A classification-based framework for trustworthiness assessment of quantitative risk analysis

Zhiguo Zeng, Enrico Zio

A classification-based framework for trustworthiness assessment of quantitative risk analysis

Zhiguo Zeng,[†] Enrico Zio (*),[†‡]

[†]Chair on System Science and the Energy Challenge, Fondation Electricite de France (EDF), CentraleSupélec, Université Paris-Saclay, Grande Voie des Vignes, 92290 Chatenay-Malabry, France

[‡]Energy Department, Politecnico di Milano, Milano, Italy

Abstract

In this paper, we develop a classification-based method for the assessment of the trustworthiness of Quantitative Risk Analysis (QRA). The QRA trustworthiness is assumed to be determined by the quality of the QRA process. Six quality criteria, i.e., completeness of documentations, understanding of problem settings, coverage of accident scenarios, appropriateness of analysis methods, quality of input data, accuracy of risk calculation, are identified as the factors most influencing the trustworthiness. The assessment is, then, formulated as a classification problem, solved by a Naive Bayes Classifier (NBC) constructed based on a set of training data, whose trustworthiness is given by experts. NBC learns the expert's assessment from the training data: therefore, once constructed, the NBC can be used to assess the trustworthiness of QRAs other than the training data. Leave-one-out cross validation is applied to validate the accuracy of the developed classifier. A stochastic hypothesis testing-based approach is also developed to check the consistency of the training data. The performance of the developed methods is tested for ten artificially generated scenarios. The results demonstrate that the developed framework is able to accurately mimic a variety of expert behaviors in assessing the trustworthiness of QRA.

Index Terms

Quantitative Risk Analysis (QRA), validity, reliability, trustworthiness, Naive Bayes classifer

Highlights

- An assessment framework is developed for trustworthiness of QRA.
- A naive Bayes classifier is developed to assess the trustworthiness of QRA.
- Consistency of training data is checked by developing a statistical hypothesis testing.

* Email of the corresponding author: enrico.zio@ecp.fr, enrico.zio@polimi.it

A classification-based framework for trustworthiness assessment of quantitative risk analysis

I. Introduction

Since its first application in nuclear power plants in 1975 [1], Quantitative Risk Analysis (QRA) has been widely applied in various fields to support safety-related decision making [2], e.g., chemical process industry [3], oil & gas industry [4], maritime transportation [5], nuclear installations [6], etc. Various methods have been developed for QRA [7]. According to Khan et al. [8], in general, QRA methods are evolving from semi-qualitative analysis to detailed quantitative analysis. For example, in the 1990s, QRA in the process industries were primarily based on semi-qualitative methods like hazard operatability (HAZOP) analysis [9], while recent QRAs are mainly based on detailed quantitative analysis methods, such as Bayesian network [10], bow-tie model [11], etc. How to compare the trustworthiness of different QRA methods, then, becomes an essential problem in QRA.

Trustworthiness of a QRA refers to the degree that a decision maker can trust the results of QRA [12–14]. This question is of paramount importance in practice: only a trustworthy QRA can be useful to support decision making. Trustworthiness assessment of QRA has been discussed by many researchers [15–17], although sometimes using different concepts and terminologies, e.g., evaluation [17], validation [18], verification [19], quality assurance [15], credibility assessment [20], etc. Goerlandt et al. present a thorough survey on the status quo of the trustworthiness assessment of QRA [18]. According to their survey, existing methods on QRA trustworthiness assessment can be broadly classified into four categories: benchmark exercise, reality check, independent peer review and quality assurance.

Benchmark exercise methods rely on the comparisons among several parallel analyses of QRA to determine its trustworthiness. Usually, two quality characteristics, i.e., reliability and validity (see [18] for a detailed discussion), are considered in the comparisons. For example, the trustworthiness of the QRA on an ammonia storage facility is assessed by comparing seven benchmark exercises in terms of their outcomes, methodologies, data and models [21]. Reality check methods assess the trustworthiness of the QRA by comparing the results with real data or operating experience of the same system or process [18]. A typical example is presented in [22], where statistical data of real accidents and incidents are compared to the risk indexes calculated by QRA, to evaluate its trustworthiness. In independent peer review methods, the process of QRA and its results are reviewed by independent experts, based on a series of predefined quality requirements, and the trustworthiness of the QRA is determined by the experts accordingly [18]. Reference [23] presents a typical example of independent review methods, where QRA is reviewed and its trustworthiness is determined by experts considering the following factors, i.e., objective and statement of purpose, project plan and scope of work, figures of merit, methodology, data base, results, implementation and application and verification of selected results. Quality assurance methods

apply quality control techniques on each phase of the QRA process, aiming at ensuring the quality of the QRA process [18]. It is assumed that a high-quality QRA process will yield trustworthy risk assessment results. For example, in [15], Suokas and Rouhiainen summarized common flaws in each phase of the QRA process and developed a check-list-based approach to ensure their quality.

Among the four methods, benchmark exercise and reality check take a retroactive perspective on trustworthiness assessment, in the sense that their assessments are primarily done by comparing the results of the analysis with either parallel analyses or field data and experience; independent peer review and quality assurance, on the other hand, are proactive in the sense that instead of directly assessing the results, these methods evaluate the capability of the QRA process (in terms of its quality) and predict the trustworthiness of the assessment results based on the quality of the QRA process. In a sense, the retroactive perspective is preferred as it is more trustable, since it directly assesses the trustworthiness of the results of the analysis [24]. However, two major shortcomings might limit its applicability. First, the retroactive methods are normally more difficult and expensive to implement, due to the requirements on field data or parallel analyses [18]. Second, the retroactive methods tell us little about the contributing factors to the trustworthiness, which, limit the ability to guide improvements of the QRA process for improved trustworthiness [25]. Since in practice, the strict requirements of the retrospective methods on field data or parallel analyses are always hard to fulfill, in this paper, we focus only on the proactive methods.

In the proactive methods, the trustworthiness is assessed (in fact, predicted) based on the capability (in terms of quality) of the QRA process. Two issues are essential when developing such methods:

- how to evaluate the capability (in terms of quality) of a QRA process?
- how to relate the trustworthiness of QRA to the capability of the QRA process?

The first issue has been addressed relatively well in literature: the major influencing factors for the quality of a QRA process have been widely discussed in literature [26–28] and various methods have been developed for the assessment [29–31]. The second issue, however, is not so well explored. In most existing researches ([29–31], for example), the relationship between the trustworthiness and the quality of the QRA process is treated as a black box and the experts are asked to directly construct a mapping from the process capability to the trustworthiness of QRA. Usually, this is done by a simple conformance/non-conformance-based framework: the conclusion of whether the QRA is trustworthy or not is made by comparing the number of the conformed quality criteria to a predefined threshold value [32]. Such a process is subject to several uncertainties, primarily due to the opacity in the elicitation process and lack of procedures to assess the accuracy of such assessments [33]. A formal and quantitative method is, therefore, needed for trustworthiness assessment.

In fact, trustworthiness assessment can be viewed as a classification problem (more broadly, supervised learning, see [34] for details): train a classifier, which is a mapping from the capability of the QRA process to the trustworthiness of its results, based on a set of training data that are pre-assessed by experts. Hence, in this paper, we develop a classification-based framework, using Naive Bayes Classifier (NBC), for a formal and quantitative assessment of the trustworthiness of QRA. NBC is a simple but effective classifier widely applied in machine learning applications, e.g., text classification [35], tumor diagnosis [36], etc. Although classification-based frameworks have been developed to assess other qualitative factors, e.g., vulnerability and safety criticality of nuclear power plants [37], prediction capability of prognostic methods [38], etc., to the best

of our knowledge, it is the first time that a classification-based framework is developed for the assessment of QRA trustworthiness and that NBC is used for that purpose. It should be noted that in this paper, the classifier is not directly used to assess the trustworthiness. Rather, it is used as a tool for constructing the evaluation criteria used for determining the trustworthiness.

The rest of the paper is organized as follows. A general classification-based assessment framework of QRA trustworthiness is developed in Section II. NBC is applied in Section III to assess the trustworthiness of QRA. In Section IV, we develop a method to check the consistency of the experts that generate the training data. Ten numerical case studies are considered and, then, an application is presented regarding a real trustworthiness assessment of QRA in Section V. The paper is concluded in Section VI, with a discussion on potential future developments.

II. Assessment Framework

In this section, we present a general framework to support classification-based trustworthiness assessment of QRA. Let T represent the trustworthiness of QRA. We take a proactive perspective on trustworthiness assessment and assume that T is determined by the quality of the QRA process. According to Rae et al. [13], a typical QRA process involves eight sub-processes, as shown in Figure 1. To ensure the quality of a QRA process, all the eight sub-processes should be conducted with high quality [13]. A framework for trustworthiness assessment is, then, developed in Figure 2 by considering the quality requirements on the eight sub-processes in Figure 1.

