Data Mining Improves Pipeline Risk Assessment

Baoqin Wang
Chongqing Communication Institute of PLA
Chongqing, China
e-mail: dyzhuo@yahoo.com.cn

Xuyang Zhou and Wenjing Zhang
Logistical Engineering University of PLA and
Chongqing Communication Institute of PLA
Chongqing, China and Chongqing, China
e-mail: smthroot@gmail.com and
zhang_zhou2010@yahoo.cn

***Abstract*—Accidents to pipelines have been recorded and they often result in catastrophic consequences for environment and society with a great deal of economic loss. Standard methods of evaluating pipeline risk have stressed index-based and conditional based data assessment processes. Data mining represents a shift from verification-driven data analysis approaches to discovery-driven methods in risk assessment.**

Key words: risk assessment; data mining; pipeline risk

I. INTRODUCTION

The consequences of the risks associated with accidents and emergencies that may occur in an oil or 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 assessment, 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.

The immense explosion in the data occasioned by developments in pipeline risk assessment models emphasizes the importance of developing data driven inductive approaches to risk analysis and modeling. Data mining methods are designed to assist the process of exploring large amounts of data in search of recurrent patterns and relationships. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). Data mining is an approach that builds upon and augments existing industry risk management practices such as indexing and panel-based methods. The result is an approach that melds effective current practices with modifications based on probabilistic and statistical trends in the science of risk management.

II. 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.

Fault Tree Analysis is an effective method to evaluate the reliability and security of large complex system. Using this way, the failure of oil and gas pipelines is discussed. The fault tree of oil and gas pipelines is established by considering two major failure modes of pipelines: leaking and rupture. All me minimal cut-sets are got through qualitative analysis. The basic risk factors are investigated that cause failure of pipelines. Fault Tree is also a model that quantitative analysis is based upon.

The Figure 1 shows the logical relation of FTA model that contributes to the probability of failure and the factors.

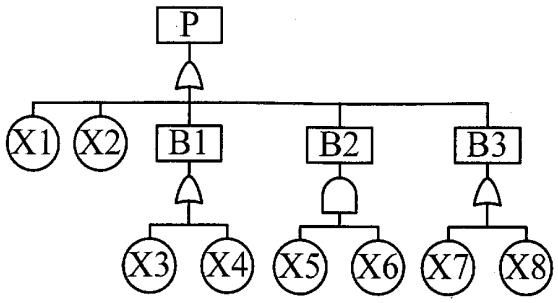

Figure 1. Fault tree of oil and gas pipelines

In conventional FTA, the failure probabilities of system components are treated as exact values. However, for many systems, it is often very difficult to estimate the precise

failure rates or probabilities of individual components or failure events in the quantitative analysis of fault trees from past occurrences. In other words the crisp approach has difficulty in conveying imprecision or vagueness nature in system modeling to represent the failure rate of a system component. This always happens under a dynamically changing environment or in systems where available data is incomplete or insufficient for statistical inferences. Therefore, in the absence of exact data, it may be necessary to work with approximate estimations of probabilities. Under these conditions, it may be inappropriate to use the conventional FTA for computing the system failure probability. Therefore, it is necessary to develop a novel formalism to capture the subjectivity and the imprecision of failure data for use in the FTA. Instead of the probability of a failure, it may be more appropriate to propose its possibility.

Data mining builds models from data, using tools that vary both by the type of model built and, within each model domain, by the type of algorithm used. At the highest level this taxonomy of data mining separates models into two classes, predictive and descriptive.

One of the most common learning tasks in data mining is classification. Classification is a supervised way of learning. The database contains one or more attributes that denote the class of a table and these are known as predicted attributes, whereas the remaining attributes are called predicting attributes. The popular algorithms of supervised learning techniques include decision trees, artificial neural networks, and so on.

When the data are unlabelled and each instance does not have a given class label the learning task is called unsupervised. If we still want to identify which instances belong together, that is, form natural clusters of instances, a clustering algorithm can be applied. Clustering techniques can be used to identify stable dependencies for risk assessment models.

Another unsupervised learning approach is association rule discovery that aims to discover interesting correlation or other relationships among the attributes. Association rule mining was originally used for risk analysis for associations among the factors.

We start from a risk assessment model for a pipeline standard featuring FTA coupled with data mining models. 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.

III. 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 are what differentiate the activity to be monitored.

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.

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.

In the field of industrial processes, there is a trend towards the use of fault 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 pipeline. The scheme of knowledge discovery process is presented in Figure 2.

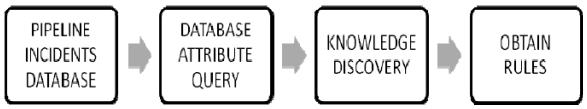

Figure 2. knowledge discovery process

The basis of the induction task 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 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. 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. 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 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.

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 pipeline network. 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.

IV. PROCESS 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 pipeline network are third party damage (THPD) and corrosion (CORR). 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). The process undertaken to discover rules of behavior from the data set available can be summarized in the following results.

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.

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 the first rule 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.

V. CONCLUSIONS

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.

Data mining can be used to enhance risk assessment. Pipeline operators of today are faced with the challenge of integrating a wide variety of design, historical and inspection related information into risk management program. Advances in data technology have overwhelmed corporations with information, and generated an urgent need for new techniques and tools that can intelligently and automatically assist in transforming this data into useful knowledge. Data mining has appeared as one of the tools to better explore engineering information and develop a risk management model.

VI. REFERENCES

[1] Muhlbauer K., Pipeline Risk Management Manual. Ideas, Techniques and Resources, Elsevier, 2004

[2] Quinlan, R. (1986). Induction of Decision Trees. Machine Learning.

[3] Quinlan, J. (1996). Learning Decision Tree Classifiers. ACM Computing Surveys,28(1) ,pp. 71-72.

[4] Joshi M,Karypis G Strategies for parallel data mining[J].IEEE Concurrency,2005,5(3) ,pp.121-130.

[5] Zaki M J.Parallel and distributed association mining: a survey [J].IEEE Concurrency,2004,7(4) ,pp.14-25.