CHEMICAL ENGINEERING TRANSACTIONS

VOL. 31, 2013

Guest Editors: Eddy De Rademaeker, Bruno Fabiano, Simberto Senni Buratti
Copyright © 2013, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-22-8; **ISSN** 1974-9791

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at: www.aidic.it/cet

DOI: 10.3303/CET1331001

Implementing the Seveso Directive Requirement on the Anticipated Presence of Dangerous Substances

Alexis Pey[a], Pablo Lerena[a]

[a] Swissi Instituto Suizo de Seguridad. c/Lope de Vega 30, Entlo. 08005 Barcelona, Spain
alexis.pey@swissi.ch

Industrial activities fall under the scope of the Seveso Directive as a function of the amount of dangerous substances present in the site. In this sense, as defined in point 12 of Art. 3 of the New Seveso III Directive, "the 'presence of dangerous substances' means the actual or anticipated presence of dangerous substances in the establishment, or of dangerous substances which it is reasonable to foresee may be generated during loss of control of the processes, including storage activities, in any installation within the establishment".

Experience shows that the requirement of declaring the presence of dangerous substances due to a loss of control of an industrial chemical process is hard to fulfil; in first place due to the complexity of defining a loss of control scenario and secondly because of the lack of information on the substances that may be generated when processes do not follow the desired reactive behaviour.

Since the behaviour of chemical substances on a chemical reaction does not follow a random path, this is an interesting and suitable field where to apply data mining techniques. These techniques intend to identify empirical regularities observed over a large data set, which are believed to be useful with substance prediction purposes.

This paper explores different possible techniques that can be applied with the purpose of predicting the substances that may be generated under loss of control conditions. Suitability of different techniques is evaluated by measuring its accuracy on the prediction of known scenarios.

1. Introduction

1.1 The problem as defined in the Seveso Directive

It is well known that the Seveso Directive traditionally classifies establishments by means of the quantities of dangerous substances present in them. For the purposes of the recently issued Seveso Directive, 'presence of dangerous substances' means the actual or anticipated presence of dangerous substances in the establishment, or of dangerous substances which it is reasonable to foresee may be generated during loss of control of the processes, including storage activities, in any installation within the establishment, in quantities equal to or exceeding the qualifying quantities set out in Part 1 or Part 2 of Annex I." [EU, 2012]

This definition means that for a full compliance of the Seveso Directive one must be able to make a guess on the substances and their quantities that can be generated under loss of control conditions.

However, no definition for "out of control condition" can be found on the Directive. Such a condition may be defined as: "The loss of control of an industrial chemical process is caused by any event that causes a deviation high enough that operators are unable to correct it and put the system back to normal operation conditions". [Cozzani 1997]

Therefore, in first place operational errors leading to an out of control condition of the system and secondly the consequences of each out of control condition must be evaluated.

In a chemical establishment, many errors may be posed in a reasonable way that may produce such an out of control condition. They may go from a mistake while setting a set point on a controller, to a mistake while loading a reactor due to a wrong order of addition, wrong quantities of reactants, use of wrong

Please cite this article as: Pey A. and Lerena P., 2013, Implementing the seveso ii directive requirement on the anticipated presence of dangerous substances, Chemical Engineering Transactions, 31, 1-6 DOI: 10.3303/CET1331001

substances, etc. Therefore not only the starting events but also its evolution must be assessed while keeping in mind that they must be identified all dangerous substances believed to be generated and their quantities.

It is clear that due to the great number of chemical substances and possible combinations in a typical chemical site, it is quite difficult to assess all possible combinations and reactions they may occur in case of accident. It is also well-known that when a process deviation takes place the risk of decomposition has to be considered, therefore, dangerous substances may be the products of decomposition reactions, which usually are not known in enough detail as to describe the potential decomposition products.

1.2 Why Artificial Intelligence techniques?

Even if the state of the art is not able to make computers learn in a similar way to that of a human being, it is clear that algorithms have been developed which achieved outstanding results on performing certain tasks which require some kind of what has been defined as learning. Success have been especially remarkable in fields were the behaviour or phenomena of study may be easily described in understandable parameters for a computer. Maybe the most important characteristic that must show a problem in order to try to solve it by using artificial intelligence techniques is that its behaviour does not follow a random pattern.