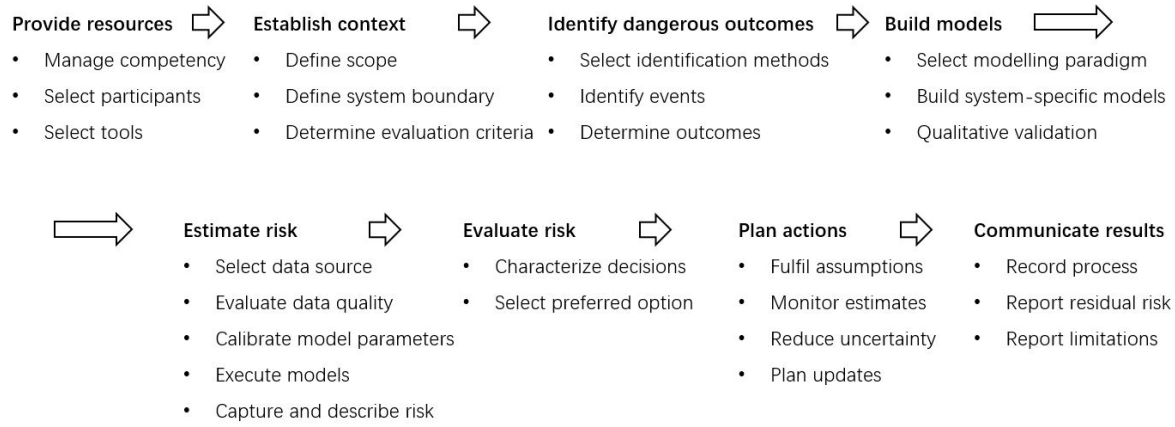


Fig. 1. A typical QRA process [13]

In Figure 2, the trustworthiness of QRA is characterized in terms of six criteria, i.e., completeness of documentations (x_1), understanding of problem settings (x_2), coverage of accident scenarios (x_3), appropriateness of analysis methods (x_4), quality of input data (x_5), accuracy of risk calculation (x_6), which reflect the quality requirements on the QRA process. Each criterion is evaluated into three grades, i.e., problematic ($x_i = 0$), acceptable ($x_i = 1$) and satisfactory ($x_i = 2$), $i = 1, 2, \cdots, 6$, based on a set of predefined scaling rules in Table A.1-A.6. Three discrete levels of T, i.e., $T \in \{0, 1, 2\}$, are considered in this paper. The levels are distinguished in Table I based on their reliability, which concerns the repeatability of the risk analysis [26]

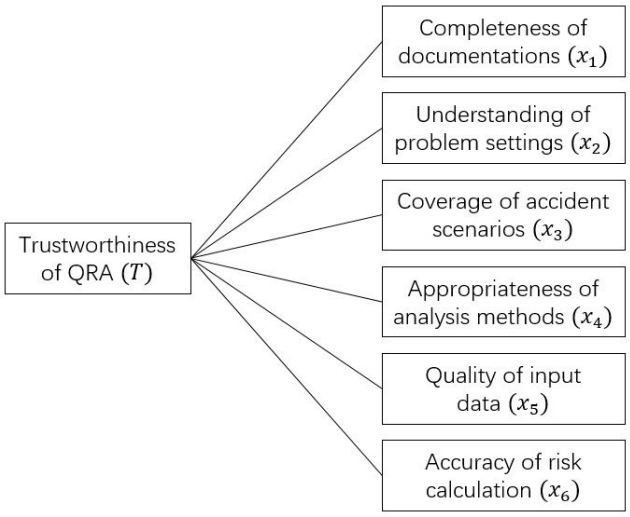


Fig. 2. Trustworthiness assessment framework

and validity, which concerns whether the risk analysis addresses the "right problem" [26]. The problem of trustworthiness assessment is, then, formulated as a classification problem: given the states of the six criteria $x_1, x_2, \cdots, x_6$, determine an appropriate category for the trustworthiness T. It should be noted that both the assessment framework in Figure 2 and the scaling rules in Table A.1-A.6 are constructed for illustrative purposes. They are defied in a general form that allows them to be adapted for capturing the problem-specific features in practical applications.

TABLE I
THREE LEVELS FOR T

Levels of trustworthiness	Descriptions
$T = 0$: Unreliable	• The result of the QRA is unrepeatable. • No further judgements can be made on the trustworthiness of the QRA. • Such QRA should not be used to support any decision making.
$T = 1$: Reliable but invalid	• The result of QRA is repeatable but • some critical hazards are not identified and analyzed by the QRA or • some important risks (and their uncertainties) are not accurately quantified by the QRA. • Such QRA can be used to support decision making, but not for safety-critical decisions.
$T = 2$: Reliable and valid	• The result of the QRA is repeatable and • all critical hazards are identified and analyzed by the QRA; • all important risks (and their uncertainties) are accurately quantified by the QRA. • Such QRA can be used to support critical decision making.

III. TRUSTWORTHINESS ASSESSMENT BASED ON NAIVE BAYES CLASSIFIER

In this section, we first review some preliminaries on NBC-based classification in Subsection III-A and, then, develop a NBC-based method to assess the trustworthiness of QRA in Subsection III-B.

A. Naive Bayes classifier

Let us define $\mathbf{x} = [x_1, x_2, \cdots, x_n] \in \mathbb{X}$ to be the input feature vector of the classification problem, where $\mathbb{X}$ is the feature space. A NBC is a function f_{NBC} that maps input feature vectors $\mathbf{x} \in \mathbb{X}$ to output class labels $T \in \{0, 1, \cdots, C\}$ [39]. Usually, the feature vector also takes discrete values, so that we have $x_i \in \{0, 1, \cdots, n_i\}, i = 1, 2, \cdots, n$. Given a feature vector $\mathbf{x}$, a NBC classifies it into the class with the maximum posterior probability [39]:

$$T = \arg\max_T Pr(T \mid \mathbf{x}). \tag{1}$$

The posterior probability in (1) is calculated using Bayes rule [39]:

$$Pr(T \mid \mathbf{x}) = \frac{Pr(\mathbf{x}, T)}{Pr(\mathbf{x})} = \frac{Pr(\mathbf{x} \mid T)Pr(T)}{\sum_{T=0}^{C} Pr(\mathbf{x} \mid T)Pr(T)}. \tag{2}$$

If we further assume that the elements $x_i, i = 1, 2, \cdots, n$ of the input feature vector $\mathbf{x}$ are independent, the nominator of (2) becomes:

$$Pr(\mathbf{x} \mid T)Pr(T) = Pr(T)\prod_{i=1}^{n} Pr(x_i \mid T). \tag{3}$$

Note that the denominator in (2) is the same for all possible values of T. Therefore, (1) can be simplified:

$$T = \arg\max_T Pr(T)\prod_{i=1}^{n} Pr(x_i \mid T). \tag{4}$$

In order to apply the NBC, the $Pr(T)$ and $Pr(x_i|T)$ in (4) should be estimated from training data. Training data are a set of samples whose correct classes are already known. Suppose we have $N_{training}$ training data, denoted by $(\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \cdots, N_{training}$. Then, the required probabilities are estimated by:

$$Pr\left(T = k\right) = \frac{\sum_{q=1}^{N_{training}} \mathbb{1}\left(T^{(q)} = k\right)}{N_{training}}, \tag{5}$$

$$Pr\left(x_i = j \mid T = k\right) = \frac{\sum_{q=1}^{N_{training}} \mathbb{1}\left(x_i^{(q)} = j, T^{(q)} = k\right)}{\sum_{q=1}^{N_{training}} \mathbb{1}\left(T^{(q)} = k\right)}, \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function and $i = 1, 2, \cdots, n, j = 0, 1, \cdots, n_i, k = 0, 1, \cdots, C$.

There is one potential problem for (5) and (6). Suppose that due to statistical variations, for some specific values of j and k, we have $\sum_{q=1}^{N_{training}} \mathbb{1}\left(x_i^{(q)} = j, T^{(q)} = k\right) = 0$. In this case, $Pr\left(x_i = j \mid T = k\right) = 0$, which, according to (3), results in $Pr(\mathbf{x} \mid T) = 0$, regardless of the posterior probabilities for other features. Misclassification often happens in such situations. To avoid such a problem, a technique called Laplacian correction is often applied when estimating $Pr\left(T = k\right)$ and $Pr\left(x_i = j \mid T = k\right)$ [39]:

$$Pr\left(T = k\right) = \frac{\sum_{q=1}^{N_{training}} \mathbb{1}\left(T^{(q)} = k\right) + \gamma}{N_{training} + (C+1)\cdot\gamma}, \tag{7}$$

$$Pr\left(x_i = j \mid T = k\right) = \frac{\sum_{q=1}^{N_{training}} \mathbb{1}\left(x_i^{(q)} = j, T^{(q)} = k\right) + \gamma}{\sum_{q=1}^{N_{training}} \mathbb{1}\left(T^{(q)} = k\right) + (n_i+1)\cdot\gamma}, \tag{8}$$

where $\gamma \in (0, 1]$ is an adjustment factor introduced to compensate for the possible zero probabilities; $C + 1$ and $n_i + 1$ are the number of possible values for T and x_i, respectively.

