Knowledge-Based Systems 88 (2015) 264–272

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A data-driven approximate causal inference model using the evidential reasoning rule

Yue Chen [a], Yu-Wang Chen [b,*], Xiao-Bin Xu [c], Chang-Chun Pan [a], Jian-Bo Yang [b], Gen-Ke Yang [a]

[a] Department of Automation, Shanghai Jiao Tong University, China
[b] Manchester Business School, The University of Manchester, Manchester M15 6PB, UK
[c] Institute of System Science and Control Engineering, School of Automation, Hangzhou Dianzi University, Hangzhou, China

ARTICLE INFO

Article history:
Received 5 January 2015
Received in revised form 17 May 2015
Accepted 21 July 2015
Available online 29 July 2015

Keywords:
Evidential reasoning
Bayesian inference
Belief distribution
Approximate causal inference

ABSTRACT

This paper aims to develop a data-driven approximate causal inference model using the newly-proposed evidential reasoning (ER) rule. The ER rule constitutes a generic conjunctive probabilistic reasoning process and generalises Dempster's rule and Bayesian inference. The belief rule based (BRB) methodology was developed to model complicated nonlinear causal relationships between antecedent attributes and consequents on the basis of the ER algorithm and traditional IF-THEN rule-based systems, and in essence it keeps methodological consistency with Bayesian Network (BN). In this paper, we firstly introduce the ER rule and then analyse its inference patterns with respect to the bounded sum of individual support and the orthogonal sum of collective support from multiple pieces of independent evidence. Furthermore, we propose an approximate causal inference model with the kernel mechanism of data-based approximate causal modelling and optimal learning. The exploratory approximate causal inference model inherits the main strengths of BN, BRB and relevant techniques, and can potentially extend the boundaries of applying approximate causal inference to complex decision and risk analysis, system identification, fault diagnosis, etc. A numerical study on the practical pipeline leak detection problem demonstrates the applicability and capability of the proposed data-driven approximate causal inference model.

1. Introduction

The evidential reasoning (ER) rule has been established recently to combine multiple pieces of independent evidence conjunctively with weights and reliabilities [34]. Through the implementation of the orthogonal sum operation on weighted belief distributions with reliabilities, the ER rule takes into account both individual and collective support from two pieces of evidence in a rational way, and it constitutes a generic conjunctive probabilistic reasoning process or a generalised Bayesian inference process [34,35]. In inheritance of the basic probabilistic properties of being associative and commutative, the ER rule can be easily used to aggregate multiple pieces of evidence recursively. It is expected that the ER rule can further extend the boundaries of existing Bayesian inference methodology and provide a scientific way of reasoning with various probabilistic uncertainties. The ER rule advances the seminal Dempster–Shafer (D–S) theory of evidence [7,16] and the original ER algorithm [29,33]. It has been proved that (1) Dempster's combination rule is a special case of the ER rule when each piece of evidence is fully reliable, and (2) the ER algorithm is also a special case when the reliability of each piece of evidence is assumed to be equal to its normalised weight. In a theoretical sense, the reliability of each piece of evidence is used to measure its inherent quality of the information source, and contrarily the normalised weight reflects its relative importance compared with other pieces of evidence [34]. Previously it was widely accepted that the D–S theory of evidence is one of the most prominent work to generalise Bayesian inference, which consists of a rigorous probabilistic reasoning process [16,34]. In the D–S theory, a frame of discernment is defined by a set of mutually exclusive and collectively exhaustive hypotheses. It is assumed that basic probabilities can be assigned to not only singleton hypotheses but also any subsets of hypotheses. As a result, each piece of evidence is profiled by a belief distribution on the power set of the frame of discernment. Correspondingly, belief distribution is a generalisation of conventional probability distribution in which basic probabilities are only assigned to singleton hypothesis. However, when combining highly or completely conflicting evidence, Dempster's rule combination was found to generate counter-intuitive results [36,27].

* Corresponding author. Tel.: +44 161 2756345.
E-mail address: yu-wang.chen@mbs.ac.uk (Y.-W. Chen).

http://dx.doi.org/10.1016/j.knosys.2015.07.026
0950-7051/

Thereafter, much work has been undertaken to resolve the issue and design new combination rules [18,9,28]. The ER algorithm was originally presented in the context of multiple criteria decision analysis. The holistic approach consists of the belief structure for modelling various types of uncertainty [31,25], the rule and utility based information transformation techniques [29], and the ER algorithm for information aggregation [33], etc. In the past twenty years, the ER algorithm has been widely applied to many system and decision analysis problems as surveyed by Xu [25]. It has also been extended to multi-criteria fuzzy decision-making problems [22,23], fuzzy failure mode and effects analysis [13], rule-base evidential reasoning [8], group decision analysis [11], medical diagnosis [24], and so on. Furthermore, the ER algorithm was introduced to extend traditional If-Then rule based systems to belief rule based (BRB) systems [30]. The BRB methodology employs the informative belief structure to represent various types of information and knowledge with uncertainties and shows superior capability of approximating complicated nonlinear causal relationships across a wide variety of application areas, including fault diagnosis, system identification, risk and decision analysis [26,1,37,6,4].

Given that the ER rule has explicitly generalised the D–S theory of evidence and the original ER algorithm, it becomes perfectly logical and also extremely important to revisit and further improve those techniques which were previously developed from the latter two methods. In this paper, we aim to conduct some exploratory research of building a data-driven approximate causal inference model using the ER Rule and sharpening the edges of the ER and BRB methodologies. The rest of the paper is organised as follows: in Section 2, the inference patterns of the ER rule with respect to the bounded sum of individual support and the orthogonal sum of collective support from multiple pieces of evidence are analysed on the basis of its fundamentals. In Section 3, an approximate causal inference model using the ER rule is explored in view of data-based causal modelling and optimal learning. A numerical study is conducted to illustrate the applicability of the proposed data-driven approximate causal inference model in Section 4. Some concluding remarks are presented in Section 5.

2. The ER rule for inference

2.1. Brief introduction of the ER rule

Suppose a frame of discernment $\Theta = \{\theta_1, \ldots, \theta_N\}$ is a set of mutually exclusive and collectively exhaustive hypotheses, with $\theta_n \cap \theta_m = \emptyset$ for any $n, m \in \{1, \ldots, N\}$ and $n \neq m$ where $\emptyset$ is an empty set. The power set of Θ, denoted by $P(\Theta)$ or 2^{Θ}, consists of 2^N subsets of Θ as follows

$$P(\Theta) = 2^{\Theta} = \{\emptyset, \theta_1, \ldots, \theta_N, \{\theta_1, \theta_2\}, \ldots, \{\theta_1, \theta_N\}, \ldots, \{\theta_1, \theta_{N-1}\}, \Theta\} \tag{1}$$

In the ER rule, a piece of evidence e_i is profiled by the following belief distribution.

$$e_i = \left\{ (\theta, p_{\theta,i}), \forall \theta \subseteq \Theta, \sum_{\theta \subseteq \Theta} p_{\theta,i} = 1 \right\} \tag{2}$$

where $p_{\theta,i}$ denotes the degree of belief to which the evidence e_i supports proposition θ being any element of $P(\Theta)$ except for the empty set. $(\theta, p_{\theta,i})$ is referred to as a focal element of e_i if $p_{\theta,i} > 0$. Specifically, the degree of belief assigned exactly to the complete set Θ reflects the degree of global ignorance, and to a smaller subset of Θ except for any singleton proposition measures the degree of local ignorance. If there is no local or global ignorance, the belief distribution reduces to a classical probability distribution [34].

