Expert Systems with Applications 38 (2011) 5061–5080

Contents lists available at ScienceDirect

Expert Systems with Applications

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

On the dynamic evidential reasoning algorithm for fault prediction

Xiao-Sheng Si [a,*], Chang-Hua Hu [a], Jian-Bo Yang [b], Qi Zhang [a]

[a] Unit 302, Department of Automation, Xi'an Institute of Hi-Tech, Hongqing Town, Xi'an, Shanxi 710025, PR China
[b] Manchester Business School, The University of Manchester, P.O. Box 88, Manchester M60 1QD, UK

ARTICLE INFO

Keywords:
Artificial intelligence
Nonlinear programming
Dynamic evidential reasoning approach
Utility
Fault prediction

ABSTRACT

In this paper, a new fault prediction model is presented to deal with the fault prediction problems in the presence of both quantitative and qualitative data based on the dynamic evidential reasoning (DER) approach. In engineering practice, system performance is constantly changed with time. As such, there is a need to develop a supporting mechanism that can be used to conduct dynamic fusion with time, and establish a prediction model to trace and predict system performance. In this paper, a DER approach is first developed to realize dynamic fusion. The new approach takes account of time effect by introducing belief decaying factor, which reflects the nature that evidence credibility is decreasing over time. Theoretically, it is show that the new DER aggregation schemes also satisfy the synthesis theorems. Then a fault prediction model based on the DER approach is established and several optimization models are developed for locally training the DER prediction model. The main feature of these optimization models is that only partial input and output information is required, which can be either incomplete or vague, either numerical or judgmental, or mixed. The models can be used to fine tune the DER prediction model whose initial parameters are decided by expert's knowledge or common sense. Finally, two numerical examples are provided to illustrate the detailed implementation procedures of the proposed approach and demonstrate its potential applications in fault prediction.

1. Introduction

Fault detection and diagnosis have been researched since 1970s and become one of the highlights in control theory. With the increasingly high requirements of system reliability and security, it is generally expected to acquire failure information before a fault is developed to cause system damage. An effective means to ensure reliability and security is to fast and effectively predict system failures so that remedies can be undertaken in time. As a result, fault prediction has been given much attention, in particular in aeronautics and astronautics engineering over the past decade. For example, a fault prediction unit was equipped in the fault diagnosis expert system of the communication system of NASA's deep-space detector (Zhang & Hu, 2008).

Fault prediction is a special area that combines fault detection with event forecast, which has been developed since 1979 (Lu & Sakes, 1979). However, little progress has been made on fault prediction so far. The main difficulties are explained as follows. The first one is how to build a prediction model for describing a systems status. The second is how to obtain the characteristic signals for expressing a fault (Zhang & Hu, 2008). The existing prediction methods can be broadly classified into two categories: fault prediction methods based on analytical models and fault prediction methods based on non-analytical models. Fault prediction methods based on analytical models exploit relationships between system input variables and output variables in conditions where the mathematic model of a system is known. The extended Kalman filtering (EKF) and strong tracking filtering and particle filtering are the most popular methods in this category (Chen, Zhou, & Liu, 2005; Hu, Danil, & Donald, 2007; Wang, Zhou, & Jin, 2004; Yang, 2002; Yang & Liu, 1999; Zhang, Chen, Zhou, & Li, 2006; Zhou & Frank, 1996). Yang and Liu (1999) demonstrated the application of EKF to estimate the state of a DC motor and predict faults based on the change of its rotation speed. As a premise, the system model must be known *a priori* and accurate for these methods to be highly effective. Unfortunately, nonlinearity and uncertainty widely exist in system models. For some complex nonlinear system, accurate models cannot be obtained. This can easily degrade the quality of output estimation and cause either missed detections or false alarms. In contrast, fault prediction methods based on non-analytical model have been equally popular in forecasting. Time series analysis, which is developed by Box and Jenkins (1970), has been applied to forecasting extensively (Box & Jenkins, 1970). The autoregressive integrated moving average models (ARIMA), pioneered by Box and Jenkins, can be a suitable alternative in fault prediction (Ho & Xie, 1998; Zhang, 2003). Grey system theory is a truly

* Corresponding author. Tel.: +86 29 84743949; fax: +86 29 84744603.
E-mail addresses: sixiaosheng@126.com, sxs09@mails.tsinghua.edu.cn (X.-S. Si), hch6603@263.net (C.-H. Hu).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.144

multidisciplinary theory that can be used to deal with grey systems characterized by both partially known and partially unknown information (Deng, 1982). As an essential part of grey system theory, grey forecasting models have gained popularity in time-series forecasting due to their simplicity and ability to characterize an unknown system by using as few as four data points. The grey prediction often applies to prediction in time-varying non-linear systems, and it has been widely used in several fields such as agriculture, industry and fault prediction (Deng, 1982; Jiang, 2004; Zhang et al., 2006, Zhang, Wang, & Zhao, 2007; Zhou & Hu, 2008). However, the grey prediction method can only be used to describe data that change exponentially. In the absence of such characteristics in time series data, applying grey prediction model to any time series generates inaccurate prediction results. Artificial neural networks (ANN) have general nonlinear mapping capabilities and so have increasingly attracted attention in the field of time series forecasting. ANN was applied to learn and predict time series simulation data generated by computer in 1987 (Lapedes & Farber, 1987). After that, it has been widely used in fault prediction field (Connor, Martin, & Atlas, 1994; Hu, Cao, & Zhang, 2005; Khoshgoftaar, 2003; Rietman & Beachy, 1998; Zhang, Patuwo, & Hu, 1998, 2005). However, a neural network is data-driven and suffers from a number of drawbacks. One of the drawbacks of ANN is the subjectivity in designing an ANN and determining its parameters. For example, one can only determine subjectively rather than objectively how many hidden layers an ANN should have, how many neurons each hidden layer should be given, how to determine the learning rate parameter, momentum factor and epoch, and how to choose the learning rule and transfer function. Different people may design different networks and choose different parameters for training and testing. Therefore, they may get different results on the same subject. Another significant drawback of ANN is that its connection weights are not easy to be explained. In addition, ANN is usually a network without any constraint conditions. If there are any constraints on inputs and/or outputs, it will be very hard for an ANN to be trained to meet such constraints. Support vector machine (SVM), which is developed by Vapnik, is a novel learning machine based on statistical learning theory (Vapnik, 1995). In recent years, SVM has been applied successfully to solve forecasting problems in many areas (Chen, 2007; Mohandes, Halawani, & Reman, 2004; Pai & Hong, 2006; Pai, 2006; Suykens & Vandewalle, 1999; Wang & Guo, 2008). However the selection of three positive parameters, σ, ε and C, of a SVM model is important for the accuracy of prediction. In addition to the models mentioned above, other methods such as rough set theory (Dimitras, Slowinski, Susmaga, & Zopounidis, 1999), Hough transformation (Flint, Ingleby, & Morton, 1992), expert system (Angeli & Chatzinikolaou, 1999) and hybrid methods (Tseng, Tseng, & Yuan, 2001; Zhang, 2003; Zhou & Hu, 2008), have also been adopted to deal with fault prediction problems.

However, for a complex nonlinear engineering system, many fault prediction problems involve both quantitative data and qualitative information, as well as various types of uncertainties such as incompleteness and fuzziness. It is difficult to obtain a complete set of historical data for developing a perfect mathematical model to simulate a system. Conventional analytical models based on pure data, such as time series analysis models and filter based models, may not always be possible to build. For example, in analyzing system safety in design and operations of a large engineering system, it is difficult to obtain its model structure or analytic model due to lack of historical data, so the time series analysis and filter based methods are not applicable. On the other hand, due to the fact that human beings hold ultimate responsibilities in most situations, their subjective judgments play an irreplaceable role in making final decisions. However, afore-mentioned prediction models are limited in dealing with both numerical data and human judgmental information under uncertainty that is likely to be incomplete and can hardly be accurate. In such cases, it is highly desirable to develop a single universal prediction model that can model and analyze prediction problems using both numerical data and human judgmental information, which is likely to be incomplete and vague.

The development of methods for dealing with uncertainty has received considerable attention in the last three decades. Several numerical and symbolic methods have been proposed for handling uncertain information. Three of the most common frameworks for representing and reasoning with uncertain knowledge are:

(a) Bayesian probability theory.
(b) Dempster–Shafer (D–S) theory of evidence.
(c) Fuzzy set theory.