If it is assumed that the behaviour of chemical substances on a chemical reaction does not follow a random pattern, the problem of predicting substances generated under out of control conditions may be a suitable problem to be solved by using artificial intelligence techniques. The basis lays on the fact that even if the rules governing this field are hard to conceive in terms of a cognitive process for a human brain, as they are not random they can be explored by using artificial intelligence techniques.

For the aim of this paper, artificial intelligence is the ability of a computer to solve a well-posed problem improving its performance through experience. To put it more precisely a definition may be used: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” [Mitchell 1997]. Taking this definition it is possible to well-define the learning problem by identifying the task (T), the performance measure (P) and the training experience (E):

- T: prediction of substances generated under out of control conditions.
- P: percent of substances rightly predicted.
- E: database of chemical systems that underwent out of control conditions with known information available on present reactants and substances generated for each example.

In any case it is necessary to have a database where data to be analysed is stored, with this purpose a database, based initially on the EUCLIDE database [Cozzani 1997], has been developed finally containing 529 accidents and 420 substances.

1.3 Define the problem in terms of Artificial Intelligence techniques

The challenge now is to be able to define this problem in understandable terms for a computer without losing its chemical background.

The target function may be called *Predict* and its notation is $Predict : R \rightarrow P$ to indicate that its inputs are sets of reactants and its outputs are sets of generated products. The representation for this target function depends on the artificial intelligence technique used.

The representation of information concerning the reactants and products will define the design of the systems including the database, which in this case will be used as the source of experience for the artificial intelligence systems. The database will contain information concerning inputs (reactants) and outputs (products generated under out of control conditions). However, in order to be used as a source of experience, an artificial intelligence system would have to be able to understand this information, thus information on database must match target function representation requirements.

Within this work, two different types of notations have been used to represent substances within the Artificial Intelligence techniques.

In first place, it may be obvious to use substances in order to represent reactants and products. The use of a substance notation would allow to uniquely identify reactants or products, but if a small part off the molecule is changed then the match would be lost and the system may loss the capacity to represent reactive behaviour. In this sense, substances have been taken mainly as output values for the system, therefore substances are identified specifically and do not have to be guessed by a representation not considering the full molecule of interest.

On the other hand, if substances were used as input variables to the systems, only substances present among reactants could be considered among those over which a prediction could be made. Therefore, it is important to be able to represent reactants in a way that even reactants not present in the database can be considered on the prediction of the system. To do so the Benson Groups [Benson 1976]&[CHETAH 2009] have been used. The use of Benson groups assumes that the functional groups are those who basically define the reactive behaviour of a system under out of control conditions.

Finally it must be explained that, data will be represented by vectors that may be defined taking into account the different representations of substances described.

The target function Predict will be given a vector with information concerning a reactive system and will output a vector containing information on generated substances. The level of information detail used on the input and output vectors is an important feature when designing the target function representation.

Use of vectors to represent data is quite common among AI techniques, but depending on the technique and the posed problem each element of a vector may represent quite different things. In this case, given a vector containing information about a reactive chemical system, the AI algorithm must output a vector containing information useful to define which substances are generated once the system undergoes out of control reaction conditions. Experience will be provided through examples contained on a database for which reactants and generated products are known.

For this reason Artificial Neural Networks are considered a suitable artificial intelligence technique, whereas other techniques such as Montecarlo or Bayesian inference are not initially selected as prediction in one case is made randomly and in the Bayesian case hypothesis on predictions would be quite complex to represent in terms of influence of a reactive mixture.

2. Training and Testing

The way techniques and algorithms performance and reliability is measured is by creating two different sets of data, the training and the testing one. Both data sets are compiled from available data on the database, with the particularity that when one example is said to belong to one of those data sets is then excluded from the other one. Thus one example may not appear both in the training and testing data sets.

Training data set, as its name shows, is the data set that is used to train the system, thus it contains the data that is showed to the system. The system is then adjusted to well represent data contained in the training data set.