B. Trustworthiness assessment

In this section, we apply the NBC to develop a classifier for the trustworthiness assessment problem in Figure 2. In this case, we have six features, i.e., $\mathbf{x} = [x_1, x_2, \cdots, x_6]^T$. Each feature has three discrete levels, i.e., $x_i \in \{0, 1, 2\}, i = 1, 2, \cdots, 6$. Hence, $\mathbb{X} = \{0, 1, 2\} \times \cdots \times \{0, 1, 2\} = \{0, 1, 2\}^6$. The trustworthiness also takes three values, i.e., $T \in \{0, 1, 2\}$. In general, three steps are involved in the development of the classifier, as shown in Figure 3.

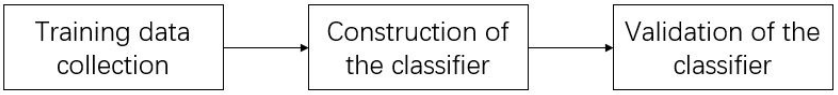


Fig. 3. Procedures of constructing the NBC for trustworthiness assessment

1) Training data collection: Since $\mathbb{X} = \{0, 1, 2\}^6$, the feature vector $\mathbf{x}$ can take $3^6 = 729$ different values. A fraction of them, denoted by $\mathbf{x}^{(q)}, q = 1, 2, \cdots, N_{training}$, are selected as training samples. The trustworthiness of these training samples, denoted by $T^{(q)}, q = 1, 2, \cdots, N_{training}$, are evaluated by experts, based on the descriptions in Table I. The training data are, then, used to construct the NBC and once constructed, the NBC is exploited to replace the expert for the assessment of trustworthiness.

Since the NBC learns the expert's evaluation rationale from the training data, it is essential that the training data are a reasonable representation of the whole feature space. On the other hand, we want to reduce the number of training data as much as possible, since collecting training data is often expensive and time-consuming. For this, in this paper, we use an experiment design technique, i.e., the row-exchange algorithm in Matlab R2015b, to design the training data collection scheme. The response model in the row-exchange algorithm is assumed to be a linear model and the resulted D-optimal design matrix is used for the collection of training data. This approximates an orthogonal design on the $\mathbf{x}^{(q)}, q = 1, 2, \cdots, N_{training}$, where the collected training data are equally distributed and can equally "represent" the entire space of $\mathbb{X}$.

Another issue that needs to be considered when designing the training data collection scheme is the sample size $N_{training}$. Apparently, a large value of $N_{training}$ would enhance the performance of the developed classifier in terms of its accuracy. On the other hand, large values of $N_{training}$ also create more difficulties in collecting the data (experts easily get impatient when asked to judge too many scenarios). Hence, a trade-off needs to be made in determining the value of $N_{training}$.

2) Construction of the classifier: The procedures for constructing the NBC is summarized in Figure 4. In the preparation phase, the sample size of the training set and the training data collection scheme are determined using the methods discussed previously. The training data $(\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \cdots, N_{training}$ are, then, collected by expert judgements following the scaling rules in Table A.1-A.6. In the training phase, the NBC is constructed by estimating $Pr(T)$ and $Pr(x_i|T)$ from the training data, using (7) and (8), respectively. In the evaluation phase, the constructed NBC is applied to replace the role of the experts and determine the trustworthiness of a new QRA. By reviewing the related documents, the value for the feature vector $\mathbf{x}$ of the QRA is determined first, based on the scaling rules defined in Table A.1-A.6. Its trustworthiness is, then, determined based on the constructed NBC using (4).

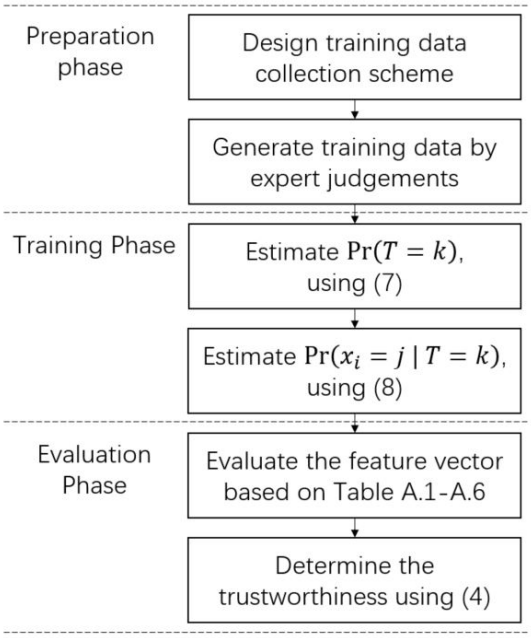


Fig. 4. NBC construction procedure for QRA trustworthiness assessment

3) Validation of the classifier: In practice, the sample size of the training data available for the construction of the NBC is always small, which might impair the accuracy of the classifier. If the NBC is not accurate enough, it might not be able to correctly "mimic" the expert's behavior in assessing the trustworthiness. It is, then, necessary for us to consider the validation of the developed NBC, i.e., to determine our confidence that the classifier can correctly represent the experts' judgement behaviors. As [37], such confidence is measured by the probability that the classifier correctly determines the trustworthiness of a QRA, denoted by CR. Leave-One-Out Cross Validation (LOOCV) is exploited in this paper to estimate CR, where one sample from the training data is left to test the model while the remaining training data are used to train the classifier [33, 34]. The procedures of implementing LOOCV is summarized in Figure 5, where $(\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \cdots, N_{training}$ are training data and CR_{CV} is the correctness rate estimated by LOOCV. The initial values for i and sum are $i = 0, sum = 0$, respectively.

IV. Checking the consistency of training data

Since the training data are empirically assessed by the experts, the consistency of the expert, therefore, is essential in the NBC-based trustworthiness assessment. Training data generated by an inconsistent expert might be self-contradicting and therefore misleading. In this section, we develop a statistical hypothesis testing to check the consistency of the training data.

Motivated by the methods for consistency checking used in the Analytical Hierarchy Process (AHP) [40], we assume that if an expert is inconsistent, he/she would classify a feature vector $\mathbf{x}$ in the training data set to a random trustworthiness level, regardless of the value of $\mathbf{x}$. Suppose we have $T \in \{0, 1, \cdots, C\}$. Then, for any feature vector $\mathbf{x}$ in the training data set, an inconsistent expert would judge it to be $T = i, i = 0, 2, \cdots, C$ with probability $1/(C + 1)$. Hence, we develop the following hypothesis testing to check the consistency of

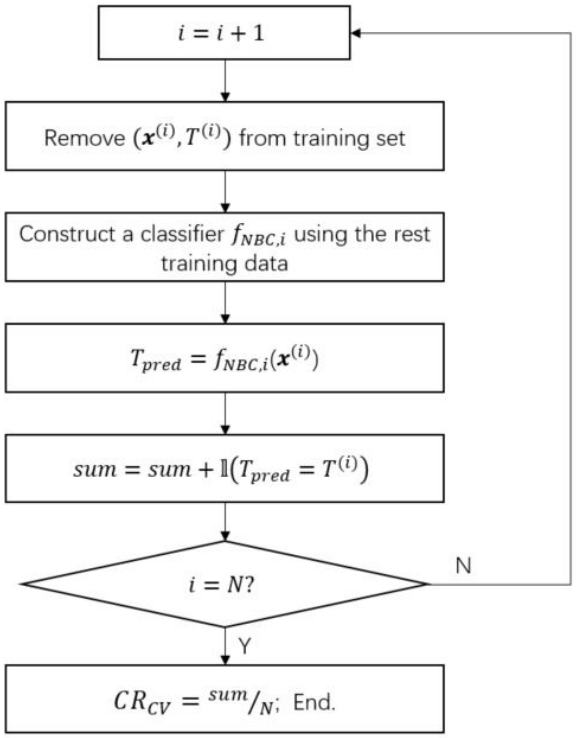


Fig. 5. Procedures of implementing LOOCV

the training data:

$$H_0 : \text{The expert is inconsistent.} \qquad H_1 : \text{The expert is consistent.} \tag{9}$$

In (9), H_0 and H_1 are the null hypothesis and alternate hypothesis, respectively. If the observed data can support us to reject the null hypothesis, the training data is believed to be consistent and can be used to train the NBC; otherwise, we cannot trust the consistency of the training data and a reevaluation of the training data is required.