Due to the power of the D–S theory in handling uncertainties (Bauer, 1997; Beynon, Cosker, & Marshall, 2001, 2002a; Chen, 1997; Yager, 1992; Yen, 1990), so far, it has found wide applications in many areas such as expert systems (Biswas, Oliff, & Sen, 1988; Beynon et al., 2001; Chen, 1997; Wallery, 1996), uncertainty reasoning (Benferhat, SafHotti, & Smets, 2000; George & Pal, 1996; Hullermeier, 2001; Ishizuka, Fu, & Yao, 1982; Jones, Lowe, & Harrison, 2002; Rakar, Juricic, & Ball, 1999), pattern classification (Binaghi & Madella, 1999; Binaghi, Gallo, & Madella, 2000; Denoeux, 1997, 1999, 2000a, 2000b; Denoeux & Zouhal, 2001; Denoeux & Masson, 2004), fault diagnosis and detection (Fan and Zuo, 2006a, 2006b; Parikh and Pont, 2001; Rakar and Juricic, 2002; Yang and Kim, 2006), information fusion (Fabre et al., 2001; Ruthven and Lalmas, 2002; Telmoudi and Chakhar, 2004), Multiple attribute decision analysis (Beynon et al., 2001; Beynon, 2002a, 2002b; Guo, Yang, & Chin, in press; Xu, Yang, & Wang, 2006; Yang & Xu, 2002a, 2002b; Yang & Sen, 1994; Yang, Liu, Wang, Sii, & Wang, 2006a; Yang, Wang, Xu, & Chin, 2006b), image processing (Bloch, 1996; Huber, 2001; Krishnapuram, 1991) and regression analysis (Monney, 2003; Pent-Renaud and Denoeux, 2004; Wang and Elhag, 2007). In the last decade, an evidential reasoning (ER) approach has been developed for MADA under uncertainty (Xu & Yang, 2003; Xu et al., 2006; Yang & Sen, 1994; Yang & Xu, 2002a, 2002b, 2004; Yang et al., 2006a, 2006b). This approach is developed on the basis of decision theory and the Dempster–Shafer (D–S) theory of evidence (Dempster, 1967; Shafer, 1976). Extensive research dedicated to the ER approach has been conducted in recent years. Firstly, the rule and utility-based information transformation techniques were proposed within the ER modeling framework (Yang, 2001). This work enables the ER approach to deal with a wide range of MADA problems having precise data, random numbers and subjective judgments with probabilistic uncertainty in a way that is rational, transparent, reliable, systematic and consistent. Then, the in-depth research into the ER algorithm has been conducted by treating the unassigned belief degree in two parts, one caused by the incompleteness and the other caused by the fact that each attribute plays only one part in the whole assessment process because of its relative weight (Yang & Xu, 2002a, 2002b). This work leads to a rigorous yet pragmatic ER algorithm that satisfies several common sense rules governing any approximate reasoning based aggregation procedures. The ER approach has thus been equipped with the desirable capability of generating the upper and lower bounds of the degree of belief for incomplete assessments, which are crucial to measure the degree of ignorance. Thirdly, the analysis process of the ER approach was fully investigated, which reveals the nonlinear features of the ER aggregation process (Yang & Xu, 2002a, 2002b). This work provides guidance on conducting sensitivity analysis using the ER approach. Fourthly, the ER approach was further developed to deal with MADA problems with both probabilistic and fuzzy uncertainty. This work leads to a new fuzzy ER algorithm that aggregates multiple attributes using the information contained in the fuzzy belief matrix, which can model precise data, ignorance and fuzziness under the unified framework of a distributed fuzzy belief structure. Fifthly, the ER approach was reanalyzed explicitly in terms of D–S theory and a general scheme of attribute aggregation was proposed for the purpose of dealing with MADA problems (Huynh, Nakamori, Ho, & Murai, 2006). This work interprets the ER approach using the discounting operator and relaxes the constraint that the ER approach need to satisfy so that four synthesis axioms proposed by Yang and Xu (2002a, 2002b) can hold. Thus, this work provides convenience to develop new aggregation schemes. In most recent, Hu, Si, and Yang (2010) develops a new prediction model based on evidential reasoning algorithm, named ER-based prediction model, in which some nonlinear optimization models are investigated to determine the parameters of the proposed model accurately. A practical case study for turbocharger engine systems' reliability forecasting is provided to demonstrate the detailed implementation procedures. The experimental results show that the prediction performance of the ER-based prediction model outperforms several existing methods in terms of prediction accuracy or speed. In addition, the ER approach has been applied to decision problems such as business performance assessment (Sonmez, Graham, Yang, & Holt, 2002; Siow, Yang, & Dale, 2001; Xu & Yang, 2003; Yang, Dale, & Siow, 2001; Yang & Xu, 2004, 2005), environmental impact assessment (EIA) (Wang, Yang, & Xu, 2006), organizational self-assessment (Siow et al., 2001; Yang et al., 2001), safety analysis (Wang and Yang, 2001; Liu, Yang, Wang, Sii, & Wang, 2004), bridge condition assessment (Wang & Elhag, 2008), etc. However, none of these efforts was directed to deal with fault prediction problems under uncertainty. This research has been conducted to fill the gap.

The current ER approach is of a static fusion style in nature and does not take account of time information in fusion. In static fusion, all belief functions are combined simultaneously, the credibility of evidence source is unchanged over time, and static combination takes no account of time effect, while time information has no influence in combination results. However, in engineering practice, system performance changes significantly with time. Moreover, system reliability is decreasing with time. As such, in order to trace system state accurately in time, there is a need to extend the original evidential approach and aggregate time information into the evidence combination process.

In this paper, firstly, the D–S theory will be further developed to take account of time information, resulting in a new Dempster's combination rule for dynamic fusion. In particular, the original evidential approach will be extended to a dynamic evidential reasoning (DER) approach based on the new Dempster's combination rule, considering time effect for evidence credibility. The new DER approach is a recursive algorithm similar to the existing ER approach. In order to facilitate parameters optimization, a new analytical DER algorithm will be investigated as well. The equivalence between the recursive DER and analytical DER is proved. Secondly, the DER approach is applied to fault prediction problems, and a fault prediction model is established based on the DER approach and utility theory. In the DER prediction model, input data, attribute weights, and belief decaying factor are combined to generate appropriate conclusions using the DER algorithm. Thirdly, due to the difficulty to accurately determine the parameters of the DER prediction model entirely subjectively, several nonlinear optimization models for training the parameters of the DER prediction model and other knowledge representation parameters in the DER approach are proposed. The new optimization models are either single or multiple-objective nonlinear optimization problems. This optimization process is formulated as nonlinear objective functions to minimize the differences between observed outputs and the simulated outputs of the DER prediction model. Parameter specific limits and partial expert judgments can be formulated as constraints. The optimization problems can be solved using existing tools such as the optimization tool box provided in Matlab.

The rest of this paper is organized as follows. Section 2 gives a brief description of the D–S theory and some theoretical issues. A new Dempster's combination rule for dynamic fusion is proposed. Furthermore, the original evidential reasoning approach will be extended to be the dynamic evidential reasoning (DER) approach. In Section 3, a fault prediction model based on the DER approach is constructed. Section 4 provides several optimal learning models for training the parameters of the DER prediction model. In Sections 5 and 6, two numerical examples are provided to demonstrate the detailed implementation procedures and the validity of the proposed approach in the areas of fault prediction. This paper is concluded in Section 7. The proof of the recursive DER approach and analytical DER approach is detailed in Appendix A.

2. Dynamic evidential reasoning algorithm

The original ER approach is characterized by a distributed modeling framework capable of modeling both precise data and ignorance, by an evidential reasoning algorithm for aggregating both complete and incomplete information and by the interval utility for characterizing incomplete assessments and for ranking alternatives (Yang & Sen, 1994; Yang & Singh, 1994; Yang & Xu, 2002a, 2002b). It is of a static combination style in essence. Although evidence is collected with time, in static fusion all belief functions are combined simultaneously and the credibility of evidence source is unchanged over time. Since static combination takes no account of time effect, time information has no influence in combination results.

In engineering practice, however, system performance changes significantly with time and also system reliability is decreasing with time. As such, to trace system state accurately in time, there is a need to extend the original evidential reasoning approach and aggregate time information in the evidence combination process. In this section, a dynamic evidential reasoning approach is developed. In the new approach, time effect to the credibility of engineering system is taken into account, and it is also considered that the credibility of evidence is diminishing over time. It is assumed in the new approach that more recent evidence has greater influence in combination results. In other words, the final combination results can trace system dynamic behavior. This is one of the distinctive features of the new dynamic evidential reasoning approach from the original evidential reasoning approach. The detailed implementation process is described in detail in the following sections. Some basic concepts of the Dempster–Shafer theory of evidence are discussed first.

2.1. Basics of the evidence theory

The evidence theory developed by Dempster (1967) was extended and refined by Shafer (1976). The evidence theory is related to Bayesian probability theory in the sense that they both can update subjective beliefs given new evidence. The major difference between the two theories is that the evidence theory is capable of combining evidence and dealing with ignorance in the evidence combination process. The basic concepts and definitions of the evidence theory relevant to this paper are briefly described as follows.

Let $\Theta = \{F_1, \ldots, F_n\}$ be a collectively exhaustive and mutually exclusive set of hypotheses, called the frame of discernment. A basic probability assignment (BPA) is a function $m: 2^{\Theta} \to [0, 1]$, called a mass function and satisfying

$$m(\emptyset) = 0, \quad 0 \leqslant m(A) \leqslant 1 \text{ and } \sum_{A \subseteq \Theta} m(A) = 1 \tag{1}$$

where $\emptyset$ is an empty set, A is any subset of Θ, and 2^{Θ} is the power set of Θ, which consists of all the subsets of Θ, i.e. $2^{\Theta} = \{\emptyset, \{F_1\}, \ldots, \{F_N\}, \{F_1, F_2\}, \ldots, \{F_1, F_N\}, \ldots, \Theta\}$. The assigned probability (also called probability mass) $m(A)$ measures the belief exactly assigned to A and represents how strongly the evidence supports A. All assigned probabilities sum to unity and there is no belief in the empty set $\emptyset$. The probability assigned to Θ, i.e. $m(\Theta)$, is called the degree of ignorance. Each subset $A \subseteq \Theta$ such that $m(A) > 0$ is called a focal element of m. All the related focal elements are collectively called the body of evidence.

Associated with each BPA are a belief measure (Bel) and a plausibility measure (Pl) which are both functions: $2^{\Theta} \to [0,1]$, defined by the following equations, respectively:

$$Bel(A) = \sum_{F \subseteq A} m(F) \text{ and } Pl(A) = \sum_{A \cap F \neq \emptyset, F \subseteq \Theta} m(F) = 1 - Bel(\overline{A}) \tag{2}$$

where A and B are subsets of Θ. Bel(A) represents the exact support to A, i.e. the belief of the hypothesis A being true; Pl(A) represents the possible support to A, i.e. the total amount of belief that could be potentially placed in A. [Bel(A), Pl(A)] constitutes the interval of support to A and can be seen as the lower and the upper bounds of the probability to which A is supported. The two functions can be connected by the following equation:

$$Pl(A) = 1 - Bel(\overline{A}) \tag{3}$$

where $\overline{A}$ denotes the complement of A. The difference between the belief and the plausibility of a set A describes the ignorance of the assessment for the set A (Shafer, 1976).

Since $m(A)$, Bel(A) and Pl(A) are in one-to-one correspondence, they can be seen as three facets of the same piece of information. There are several other functions such as commonality function, doubt function, and so on, which can also be used to represent evidence. They all represent the same information and provide flexibility in a variety of reasoning applications.