2.1 Training

Training process consists in showing data on the training data set to an algorithm and allow it to make a prediction thus being able to measure its own results by comparing them with those on the training data set, as a result an adjustment of its parameters is made, then data on the training data set is presented again and a new prediction is done, which is again measured by comparing with available data on the training data set, result is another adjustment of algorithm parameters. Each time data from training data set is showed to the system it modifies its parameters. Modification is done according to a certain error measurement criterion. Then the more times data is presented the better data is represented and the better performance is achieved. Each time data is presented and parameters are modified is known as a training step or training pass. Next a figure is presented to illustrate ANN tuning, so-called training, process:

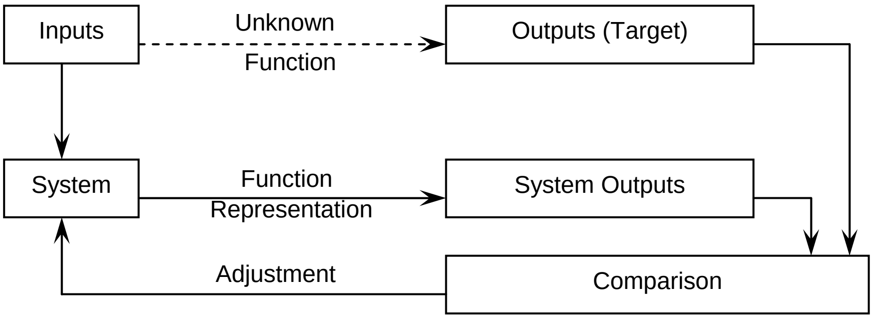

Figure 1. Training & Testing Scheme.

2.2 Testing

Testing data set contains data that is not shown to the system with training, or adjustment, purposes. The system is ask to use data in the testing data set to make predictions. In fact, when using this data as input, the system is making predictions over previously unseen data. By measuring reliability of predictions with data contained on testing data set it is assumed that results show expectable reliability with data not contained in the database.

2.3 Overfitting

At this point it is interesting to introduce the concept of overfitting, although this phenomena is quite specific when working with ANN.

In the former point, data set concept for training and testing was introduced. Basic idea is that when showing training data set to an AI algorithm it will learn how to well represent this data, and if the data all over the data base is homogeneous, then it will be learning as well, how to represent unseen data from testing data set. But this ideal situation is not usually given in a real problem.

When working with an algorithm to measure its performance, results obtained with training data and results obtained with testing data are compared. Behaviour showing overfitting is that which at the beginning, on initial passes, both training and testing performances do usually improve, but one point is reached where performance measured over training data set continues to improve but performance measured over testing data set decreases.

Overfitting occurs because of the network continues to adjust its parameters to better represent known data from the training data set, thus improving its performance, but this well-adjustment to known data decreases the system capacity to generalize over unseen examples, thus performance over test data set decreases.

3. Performance Measurement

In this point it will be discussed how to finally define performance measurement together with all the considerations that one must take into account.

3.1 Hits and Errors

In first place they must be defined and discussed different prediction situations and how they will be considered when measuring prediction performance. It is clear that while making predictions 4 possible cases are expectable as a result:

a) A substance is predicted, and it is generated in the reaction. When this situation occurs it will be referred to as a Hit.

b) A substance is predicted, but it is not generated in the reaction. This type of error will be referred to as β Error.

c) A substance is not predicted, but it is generated in the reaction. This type of error will be referred as α Error.

d) A substance is not predicted, and it is not generated in the reaction. From the point of view of a AI prediction system, this is not an error thus is a right prediction, but from the point of view of a chemist or an engineer this information is not useful at all, since for them it will be quite easy to elaborate a long list with substances that surely will not be generated from a certain chemical system. Even more, it must be taken into account that an AI algorithm will always predict formation or not of all substances appearing as products in the database. In this way the higher number of accidents contained in the database the higher number of possible substances to be predicted. Then performance, if measured in this way, may increase without increasing well-done predictions. When a prediction is a case d) prediction, it will not be considered an error nor a hit. It will simply not be taken into account to measure performance.