Each piece of evidence e_i is also associated with a weight and a reliability, denoted by w_i and r_i respectively. It is worth noting that weight and reliability are not differentiated clearly in many information aggregation methods [17,34]. In the ER framework, the weight is used to reflect the relative importance of a piece of evidence in comparison with other evidence, and nevertheless the reliability is the inherent property of the evidence and sets the degree of support for a proposition.

As a result, there are mainly three elements to be taken into account when combining a piece of evidence with other evidence: its belief distribution, weight and reliability. The reasoning process in the ER rule is achieved by defining a weighted belief distribution with reliability [34].

$$m_i = \left\{ \left(\theta, \widetilde{m}_{\theta,i}\right), \forall \theta \subseteq \Theta; \left(P(\Theta), \widetilde{m}_{P(\Theta),i}\right) \right\} \tag{3}$$

where $\widetilde{m}_{\theta,i}$ measures the degree of support for θ from e_i with taking into account all the three elements.

$$\widetilde{m}_{\theta,i} = \begin{cases} 0, & \theta = \emptyset \\ c_{rw,i} m_{\theta,i}, & \theta \subseteq \Theta, \theta \neq \emptyset \\ c_{rw,i}(1 - r_i), & \theta = P(\Theta) \end{cases} \text{ or } \widetilde{m}_{\theta,i} = \begin{cases} 0, & \theta = \emptyset \\ \widetilde{w}_i p_{\theta,i}, & \theta \subseteq \Theta, \theta \neq \emptyset \\ 1 - \widetilde{w}_i, & \theta = P(\Theta) \end{cases} \tag{4}$$

where $m_{\theta,i} = w_i p_{\theta,i}$ and $c_{rw,i} = 1/(1 + w_i - r_i)$. The normalisation factor $c_{rw,i}$ determines $\sum_{\theta \subseteq \Theta} \widetilde{m}_{\theta,i} + \widetilde{m}_{P(\Theta),i} = 1$. Implicitly, a new hybrid weight $\widetilde{w}_i = c_{rw,i} w_i = w_i/(1 + w_i - r_i)$ is used to calculate $\widetilde{m}_{\theta,i}$ from the original belief degree $p_{\theta,i}$, and $\widetilde{m}_{P(\Theta),i} = 1 - \widetilde{w}_i$. The residual support $\widetilde{m}_{P(\Theta),i} = 0$, when $r_i = 1$.

Given the definition of the weighted belief distribution with reliability, the new ER rule can then be used to combine multiple pieces of evidence recursively. Without loss of generality, the combined degrees of belief to which two pieces of independent evidence e_i and e_j jointly support proposition θ, denoted by $p_{\theta,e(2)}$, can be generated by the orthogonal sum of the weighted belief distributions with reliability (i.e., m_i and m_j) as follows

$$p_{\theta,e(2)} = \begin{cases} 0, & \theta = \emptyset \\ \dfrac{\widehat{m}_{\theta,e(2)}}{\sum_{D \subseteq \Theta} \widehat{m}_{D,e(2)}}, & \theta \subseteq \Theta, \theta \neq \emptyset \end{cases} \tag{5}$$

$$\widehat{m}_{\theta,e(2)} = \left[(1 - r_j) m_{\theta,i} + (1 - r_i) m_{\theta,j}\right] + \sum_{B \cap C = \theta} m_{B,i} m_{C,j}, \quad \forall \theta \subseteq \Theta \tag{6}$$

There are mainly two terms in the equation above. The first square bracket term is regarded as the bounded sum of individual support on proposition θ from each of the two pieces of evidence e_i and e_j. $(1 - r_i)$ reflects the unreliability of evidence e_i, and it sets a bound within which e_j can play a limited role. Here we take two extreme cases as examples. When evidence e_i is fully reliable, i.e., $(1 - r_i) = 0, (1 - r_i) m_{\theta j} = 0$ and the individual support from evidence e_j will not be counted at all. When evidence e_i is fully unreliable, i.e., $(1 - r_i) = 1, (1 - r_i) m_{\theta j} = m_{\theta j}$ and the individual support from evidence e_j will be counted completely. The second term is regarded as the orthogonal sum of collective support from both pieces of evidence e_i and e_j, measuring the degree of all intersected support on proposition θ.

2.2. Inference analysis of the ER rule

As introduced previously, the ER rule generalises a few special cases, which can essentially be characterised by the three elements of evidence. Firstly, the ER rule reduces to Bayesian inference given that each piece of evidence is formulated by a probability distribution, or a so-called belief distribution without local or global ignorance. Secondly, with regard to evidence weight and reliability, there are two possible scenarios: (i) The ER rule turns into the

original ER algorithm, when the reliability of evidence is equal to its weight which is usually normalised to reflect its relative importance. However, the relative importance of each piece of evidence is often influenced by the decision maker's subjective preference; (ii) the ER rule reduces to Dempster's rule, when each piece of evidence is regarded as fully reliable, mathematically the bounded sum of individual support equals to zero. The inference mechanism of these special cases has been widely analysed in previous research [33,5]. Whereas, the ER rule explicitly reveals that the combined degree of joint support on a proposition consists of two parts: the bounded sum of individual support and the orthogonal sum of collective support from multiple pieces of independent evidence [34]. Thus it is useful to analyse how the two parts play a role in the reasoning process and in what capacity under different cases.

First of all, suppose that each piece of evidence e_i is profiled by the following belief distribution with global ignorance only, and its reliability is equal to its weight w_i.

$$e_i = \{(\theta_n, p_{n,i}), n = 1, \ldots, N; (\Theta, p_{\Theta,i})\} \tag{7}$$

where $0 \leqslant p_{n,i} \leqslant 1 (n = 1, \ldots, N), \sum_{n=1}^{N} p_{n,i} \leqslant 1$ and $p_{\Theta,i} = 1 - \sum_{n=1}^{N} p_{n,i}$ being the degree of global ignorance.

The combined degrees of belief $p_{n,e(2)}$ and $p_{\Theta,e(2)}$ from two pieces of independent evidence e_i and e_j can be reduced from Eqs. (5) and (6) as follows

$$p_{n,e(2)} = k_1 \hat{m}_{n,e(2)}, \ n = 1, \ldots, N \quad \text{and} \quad p_{\Theta,e(2)} = k_1 \hat{m}_{\Theta,e(2)} \tag{8a}$$

$$k_1 = \left(\sum_{n=1}^{N} \hat{m}_{n,e(2)} + \hat{m}_{\Theta,e(2)} \right)^{-1} \tag{8b}$$

$$\hat{m}_{n,e(2)} = \left[(1 - w_j) m_{n,i} + (1 - w_i) m_{n,j} \right] + \left[m_{n,i} m_{n,j} + m_{n,i} m_{\Theta,j} + m_{\Theta,i} m_{n,j} \right] \tag{8c}$$

$$\hat{m}_{\Theta,e(2)} = \left[(1 - w_j) m_{\Theta,i} + (1 - w_i) m_{\Theta,j} \right] + \left[m_{\Theta,i} m_{\Theta,j} \right] \tag{8d}$$

where $m_{n,i} = w_i p_{n,i}$ and k_1 acts as a normalisation factor. The weights are not necessarily normalised to unity.