The kernel of the evidence theory is the Dempster's rule of combination by which the evidence from different sources is combined. The rule assumes that the information sources are independence and use the orthogonal sum to combine multiple belief structures $m = m_1 \oplus m_2 \oplus \cdots \oplus m_n$, where $\oplus$ represents the operator of combination. With two belief structuresm_1 and m_2, the Dempster's rule of combination is defined as follows:

$$m_1 \oplus m_2(C) = \begin{cases} 0, & C = \emptyset \\ \frac{\sum_{A \cap B = C} m_1(A) m_2(B)}{1 - \sum_{A \cap B = \emptyset} m_1(A) m_2(B)}, & C \neq \emptyset \end{cases} \tag{4}$$

where A and B are both focal elements and $[m_1 \oplus m_2](C)$ itself is a BPA. The denominator, $1 - \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$ is called the normalization factor, and $\sum_{A \cap B = \emptyset} m_1(A) m_2(B)$ is called the degree of conflict, which measures the conflict between the pieces of evidence (George & Pal, 1996). Several researchers have investigated the combination rules of evidence theory and fuzzy sets. Note that the crude application of the D–S theory and the combination rule can lead to irrational conclusions in the aggregation of multiple pieces of evidence in conflict (Murphy, 2000).

Data fusion (or combination) can be seen as a static or as a dynamic problem. In static fusion, all belief functions are combined simultaneously. In dynamic fusion, belief functions are collected sequentially one by one (Smets, 2007). This fusion process results in a revision or an update of the current belief function. The distinction between static fusion and dynamic fusion is described in detail in the following sections.

2.2. Static fusion

In static fusion, all belief functions are combined simultaneously. Associativity and commutativity of combination rule should be satisfied, which is described in detail as follows:

Definition 1 (*Associativity*). The operator $\oplus$ is associativity if and only if for any three BPA functions m_1, m_2, m_3,

$$(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3) \tag{5}$$

Definition 2 (*Commutativity*). The operator $\oplus$ is commutativity if and only if for any two BPA functions m_1, m_2,

$$m_1 \oplus m_2 = m_2 \oplus m_1 \tag{6}$$

Definition 3. Let $m_1, m_2, \ldots, m_n$ be n BPA functions defined on Θ, let:

$$\bigoplus_{i=1,2\ldots,n} m_i = f_n(m_1, m_2, \ldots m_n) \tag{7}$$

where f_n maps n BPA functions into a BPA function.

So the requirements in static fusion can be denoted as follows:

$$f_n(m_1, m_2, \ldots m_n) = f_n(m_{i_1}, m_{i_2}, \ldots m_{i_n}) \tag{8}$$

where the index i_j are any permutation of the $i = 1, 2, \ldots, n$, indexes.

Associativity and commutativity of combination rule imply that the order under which the data are combined with f_n is irrelevant as the outcome is the same. In fact, associativity and commutativity of Dempster's combination rule form the basis for the ER algorithm (Huynh et al., 2006).

2.3. Dynamic fusion

When evidence is collected with time sequentially and time is meaningful, we may not permute the combination order with which evidence is collected. In this case, associativity is not an adequate requirement. All we need is that we can compute m_{123} from m_{12} and m_3, and we do not have to store the BPA functions that were at the origin of m_{12}. This requirement is called Markovian requirement (Smets, 2007). It just means that the present belief state summarizes the whole past. Markovian requirement just means that we can combine evidence at time t with the combination result at time $t-1$ to obtain final combination result as following definition.

Definition 1 (*Markovian requirement without memory decaying*). Let the BPA functions m_i, $i = 1, \ldots, n$, all be defined on the frame of discernment Θ and collected at time t_i, $i = 1, \ldots, n$ with $t_i > t_{i-1}$, for $i = 2, \ldots, n$. Let $f_n(m_1, m_2, \ldots, m_n)$ be the BPA function derived by combining all the BPA functions $m_1, m_2, \ldots, m_n$. The combination represented by f_n is qualified as Markovian if and only if there exists a g function that combines two BPA functions on Θ into a BPA function on Θ, such that:

$$f_n(m_1, m_2, \ldots, m_n) = g(g(\ldots g(g(m_1, m_2), m_3), \ldots m_{n-1}), m_n) \tag{9}$$

Considering that the credibility of evidence themselves is decaying over time, this decaying concept can be translated by claiming that every BPA function is discounted with time. In other words, the longer the time since the BPA functions were collected, the stronger the discounting is. So if a BPA function m_i has been collected since t unites of time, the BPA function is discounted by

$$m_i^{\alpha(t)}(A) = \alpha(t) m_i(A), \ \forall A \subseteq \Theta \tag{10}$$

$$m_i^{\alpha(t)}(\Theta) = 1 - \alpha(t) + \alpha(t) m_i(\Theta), \ \forall A \subseteq \Theta \tag{11}$$

where $1-\alpha(t)$ denotes the discount rate. $\alpha(t)$, the reliability factor, is a decreasing function of time with $\alpha(0)=1$ and $\lim_{t\to\infty}\alpha(t)=0$. In the following sections, $\alpha(t)=e^{-\gamma t}$ is selected as the reliability factor. We present one property of discounting. That is, iterated discounting is still regarded as discounting and the final reliability coefficient is the product of the individual reliability coefficients. This property is represented as follows:

Lemma 1. *Suppose there is a BPA function m and parameters* α_1, $\alpha_2 \in [0,1]$. *Then* $(m^{\alpha_1})^{\alpha_2} = m^{\alpha_1\alpha_2}$.

Proof. We have $(m^{\alpha_1})^{\alpha_2} = \begin{cases} \alpha_1\alpha_2 m(A), A \neq H \\ 1-\alpha_2+\alpha_2[1-\alpha_1+\alpha_1 m(H)] \end{cases} = \begin{cases} \alpha_1\alpha_2 m(A), A\neq H \\ 1-\alpha_1\alpha_2+\alpha_1\alpha_2 m(H) \end{cases} = m^{\alpha_1\alpha_2}$. Then we discount m^{α_1} at rate $1-\alpha_2$. The result is given by

$$(m^{\alpha_1})^{\alpha_2} = \begin{cases} \alpha_1\alpha_2 m(A), & A\neq H \\ 1-\alpha_2+\alpha_2[1-\alpha_1+\alpha_1 m(H)] \end{cases} = m^{\alpha_1\alpha_2}.$$

In such a case, we compute iteratively:

$$f_2(m_1,m_2) = g\left(m_1^{\alpha(dt_2)}, m_2\right) \tag{12}$$

$$f_2(m_1,m_2,m_3) = g\left(\left(m_1^{\alpha(dt_2)}, m_2\right)^{\alpha(dt_3)}, m_3\right) \tag{13}$$

$$f_2(m_1,m_2,m_3,m_4) = g\left(\left(\left(m_1^{\alpha(dt_2)}, m_2\right)^{\alpha(dt_3)}, m_3\right)^{\alpha(dt_4)}, m_4\right) \tag{14}$$

$\vdots$

$$f_n(m_1,m_2,\ldots,m_n) = g\left(g\left(\ldots g\left(g\left(\left((m_1)^{\alpha(dt_2)}, m_2\right)^{\alpha(dt_3)}, m_3\right)^{\alpha(dt_4)}, \ldots m_{n-1}\right)^{\alpha(dt_n)}, m_n\right)\right) \tag{15}$$

where $m_1^{\alpha(dt_2)}$ is the BPA that results from the discounting of m_1 by a rate $\beta(dt_2)=1-\alpha(dt_2)$, and $g(m_1^{\alpha(dt_2)}, m_2)^{\alpha(dt_3)}$ is the BPA obtained by discounting the BPA $g(m_1^{\alpha(dt_2)}, m_2)$ by a rate $\beta(dt_3)=1-\alpha(dt_3)$, etc. This property is defined as follows. □

Definition 2 (*Markovian requirement with memory decaying*). Let the BPA functions m_i, $i=1,\ldots,n$, be all defined on the frame of discernment Θ and collected at time t_i, $i=1,\ldots,n$ with $t_i > t_{i-1}$, for $i=2,\ldots,n$. Let $f_n(m_1, m_2,\ldots,m_n)$ be the BPA function derived by combining all the BPA functions $m_1,m_2,\ldots,m_n$. Let $dt_i = t_i - t_{i-1}$, for $i=2,\ldots,n$. There exists a g function that combines two BPA functions on Θ into a BPA function on Θ, such that

$$f_n(m_1,m_2,\ldots,m_n) = g\left(g\left(\ldots g\left(g\left(\left((m_1)^{\alpha(dt_2)}, m_2\right)^{\alpha(dt_3)}, m_3\right)^{\alpha(dt_4)}, \ldots m_{n-1}\right)^{\alpha(dt_n)}, m_n\right)\right) \tag{16}$$

According to Definition 2, we can obtain Dempster evidence combination rule for dynamic fusion, which takes account of time information and introduces the reliability factor $\alpha(t)$, the new combination rule is defined below.

Definition 4 (*Dempster evidence combination rule for dynamic fusion*). Let $m_1(t_1)$ and $m_2(t_2)$ be evidence collected at time t_1 and t_2, respectively. $m_1\oplus m_2(C,t)$ denotes the combination result of $m_1(t_1)$ and $m_2(t_2)$ at time t, $t\geqslant t_1$ and $t > t_2$. When applying the g function on the BPA functions $m_1(t_1)$ and $m_2(t_2)$, we first discount $m_1(t_1)$ by a rate $\beta(dt_1)=1-\alpha(dt_1)$ where$dt_1 = t - t_1$, and discount $m_2(t_2)$ by a rate $\beta(dt_2)=1-\alpha(dt_2)$ where $dt_2 = t-t_2$. Then, the discounted BPA functions are combined at time t. The way of combining $m_1(t_1)$ and $m_2(t_2)$ at time t is produced by

$$m_1\oplus m_2(C,t) = \begin{cases} 0, & C=\emptyset \\ \frac{\sum_{A\cap B=C} m_1^{\alpha(dt_1)}(A,t_1)m_2^{\alpha(dt_2)}(B,t_2)}{1-K(t)}, & C\neq\emptyset \end{cases} \tag{17}$$

where $K(t)=\sum_{A\cap B=\emptyset} m_1^{\beta(dt_1)}(A,t_1)m_2^{\beta(dt_2)}(B,t_2)$ denotes the degree of conflict between the pieces of evidence $m_1(t_1)$ and $m_2(t_2)$ at time t.

2.4. Dynamic evidential reasoning approach

To begin with, suppose there are L basic attributes $e_i(i=1,\ldots,L)$ associated with system fault state. Define a set of L basic attributes as evidence source as follows:

$$E = \{e_1,\ldots,e_L\} \tag{18}$$

Suppose the weights of the attributes are given by $w=\{w_1,\ldots,w_i,\ldots,w_L\}$ where w_i is the relative weight of the ith basic attribute e_i, and the weights of the attributes are normalized to satisfy the following constraints:

$$0\leqslant w_i\leqslant 1 \text{ and } \sum_{i=1}^{L} w_i = 1 \tag{19}$$

Define N distinctive fault state evaluation grades as represented by

$$F=\{F_1,\ldots,F_N\} \tag{20}$$

where F_n is the nth fault state evaluation grade. It is worth noting that F provides a complete set of standards for assessing attributes.