The CR_{CV} estimated by LOOCV is chosen as the test statistic:

$$CR_{\text{CV}} = \frac{\sum_{i=1}^{N_{training}} \mathbb{1}\left(T_{pred}^{(i)} = T^{(i)}\right)}{N_{training}}, \tag{10}$$

where $T_{pred}^{(i)}$ and $T^{(i)}$ are the predicted and true classes of the ith cross validations, respectively. The empirical distribution of the test statistic under null hypothesis can be approximated using randomly generated training data. Suppose for a given significance level α, $p_{1-\alpha}$ denotes the $(1-\alpha)$ percentile of the empirical distribution of the CR_{CV} calculated using the randomly generated training data. Then, decisions on the consistency of the training data can be made by comparing $p_{1-\alpha}$ and CR_{data}, which is calculated using the real training data [41, 42]:

- If $CR_{data} > p_{1-\alpha}$, reject H_0, the expert is consistent;
- otherwise, cannot reject H_0, the expert is inconsistent.

The physical meaning of the significance level α is the probability that an inconsistent expert is mistakenly judged consistent by the test [42]. Depending on the confidence requirements, two values of α are commonly used: $\alpha = 0.01$ and $\alpha = 0.05$.

Algorithm 1 below presents a pseudo-code for the developed consistency-checking method. In Algorithm 1, n_{random} is the sample size used to approximate the distribution of the test statistic. In principle, we prefer a large value of n_{random} since it helps to reduce the uncertainties in the decision caused by the approximation errors in estimating $p_{1-\alpha}$. However, balance is needed between the accuracy of the estimation and the required computational costs. If Algorithm 1 returns IsConsist $= 1$, we conclude that the training data under evaluation is consistent under the significance level α; otherwise, we cannot reach the conclusion that the training data are consistent and a re-evaluation of the training data is required.

Algorithm 1 Consistency verification based on hypothesis testing

Inputs: $\alpha, n_{\text{random}}, \left(\mathbf{x}^{(q)}, T^{(q)}\right), q = 1, 2, \cdots, N_{training}$

Onputs: IsConsist

for $i = 1 : n_{\text{random}}$ **do**
  $T^{(q)}_{\text{random}} \leftarrow$ GENRNDSAMPLE, $q = 1, 2, \cdots, N_{training}$;
  $CR^{(i)}_{\text{random}} \leftarrow$ Do LOOCV using the randomly generated training data $(\mathbf{x}^{(q)}, T^{(q)}_{\text{random}}), q = 1, 2, \cdots, N_{training}$;
end for
$CR_{\text{sort}} \leftarrow$ Sort CR_{random} in ascending order;
$p_{1-\alpha} \leftarrow CR^{(n_{\text{random}} \cdot (1-\alpha))}_{\text{sort}}$;
$CR^{(i)}_{\text{data}} \leftarrow$ Do LOOCV using real training data $(\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \cdots, N_{training}$;
if $CR_{\text{data}} > p_{1-\alpha}$ **then**
  return IsConsist = 1;
else
  return IsConsist = 0;
end if

Inputs: NULL

Onputs: y_{random}

function GENRNDSAMPLE
  Generate a random number r;
  if $r \leq 1/3$ **then**
    return $y_{\text{random}} = 0$;
  else if $r \leq 2/3$ **then**
    return $y_{\text{random}} = 1$;
  else
    return $y_{\text{random}} = 2$;
  end if
end function

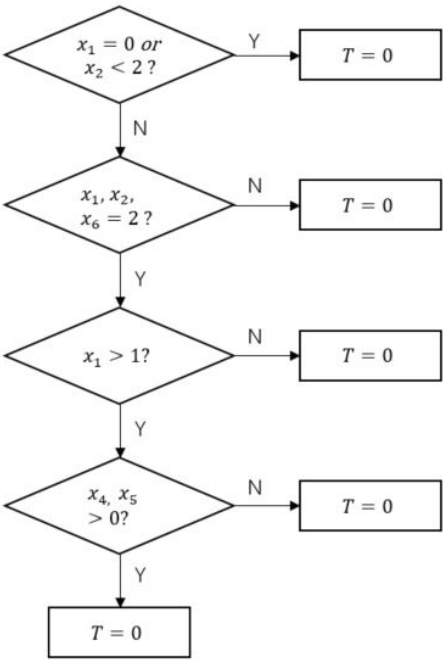


Fig. 6. Illustration of the assumed expert rationale

V. Applications

In this section, we first test the performance of the developed methods under ten artificially generated scenarios in Subsection V-A. Then, the methods are applied in Subsection V-B to assess the trustworthiness of a real-world methanol QRA.

A. A numerical case study

Ten scenarios are artificially generated to test the performance of the developed algorithms. For each scenario, we assume that the behavior of the experts when assessing the trustworthiness is consistent and known to us. Ten different expert behaviors are used to generate the testing scenarios. We present one of the assumed expert behaviors in Figure 6 for illustration purposes.

Let $\mathbf{x}_{full}^{(i)}$ represent all the possible states in the feature space $\mathbb{X}$, where in this case, we have $i = 1, 2, \cdots, 729$. The true trustworthiness for each $\mathbf{x}_{full}^{(i)}$, denoted by $T_{full}^{(i)}$, can, then, be determined by the same known assessment behavior. According to the designed training data collection scheme, a fraction of the $\left(\mathbf{x}_{full}^{(i)}, T_{full}^{(i)}\right)$ is selected as the training data and the remaining elements are used as test data to assess the performance of the developed NBC. Let the test data be $\left(\mathbf{x}_{test}^{(i)}, T_{test}^{(i)},\right), i = 1, 2, \cdots, N_{test}$. For each artificially generated scenario, two numerical metrics are calculated:

- CR_{CV}, calculated by LOOCV using the training data set, based on (10);
- CR_{test}, the classification correctness rate, calculated using the test data set:

$$CR_{test} = \frac{\sum_{i=1}^{N_{test}} \mathbb{1}\left(T_{pred}^{(i)} = T_{test}^{(i)}\right)}{N_{test}}, \tag{11}$$

The influence of the size of the training data set on the performance of trustworthiness assessment is also investigated. For this, nine levels of $N_{training}$, i.e., $N_{training} = 9, 18, 27, 36, 45, 54, 63, 72, 81$ are considered and the row-exchange algorithm in Matlab R2015b is used to design the training data collection scheme for each value of $N_{training}$.

Figure 7 and Figure 8 present the values of CR_{CV} and CR_{test}, evaluated under different scenarios, using different numbers of training data. It can be seen from the two Figures that both CR_{CV} and CR_{full} are improved as the number of training data $N_{training}$ increases. Hence, we can improve the accuracy of the developed classifier by choosing a larger $N_{training}$.

Although the classification accuracies are affected by uncertainties arising from the difference in the expert judgement behaviors adopted in different testing scenarios, the developed NBC can, in general, achieve satisfactory accuracies even for small values of $N_{training}$. As demonstrated in Figure 7: the average CR_{test} exceeds 0.9 for $N_{training}$ greater than 18. However, when $N_{training} = 9$, the classification accuracy is relatively poor. This is because when $N_{training} = 9$, the training data set is too small for accurate estimation of the posterior probabilities using (7) and (8). To avoid such a problem, it is suggested to ensure that $N_{training} \geq 18$ in practical applications.

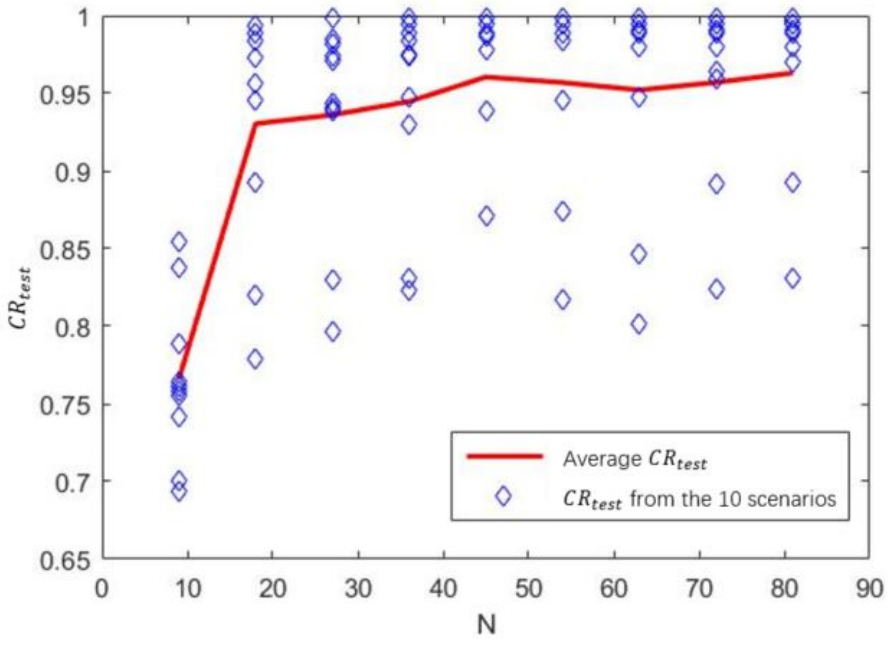


Fig. 7. CR_{test} evaluated under different scenarios, with different numbers of training data

A comparison is made between CR_{CV} and CR_{test} in Figure 9. It can be seen that although CR_{CV} is estimated by cross-validation using only the training data, it shows the same tendency as CR_{full}, which is estimated using the true data outside the training data set. Therefore, although sometimes in practical applications, we can only do cross-validation using the training data, a reasonable estimate of the classification accuracy can still be achieved.