It is not straightforward to analyse how the bounded sum of individual support and the orthogonal sum of collective support for proposition θ_n change with respect to the change of the weights of w_i and w_j, as they also heavily depend on the degrees of belief $p_{n,i}$ and $p_{n,j}$. As a compromise, we calculate the sum total of the bounded sums of individual support for all propositions and that of the orthogonal sums of collective support, denoted by $\widehat{bs}_{e(2)}$ and $\widehat{os}_{e(2)}$ respectively.

$$\begin{aligned} \widehat{bs}_{e(2)} &= \sum_{n=1}^{N} \left[(1 - w_j) m_{n,i} + (1 - w_i) m_{n,j} \right] \\ &\quad + \left[(1 - w_j) m_{\Theta,i} + (1 - w_i) m_{\Theta,j} \right] \\ &= w_i (1 - w_j) + (1 - w_i) w_j = w_i + w_j - 2 w_i w_j \end{aligned} \tag{9}$$

$$\begin{aligned} \widehat{os}_{e(2)} &= \sum_{n=1}^{N} \left[m_{n,i} m_{n,j} + m_{n,i} m_{\Theta,j} + m_{\Theta,i} m_{n,j} \right] + \left[m_{\Theta,i} m_{\Theta,j} \right] \\ &= w_i w_j \left[\sum_{n=1}^{N} (p_{n,i} p_{n,j}) + 1 - \sum_{n=1}^{N} p_{n,i} \sum_{n=1}^{N} p_{n,j} \right] \end{aligned} \tag{10}$$

A mesh graph with a contour plot can be used to illustrate the relationship between the bounded sum of individual support and the weights of evidence in an indirect way.

It can be observed in Fig. 1 that the orthogonal sum of individual support plays a more and more important role, as the weights

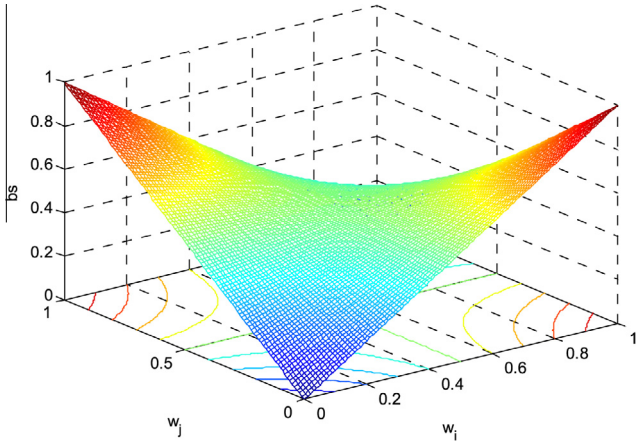


Fig. 1. The sum total of the bounded sums of individual support on all propositions.

w_i and w_j become more imbalanced. There is an extreme case $\widehat{bs}_{e(2)} = 1$, when $w_i = 0$ and $w_j = 1$, or vice versa.

Similarly, a contoured mesh graph can be drawn to illustrate the relationship between the orthogonal sum of collective support and the weights of evidence.

It is evident in Fig. 2 that the orthogonal sum of collective support plays an increasingly important role with the increase of either w_i or w_j. It takes the highest value $\left[\sum_{n=1}^{N} (p_{n,i} p_{n,j}) + 1 - \sum_{n=1}^{N} p_{n,i} \sum_{n=1}^{N} p_{n,j}\right]$, when both w_i and w_j are equal to 1. The sum total is no more than $w_1 w_2$, as $\sum_{n=1}^{N} (p_{n,i} p_{n,j}) \leqslant \sum_{n=1}^{N} p_{n,i} \sum_{n=1}^{N} p_{n,j}$, where two sides are equal only if the degrees of belief are non-zero only for one proposition from both pieces of evidence.

In the ER rule, weight and reliability need to be considered simultaneously in order to obtain the hybrid weight $\tilde{w}_i = w_i / (1 + w_i - r_i)$. Therefore, it is also necessary to investigate how the original weight w_i and reliability r_i affect the new hybrid weight $\tilde{w}_i$ which is implicitly used in the reasoning process.

The implementation of the formula $\tilde{w}_i = w_i / (1 + w_i - r_i)$ results in that $\tilde{w}_i < w_i$ if $r_i < w_i$, $\tilde{w}_i = w_i$ if $r_i = w_i$, and $\tilde{w}_i > w_i$ if $r_i > w_i$. There are two special cases: (1) when evidence e_i is fully reliable, or $r_i = 1$, there will be $\tilde{w}_i = 1$ even if w_i is close to zero. It may not be necessary to take into account a neutral piece of evidence with zero weight. (2) when evidence e_i is regarded as having a

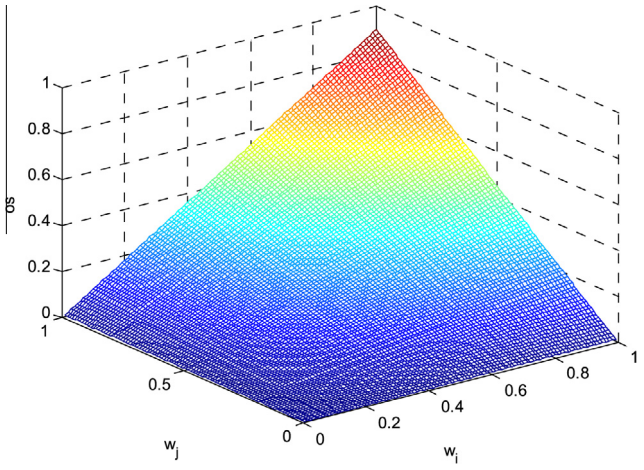


Fig. 2. The sum total of the orthogonal sums of collective support on all propositions.

dominating importance, or $w_i = 1, r_i$ regulates the hybrid weight in a gradual way. It can be observed in Fig. 3 that r_i plays a crucial regulative effect, especially when $r_i > w_i$, with the extreme case of $r_i \to 1$ and $w_i \to 0$. The above formula provides a rigorous way to obtain the hybrid weight. However, it can be conjectured that there may exist different ways to combine weight with reliability in real-world applications, depending on the physical definitions of weight and reliability and also the decision makers' understanding.

Next, a simple numerical example is used to illustrate the inference patterns of the ER rule, including the support for the parts of the bounded sum of individual support and the orthogonal sum of collective support. Suppose that there are two pieces of independent evidence $e_i = \{(\theta_1, 0.3), (\theta_2, 0.3), (\theta_3, 0.2); (\Theta, 0.2)\}$ and $e_j = \{(\theta_1, 0.1), (\theta_2, 0.5), (\theta_3, 0.4); (\Theta, 0)\}$ on the frame of discernment $\Theta = \{\theta_1, \theta_2, \theta_3\}$. First of all, we consider that weights are normalised, i.e., $w_i + w_j = 1$, and also $w_i = r_i, w_j = r_j$. The following figures are generated to illustrate the inference patterns of both the bounded sums and the orthogonal sums before and after multiplying with the normalisation factor.

It is shown in Fig. 4 that the inference is nonlinear on both the bounded sum of individual support (marked by down triangle) and the orthogonal sum of collective support (marked by up triangle). The combined degrees of belief (marked by star) are proportionally amplified after normalisation in Fig. 4(b), as the normalisation factor $k_1 \leqslant 1$. Given that the weights are normalised, it appears that the bounded sum of individual support plays a more important role in the reasoning process.