Mathematically, a given assessment for e_i $(i=1,\ldots,L)$ may be represented as the following distribution:

$$S(e_i(t_i)) = \{(F_n, \beta_{n,i}(t_i)), n=1,\ldots,N\}, \quad i=1,\ldots,L \tag{21}$$

where $\beta_{n,i}(t_i)\geqslant 0$, $\sum_{n=1}^{N}\beta_{n,i}(t_i)\leqslant 1$, and $\beta_{n,i}(t_i)$ denotes a degree of belief at time t_i. The above distributed assessment reads that the attribute e_i is assessed to the grade F_n with the degree of belief $\beta_{n,i}(t_i)$ at time t_i, $n=1,\ldots,N$. A assessment $S(e_i(t_i))$ is complete if $\sum_{n=1}^{N}\beta_{n,i}(t_i)=1$ and incomplete of $\sum_{n=1}^{N}\beta_{n,i}(t_i)<1$. $\beta_{n,i}(t_i)$ could be generated using various ways, depending on the nature of the attribute and data available such as a quantitative attribute using numerical values or a qualitative attribute using linguistic values (Yang, Liu, Xu, Wang, & Wang, 2007). To facilitate data collection, a scheme for handling various types of input information is summarized for the following cases (Wang et al., 2006; Yang, 2001; Yang et al., 2006a, 2006b):

(1) Quantitative attribute that is assessed using referential terms:
   (a) Rule-based or utility-based equivalence transformation techniques for quantitative data;
   (b) Transformation based on fuzzy membership function.
(2) Quantitative attributes that are assessed using interval;
(3) Qualitative attributes that are assessed using subjective judgments;
(4) Symbolic attributes that are assessed using subjective judgments.

Let $\beta_n(t)$ be a degree of belief to which the system fault state is assessed to the grade F_n at time t. The aggregation problem is to generate $\beta_n(t)(n=1,\ldots,N)$ by aggregating the assessments for all the associated attributes $e_i(i=1,\ldots,L)$ at time t_i. To this end, the following dynamic evidential reasoning approach can be used.

Let $m_{n,i}(t)$ be a basic probability mass representing the degree to which the ith basic attribute e_i at time t_i supports the hypothesis that the system fault state is assessed to the nth grade at time t. Let $m_{F,i}(t)$ be a remaining probability mass unassigned to any individual grade after all the N grades have been considered for assessing the system fault state as far as e_i is concerned at time t. According to Lemma 1, the basic probability mass can be calculated at time t as follows:

$$m_{n,i}(t) = w_i\alpha(dt_i)\beta_{n,i}(t_i) \quad n = 1,\ldots,N \; i = 1,\ldots,L \tag{22}$$

$$m_{F,i}(t) = 1 - \sum_{n=1}^{N} m_{n,i}(t) = 1 - w_i\alpha(dt_i)\sum_{n=1}^{N}\beta_{n,i}(t_i), \quad i = 1,2,\ldots,L \tag{23}$$

$$\bar{m}_{F,i}(t) = 1 - w_i\alpha(dt_i), \quad i = 1,2,\ldots,L \tag{24}$$

$$\tilde{m}_{F,i}(t) = w_i\alpha(dt_i)\left(1 - \sum_{i=1}^{N}\beta_{n,i}(t_i)\right), \quad i = 1,2,\ldots,L \tag{25}$$

$$m_{F,i}(t) = \bar{m}_{F,i}(t) + \bar{m}_{F,i}(t) \tag{26}$$

where $dt_i = t - t_i$. γ is the belief decaying coefficient and need be estimated using expert knowledge or optimization model.

Note that the probability mass assigned to the whole set $F, m_{F,i}(t)$ which is currently unassigned to any individual grades, is split into parts: $\bar{m}_{F,i}(t)$ and $\tilde{m}_{F,i}(t)$, where $\bar{m}_{F,i}(t)$ is caused by the relative importance of the attribute e_i and $\tilde{m}_{F,i}(t)$ by the incompleteness of the assessment on e_i for fault state at time t. As the original ER approach, all the basic probability assignments above are added to one based on the discounting operation (Huynh et al., 2006).

The second step is to aggregate the attributes by combining the basic probability masses generated above at time t, or reasoning based on the given evidence (Yang & Sen, 1994; Yang & Singh, 1994). Due to the assumptions that the evaluation grades are mutually exclusive and collectively exhaustive and that assessments on a basic attribute are independent of assessments on other attributes, the Dempster's combination rule can be directly applied to combine the basic probability masses in a recursive fashion. According to Definition 4, under the Markovian requirement with memory decaying, combination process can be developed into the following recursive DER algorithm.

$$\{F_j\} : m_{j,I(k+1)}(t) = K_{I(k+1)}(t)\big(m_{j,I(k)}(t)m_{j,k+1}(t) + m_{j,I(k)}(t)m_{F,k+1}(t) + m_{F,I(k)}(t)m_{j,k+1}(t)\big) \tag{27}$$

$$m_{F,I(k)}(t) = \bar{m}_{F,I(k)}(t) + \tilde{m}_{F,I(k)}(t), \quad k = 1,\ldots,L \tag{28}$$

$$\{F\} : \tilde{m}_{F,I(k+1)}(t) = K_{I(k+1)}(t)\big(\tilde{m}_{F,I(k)}\tilde{m}_{F,k+1}(t) + \tilde{m}_{F,I(k)}(t)\bar{m}_{F,k+1}(t) + \bar{m}_{F,I(k)}(t)\tilde{m}_{F,k+1}(t)\big) \tag{29}$$

$$\{F\} : \bar{m}_{F,I(k+1)}(t) = K_{I(k+1)}(t)\big(\bar{m}_{F,I(k)}(t)\bar{m}_{F,k+1}(t)\big) \tag{30}$$

$$K_{I(k+1)}(t) = \left[1 - \sum_{n=1}^{N}\sum_{m=1, m\neq n}^{N} m_{n,I(k)}(t)m_{m,k+1}(t)\right]^{-1}, \quad k = 1,\ldots,L-1 \tag{31}$$

$$\{F_j\} : \beta_j(t) = \frac{m_{j,I(L)}(t)}{1 - \bar{m}_{F,I(L)}(t)} \tag{32}$$

$$\{F\} : \beta_F(t) = \frac{\tilde{m}_{F,I(L)}(t)}{1 - \bar{m}_{F,I(L)}(t)} \tag{33}$$

where $m_{j,I(k)}(t)$ denotes the combined probability mass generated by aggregating k attributes at time t; $m_{j,I(k)}(t)m_{j,k+1}(t)$ measures the relative support to the hypothesis that the general attribute should be assessed to the grade F_j by both the first k attributes and the (k+1) th attribute at time t; $m_{j,I(k)}(t)m_{F,k+1}(t)$ measures the relative support to the hypothesis by the first k attributes only at time t; $m_{F,I(k)}(t)m_{j,k+1}(t)$ measures the relative support to the hypothesis by the $(k+1)$th attribute only at time t. It is assumed in the above equations that $m_{j,I(1)}(t) = m_{j,1}(t)(j = 1,2,\ldots,N)$, $m_{F,I(1)}(t) = m_{F,1}(t), \bar{m}_{F,I(1)}(t) = \bar{m}_{F,1}(t)$ and $\tilde{m}_{F,I(1)}(t) = \tilde{m}_{F,1}(t)$.

$\beta_j(t)$ and $\beta_F(t)$ represent the belief degrees of the aggregated assessment, to which the general attribute is assessed to the grades F_n and F at time t, respectively. The combined assessment can be denoted by $S(y) = \{(F_j, \beta_j(t)), j = 1,2,\ldots,N\}$. In the above DER algorithm, Eqs. (27)–(31) are the direct implementation of the Dempster's evidence combination rule for dynamic fusion; the weight normalization process shown in Eq. (19), the assignment of the basic probability masses shown in Eqs. (22)–(26) and the normalization of the combined probability masses shown in Eqs. (32) and (33) are developed to ensure that the DER algorithm can process conflicting evidence rationally and satisfy common sense rules for attribute aggregation in fault prediction (Huynh et al., 2006; Yang & Xu, 2002a, 2002b). Yang and Xu put forward the synthesis axioms for which any rational aggregation process should grant. Similarity, the following theorems that are taken for granted to develop the new DER algorithm above are due to Yang & Xu (2002a, 2002b).

Theorem 1. *The basic synthesis theorem. The degrees of belief defined by* (32) and (33) *satisfy*

$$0 \leqslant \beta_n(t), \quad \beta_F(t) \leqslant 1, \; n = 1,\ldots,N$$

$$\sum_{n=1}^{N}\beta_n(t) + \beta_F(t) = 1$$

Theorem 2. *The independent synthesis theorem. If $\beta_{n,i}(t_i) = 0$ for all $i = 1,\ldots,L$, then $\beta_n(t) = 0$.*

Theorem 3. *The consensus synthesis theorem. If $\beta_{n,i}(t_i) = 1$ and $\beta_{k,i}(t_i) = 0$ for all $k = 1,\ldots,N$ with $k \neq n$, $i = 1,\ldots,L$, then $\beta_n(t) = 1$ and $\beta_k(t) = 0$ for all $k = 1,\ldots,N$ with $k \neq n$ and $\beta_F(t) = 0$.*

Theorem 4. *The complete synthesis theorem. Assume $F^+ \subset F$ and $F^- = F \backslash F^+$ and denote $I^+ = \{n|F_n \in F^+\}$ and $I^- = \{k|F_k \in F^-\}$. If $\beta_{n,i}(t_i) > 0$ $(n \in I^+)$ and $\sum_{n\in I^+}\beta_{n,i}(t_i) = 1$, and $\beta_{k,i}(t_i) = 0$ for all $i = 1,\ldots,L$, then $\sum_{n\in I^+}\beta_n(t) = 1$ and $\beta_k(t) = 0$ for all $k \in I^-$.*

Theorem 5. *The incomplete synthesis theorem. If there exists $i \in \{1,\ldots,L\}$ such that $\sum_{n=1}^{N}\beta_{n,i}(t_i) < 1$, then $\sum_{n=1}^{N}\beta_n(t) < 1$.*

The independent synthesis theorem is naturally followed in the new DER algorithm as from (22)–(33) we have $\beta_n(t) = 0$ if $m_{n,i}(t) = 0$ for all $i = 1,\ldots,L$. Fortunately, the basic synthesis theorem, the consensus synthesis theorem, the complete synthesis theorem and the incomplete synthesis theorem also hold for the DER algorithm. The detailed proof will be given in Appendix A.