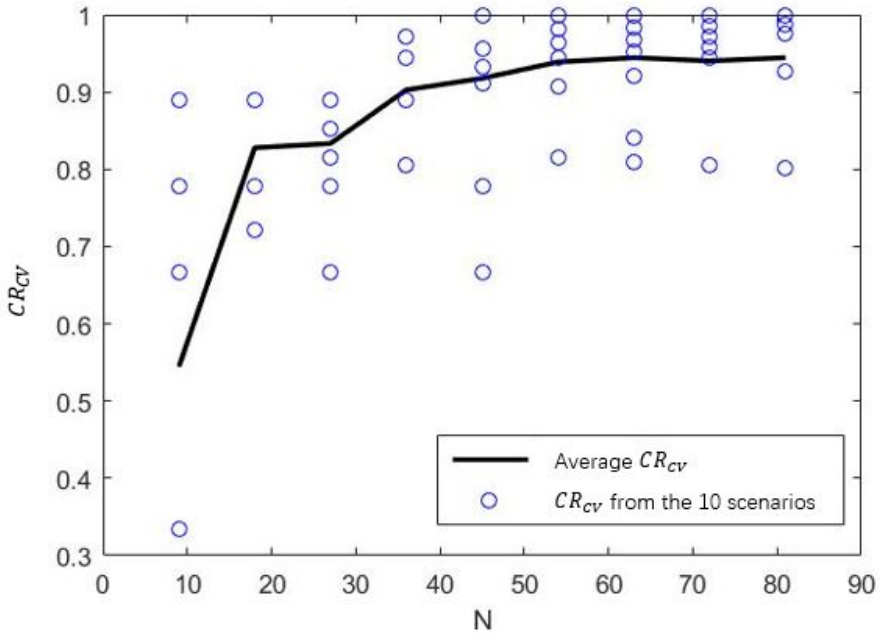


Fig. 8. CR_{CV} evaluated under different scenarios, with different numbers of training data

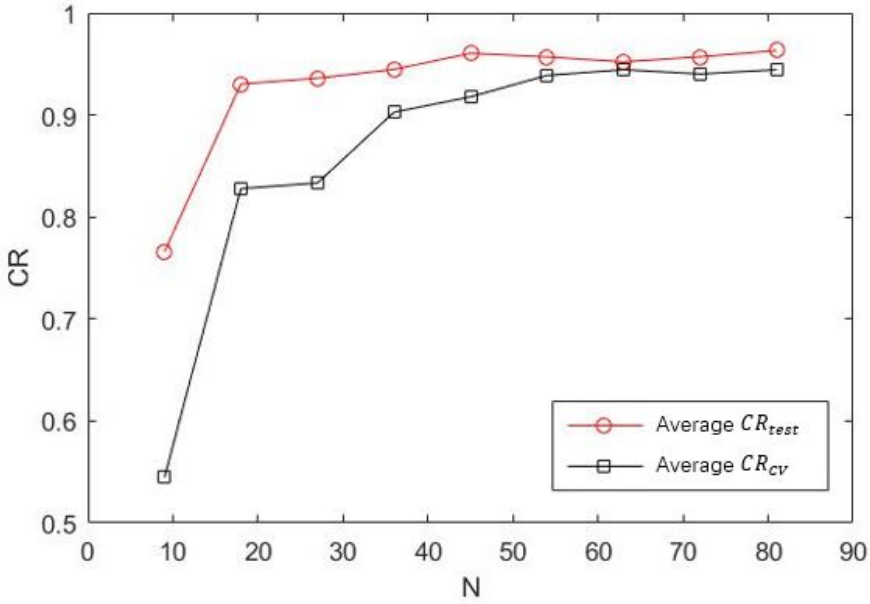


Fig. 9. A comparison of CR_{CV} and CR_{test}

B. Application

In this section, we show how to apply the developed framework to assess the trustworthiness of a real-world methanol plant, wherein the associated individual and social risks are assessed by a systematic QRA process, in terms of risk contours and F-N curve, respectively [43]. The training data used for the construction of the NBC are generated by asking an expert to assess the trustworthiness of a set of artificially generated "pseudo"

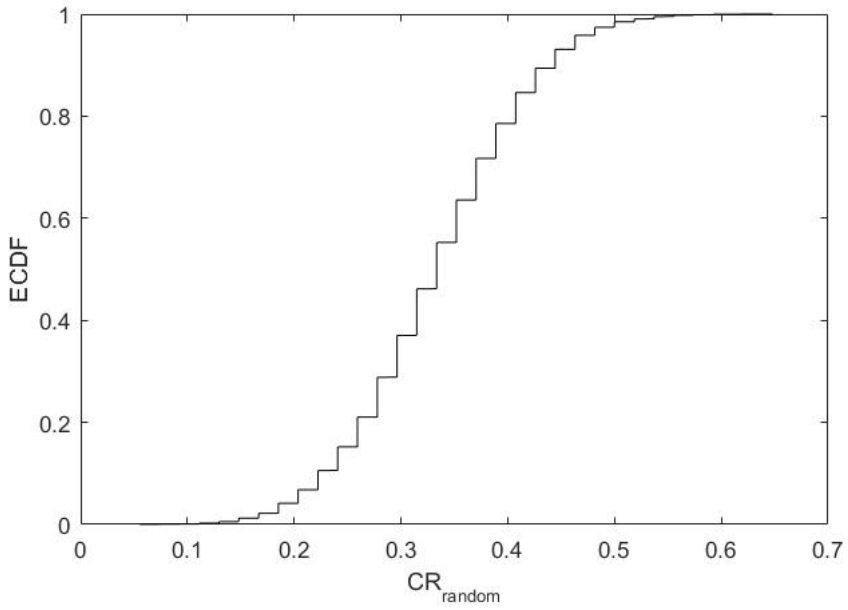


Fig. 10. ECDF of CR_{random}

QRAs. The quality criteria of the methanol QRA is evaluated by reviewing its final report, which is available online from [43]. The application follows the procedures in Figure 4 and the main results are summarized as follows.

1) Training data collection scheme: In this step, we design the training data collection scheme. From the discussion in Subsection V-A, we can see that $N_{training} = 54$ can, in general, yield good classification accuracy. Therefore, we choose $N_{training} = 54$. The row-exchange algorithm in Matlab R2015b is used to design the training data collection scheme. The resulting $\mathbf{x}^{(q)}, q = 1, 2, \cdots, N_{training}$ are listed in Table II. It can be verified that the training data collection scheme in Table II is an orthogonal design. The values of $\mathbf{x}^{(q)}, q = 1, 2, \cdots, N_{training}$ correspond to the levels of the quality criteria in Table A.1-A.6.

2) Training data collection: Each row in Table II represents a pseudo QRA, characterized by specific quality criteria. An expert is asked to assess the trustworthiness for these pseudo QRAs, for generating the training data. Take the first row in Table II as an example. To generate the training data, the expert is asked the following question: if the quality of a QRA process is as depicted in Table III, which level of trustworthiness in Table I do you think the QRA has? Table III is generated by relating the values of $\mathbf{x}^{(q)}, q = 1, 2, \cdots, N_{training}$ to the corresponding quality criteria in Table A.1-A.6. The procedures are repeated for the other rows in Table II. The training data generated by the expert are also listed in Table II.

3) Consistency verification: The consistency of the expert is checked using Algorithm 1. In this case, we choose $\alpha = 0.01$ and $n_{\text{random}} = 10^4$. The Empirical Cumulative Distributive Function (ECDF) of CR_{random} is presented in Figure 10, where $p_{1-\alpha} = 0.519$. The procedures in Figure 5 are used to calculate CR_{data} using the real training data. Since $CR_{data} = 0.852 > p_{1-\alpha}$, according to Algorithm 1, we can conclude that the expert provided the training data consistently under the significance level $\alpha = 0.01$.