Secondly, we consider the original weights which are not normalised, i.e., $0 < w_i, w_j \leqslant 1$, and also $w_i = r_i, w_j = r_j$. Similarly, the following figures can be used to illustrate the inference patterns of the two parts before and after normalisation.

It can be seen in Fig. 5 that the orthogonal sum of collective support plays a more and more important role with the increase of evidence weights. It has no practical meaning when the weights are equal to zero for both pieces of evidence. When the weights are close to zero, it is obvious that the bounded sum of individual support plays a dominating role. Nevertheless when the weights approach one, the orthogonal sum of collective support tends to play a leading role. As discussed previously, the ER rule reduces to Dempster's rule with the weights of $w_i = w_j = 1$, under which the combined support on the complete set Θ (marked by x-mark) is ruled out due to its non-support from the second piece of evidence. The combined degree of belief on proposition θ_2 (denoted by dotted line) is further emphasised in a nonlinear pattern, as it has relatively high degrees of support from both pieces of evidence.

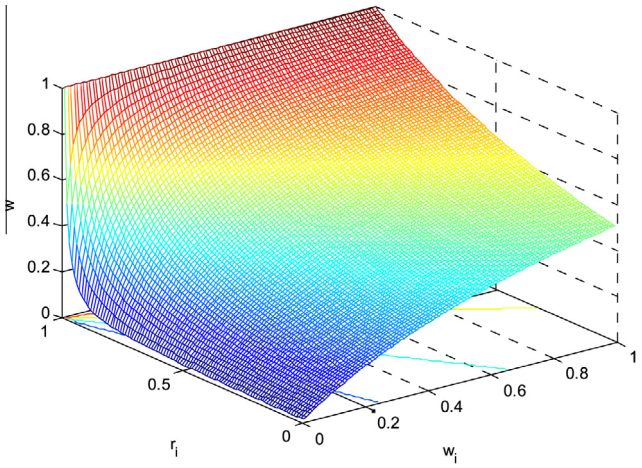


Fig. 3. The hybrid weight with respect to evidence weight and reliability.

3. A data-driven approximate causal inference model

Bayesian inference, the original ER algorithm and their derivatives, Bayesian Network (BN) and BRB methodology have been applied to a wide range of application areas, such as system identification, risk and decision analysis [19,25,34,4]. In the perspective that the ER rule is developed as a more generic probabilistic reasoning process and has rigorous inference patterns as analysed previously, the section aims to explore potential ways of developing a data-driven approximate causal inference model with reference to the merits of BN and BRB methodology.

3.1. Representation of attributes

We consider a typical basic reasoning fragment as shown in Fig. 6. There are M antecedent attributes or variables $\boldsymbol{x} = \{x_i; i = 1, \ldots, M\}$ which are used to reason about the consequent attribute or output variable y.

It is well-known that BN is usually formulated as a directed acyclic graph with a set of nodes and arcs connecting a parent node to a child node. However, any complex BNs can always be decomposed into a set of the above basic fragments, each of which has one child node with its parent(s). The relation amongst parent node(s) and a child node in BN can be modelled by a Conditional Probability Table (CPT). The above reasoning fragment can also be naturally translated into a set of belief rules in the BRB methodology, as each If-Then belief rule formulates a causal relationship between the packet antecedent attributes $\{x_1, \ldots, x_M\}$ and the consequent y. Specifically, each antecedent attribute is corresponding to a parent node, the packet antecedent in each belief rule is corresponding to a state combination of the parent nodes, and the consequence is corresponding to the child node in the basic BN model. In both BN and BRB models, one important capability is that prior information or knowledge can be updated whenever new evidence or observations are available from any attribute or on causal mappings [10,5].

To formulate the reasoning process in the above fragment, it is firstly important to consider the representation of antecedent attributes. Antecedent attributes can usually be categorised as qualitative or quantitative, or as discrete or continuous in a mathematical sense. Generally, a qualitative attribute can be characterised by a set of mutually exclusive and exhaustive linguistic terms, as implemented in a wide range of decision and risk analysis problems [14,25]. In this case, the belief degree assigned to each subjective grade in the ER rule is corresponding to the probability of each state of the parent node in Bayesian inference. In this paper, we mainly focus on dealing with quantitative attributes. In BNs, discretisation is the most commonly used way to capture the rough characteristics of the distribution of a quantitative attribute without making explicit assumptions [2,21]. Under discretisation any continuous value will be associated with a discretised interval, within which discretisation loses the ability to differentiate values and might suffer information loss [32]. The way of representing a quantitative attribute x_i in the BRB methodology is technically different, where a set of referential values $\boldsymbol{A}_i = \{A_{i,j}; j = 1, \ldots, J_i\}$ is defined to characterise its distribution of belief. Thus each belief rule can be interpreted by stating that the consequent y is believed to be $\theta_n (n = 1, \ldots, N)$ with the belief degree $\beta_{n,i}$, on the condition that each antecedent attribute $x_i (i = 1, \ldots, M)$ takes a certain referential value of $A_{i,j} (j = 1, \ldots, J_i)$. It is worth noting that each consequent element θ_n can also be associated with a referential value if y is a numerical output. Through constructing a base of belief

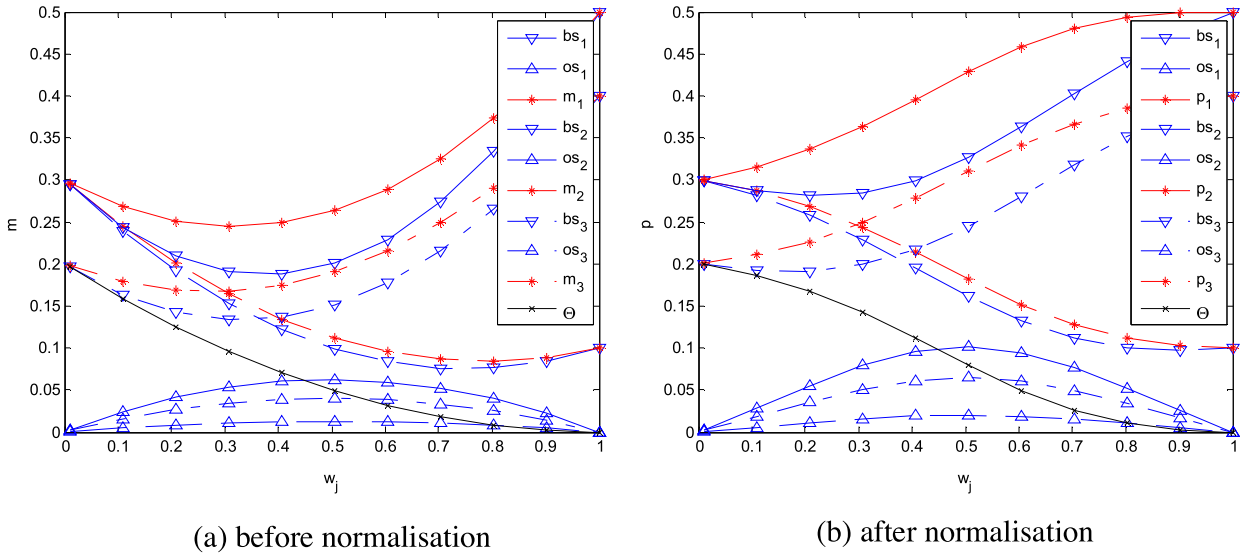


Fig. 4. Illustration of nonlinear inference patterns with normalised weights.