The recursive DER algorithm combines various piece of evidence on a one-by-one basis. The advantage of doing so lies in its clarity in concept. In situations where an explicit DER aggregation function is required such as in optimization, an analytical DER algorithm will be desirable (Wang et al., 2006). In view of this, the following analytical DER algorithm is developed:

$$\{F_j\} : m_j(t) = K_L(t)\left[\prod_{i=1}^{L}\big(m_{j,i}(t) + \bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)\big) - \prod_{i=1}^{L}\big(\bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)\big)\right] \tag{34}$$

$$\{F\} : \tilde{m}_F(t) = K_L(t)\left[\prod_{i=1}^{L}(\bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)) - \prod_{i=1}^{L}\bar{m}_{F,i}(t)\right] \tag{35}$$

$$\{F\} : \bar{m}_F(t) = K_L(t)\prod_{i=1}^{L}\bar{m}_{F,i}(t), \quad j = 1,\ldots N \tag{36}$$

$$K_L(t) = \left[\sum_{j=1}^{N} \prod_{i=1}^{L} (m_{j,i}(t) + \bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)) - (N-1) \prod_{i=1}^{L} (\bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)) \right]^{-1} \tag{37}$$

$$\{F_j\} : \beta_j(t) = \frac{m_j(t)}{1 - \bar{m}_F(t)} \tag{38}$$

$$\{F\} : \beta_F(t) = \frac{\tilde{m}_F(t)}{1 - \bar{m}_F(t)} \tag{39}$$

The equivalence between the recursive and analytical DER algorithms is proved in Appendix A. The analytical DER algorithm offers the DER algorithm more flexibility in aggregating a large number of basic attributes. The original ER approach has the nonlinear features (Yang & Xu, 2002a, 2002b). The analytical DER algorithm clearly shows nonlinear features and provides a straightforward way to conduct sensitivity analysis for the parameters of the DER algorithm such as attributes weights and belief decaying factors. It also facilitates the estimation and optimization of these parameters.

In the above equations, different from the DER algorithm that is of a recursive nature, Eqs. (34)–(37) provide an analytical means for combining all BPA function in one go without iteration, thus facilitating parameter optimization.

Note that the original evidential reasoning algorithm is a special case of dynamic evidential reasoning algorithm with $\alpha(t) = 1$, that is $\gamma = 0$. In this case, time factor has no influence on belief decaying and the reliability of evidence source is not decreasing with time.

3. Fault prediction model based on dynamic evidential reasoning algorithm

3.1. Fault prediction model structure and representation

In this section, fault prediction model is investigated in the DER framework. It is assumed that a set of observed data is provided in the form of input–output pairs $(\mathbf{x}(t_m), y(t_m))$, $m = 1,\ldots,M$, with $\mathbf{x}(t_m)$ being an input vector of the actual system at time t_m and $y(t_m)$ being a scalar representing the corresponding output value or subjective distribution value of the actual system at time t_m.

For prediction problems, the inputs used in a prediction model are the past input vector and the lagged observations of the current time, while the outputs are the future values. Each set of input patterns is composed of any moving fixed-length window within the time series of the input data. The general fault prediction model can be represented as

$$\hat{y}(t+k-1) = f(\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_{t-p}) \tag{40}$$

where $\hat{y}(t+k-1)$ is a scalar representing the predicted value at time $t-k-1$, $(\mathbf{x}_{t-1},\mathbf{x}_{t-2},\ldots,\mathbf{x}_{t-p})$ is a vector of lagged variables, and p represents the dimensions of the input vector (number of input nodes) or the number of past inputs related to the future value. In fault prediction, it is reasonable to assume that the current output value y_t is related to the most recent input vector $\mathbf{x}_{t-1}$ or even extended into the past values $\mathbf{x}_{t-2},\ldots,\mathbf{x}_{t-p}$. This is because the next output value is dependent on the current output value to certain extent which is in turn related to the current inputs. For simplicity, let $\mathbf{X}(t) = (\mathbf{x}_{t-1},\mathbf{x}_{t-2},\ldots,\mathbf{x}_{t-p})$ denote the input vector of the prediction model.

Table 1
The pattern of training data set.

X				Y
$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_p$	y_{p+1}
$\mathbf{x}_2$	$\mathbf{x}_3$	$\cdots$	$\mathbf{x}_{p+1}$	y_{p+2}
$\mathbf{x}_3$	$\mathbf{x}_4$	$\cdots$	$\mathbf{x}_{p+2}$	y_{p+3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{M-p}$	$\cdots$	$\cdots$	$\mathbf{x}_M$	y_{M+1}

3.2. Fault prediction model using DER approach

In order to apply the DER approach, suppose $(\mathbf{x}_{t-1},\mathbf{x}_{t-2},\ldots,\mathbf{x}_{t-p})$ are p basic attributes associated with system prediction output y_t, and the DER approach attempts to identity the appropriate internal representation between $(\mathbf{x}_{t-1},\mathbf{x}_{t-2},\ldots,\mathbf{x}_{t-p})$ and $\hat{y}_t$. The key for solving the prediction problem is how to approximate the function f.

Table 1 shows how training patterns can be designed in the DER prediction model. In Table 1, p denotes the number of lagged variables, while $(t\text{-}p)$ represents the total number of data samples. Moreover, X is the input nodes and Y denotes the predicted output node. Following successful training, the DER prediction model can predict future outcomes y_{t+k} at k different time steps. If $k = 1$, the prediction is a one-step-ahead prediction, if $k > 1$, the prediction is a multi-step prediction. Although multi-step forecasting may capture some system dynamics, the performance may be poor due to the accumulation of errors. In practice, one-step-ahead prediction results are more useful since they provide timely information for preventive and corrective maintenance plans (Xu, Xie, Tang, & Ho, 2003). In addition, according to Wang (2007), the more the step ahead is, the less reliable the forecasting operation is because multi-step prediction is associated with multiple one-step prediction operation. Thus, this study only considers one-step-ahead prediction fault prediction.

In fault prediction problems, DER can be trained first to learn relationships between past historical drift data and the corresponding targets, and then future output values can be predicted. Due to the fact that the input data $\mathbf{x}(t_m)$ may be a numerical value or a subjective distribution, there is a need to transform such data into the belief structure. Rule or utility equivalence transformation techniques can be used in this case, more discussions on this issue can be found in Yang (2001). As a result, each input can be represented as a distribution on referential values using a belief structure. The main advantage of doing so is that precise data, random numbers, and subjective judgments with uncertainty can be consistently modeled under the same framework (Yang, 2001; Yang & Xu, 2002a, 2002b; Yang et al., 2007).

Having represented each attribute using Eq. (21), the DER approach can be directly applied to combine all attributes and generate final conclusions. Using the DER analytical algorithm (34)–(39), the final prediction result $O(\hat{y}(t))$ that is generated by aggregating all attributes $(\mathbf{x}_{t-1}0,\mathbf{x}_{t-2},\ldots,\mathbf{x}_{t-p})$ can be represented as follows:

$$O(\hat{y}(t)) = \{(F_j, \hat{\beta}_j(t)), (F, \hat{\beta}_F(t))j = 1, \ldots, N\} \tag{41}$$

where $\hat{\beta}_j(t)$ can be obtained by the analytical DER algorithm as follows:

$$\beta_j(t) = \frac{\prod_{i=1}^{L}(w_i\alpha(dt_i)\hat{\beta}_{j,i}(t_i) + 1 - \alpha(dt_i)w_i + w_i\alpha(dt_i)\beta_{F,i}(t_i)) - \prod_{i=1}^{L}(1 - \alpha(dt_i)w_i + w_i\alpha(dt_i)\beta_{F,i}(t_i))}{D(t)} \tag{42}$$

$$\hat{\beta}_F(t) = \frac{\prod_{i=1}^{L}(1-\alpha(dt_i)w_i + w_i\alpha(dt_i)\beta_{F,i}(t_i)) - \prod_{i=1}^{L}(1-\alpha(dt_i)w_i)}{D(t)} \tag{43}$$

$$\begin{aligned} D(t) = \sum_{q=1}^{N}\prod_{i=1}^{L}(w_i\alpha(dt_i)\beta_{q,i}(t_i) + 1 - \alpha(dt_i)w_i \\ + w_i\alpha(dt_i)\beta_{F,i}(t_i)) - (N-1)(1-\alpha(dt_i)w_i \\ + w_i\alpha(dt_i)\beta_{F,i}(t_i)) - \prod_{i=1}^{L}(1-\alpha(dt_i)w_i) \end{aligned} \tag{44}$$

The aggregated distributed assessment $O(\hat{y}(t)) = \{(F_j, \hat{\beta}_j(t)), (F, \hat{\beta}_F) j = 1, \ldots, N\}$ represents the overall assessment of fault. It provides a panoramic view about the fault state at time t, from which one can tell which assessment grades the fault is assessed to, and what belief degrees are assigned to the defined fault grades F_j, $j = 1,\ldots,N$. However, if the output value $\hat{y}(t)$ is a numerical value, it is desirable to generate numerical values equivalent to the distributed assessments Eq. (41) in a sense. The concept of expected utility is used to define such values (Xu et al., 2007; Yang & Xu, 2002a, 2002b). Suppose $u(F_i)$, $i = 1,\ldots,N$ are the utility of the fault grade F_i, $i = 1,\ldots,N$, respectively, with $u(F_i) < u(F_j)$. If for fault prediction problem, F_j is preferred to F_i. $u(F_i)$, $i = 1,\ldots,N$ may be estimated using prior expert knowledge or by constructing optimization models, in this paper, we select the later.