TABLE II

TRAINING DATA

Runs	$x_1^{(q)}$	$x_2^{(q)}$	$x_3^{(q)}$	$x_4^{(q)}$	$x_5^{(q)}$	$x_6^{(q)}$	$T^{(q)}$	Runs	$x_1^{(q)}$	$x_2^{(q)}$	$x_3^{(q)}$	$x_4^{(q)}$	$x_5^{(q)}$	$x_6^{(q)}$	$T^{(q)}$
1	0	0	0	0	1	0	0	28	1	1	1	2	1	1	1
2	0	0	0	0	2	2	0	29	1	1	2	1	0	1	1
3	0	0	1	2	1	1	0	30	1	1	2	2	1	2	1
4	0	0	1	2	2	2	0	31	1	2	0	1	2	0	1
5	0	0	2	1	0	1	0	32	1	2	0	2	0	1	1
6	0	0	2	1	1	1	0	33	1	2	1	0	1	0	1
7	0	1	0	1	0	0	0	34	1	2	1	0	1	2	1
8	0	1	1	0	2	1	0	35	1	2	2	0	2	0	1
9	0	1	1	2	0	0	0	36	1	2	2	2	0	2	2
10	0	1	2	0	0	0	0	37	2	0	0	0	0	2	0
11	0	1	2	0	1	2	0	38	2	0	0	2	1	0	0
12	0	1	2	2	2	0	0	39	2	0	1	1	0	2	1
13	0	2	0	1	1	1	0	40	2	0	1	2	2	0	1
14	0	2	0	1	1	2	0	41	2	0	2	0	0	0	0
15	0	2	0	2	2	0	0	42	2	0	2	2	2	1	2
16	0	2	1	0	2	1	0	43	2	1	0	0	0	1	0
17	0	2	1	2	0	2	0	44	2	1	0	1	2	2	1
18	0	2	2	1	0	2	0	45	2	1	0	2	1	0	1
19	1	0	0	0	1	2	0	46	2	1	1	1	1	1	1
20	1	0	0	2	0	1	0	47	2	1	1	1	2	2	1
21	1	0	1	1	0	2	1	48	2	1	2	0	1	2	1
22	1	0	1	1	1	0	1	49	2	2	0	0	0	1	0
23	1	0	2	0	2	1	1	50	2	2	1	0	2	1	1
24	1	0	2	1	2	0	1	51	2	2	1	1	0	0	1
25	1	1	0	1	2	1	1	52	2	2	2	1	1	0	1
26	1	1	0	2	2	2	1	53	2	2	2	2	1	1	2
27	1	1	1	0	0	0	0	54	2	2	2	2	2	2	2

4) Classifier construction: The training data are used to construct the NBC, following the procedures in Figure 4. The estimated $Pr(T=k)$ and $Pr(x_i=j \mid T=k)$ are presented in Table IV and Figure 11, respectively. The accuracy of the constructed classifier is evaluated by the correct classification rate and we have $CR = 0.944$. Therefore, the constructed NBC can be used to represent the expert judgements and provide reasonable assessment of the trustworthiness of QRA.

The constructed NBC can also help to explain the expert's behavior in assessing the trustworthiness. For example, from Figure 11, we notice that $Pr(x_1=0 \mid T=0) = 0.6882$, $Pr(x_1=0 \mid T=1) = 0.0041$, $Pr(x_1=2 \mid T=0) = 0.0233$. From Bayes theorem,

TABLE III

QUALITY OF THE FIRST PSEUDO QRA

Criteria	Level
Completeness of documentation	Some the following elements are missing in the documentations: • scopes and objectives of the QRA; • descriptions of the system under investigation and related references; • accounts of the adopted analysis methods; • presentation of source data needed for the analysis; • report of the analysis results.
Understanding of problem settings	The analysts are unaware of the problem settings of the QRA due to the presence of all the following flaws: • the purposes of the QRA are not clearly understood; • the systems of interests are not well defined; • the resources constraints (e.g., time, computational resources, etc) are not clearly defined.
Coverage of accident scenarios	Some critical accident scenarios are highly likely to be missed by the identification process: • the coverage of the identified accident scenarios is not validated; • the validation shows that some critical accident scenarios might be missing.
Appropriateness of analysis methods	• The features of the selected analysis method satisfy the requirements of the problem and • successful applications in similar problems can justify the choice of the method.
Quality of input data	• There is no sufficient statistical data and the input data is purely based on expert judgements; • epistemic uncertainty in the expert-generated input data is not considered.
Accuracy of risk calculation	• Only errors from the calculation process itself (e.g., the accuracy of Monte Carlo simulations) might exist and • the uncertainties caused by the errors are properly modeled.

TABLE IV

ESTIMATED $Pr\,(T=k)$

k	$Pr\,(T=k)$
0	0.4807
1	0.4438
2	0.0755

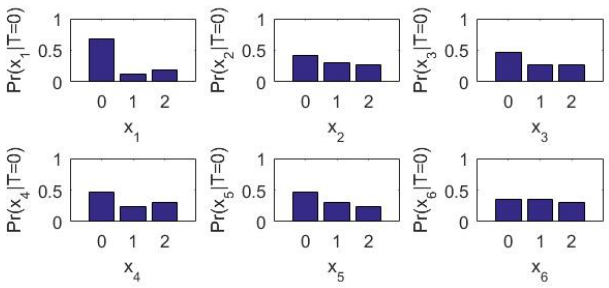
$$Pr\,(T=0\,|\,x_1=0) = \frac{Pr\,(x_1=0\,|\,T=0)\cdot Pr(T=0)}{Pr(x_1=0)}$$

$$= \frac{Pr\,(x_1=0\,|\,T=0)\cdot Pr(T=0)}{\sum_{i=1}^{3} Pr\,(x_1=0\,|\,T=i)\cdot Pr(T=i)} \tag{12}$$

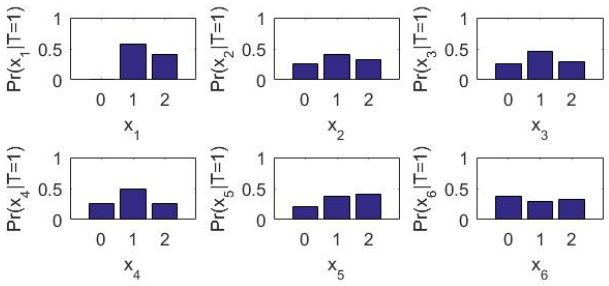
$$= 0.9896 \tag{13}$$

That is, if x_1 equals to zero, the expert tends to judge the QRA as unreliable. This is a natural result, since x_1 denotes the completeness of documentations. If the QRA process is not well-documented, it is unlikely to be repeatable: therefore, the associated QRA is unreliable according to the criteria in Table I.

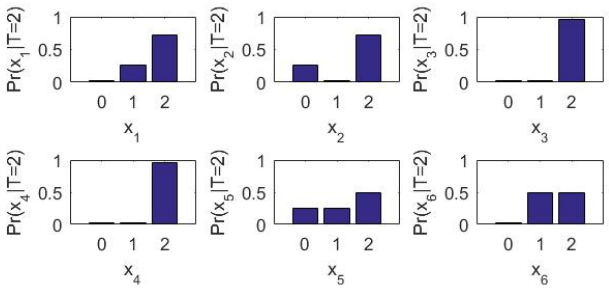
5) Comparison to existing methods: In traditional proactive trustworthiness assessment methods, e.g., [32], expert knowledge is elicited to develop a simple conformance/non-conformance-based framework that relates the quality criteria to the trustworthiness of the QRA. That is, the conclusion of whether the QRA is trustworthy



(a) $T = 0$



(b) $T = 1$



(c) $T = 2$

Fig. 11. Estimated $Pr\left(x_i = j \mid T = k\right)$

or not is made by comparing the number of the conformed quality criteria to a predefined threshold value n_{th}. In this paper, we assume that a quality criterion i is conformed when $x_i = 2$. Table V shows a comparison between the classification-based framework and the conformance/non-conformance-based framework, using the training data in Table II. It can be seen that in general, the existing conformance/non-conformance-based framework cannot accurately model the complex expert knowledge expressed in the empirical data in Table II. The developed method, on the other hand, is capable of capturing the complex behavior of expert judgement in assessing the trustworthiness of the QRA.