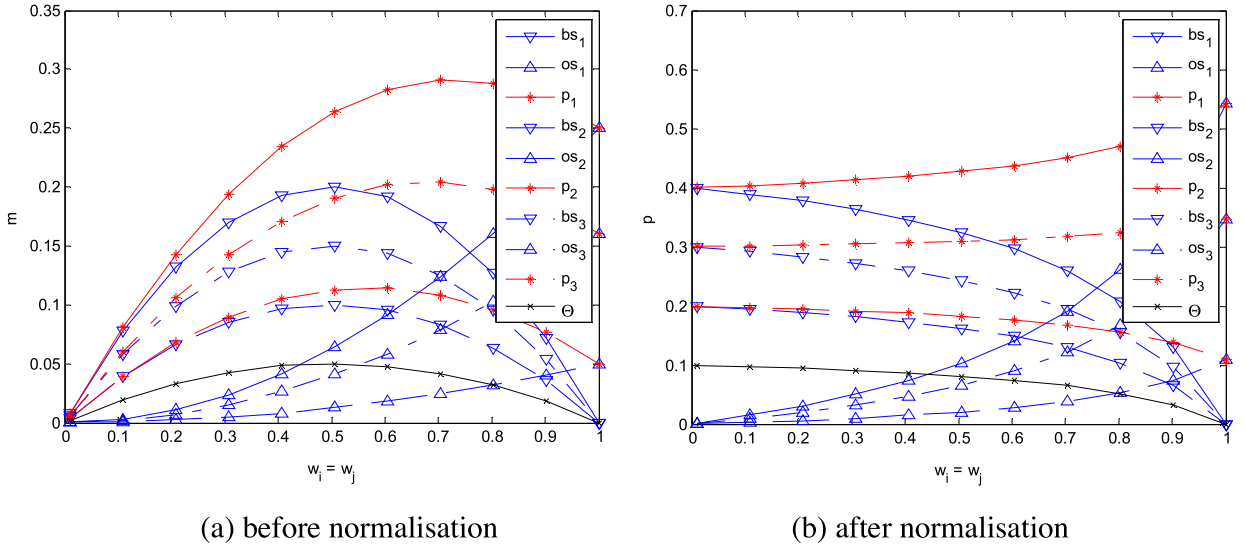


Fig. 5. Illustration of nonlinear inference patterns with un-normalised weights.

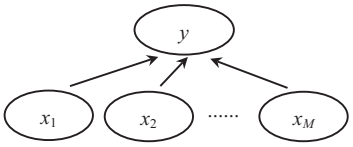


Fig. 6. A typical basic reasoning fragment.

rules, the consequent y can be reasoned in a causal way, whatever value each antecedent attribute takes, either a referential value itself or any value between two consecutive referential values. This capability extends the general boundary of approximate reasoning and guarantees the applicability of the BRB methodology to complex nonlinear system modelling and identification problems [30,6,4].

With the definition of referential values $\boldsymbol{A}_i = \{A_{i,j}; j = 1, \ldots, J_i\}$, the following information transformation technique can be used to generate the corresponding belief distribution for the quantitative attribute x_i.

$$S(x_i) = \{(A_{i,j}, \alpha_{i,j}); j = 1, \ldots, J_i\} \tag{11a}$$

where

$$\alpha_{i,j} = \frac{A_{i,j+1} - x_i}{A_{i,j+1} - A_{i,j}} \quad \text{and} \quad \alpha_{i,j+1} = 1 - \alpha_{i,j}, \ \text{if } A_{i,j} \leqslant x_i \leqslant A_{i,j+1} \tag{11b}$$

$$\alpha_{i,j'} = 0, \quad \text{for } j' = 1, \ldots, J_i \text{ and } j' \neq j, j+1 \tag{11c}$$

Here, $\alpha_{i,j}$ represents the similarity degree to which the antecedent attribute x_i matches the referential value $A_{i,j}$.

3.2. Approximate causal modelling

After both antecedent and consequent attributes are represented properly by the above belief structure, it will be therefore necessary to model the causal relationship between the antecedent attributes and the consequent in order to perform further inference. There may not exist priori information about how each attribute influences others in general. In BNs, the CPT on the complete

combinations of discrete states must generally be known to correctly perform probabilistic reasoning. For a real-world reasoning model, the CPT is usually very large and is difficult or impossible to be directly obtained. To simplify the causal modelling process of complex BNs, the joint conditional probability is sometimes approximated by $P(\theta_n|x_1,\ldots,x_M)=\alpha\prod_{i=1}^{M}P(\theta_n|x_i)$, where α is a normalisation factor to make sure $\sum_{n=1}^{N}P(\theta_n|x_1,\ldots,x_M)=1$ [12,15,20]. This approximation makes the construction of CPTs less restrictive. There is a similar issue of complexity that could arise from BRB models, as we may need to consider a multiplicative combination of referential values potentially for all antecedent attributes in constructing a complete belief rule base, although not all combinations have to be taken into account for reasoning in the BRB methodology.

Similar to the strategy of approximating the joint conditional probability in BNs, we attempt to model the causal relationship between each antecedent attribute and the consequent separately, and then combine the consequent belief distribution supported by individual antecedent attribute together using the ER rule. Firstly, a $J_i\times(N+1)$ dimensional causal belief matrix, denoted as $\boldsymbol{\beta}_i$, can be constructed to characterise the causal relationship between the antecedent attribute x_i and the consequent y. Without loss of generality, here we assume that there is no local ignorance in the causal relationship, and the belief degree on a smaller subset of the complete set Θ except for any singleton θ_n is equal to zero.

In the above matrix, each element means that the consequent y is believed to be $\theta_n(n=1,\ldots,N)$ with the belief degree $\beta_{n,i,j}$, given that the antecedent attribute x_i takes a certain referential value of $A_{i,j}(j=1,\ldots,J_i)$ without considering all the other antecedent attributes. $\beta_{\Theta,i,j}$ denotes the global ignorance and $\beta_{\Theta,i,j}=1-\sum_{n=1}^{N}\beta_{n,i,j}(j=1,\ldots,J_i)$. Various relevant techniques, such as utility-based or rule-based approaches with domain knowledge [29], simulation and proportional allocation method with numerical data [3], can be used to construct the initial causal belief matrices.

Following the philosophy of probabilistic inference in the ER rule, the belief degree $p_{n,i}$ to which the consequent y is believed to be $\theta_n(n=1,\ldots,N)$ with support of the antecedent attribute x_i is calculated as follows

$$p_{n,i}=\sum_{j=1}^{J_i}\alpha_{i,j}\beta_{n,i,j} \tag{12}$$

The following matrix operation can be used to reason the consequent belief distribution $\boldsymbol{p}_i$ supported by the antecedent attribute x_i

$$\boldsymbol{p}_i=\boldsymbol{\alpha}_i\times\boldsymbol{\beta}_i \tag{13}$$

where the vector $\boldsymbol{\alpha}_i=[\alpha_{i,1},\ldots,\alpha_{i,J_i}]$, and the matrix $\boldsymbol{\beta}_i$ takes elements from the causal belief matrix in Table 1.