In order to generate numerical output values equivalent to the distributed assessments Eq. (41), maximum, minimum and average utilities are introduced. Suppose the utility of an evaluation grade F_n is $u(F_n)$. Then the expected utility of the aggregated assessment $O(\hat{y}(t))$ is defined as follows:

$$u(O(\hat{y}(t))) = \sum_{n=1}^{N}\hat{\beta}_n(t)u(F_n) \tag{45}$$

The belief degree $\hat{\beta}_n(t)$ stands for the lower bound of the likelihood that $\hat{y}_t$ is assessed to H_n, whilst the corresponding upper bound of the likelihood is given by $(\hat{\beta}_n(t) + \hat{\beta}_H(t))$ (Yang, 2001; Yang & Xu, 2002a, 2002b), which leads to the establishment of a utility interval if the assessment is incomplete. Without loss of generality, suppose the least preferred assessment grade having the lowest utility is F_1 and the most preferred assessment grade having the highest utility is F_n. Then the maximum, minimum and average utilities of $\hat{y}(t)$ can be calculated by

$$u_{\max}(O(\hat{y}(t))) = \sum_{n=1}^{N-1}\hat{\beta}_n(t)u(F_n) + (\hat{\beta}_N(t) + \hat{\beta}_F(t))u(F_N) \tag{46}$$

$$u_{\min}(O(\hat{y}(t))) = (\hat{\beta}_1(t) + \hat{\beta}_F(t))u(F_1) + \sum_{n=2}^{N}\hat{\beta}_n(t)u(F_n) \tag{47}$$

$$u_{avg}(O(\hat{y}(t))) = \frac{u_{\max}(O(\hat{y}(t))) + u_{\min}(O(\hat{y}(t)))}{2} \tag{48}$$

If original distribution assessments $O(\hat{y}(t))$ in Eq. (41) are all complete, then $\hat{\beta}_F(t) = 0$ and $u_{\min}(\hat{y}(t)) = u_{\max}(\hat{y}(t)) = u_{avg}(\hat{y}(t))$. Having obtained the outcome shown in Eq. (41), the prediction outcome $\hat{y}(t)$is calculated as follows:

$$\hat{y}(t) = u_{avg}(O(\hat{y}(t))) = \sum_{n=1}^{n=N}\hat{\beta}_n(t)u(F_n) + \frac{u(F_1)+u(F)}{2}\hat{\beta}_F(t) \tag{49}$$

where $\hat{y}(t)$ represents the prediction outcome. If the distributed fault assessments are complete and precise, there will be $\hat{\beta}_F(t) = 0$ and the prediction outcome $\hat{y}(t)$ is reduced to

$$\hat{y}(t) = \sum_{j=1}^{N}\hat{\beta}_j(t)u(F_j). \tag{50}$$

The ER approach does not need to specify the functional relationship between outputs and inputs. The ER approach is a nonlinear aggregation method in nature (Yang & Xu, 2002b). So the DER approach is also a nonlinear aggregation method. Moreover, the DER algorithm considers that time information and belief are decaying with time. Therefore, the functional relationship between output and inputs in DER model is nonlinear in nature, and DER approach provides scope and flexibility for on-line prediction.

From Eqs. (42), (25) and (49), it can be seen that the attribute weights w_i, belief decaying coefficient$\gamma$and the grade utilities $u(F_j)$ play a significant role in the final conclusion $O(\hat{y}(t))$ or $\hat{y}(t)$. The degree to which the final output can be affected is determined by the magnitude of the attribute weights, the belief decaying coefficient and the grades utilities. On the other hand, if the parameters of a DER prediction model such as w_i, γ, $u(F_j)$, $j = 1,\ldots,N$ are not given *a priori* or only known partially or imprecisely, they could be trained using observed input and output information.

However, it is difficult to determine the parameters of the DER model accurately. In addition, a change in attribute weight or belief decaying factor may lead to changes in the performance of the DER prediction model. As such, there is a need to develop a method that can optimally learn parameters using observed input and output information. This is exactly the topic for the rest of this paper.

4. Optimal learning algorithm of DER prediction model

In this section, optimization models and procedures are investigated in the DER framework to help search for optimally trained belief decaying factor and weights and expected utilities simultaneously. Fig. 1 shows the process of training a DER, where $\mathbf{X}(t)$ is a given input; $y(t)$ is the corresponding actual observed output, either measured using instruments or assessed by experts; $\hat{y}(t)$ is the simulated output that is generated by the DER prediction model; and $\xi(P)$ is the difference between $y(t)$ and $\hat{y}(t)$, as defined later.

It is desirable that $\xi(P)$ is as small as possible where P is the vector of training parameters including w_i, $i = 1, \ldots, L$, $u(F_j)$, $j = 1,\ldots,N$. and γ. This objective is difficult to achieve if the DER prediction model is constructed using expert judgments only. Several learning models are designed to adjust the parameters in order to minimize the difference between the observed output $y(t)$ and the simulated output $\hat{y}(t)$, i.e.,$\xi(P)$. Such an optimally trained DER model may then be used to predict the behavior of the system. In general, the optimal learning problem can be represented as the following nonlinear multiple-objective programming problem:

$$\begin{aligned} &\min\ f(P) \\ &\text{s.t.}\quad A(P) = 0,\ B(P) > 0 \end{aligned} \tag{51}$$

where $f(P)$ is the objective function, P is the training parameter vector, $A(P)$ is the equality constraint functions, and $B(P)$ is the inequality constraint functions.

In the learning process, a set of observations on the system inputs and outputs is required. In the following, we assume that a set of observation pairs $(\mathbf{x}(t_m), y(t_m))$, $m = 1,\ldots,M$ is available, where

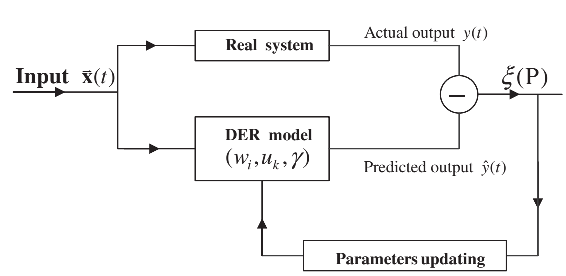

Fig. 1. Illustration of optimal learning process for DER prediction model.

$\mathbf{x}(t_m)$ is an input vector and $y(t_m)$ is the corresponding output. Both $\mathbf{x}(t_m)$ and $y(t_m)$ can be either numerical, judgmental, or both. The format of the objective function is important for the parameter optimization. Depending on the types of input and output, the optimal learning models can be constructed in different ways, as discussed in detail in the next sections.

For fault prediction problem, the relative weights of attributes w_i, $i = 1,\ldots,L$, the utility of assessment grades $u(F_j)$, $j = 1,\ldots,N$. and the coefficient of belief decaying γ are all unknown. In order to train parameters from input–output pairs, the optimization model needs to be constructed.

4.1. Optimal learning model based on numerical output

In this case, it is assumed that a set of observed training data is provided in the form of input–output pairs $(x(t_m), y(t_m))$, $m = 1,\ldots,M$, with $y(t_m)$ being a numerical output value of the actual system at time t_m. The output that is shown in (41) is represented as a distribution, and its average utility is given by

$$\hat{y}(t) = \sum_{n=1}^{n=N} \hat{\beta}_n(t)u(F_n) + \frac{u(F_1)+u(F)}{2}\hat{\beta}_F(t) \tag{52}$$

where $\hat{y}(t)$ is considered to be the predicted output at time t from DER prediction approach.

For a fault prediction problem, the relative weight of attribute w_i, the utilities of assessment grades $u(F_j)$, $j = 1,\ldots,N$ and the coefficient of belief decaying γ are all unknown. In order to train and learn these parameters from input–output pairs, the optimization model is constructed as follows:

$$\min J = \xi(P) = \frac{1}{M}\sum_{i=1}^{M}(y(t_i) - \hat{y}(t_i))^2 \tag{53}$$

where $y(t_i)$ denotes the actual output data of a system at time t_i and $\hat{y}(t_i)$ denotes the predicted output data of the system at time$t_i$from the DER prediction model, which can be obtained from Eq. (49). w_i, γ, $u(F_j)$, $j = 1,\ldots,N$ are the relative weights of attributes, the belief decaying factor and the utilities of fault assessment grades to be estimated, respectively. P is the parameter vector including w_i, $u(F_j)$, $j = 1,\ldots,N$ and $\boldsymbol{\gamma}$. $\boldsymbol{M}$ is the number of training data in the input–output pairs. $(y(t_i) - \hat{y}(t_i))$ is the residual of training data at time t_i.

The construction of the constraints of the learning model is given as follows.

(1) An attribute weight is normalized, so that it is between zero and one and the total weights will be equal to one, e.g.,

$$0 \leqslant w_i \leqslant 1, \quad i = 1,\ldots,L \tag{54}$$

$$\sum_{i=1}^{L} w_i = 1 \tag{55}$$

(2) The belief decaying factor is a nonnegative number, e.g.,

$$\gamma \geqslant 0 \tag{56}$$

(3) For numerical data, the utility of the fault state evaluation grade is a nonnegative number, e.g.,

$$u(F_j) \geqslant 0, \quad j = 1,\ldots,N \tag{57}$$

And if F_j is preferred to F_i, then

$$u(F_j) > u(F_i) \tag{58}$$

Such a nonlinear optimization model can be solved using MATLAB optimization toolbox. Once the weights, the scores of grades and the belief decaying factor are learned, the DER prediction model can be used for fault prediction.

4.2. Optimal learning model based on belief distribution output

In this case, a set of observed training data is assumed to be composed of M input–output pairs $(x(t_m), y(t_m))$, $m = 1,\ldots,M$, with $y(t_m)$ being a belief distribution structure and represented using a distributed assessment with different degrees of belief as follows:

$$y(t_m) = \{(F_j, \beta_j(t_m)), \quad j = 1,\ldots,N\} \tag{59}$$

where F_j is a fault state evaluation grade of system obtained by DER prediction model, and $\beta_j(t_m)$ is the degree of belief to which F_j is assessed by the mth pair of observed data at time t_m.