From a safety point of view having an α Error on a prediction is a worse case than having a β Error. But if algorithms are pushed to predict a lot of substances in order to prevent having any α Error, then many β Errors may appear, which is neither a desirable situation. Too many wrong predictions, even if they are β Errors, will cause prediction algorithms being useless.

3.2 Percentage Performance Measurement

As a consequence of what has been exposed until now, performance measure for the problem of substance generation prediction by using AI techniques is defined as:

$$Performance\,\% = \frac{Hits}{Hits + \alpha\ Errors + \beta\ Errors}\cdot 100 = \frac{Hits}{Hits + Total\ Errors}\cdot 100 \qquad (1)$$

By using this definition as a starting point, error percentage is defined as:

$$Error\ \% = 100 - Performance\ \% \tag{2}$$

As can be seen from definitions above, when using Performance % to measure how good a prediction is, it is possible that a lower value is obtained on a prediction with a higher number of hits if, at the same time, this prediction is generating more α or β Errors than another prediction which is well-predicting a lower number of substances but, at the same time, is committing a lower number of errors.

Performance % may be used as a reference point but results must be analysed by taking into account not only the number of hits but also the number of errors, and within the errors, it must be remembered that α Errors have, from safety point of view, worse consequences than β Errors.

In order to obtain percentage values as previously defined, the number of Hits and α and β Errors must be calculated.

4. Artificial Neural Networks (ANN)

4.1 Estructure of ANN

ANN are originally inspired by biological nervous systems, a brief example is given [Mitchell 1997] in order to illustrate the origins for this biological interest: a human brain is estimated to contain about 10^{11} neurons, each one connected to other 10^{4} neurons on average. Neuron activity is typically excited or inhibited through connections to other neurons. Fastest neuron switching times are on the order of 10^{-3} seconds. Since it is estimated that takes about 10^{-1} seconds to visually recognise ones' mother, it is clear that only few neuron may switch during this time.

ANN structure is constituted by single elements, so-called neurons, which are connected at a time with other neurons. A neuron receives one or more inputs, which once processed by its internal function, generates an output. This output may be a final result or an input for one or more other neurons. The way neurons are connected with each other and the values of those connections, so-called weights, determine the function that an ANN is representing. For this reason an ANN is not able to give any proper output until it is trained. Making predictions with ANNs is generally a low time consuming process, being the training process the one which is high time consuming.

Results presented in this paper were obtained with Artificial Neural Networks consisting in 3 layers with two hidden layers and an output layer. All ANN were generated using sigmoid transfer function neurons on all layers, as well, all layers had the same number of neurons as the output layer, which equals the maximum number of products able to be predicted.

Tests were run with all chemical levels of representation initially with only 5 data sets and then results with Boolean Benson Groups were obtained at extreme data set numbers, that is being each instance a data set itself and using all other remaining instance to train the network.

4.2 Training ANN

When talking about multilayer networks one must talk about the learning algorithm that is probably most widely used in this type of ANNs, this is the Backpropagation learning algorithm.

Standard Backpropagation algorithm uses gradient descent to minimize the squared error measured between ANN outputs and target known values. But many variations on the standard Backpropagation algorithm were made on the basis of other optimization techniques, such as conjugate gradient descent or Newton methods. One of these variations is called Resilient Backpropagation. This algorithm is quite robust in front of net training parameter changes and data errors and shows a quite outstanding performance compared to other training algorithms [Hagan 1996].

Due to the weight and bias adjustment algorithms, most of multilayer ANNs require having differentiable neuron transfer functions. In this case, functions such as log-sigmoid, tan-sigmoid or linear transfer functions are commonly used. Generally speaking, it can be said that a network with a sigmoid layer, biases and a linear output layer is able to approximate any function with a finite number of discontinuities, providing that sufficient number of neurons is given in the hidden sigmoid layer [Demuth 2000].