Furthermore the ER rule formulated in Eqs. (8a)–(8d) can be applied to combine the consequent belief distribution $\boldsymbol{p}_i$ supported by individual antecedent attribute. The weight w_i of each antecedent attribute x_i is not necessarily normalised and can be initially assigned with domain knowledge or through preliminary analysis of historical data. The consequent reasoning output $\hat{y}$ from the above approximate causal inference model can therefore be represented by

$$S(\hat{y})=\{(\theta_n,p_n),n=1,\ldots,N;(\Theta,p_\Theta)\} \tag{14}$$

If there is also no global ignorance $\beta_{\Theta,i,j}=0(i=1,\ldots,M;j=1,\ldots,J_i)$, $p_\Theta=0$. Then the numerical reasoning output can be calculated as

$$\hat{y}=\sum_{n=1}^{N}\theta_n p_n \tag{15a}$$

Otherwise, the uncertain reasoning output can be characterised by the interval $[\hat{y}_{\min},\hat{y}_{\max}]$.

$$\hat{y}_{\min}=\theta_1(p_1+p_\Theta)+\sum_{n=2}^{N}\theta_n p_n \tag{15b}$$

$$\hat{y}_{\max}=\sum_{n=1}^{N-1}\theta_n p_n+\theta_N(p_N+p_\Theta) \tag{15c}$$

where the referential values of the consequent y are assumed to satisfy $\theta_{n+1}\geqslant\theta_n(n=1,\ldots,N-1)$.

3.3. Optimal training

As discussed in Section 3.2, different types of prior information or knowledge are used to configure the initial parameters of the approximate causal inference model, including attribute weights $w_i(i=1,\ldots,M)$, referential values of antecedent attributes $A_{i,j}(i=1,\ldots,M;j=1,\ldots,J_i)$, belief degrees in causal belief matrices $\boldsymbol{\beta}_i(i=1,\ldots,M)$ and referential values of consequent attribute $\theta_n(n=1,\ldots N)$. However the initial parameters $\boldsymbol{P}=\langle w_i,A_{i,j},\beta_{n,i,j},\theta_n\rangle$ may not capture the complex causal relationship between all antecedent attributes and the consequent. Therefore, it is extremely important to train these parameters when historical data are available. According to their theoretical definitions, the parameters must satisfy the following basic bound, equality and inequality constraints.

- Lower and upper bound constraints

$$0\leqslant w_i\leqslant 1;\quad i=1,\ldots,M \tag{16a}$$

$$0\leqslant\beta_{n,i,j}\leqslant 1;\ \ n=1,\ldots,N;\ i=1,\ldots,M;\ j=1,\ldots,J_i \tag{16b}$$

- Equality and inequality constraints

$$\sum_{n=1}^{N}\beta_{n,i,j}\leqslant 1;\quad i=1,\ldots,M;\ j=1,\ldots,J_i \tag{16c}$$

As discussed in Section 3.2, the above inequality turns into an equality constraint if there is no global ignorance. In addition, some other constraints may need to be taken into consideration depending on the practical requirements, such as the bound constraints of referential values and the unity constraint of normalised weights.

If there are a set of historical dataset $(\boldsymbol{x}_t,y_t),t=1,\ldots T$, various traditional accuracy measures, such as mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), or interval accuracy measures, such as l_∞-norm, can be employed to define the optimal training objective. For example, the objective of minimising MSE can be simply defined as follows

$$\min_{\boldsymbol{P}}\xi(\boldsymbol{P})=\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t-y_t)^2. \tag{17}$$

Table 1
A causal belief matrix.

x_i	y			
	θ_1	$\cdots$	θ_N	Θ
$A_{i,1}$	$\beta_{1,i,1}$	$\cdots$	$\beta_{N,i,1}$	$\beta_{\Theta,i,1}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_{i,J_i}	β_{1,i,J_i}	$\cdots$	β_{N,i,J_i}	β_{Θ,i,J_i}

4. A numerical study

In this section, we re-investigate the practical pipeline leakage detection problem studied in previous research [26,5]. A series of 25% leak trial data was sampled from the mass flow metres at the inlet and outlet and the pressure metres at the inlet, outlet and eight middle points along the 100 km pipeline at the rate of 10 s. Totally 2008 data samples were collected in an experiment. Two antecedent attributes which are the difference between outlet flow and inlet flow and the average pipeline pressure change over time, denoted as *FlowDiff* and *PressureDiff* respectively, were introduced to estimate the consequent leak rate, denoted by *LeakSize*.

For comparison purpose, we use the same number of initial referential values for both antecedent attributes and the consequent attribute as in previous research [5], specifically eight referential values $A_1 = \{-10, -5, -3, -1, 0, 1, 2, 3\}$ for antecedent *FlowDiff*, seven referential values $A_2 = \{-0.015, -0.005, -0.002, 0, 0.002, 0.005, 0.015\}$ for *PressureDiff*, and five referential points for consequent *LeakSize* $D = \{0, 2, 4, 6, 8\}$. With the definition of referential values, we can formulate the causal relationship between each antecedent attribute and the consequent respectively. The scatters of consequent *LeakSize* to each antecedent attribute are visualised in Fig. 7.

It is evident in Fig. 7 that antecedent *FlowDiff* can be used to estimate consequent *LeakSize* roughly, and however there is no directly observable causal relationship between *PressureDiff* and *LeakSize*. With the preliminary analysis of historical data, the causal belief matrices with respect to *FlowDiff* and *PressureDiff* can be constructed empirically in Table 2.

In order to improve the causal modelling capability of the above inference model, the same amount of 500 training data are randomly selected to train the initial parameters, including the initial attribute weight $w_1 = 0.8$ for *FlowDiff* and $w_2 = 0.2$ for *PressureDiff*. The optimal training model discussed in Section 3.3 can be easily solved by the nonlinear optimiser *fmincon* in Matlab. The trained causal belief matrices are listed in Table 3. The updated attribute weights $w_1 = 0.8782$ and $w_2 = 0.1218$, which means that antecedent *FlowDiff* plays a more important role in estimating consequent *LeakSize*.

As shown in Fig. 8(a), the values of the reasoned *LeakSize* from the initial inference model are a bit far away from the observed on the training dataset. However, in Fig. 8(b), the trained inference model can replicate closely the causal relationship between antecedent *FlowDiff* and *PressureDiff* and consequent *LeakSize* on the complete testing dataset.

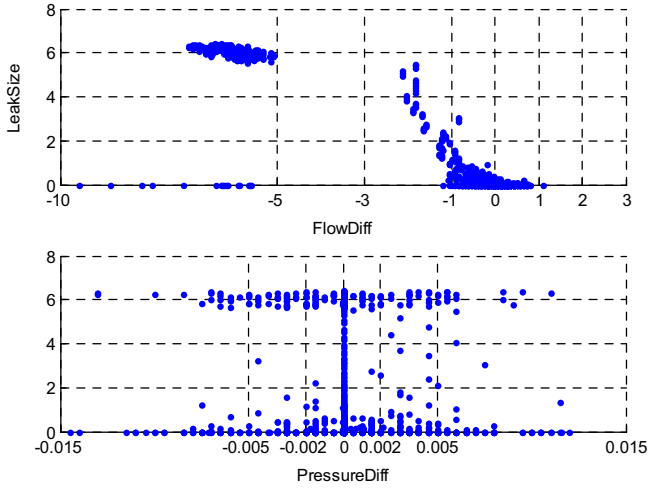


Fig. 7. The scatters of consequent *LeakSize* to two antecedent attributes.