Using the same referential terms as for the observed output $y(t_m)$, a belief distribution conclusion that is generated by aggregating all the BPA functions can also be represented as follows:

$$\hat{y}(t_m) = \{(F_j, \hat{\beta}_j(t_m)), j = 1,\ldots,N\} \tag{60}$$

where $\hat{\beta}_j(t_m)$ is generated by DER prediction model using Eqs. (38) and (39) for a given input. It is desirable that, for a given input $\mathbf{x}(t_m)$, the DER prediction model can generate an output, which is represented as $\hat{y}(t_m) = \{(F_j, \hat{\beta}_j(t_m)), j = 1,\ldots,N\}$ and can be as close to $y(t_m)$ as possible. In other words, for the mth pair of the observed data $(x(t_m), y(t_m))$, the DER prediction model is trained to minimize the difference between the observed belief $\beta_j(t_m)$ and the belief $\hat{\beta}_j(t_m)$ that is generated by the DER prediction model for each referential term. Such a requirement is true for all pairs of the observed data. This leads to the definition of the objectives for all the referential output terms as follows:

$$\min_{Q}\{\varepsilon_j(Q), j = 1,\ldots,N\} \tag{61}$$

s.t. Eqs. (34)–(36)

where

$$\varepsilon_j(Q) = \frac{1}{M}\sum_{m=1}^{M}\left(\beta_j(t_m) - \hat{\beta}_j(t_m)\right)^2, \quad j = 1,\ldots,N \tag{62}$$

$(\hat{\beta}_j(t_m) - \beta_j(t_m))$ is the residual at the mth data point, and $Q = \{w_1,\ldots,w_L,\gamma\}$ is the training parameter vector without $u(F_j)$ because F_j does not need to be quantified in this case, that is, the training parameters include attribute weights and belief decaying factor only. The constraints are given by (34)–(36).

The optimal training problem, formulated in Section 4.3, is to minimize the numerical difference between the observed output and the simulated output. It is therefore a single objective nonlinear optimization problem. The training problem for the belief distribution output is a multiple objective nonlinear optimization problem with N objective that are defined as in (41) and (42), $L + 1$ training parameters as given (w_i, γ) and $2L + 2$ constraints, as defined in Eqs. (54)–(56).

This multiple objective nonlinear optimization problem can be recast as follows:

$$\begin{aligned} &\min_{Q}\max_{\{\varepsilon_j\}}\{\varepsilon(Q), j = 1,\ldots,N\} \\ &\varepsilon_j(Q) = \frac{1}{M}\sum_{m=1}^{M}\left(\beta_j(t_m) - \hat{\beta}_j(t_m)\right)^2 \end{aligned} \tag{63}$$

s.t. Eqs. (54)–(56)

This can be solved in MATLAB optimization toolbox.

4.3. Optimal learning model based on mixed output

In the mixed case, a set of observed training data is assumed to be composed of M input–output pairs $(x(t_m), y(t_m))$, $m = 1,\ldots,M$, with $y(t_m)$ being either a numerical value or a subjective output that is represented using a distribution with different degrees of belief. Without loss of generality, suppose that the first M_1 pairs of training data are subjective outputs, and the last $M_2 = M - M_1$ pairs of training data are numerical values. In this case, the optimization problem can be formulated as the following multiple objective optimization problem for minimizing the differences between the outputs of the DER prediction model and the corresponding observed outputs in both the numerical and subjective formats:

$$\min_{P}\{\varepsilon_1(Q), \varepsilon_2(Q), \ldots, \varepsilon_N(Q), \varepsilon(P)\} \tag{64}$$

s.t. Eqs. (54), (56)–(58)

where P is the training parameter vector, and ε_j is the total mean squared error for the jth referential term that is given as follows:

$$\varepsilon_j = \frac{1}{M_1}\sum_{m_1=1}^{M_1}(\beta_j(t_{m_1}) - \hat{\beta}_j(t_{m_1}))^2 \tag{65}$$

$\beta_j(t_{m_1})$ is the observed belief degree for the jth consequent F_j; corresponding to the m_1th observed output in the training data set ($m_1 = 1,\ldots,M_1$), and $\hat{\beta}_j(t_{m_1})$ is the belief degree for F_j that is given by Eq. (38) corresponding to the m_1th input of the training data set. ε is the total mean squared error for all the numerical data set given as follows:

$$\varepsilon = \frac{1}{M_2}\sum_{m_2=M_1+1}^{M_2}(y(t_{m_2}) - \hat{y}(t_{m_2}))^2 \tag{66}$$

where $\hat{y}(t_{m_2}) = \sum_{j=1}^{N}u(F_j)\hat{\beta}_j(t_{m_2}) + \frac{u(F_1)+u(F)}{2}\hat{\beta}_F(t_{m_2})$, $\hat{\beta}_j(t_{m_2})$ is given by (38) for the m_2th input in the numerical training data set, and $y(t_{m_2})$ is the observed numerical output ($m_2 = M_1 + 1,\ldots,M$).

The objective that is given by (64) has $N + 1$ nonlinear objective functions and can be solved using the following minimax formulation to generate efficient solutions:

$$\min_{P}\max_{\{\varepsilon_j,\varepsilon\}}\{\varepsilon_1(Q), \varepsilon_2(Q), \ldots \varepsilon_n(Q), \varepsilon(P)\}$$
$$\begin{cases} \varepsilon_j = \frac{1}{M_1}\sum\limits_{m_1=1}^{M_1}(\beta_j(m_1,t) - \hat{\beta}_j(m_1,t))^2 \\ \varepsilon = \frac{1}{M_2}\sum\limits_{m_2=1}^{M_2}(y_{m_2}(t) - \hat{y}_{m_2}(t))^2, y_{m_2}(t) = \sum\limits_{j=1}^{N}u(F_j)\beta_j(m_2,t) \end{cases} \tag{67}$$

s.t. Eqs. (54)–(58)

By using these nonlinear optimization models, once the input–output data become available, the related parameters to DER prediction model can be obtained rationally. Then the system fault state can be predicted when the new information becomes available. Fig. 2 shows the basic architecture of DER prediction model.

Complex engineering system such as missile has multiform faults that are dangerous and have a short latent period. Experimented with fault can be a great expense and damage people's life. So the data or prior knowledge of faults, which are required for fault prediction, could hardly be acquired. In other words, the data of normal situation can be acquired only. Therefore, in this study, fault prediction only with data of normal condition is provided to demonstrate the prediction performance of the new prediction methods.

The implementation procedures of the new techniques based on the dynamic evidential reasoning approach will be detailed in the next two sections.

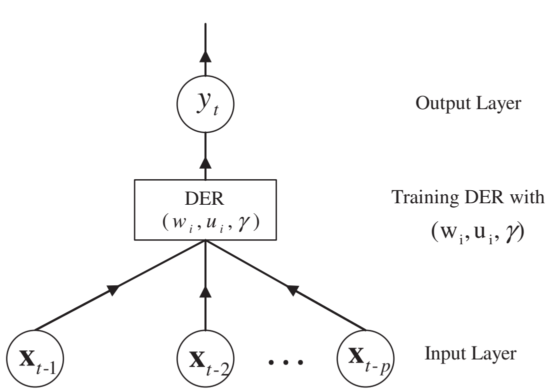

Fig. 2. The basic architecture of DER prediction model.

5. An simulation case study—tank fault prediction

In this section, a numerical example is given to demonstrate the implementation process of the new fault prediction model based on DER approach.

5.1. Problem description of the two tanks fault prediction problem

Here we choose to analyze a system with two tanks as shown in Fig. 3. Water flows into the first tank at the rate Q_1 (m^3/s), then flows to the second tank at the rate Q_{12} (m^3/s) and finally flows out of the second tank at the rate Q_{20} (m^3/s). Suppose that a slow jam fault happens at the output of tank 2.

The dynamic model of the two tanks is given as follows:

$$\begin{cases} A\dot{h}_1 = Q_1 - Q_{12} \\ A\dot{h}_2 = Q_{12} - Q_{20} \end{cases} \tag{68}$$

With

$$\begin{cases} Q_{12} = a_1 S sign(h_1 - h_2)\sqrt{2g|h_1 - h_2|} \\ Q_{20} = a_2 S\sqrt{2gh_2} \end{cases} \tag{69}$$

Where $a_i, h_i (i = 1, 2)$ are the output coefficients and the liquid level of the two tanks, *sign(z)* is the sign of the argument z, S is the section area of the connection pipe (m^2), A is the section area of the two tanks which are of the same size, and T_s is the sampling period. The nominal technical data of the setup is shown in Table 2.

Suppose that a_2 decreases along the time as follows:

$$\varphi(t) = a_2 - 0.001t \tag{70}$$

where $\varphi(t)$ is the outflow coefficient of the second tank when the system is in the slow jam fault.

In order to verify the proposed fault prediction method, we will use the Level 1 h_{t-1} and Level 2 h_{t-2} at sample time $t - 1$ as the inputs of the DER prediction model, and the outflow coefficient $a_2(t)$ which is denoted by OC at time t as the output. In other words, the two levels are considered as the basic attributes and the outflow

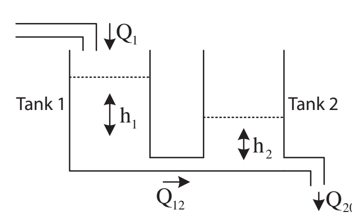

Fig. 3. The layout of the two tanks.

Table 2
Some technical parameters of the double tanks.

$A = 0.15\ m^2$	$S = 0.00005\ m^2$	$Q_1 = 0.00005\ m^3/s$	$T_s = 1$ s
$g = 9.8\ cm^2/s$	$a_1 = 0.4$	$a_2 = 0.5$	

coefficient is the consequent in the DER prediction model. In order to generate the simulated fault data, we let $A = 0.15\ m^2$, $S = 0.00005\ m^2$, $T_s = 1$ s, $Q_1 = 0.00005\ m^3/s$, $g = 9.8\ cm/s^2$, $a_1 = 0.4$, the initial value of $a_2 = 0.5$, and the sampling time $T_s = 1$ s. The simulated data in the slow jam fault is obtained through Matlab, and we select the simulated data from time zero second to 325 s as experimental data to demonstrate the proposed method. The levels of the two tanks during the slow jam fault are given in Fig. 4.