4.3 Results

Next figure shows results obtained with a three-layered ANN when Boolean Benson Groups are used, and each instance being considered as a data set itself:

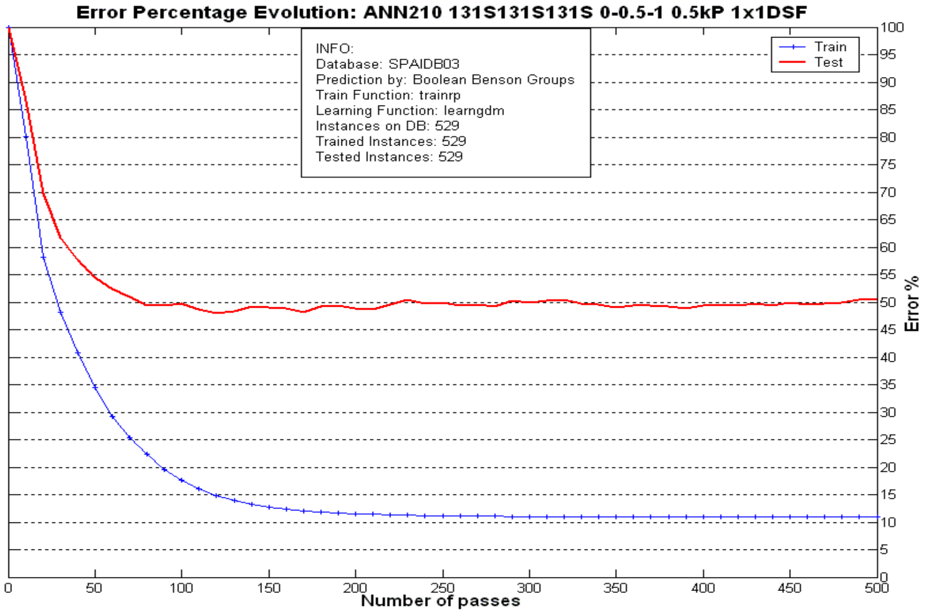

Figure 2. Multilayer ANN Test and Train evolution. Boolean BG. 3-Layered ANN.

5. Conclusions

The problem posed by the Seveso II Directive has been defined into an understandable way to a computer and suitable to be solved by Artificial Intelligence systems.

A database, has been developed with prediction purposes containing 529 accidents and 420 substances, from which 131 appear at least 1 time as a generated product, thus able to be predicted by Artificial Intelligence algorithms.

Best prediction results are obtained by using Multilayer Artificial Neural Networks. Over a set of unseen instances, best results are achieved with 3 layered Artificial Neural Networks, which are able to reach a prediction performance around 55%.

Prediction performance had been initially assessed by random prediction, as a reference it can be said that a random prediction never reached a prediction performance higher than 2.5%.

Artificial Intelligence algorithms had been proven useful on helping expertise to analyse potentially generated substances under out of control conditions, thus helping to fulfil the Seveso II Directive requirements.

6. References

Arbib M.A. 1997, Brains, Machines and Mathematics. Springer-Verlag. Berlin.

Affolter Ch., Clerc J.T 1993., Chemometrics and Intelligent Laboratory Systems. Laboratory Information Management, 1993, 21, 151-157.

CHETAH, 2009, The ASTM chemical thermodynamic and energy release evaluation program version 9.0.

Benson, S. W., 1976, Thermochemical kinetics. Methods for the estimation of thermochemical data and rate parameters (2nd edition), John Willey & Sons Inc. New York

Bhat N., McAvoy T.J., 1990, Computers and Chemical Engineering, 14, 573-583.

Cozzani V., Zanelli S., 1997, EUCLID a study on emission of unwanted compounds linked to industrial disasters, European Commission Joint Research Centre. Ispra, Italy.

Demuth H., Beale M., 2000, Neural Network Toolbox. The MathWorks, Inc. United States of America.

EU, 2012, Directive 2012/18/EU Of The European Parliament And Of The Council of 4 July 2012 on the control of major-accident hazards involving dangerous substances, amending and subsequently repealing Council Directive 96/82/EC

Hagan M. T., Demuth H. B., Véale, M. H., 1996, Neural network design. PWS Publishing. Boston

Mitchell T.M., 1997, Machine Learning, McGraw-Hill. United States of America.