Table 2
Initial causal belief matrices.

Antecedent attributes	Referential values	Consequent *LeakSize* $\Theta = \{0, 2, 4, 6, 8\}$
FlowDiff	-10	$\{0, 0, 0, 0, 1\}$
	−5	$\{0, 0, 0, 1, 0\}$
	−3	$\{0, 0, 1, 0, 0\}$
	−1	$\{0, 0.5, 0.5, 0, 0\}$
	0	$\{0, 0.25, 0.75, 0, 0\}$
	1	$\{0.25, 0.75, 0, 0, 0\}$
	2	$\{0.5, 0.5, 0, 0, 0\}$
	3	$\{1, 0, 0, 0, 0\}$
PressureDiff	-0.015	$\{0, 0, 0, 0.5, 0.5\}$
	−0.005	$\{0, 0, 0, 1, 0\}$
	−0.002	$\{0, 0, 0.5, 0.5, 0\}$
	0	$\{0, 0, 1, 0, 0\}$
	0.002	$\{0, 0.5, 0.5, 0, 0\}$
	0.005	$\{0, 1, 0, 0, 0\}$
	0.015	$\{1, 0, 0, 0, 0\}$

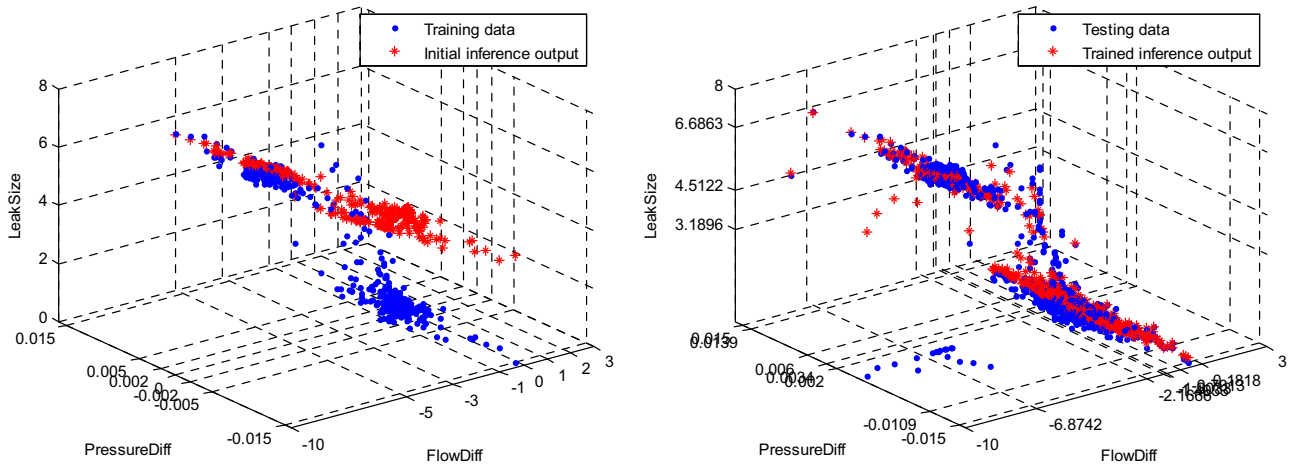
Table 3
Trained causal belief matrices.

Antecedent attributes	Referential values	Consequent *LeakSize* $\Theta = \{0, 3.1896, 4.5122, 6.6863, 8\}$
FlowDiff	−10	$\{0.2400, 0.1332, 0.1159, 0.0589, 0.4520\}$
	−6.8742	$\{0.0000, 0.1094, 0.0360, 0.7011, 0.1535\}$
	−2.1666	$\{0.3290, 0.0120, 0.0558, 0.0019, 0.6012\}$
	−1.4033	$\{0.5511, 0.0000, 0.3070, 0.0000, 0.1419\}$
	−1.3033	$\{0.5812, 0.1211, 0.2656, 0.0281, 0.0040\}$
	−0.7013	$\{0.9259, 0.0455, 0.0016, 0.0253, 0.0017\}$
	0.1818	$\{0.9983, 0.0003, 0.0000, 0.0013, 0.0000\}$
	3	$\{0.9916, 0.0012, 0.0000, 0.0035, 0.0038\}$
PressureDiff	−0.015	$\{0.3587, 0.0532, 0.0019, 0.5718, 0.0144\}$
	−0.0109	$\{1.0000, 0.0000, 0.0000, 0.0000, 0.0000\}$
	0.0020	$\{1.0000, 0.0000, 0.0000, 0.0000, 0.0000\}$
	0.0034	$\{0.4421, 0.0501, 0.0715, 0.0000, 0.4363\}$
	0.0060	$\{0.0000, 0.2067, 0.5065, 0.0024, 0.2845\}$
	0.0139	$\{0.1781, 0.0054, 0.0009, 0.0053, 0.8103\}$
	0.015	$\{0.0082, 0.0052, 0.0005, 0.1351, 0.8511\}$

It was pointed out that those outlying data points associated with high *FlowDiff* and zero *LeakSize* were sampled from the confirmation period of leakage [26]. It seems in Fig. 8(b) that some trained referential values for antecedent attributes converge to a same value. The phenomenon can be interpreted from two different perspectives. On one hand, two convergent referential values may possibly be merged together to further simplify the initial inference model. One the other hand, there may be a large number of training data in a small area, and it requires sampling multiple referential values in order to improve the inference accuracy.

In Table 4, the approximate causal inference model proposed in this paper is compared with the local and adaptive training BRB models where a large number of 8×7 belief rules were used [5]. The measures of MAE, root MSE (RMSE) and Pearson's correlation coefficient (PCC) between observed *LeakSize* and reasoned *LeakSize* are used to evaluate the inference accuracy.

It is evident that the proposed approximate causal inference model using the ER rule can also provide superior inference performance, with even lower MAE, RMSE than local and adaptive training BRB models, and higher PCC than local training BRB model. However it is worth emphasising that there is only an additive complexity on the number of referential values of antecedent attributes in contrast to the multiplicative complexity in the BRB methodology. Specifically, there are only 7 + 8 causal belief distributions in the proposed ER rule-based causal inference model,



(a) Initial inference output

(b) Trained inference output

Fig. 8. Reasoned outputs from initial and trained inference models.

Table 4
Comparison under different accuracy measures.

	MAE	RMSE	PCC
Local training BRB	0.22229	0.63791	0.96102
Adaptive training BRB	0.20643	0.63170	0.96697
Approximate causal inference model using the ER rule	0.20142	0.60904	0.96622

while there are 7×8 belief rules in both local and global training models. As the complexity is significantly reduced from rule-based inference models, the proposed approximate causal inference model can be easily implemented to complex or large-scale causal inference problems, such as decision and risk analysis, system identification, fault diagnosis, etc.