6) Trustworthiness assessment: To assess the trustworthiness of the methanol QRA using the developed NBC, its six quality criteria are first evaluated based on the QRA report [43] and following the scaling rules in Table A.1-A.6:

- The scaling rule for completeness of documentation (x_1) is listed in Table A.1. In general, the methanol QRA report contains sufficient information on the scope and objective of the analysis, the system under

TABLE V

A COMPARISON TO EXISTING METHODS

Methods	Correct classification rate
Classification-based method	$CR = 0.944$
Conformance-based method	$CR = 0.130$, for $n_{th} = 0$
	$CR = 0.315$, for $n_{th} = 1$
	$CR = 0.463$, for $n_{th} = 2$
	$CR = 0.556$, for $n_{th} = 3$
	$CR = 0.500$, for $n_{th} = 4$
	$CR = 0.500$, for $n_{th} = 5$
	$CR = 0.482$, for $n_{th} = 6$

investigation and the adopted analysis methods. However, according to Table A.1, the presentation of the analysis results is incomplete, since no accounts of uncertainty are given in the report. Therefore, we have $x_1 = 1$.

- The scaling rule for understanding of problem setting (x_2) is listed in Table A.2. As reflected in the report, the analyst understood well the purposes of the QRA, the system under investigation and the resources constraints of the analysis. Therefore, we have $x_2 = 2$.
- The scaling rule for coverage of accident scenario (x_3) is listed in Table A.3. The methanol QRA identifies accident scenarios by an initial HAZard IDentification (HAZID) workshop. The identified accident scenarios are verified by expert reviews, conducted by experts from the QRA analysis team, the operator of the methanol plant and other organizations in related fields [43]. However, no field data are used to verify the accident scenarios. Therefore, we have $x_3 = 1$.
- The scaling rule for appropriateness of analysis method (x_4) is listed in Table A.4. The QRA is mainly based on Event Tree Analysis (ETA), which assumes that the occurrence probabilities of the intermediate events dop not change with time. In practice, however, the safety barriers of the methanol plant might be time-dependent due to the degradation of critical components or systems. ETA is not able to capture such time-dependence. Therefore, we have $x_4 = 0$.
- The scaling rule for quality of input data (x_5) is listed in Table A.5. The input data used in the QRA come from expert judgements based on handbooks (parts count reliability prediction standards and TNO "purple book" [43]). Epistemic uncertainty in the expert judgements on the input data is not considered. Therefore, we have $x_5 = 0$.
- The scaling rule for accuracy of risk calculation (x_6) is listed in Table A.6. The risk calculation is based on ETA, which is an analytical method. Therefore, its accuracy can be ensured and we have $x_6 = 2$.

By running the NBC with the input feature vector $\mathbf{x} = [1, 2, 1, 0, 0, 2]$, we can calculate the posterior probabilities from (4), as shown in Figure 12. We can conclude that $T = 1$ for the QRA of the methanol plant, which means, according to Table I, that the QRA of the methanol plant is reliable but invalid. Such a QRA can be used to support decision making, but not for safety-critical decisions.

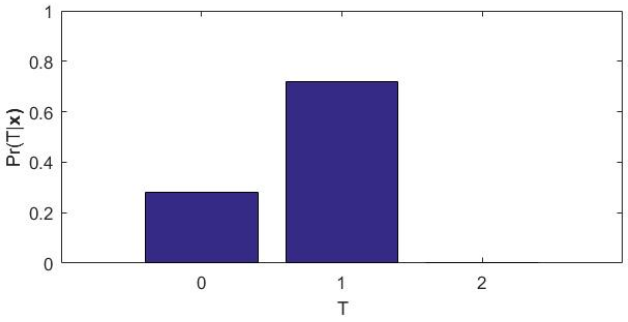


Fig. 12. Posterior probabilities for each value of T

VI. Conclusion

In this paper, a classification-based framework is developed for the trustworthiness assessment of QRA. In the developed framework, trustworthiness is assessed in terms of six criteria, i.e., completeness of documentations, understanding of problem settings, coverage of accident scenarios, appropriateness of analysis methods, quality of input data, accuracy of risk calculation. For each criteria, three levels are distinguished and corresponding scaling rules are presented for the assessment. The trustworthiness of QRA is also divided into three discrete levels, based on its reliability and validity. A Naive Bayes Classifier (NBC) is constructed for trustworthiness assessment given the values of the six criteria. The developed NBC is able to learn from the training data the expert's behavior when assessing the trustworthiness. Therefore, once accurately constructed, the NBC is able to reasonably approximate the expert's assessment of the trustworthiness. A stochastic hypothesis testing-based approach is also developed to check the consistency of the training data. The performance of the developed methods are tested by ten simulated case studies. The results demonstrate that the developed framework can accurately approximate various expert assessment behaviors. Finally, a real application has been considered.

An inherent assumption of NBC is that the features (criteria) are independent among one another. In practice, however, various dependencies might exist among the features. How to consider these dependencies needs to be explored in future researches in order to derive a more accurate classifier. Furthermore, uncertainties might affect the evaluation of the criteria $x_1, x_2, \cdots, x_6$, since their assessments involve a lot of subjective judgements. Future researches can be carried out to integrate these uncertainties in the assessment framework, with mathematical theories dealing with subjective uncertainties, e.g., evidence theory [44], possibility theory [45], uncertainty theory [46], etc.

References

[1] G. E. Apostolakis, "How useful is quantitative risk assessment?," *Risk analysis*, vol. 24, no. 3, pp. 515–520, 2004.

[2] E. Zio and T. Aven, "Industrial disasters: Extreme events, extremely rare. some reflections on the treatment of uncertainties in the assessment of the associated risks," *Process Safety and Environmental Protection*, vol. 91, no. 1, pp. 31–45, 2013.

[3] N. Khakzad, G. Reniers, R. Abbassi, and F. Khan, "Vulnerability analysis of process plants subject to domino effects," *Reliability Engineering & System Safety*, vol. 154, pp. 127–136, 2016.

[4] M. Abimbola, F. Khan, and N. Khakzad, "Risk-based safety analysis of well integrity operations," *Safety Science*, vol. 84, pp. 149–160, 2016.

[5] F. Goerlandt and J. Montewka, "Maritime transportation risk analysis: Review and analysis in light of some foundational issues," *Reliability Engineering & System Safety*, vol. 138, pp. 115–134, 2015.

[6] E. Zio and N. Pedroni, "Estimation of the functional failure probability of a thermalhydraulic passive system by subset simulation," *Nuclear Engineering and Design*, vol. 239, no. 3, pp. 580–599, 2009.

[7] F. Khan, S. J. Hashemi, N. Paltrinieri, P. Amyotte, V. Cozzani, and G. Reniers, "Dynamic risk management: a contemporary approach to process safety management," *Current Opinion in Chemical Engineering*, no. 14, pp. 9–17, 2016.

[8] F. Khan, S. Rathnayaka, and S. Ahmed, "Methods and models in process safety and risk management: Past, present and future," *Process Safety and Environmental Protection*, vol. 98, pp. 116–147, 2015.

[9] F. I. Khan and S. A. Abbasi, "Opthazopan effective and optimum approach for hazop study," *Journal of Loss Prevention in the Process Industries*, vol. 10, no. 3, pp. 191–204, 1997.

[10] H. Yu, F. Khan, and B. Veitch, "A flexible hierarchical bayesian modeling technique for risk analysis of major accidents," *Risk analysis*, 2017.

[11] N. Khakzad, F. Khan, and P. Amyotte, "Dynamic safety analysis of process systems by mapping bow-tie into bayesian network," *Process Safety and Environmental Protection*, vol. 91, no. 1-2, pp. 46–53, 2013.

[12] T. Rosqvist, "On the validation of risk analysis: A commentary," *Reliability Engineering & System Safety*, vol. 95, no. 11, pp. 1261–1265, 2010.

[13] A. Rae, R. Alexander, and J. McDermid, "Fixing the cracks in the crystal ball: A maturity model for quantitative risk assessment," *Reliability Engineering & System Safety*, vol. 125, pp. 67–81, 2014.

[14] H. J. Pasman, S. Jung, K. Prem, W. J. Rogers, and X. Yang, "Is risk analysis a useful tool for improving process safety?," *Journal of Loss Prevention in the Process Industries*, vol. 22, no. 6, pp. 769–777, 2009.

[15] J. Suokas and V. Rouhiainen, "Quality control in safety and risk analysis," *Journal of Loss Prevention in Process Industries*, vol. 2, pp. 67–77, 1989.

[16] T. Aven and E. Zio, "Foundational issues in risk assessment and risk management," *Risk Analysis*, vol. 34, no. 7, pp. 1164–1172, 2014.

[17] F. Goerlandt and P. Kujala, "On the reliability and validity of shipship collision risk analysis in light of different perspectives on risk," *Safety Science*, vol. 62, pp. 348–365, 2014.

[18] F. Goerlandt, N. Khakzad, and G. Reniers, "Validity and validation of safety-related quantitative risk analysis: A review," *Safety Science*, 2016.

