeline Risk Management Manual. Ideas, Techniques and Resourses*, Elsevier, 2004.

[3] Bier V., *An Overview of Probabilistic Risk Analysis for Complex Engineered Systems* in Fundamentals of Risk Analysis and Risk Management, CRC Press, 1997.

[4] Carbonell, J., Michalski, R., & Mitchell, T. (1983). *An Overview of Machine Learning*. In R. Michalski, J. Carbonell and T.Mitchell, (Eds.), Machine Learning: An Artificial Intelligence Approach. Tioga Publishing Company.

[5] Quinlan, R. (1986). *Induction of Decision Trees*. Machine Learning 1: 81-106.

[6] Quinlan, J. (1979). *Discovering rules by induction from large collections of examples*. In D. Michie (Ed.), Expert systems in the micro electronic age. Edinburgh University Press.

[7] Quinlan, J. (1990). *Learning Logic Definitions from Relations*. Machine Learning, 5:239-266

[8] Quinlan, J. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

[9] Quinlan, J. (1996a). *Improved Use of Continuous Attributes in C4.5*. Journal of Artificial Intelligence Research, 4: 77-90.

[10] Quinlan, J. (1996b). *Learning Decision Tree Classifiers*. ACM Computing Surveys, 28(1): 71-72.

[11] Quinlan, J.R. (1999). *Simplifying decision trees.* International Journal of Man-Machine Studies 51(2): 497-510.

[12] Britos, P. (2008). *Procesos de Explotación de Información Basados en Sistemas Inteligentes*. Tesis de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. http://laboratorios.fi.uba.ar/lsi/td-pb-fi-unlp.pdf.