5. Concluding remarks

In this paper, we analysed the inference patterns of the newly-developed ER rule and then proposed an approximate causal inference model with data-based causal modelling and optimal learning. The proposed inference model holds an additive complexity on the number of antecedent attributes and it has great potential to be applied to complex causal inference problems, where BN, BRB and relevant methodologies have limited applicability. A numerical study on the practical pipeline leak detection problem was also conducted to validate the capability of the proposed inference model. It is worth noting that the approximate causal inference model under study in this paper is not a final solution, and this exploratory work serves as a promoting further development of generic probabilistic inference models with the utilisation of big data. In addition, in this paper the antecedent attributes are implicitly assumed to be mutually independent. Actually, the independency amongst attributes needs to be investigated in real-world applications. In further research, the reliability of prior or updating information should also be taken into consideration.

Acknowledgements

This work is partially supported by the European Commission FP7 Marie Curie IRSES – REFERENCE project, and the Natural Science Foundation of China under Grant Nos. 61203178, 61304214 and 61433001.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.knosys.2015.07.026.

References

[1] J.C. Augusto, J. Liu, P.J. McCullagh, H. Wang, J.B. Yang, Management of uncertainty and spatio-temporal aspects for monitoring and diagnosis in a smart home, Int. J. Comput. Intell. Syst. 1 (4) (2008) 361–378.
[2] M. Boullé, MODL: a Bayes optimal discretization method for continuous attributes, Mach. Learn. 65 (2006) 131–165.
[3] Y.W. Chen, S.H. Poon, J.B. Yang, D.L. Xu, D. Zhang, S. Acomb, Belief rule-based system for portfolio optimisation with nonlinear cash-flows and constraints, Eur. J. Oper. Res. 223 (3) (2012) 775–784.
[4] Y.W. Chen, J.B. Yang, C.C. Pan, D.L. Xu, Z.J. Zhou, Identification of uncertain nonlinear systems: constructing belief rule-based models, Knowl.-Based Syst. 73 (2015) 124–133.
[5] Y.W. Chen, J.B. Yang, D.L. Xu, Z.J. Zhou, D.W. Tang, Inference analysis and adaptive training for belief rule based systems, Expert Syst. Appl. 36 (10) (2011) 12845–12860.
[6] Y.W. Chen, J.B. Yang, D.L. Xu, S.L. Yang, On the inference and approximation properties of belief rule based systems, Inf. Sci. 234 (2013) 121–135.
[7] A.P. Dempster, A generalization of Bayesian inference, J. Roy. Stat. Soc. Ser. B 30 (2) (1968) 205–247.
[8] L. Dymova, P. Sevastjanov, A new approach to the rule-base evidential reasoning in the intuitionistic fuzzy setting, Knowl.-Based Syst. 61 (2014) 109–117.
[9] T. Denoeux, Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence, Artif. Intell. 172 (2–3) (2008) 234–264.
[10] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, Mach. Learn. 20 (1995) 197–243.
[11] C. Fu, M. Huhns, S. Yang, A consensus framework for multiple attribute group decision analysis in an evidential reasoning context, Inf. Fusion 17 (2014) 22–35.
[12] J.H. Kim, J. Pearl, A computational model for combined causal and diagnostic reasoning in inference systems, in: Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Germany, 1983, pp. 380–385.
[13] H.C. Liu, L. Liu, Q.L. Lin, Fuzzy failure mode and effects analysis using fuzzy evidential reasoning and belief rule-based methodology, IEEE Trans. Reliab. 62 (1) (2013) 23–36.
[14] N. Mays, C. Pope, J. Popay, Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field, J. Health Serv. Res. Policy 10 (S1) (2005) 6–20.
[15] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, Cambridge, 2000.
[16] G. Shafer, Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.
[17] F. Smarandache, J. Dezert, J.M. Tacnet, Fusion of sources of evidence with different importances and reliabilities, in: The 2010 13th IEEE Conference on Information Fusion (FUSION), 2010, pp. 1–8.

[18] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, Int. J. Approx. Reason. 9 (1) (1993) 1–35.
[19] J.Q. Smith, Bayesian Decision Analysis: Principles and Practice, Cambridge University Press, 2010.
[20] D. Tang, J.B. Yang, K.S. Chin, Z.S.Y. Wong, X. Liu, A methodology to generate a belief rule base for customer perception risk analysis in new product development, Expert Syst. Appl. 38 (2011) 5373–5383.
[21] L. Uusitalo, Advantages and challenges of Bayesian networks in environmental modelling, Ecol Model. 203 (2007) 312–318.
[22] J.Q. Wang, R.R. Nie, H.Y. Zhang, X.H. Chen, Intuitionistic fuzzy multi-criteria decision-making method based on evidential reasoning, Appl. Soft Comput. 13 (2013) 1823–1831.
[23] J.Q. Wang, H.Y. Zhang, Multicriteria decision-making approach based on Atanassov's intuitionistic fuzzy sets with incomplete certain information on weights, IEEE Trans. Fuzzy Syst. 21 (3) (2013) 510–515.
[24] Y. Wang, Y. Dai, Y.W. Chen, F. Meng, The evidential reasoning approach to medical diagnosis using intuitionistic fuzzy Dempster–Shafer theory, Int. J. Comput. Intell. Syst. 8 (1) (2015) 75–94.
[25] D.L. Xu, An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis, Ann. Oper. Res. 195 (1) (2012) 163–187.
[26] D.L. Xu, J. Liu, J.B. Yang, G.P. Liu, J. Wang, I. Jenkinson, J. Ren, Inference and learning methodology of belief-rule-based expert system for pipeline leak detection, Expert Syst. Appl. 32 (1) (2007) 103–113.
[27] R.R. Yager, On the Dempster–Shafer framework and new combination rules, Inf. Sci. 41 (2) (1987) 93–137.
[28] R.R. Yager, L. Liu (Eds.), Classic Works of the Dempster–Shafer Theory of Belief Functions, Springer, Heidelberg, 2008.
[29] J.B. Yang, Rule and utility based evidential reasoning approach for multiple attribute decision analysis under uncertainty, Eur. J. Oper. Res. 131 (1) (2001) 31–61.
[30] J.B. Yang, J. Liu, J. Wang, H.S. Sii, H.W. Wang, A belief rule-base inference methodology using the evidential reasoning approach – RIMER, IEEE Trans. Syst. Man Cybern. – Part A 36 (2) (2006) 266–285.
[31] J.B. Yang, M.G. Singh, An evidential reasoning approach for multiple attribute decision making with uncertainty, IEEE Trans. Syst. Man Cybern. 24 (1) (1994) 1–18.
[32] Y. Yang, G.I. Webb, Discretization for naive-Bayes learning: managing discretization bias and variance, Mach. Learn. 74 (2009) 39–74.
[33] J.B. Yang, D.L. Xu, On the evidential reasoning algorithm for multiattribute decision analysis under uncertainty, IEEE Trans. Syst. Man Cybern. Part A. Syst. Hum. 32 (3) (2002) 289–304.
[34] J.B. Yang, D.L. Xu, Evidential reasoning rule for evidence combination, Artif. Intell. 205 (2013) 1–29.
[35] J.B. Yang, D.L. Xu, A study on generalising Bayesian inference to evidential reasoning, in: Belief Functions: Theory and Applications, Lect. Notes Comput. Sci., vol. 8764, 2014, pp. 180–189
[36] L.A. Zadeh, Review of Shafer's a mathematical theory of evidence, AI Mag. 5 (1984) 81–83.
[37] Z.J. Zhou, C.H. Hu, D.L. Xu, J.B. Yang, D.H. Zhou, New model for system behaviour prediction based on belief-rule-based systems, Inf. Sci. 180 (2010) 4834–4864.