[19] J. Graham, "Verifiability isn't everything," *Risk Analysis*, vol. 15, p. 109, 1995.

[20] J. S. Busby, R. E. Alcock, and E. J. Hughes, "Credibility in risk assessment," in *Probabilistic Safety Assessment and Management*, Probabilistic Safety Assessment and Management, pp. 2809–2814, Springer, 2004.

[21] K. Lauridsen, M. Christou, A. Amendola, F. Markert, I. Kozine, and M. Fiori, "Assessing the uncertainties in the process of risk analysis of chemical establishments: Part 1 and 2," in *Proceedings of European*

Conference on Safety and Reliability (ESREL), European Conference on Safety and Reliability (ESREL), (Torino, Italy), pp. 592–606, 2001.

[22] D. Sornette, T. Maillart, and W. Krger, "Exploring the limits of safety analysis in complex technological systems," *International Journal of Disaster Risk Reduction*, vol. 6, pp. 59–66, 2013.

[23] B. J. Garrick, "On pra quality and use," tech. rep., School of Engineering and Applied Science, University of Colifornia at Los Angeles, 1982.

[24] T. Aven and E. Zio, "Model output uncertainty in risk assessment," *International Journal of Performability Engineering*, vol. 9, no. 5, pp. 475–486, 2013.

[25] T.-R. Wang, N. Pedroni, and E. Zio, "Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent acts: a sensitivity-based decision-making approach," *Reliability Engineering & System Safety*, vol. 147, pp. 9–18, 2016.

[26] T. Aven and B. Heide, "Reliability and validity of risk analysis," *Reliability Engineering & System Safety*, vol. 94, no. 11, pp. 1862–1868, 2009.

[27] P. A. Fenner-Crisp and V. L. Dellarco, "Key elements for judging the quality of a risk assessment," *Environmental health perspectives*, 2016.

[28] V. Rouhiainen, "Importance of the quality management of safety analysis," *Reliability Engineering & System Safety*, vol. 40, no. 1, pp. 5–16, 1993.

[29] A. Pinto, R. A. Ribeiro, and I. L. Nunes, "Ensuring the quality of occupational safety risk assessment," *Risk analysis*, vol. 33, no. 3, pp. 409–419, 2013.

[30] E. Vergison, "A quality-assurance guide for the evaluation of mathematical models used to calculate the consequences of major hazards," *Journal of Hazardous Materials*, 1996.

[31] V. Rouhiainen, "Quasa: A method for assessing the quality of safety analysis," *Safety Science*, vol. 15, no. 3, pp. 155–172, 1992.

[32] "Regulatory guide 1.200: An approach for determining the technical adequacy of probabilistic risk assessment results for risk-informed activities," tech. rep., US Nuclear Regulatory Commission, 2009.

[33] T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio, "Assessing the performance of a classificationbased vulnerability analysis model," *Risk Analysis*, vol. 35, no. 9, pp. 1674–1689, 2015.

[34] C. Bishop, *Pattern Recognition and Machine Learning*. London: Springer, 2006.

[35] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752 of *AAAI-98 workshop on learning for text categorization*, pp. 41–48, Citeseer, 1998.

[36] I. Kononenko, "Inductive and bayesian learning in medical diagnosis," *Applied Artificial Intelligence an International Journal*, vol. 7, no. 4, pp. 317–337, 1993.

[37] T.-R. Wang, V. Mousseau, N. Pedroni, and E. Zio, "An empirical classification-based framework for the safety criticality assessment of energy production systems, in presence of inconsistent data," *Reliability Engineering & System Safety*, vol. 157, pp. 139–151, 2017.

[38] Z. Zeng, F. Di Maio, E. Zio, and R. Kang, "A hierarchical decision making framework for the assessment of the prediction capability of prognostic methods," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 2016. Accepted.

[39] "Naive bayes, guassian distributions and practical applicationss." http://http://www.cs.cmu.edu/~tom/10601_sp09/lectures/NBayes2_2-2-2009-ann.pdf. Accessed: 2016-10-24.

[40] T. L. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of mathematical psychology*, vol. 15, no. 3, pp. 234–281, 1977.

[41] Y. Chen, Z. Zeng, and R. Kang, "Validation methodology for distribution-based degradation model," *Systems Engineering and Electronics, Journal of*, vol. 23, no. 4, pp. 553–559, 2012.

[42] J. L. Devore, *Probability and statistics for engineering and the sciences*. Boston, MA :: Cengage Learning,, 8th ed. ed., 2010.

[43] "Quantitative risk assessment: Final report (prepared for nw innovation works, by acutech consulting group)." http://http://kalamamfgfacilitysepa.com/wp-content/uploads/2016/09/FEIS-Appendix-G1-Quantitative-Risk-Assessment.pdf. Accessed: 2016-11-08.

[44] G. Shafer, *A mathematical theory of evidence*, vol. 1. Princeton university press Princeton, 1976.

[45] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, pp. 3–28, 1978.

[46] B. Liu, *Uncertainty Theory, Fifth Edition*. Springer, 2015.

APPENDIX

SCALING RULES FOR x_1-x_6

TABLE A.1

SCALING RULES FOR x_1

Levels	Descriptions
$x_1 = 0$	Some the following elements are missing in the documentations: • scopes and objectives of the QRA; • descriptions of the system under investigation and related references; • accounts of the adopted analysis methods; • presentation of source data needed for the analysis; • report of the analysis results.
$x_1 = 1$	At least one of the following flaws present in the documentations: • descriptions of scopes and objectives are incomplete or ambiguous; • descriptions of the system under investigation are unclear; • no sufficient references on the system under investigation are given; • descriptions of the adopted methods are unclear; • presentations of the results are incomplete (e.g., no uncertainty is considered) or ambiguous.
$x_1 = 2$	The documentation of the QRA process contains sufficient information for its repetition: • the documentation contains all the necessary parts; • no flows in level $x_1 = 1$ present.

TABLE A.2

SCALING RULES FOR x_2

Levels	Descriptions
$x_2 = 0$	The analysts are unaware of the problem settings of the QRA due to the presence of all the following flaws: • the purposes of the QRA are not clearly understood; • the systems of interests are not well defined; • the resources constraints (e.g., time, computational resources, etc) are not clearly defined.
$x_2 = 1$	The analysts misunderstand part of the problem settings due to some of the following flaws: • the analysts fail to clearly understand the purposes of the QRA; • the analysts fail to clearly understand the systems of interests; • the analysts fail to clearly understand the resources constraints.
$x_2 = 2$	The analysts clearly understand the problem settings: no flaws in levels $x_2 = 0$ and $x_2 = 1$ occur.

TABLE A.3

SCALING RULES FOR x_3

Levels	Descriptions
$x_3 = 0$	Some critical accident scenarios are highly likely to be missed by the identification process: • the coverage of the identified accident scenarios is not validated; • the validation shows that some critical accident scenarios might be missing.
$x_3 = 1$	• Validation reveals that most critical accident scenarios are covered by the identification process but • the validation is conducted based on peer review rather than using real data.
$x_3 = 2$	• Validation reveals that most critical accident scenarios are covered by the identification process and • the validation is reliable by using field data.

TABLE A.4

SCALING RULES FOR x_4

Levels	Descriptions
$x_4 = 0$	The features of the selected analysis method cannot satisfy the requirements of the problem: • the methods cannot capture some features of the problem (e.g., time-dynamics, dependencies, etc.) or • the methods require more resources (e.g., data, computational power, etc.) than that can be provided.
$x_4 = 1$	• The features of the selected analysis method satisfy the requirements of the problem but • the conclusion is drawn based on expert experience.
$x_4 = 2$	• The features of the selected analysis method satisfy the requirements of the problem and • successful applications in similar problems can justify the choice of the method.

TABLE A.5

SCALING RULES FOR x_5

Levels	Descriptions
$x_5 = 0$	• There is no sufficient statistical data and the input data is purely based on expert judgements; • epistemic uncertainty in the expert-generated input data is not considered.
$x_5 = 1$	• There is no sufficient statistical data; • the input data are based on expert judgements with fully consideration of epistemic uncertainty.
$x_5 = 2$	• Sufficient statistical data can be used for risk analysis.

TABLE A.6

SCALING RULES FOR x_6

Levels	Descriptions
$x_6 = 0$	• The process of risk calculation contains major flaws; • large errors might exist in the calculated risks.
$x_6 = 1$	• The process of risk calculation does not contain major flaws; • only errors from the calculation process itself (e.g., the accuracy of Monte Carlo simulations) might exist but • the uncertainties caused by the errors are not modeled.
$x_6 = 2$	• Only errors from the calculation process itself (e.g., the accuracy of Monte Carlo simulations) might exist and • the uncertainties caused by the errors are properly modeled.