5.2. Prediction model based on DER approach for two tanks

When artificial intelligence technology is applied to the forecasting of time series, the number of input nodes (order of autoregressive terms) critically affects the forecasting performance [30], since the number of input nodes precisely describes the correlation between lagged-time observations and future values. Thus, the number of input nodes is important in determining the structure of autoregressive model within a time series. For a linear time series forecasting, many previous research articles demonstrate that the most common order of autoregressive terms is 1 or 2, but only a few particular problems have a higher order of 3 or more (Box & Jenkins, 1970; Pankratz, 1983). For nonlinear time series problems, the number of autoregressive terms typically used for real life applications has not been reported. Therefore, this study experimented with a relatively smaller number 1 for the order of autoregressive terms. Therefore, we can transform 326 observation values to 325 input patterns. So the prediction model can be constructed as

$$y_t = a_2(t) = f(\mathbf{x}_{t-1}) \tag{71}$$

where the elements of the input vector $\mathbf{x}_{t-1}$ is $h_1(t-1)$ and $h_2(t-1)$, and the relationship among $h_1(t-1)$, $h_2(t-1)$ and $a_2(t)$ is shown in Fig. 5, from which we can see the nonlinear relationship among $h_1(t-1)$, $h_2(t-1)$ and $a_2(t)$. In what follows, we apply the DER approach to approximate the function f.

5.3. Transformation of original information using the rule based method

As the input vector $\mathbf{x}_{t-1}$ is numerical values, there is a need to transform them to the belief distribution structure. For Level 1 and Level 2, the same referential points are used and they are 'Small' (F_1), 'Medium' (F_2), and 'Large' (F_3). Define the following discernment framework of tanks fault state:

$$F = \{F_j, j = 1,2,3\} = \{Small, Medium, Large\} \tag{72}$$

All factors related to the tank fault may then be assessed with reference to this framework using the rule-based information transformation technique (Yang, 2001).

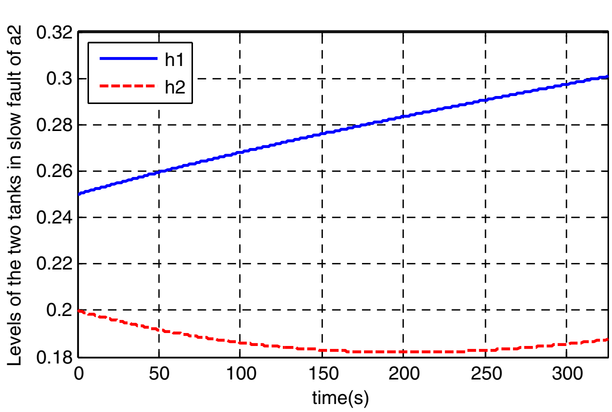

Fig. 4. The levels of the two tanks in slow fault.

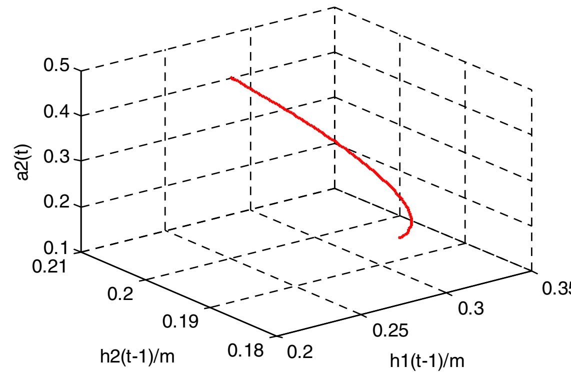

Fig. 5. The nonlinear relationship among $h_1(t-1)$, $h_2(t-1)$ and $a_2(t)$.

Table 3
The referential points of level 1.

Linguistic terms	Small (F_1)	Medium (F_2)	Large (F_3)
Numerical values (°/h)	0.2	0.28	0.32

Table 4
The referential points of level 2.

Linguistic terms	Small (F_1)	Medium (F_2)	Large (F_3)
Numerical values (°/h)	0	0.2	0.25

For Level 1, it is assumed that equivalence rules can be acquired. The first equivalence rule reads as follows:

1. If Level 1 is 0.2 m, then the probability that the tank fault could occur is estimated at a "Small" level. This rule may be represented by a simple statement 'If Level 1 is 0.2 m, then fault probability is small' (or $F_1 = 0.2$).
   Other equivalence rules could be acquired in a similar way as follows:
2. If Level 1 is 0.28 m, then fault probability is medium (or $F_2 = 0.28$).
3. If Level 1 is 0.32 m, then fault probability is large (or $F_3 = 0.32$).

The quantified results above are listed in Table 3 as follows:

The referential points of Level 2 can be quantified in a similar way, and the results are listed in Table 4.

The given values for the input vector $\mathbf{x}_t = [h_1(t-1), h_2\ (t-1)]$ can be transformed into the belief structures with respect to the defined three referential points using the information transformation technique. Take the input vector at time 20 s ($h_1(20) = 0.2539$, $h_2(20) = 0.1963$) for example. For $h_1(20) = 0.2539$, this can be equivalently transformed to belief distribution structure as follows:

$$S(h_1(20)) = \{(F_1, 0.3283), (F_2, 0.6717), (F_3, 0)\}$$

Similarly, $h_2(20)$ can be equivalently transformed to

$$S(h_2(20)) = \{(F_1, 0.0176), (F_2, 0.9824), (F_3, 0)\}$$

The other given input training data sets can be transformed in a similar way.

6. Simulation results using DER prediction model based on numerical output

To validate the DER prediction model, the available data are partitioned into a training data set and a test data set. The training data set is used for training the DER prediction model parameters. The trained DER prediction model is then used to predict outputs for the test input data.

For illustration purpose, the first 250 sets of data sampled at time 1 (the 1st second) to time 250 (the 250th second) are used as the training data for parameter estimation. After training, the remaining 75 sets of data are used for testing the trained DER prediction model. The initial parameter vector P is given by experts. The initial attribute weights are all set to be equivalent. The initial parameter vector P is set to

$$P = [w_1, w_2, u(F_1), u(F_2), u(F_3), \gamma] = [0.5, 0.5, 0.5, 0.25, 0.175, 0.1].$$

Due to the numerical output, there is a need to solve the nonlinear optimization that is given by Eq. (53). To solve the single objective model given by Eq. (53), we use FMINCON function in Matlab. In this example, the error tolerance is set to 0.000001, and the maximum iteration is set to 60 to avoid dead loop in the optimal learning process. After training, the trained parameter vector P can be obtained as follows:

$$P = [0.9007, 0.0993, 0.8477, 0.3241, 0.0018, 0.2485],$$

where $w_1 = 0.9007 > w_2 = 0.0993$ denotes that Level 1 has more influence than Level 2, and the utilities $u(F_1) > u(F_2) > u(F_3)$ satisfy constraint (58) showing that F_1 is preferred to F_2 and F_2 is preferred to F_3. The belief decaying coefficient γ is adjusted to 0.2485 after training.

The test results are illustrated in Fig. 6, where the comparisons are shown between the observed output and the simulated output that is generated using the trained DER prediction model and the DER prediction model under the initial parameter vector P. It is evident that there is great difference between the initial system output and the actual observed output that is shown in Figs. 6 and 7. This difference is caused by the fact that the initial parameter vector P is subjectively generated. It is evident that the predicted outcomes generated by the trained DER prediction model match the actual observed ones closely. It demonstrates that the DER prediction model can simulate a real system with great accuracy and confidence.

6.1. Comparison with existing methods

It is of interest to compare the results generated by the DER prediction model with the previous methods. In this section, the least squares support vector machine (LSSVM) method (Suykens & Vandewalle, 1999; Wang & Guo, 2008), the RBFNN (Zhang, Huang, & Quan, 2005) method and ER-based (Hu et al., 2010) method are used for the comparative study.

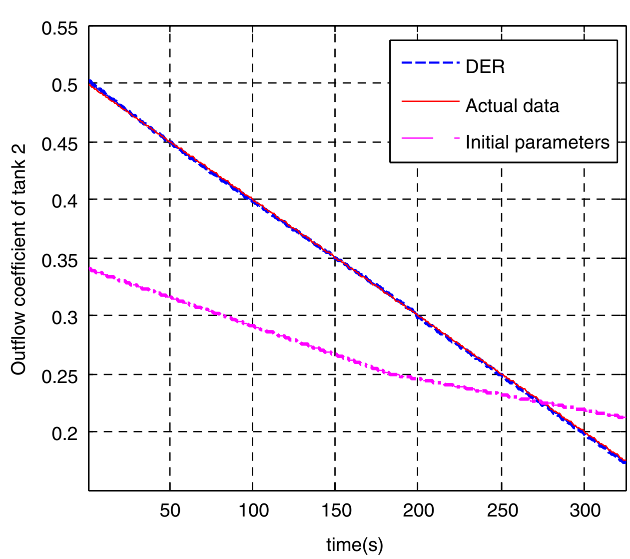

Fig. 6. Comparisons of test results.

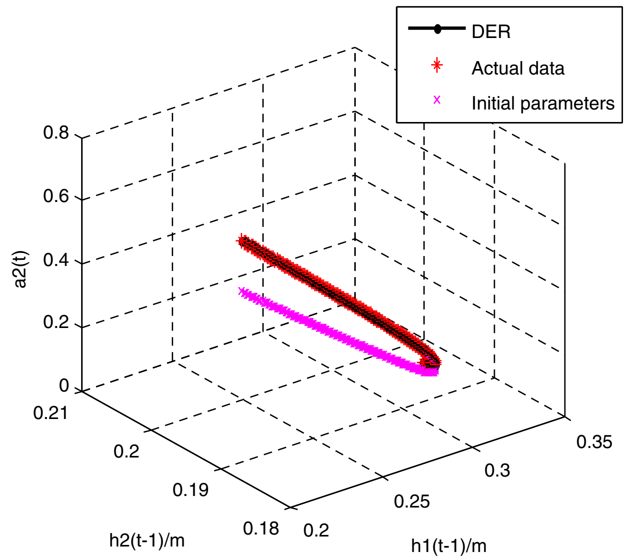

Fig. 7. Nonlinear relationship.

When the above two methods are applied to deal with the tank fault prediction problem, for LSSVM method, we select the radial basis function as the kernel function of LSSVM, in which the regularization parameter is set to 100, and the bandwidth of the RBF kernel function is set to 30. For RBFNN method, the RBF kernel center and width are set to 0.3 and 20, respectively.

The above three methods have been applied to deal with tank fault prediction problem. The training samples and the testing samples are the same as the DER prediction model in Section 5.3. The only difference is that there is no need to transform input vector $\mathbf{x}_{t-1}$ to a belief distribution structure while applying LSSVM and RBFNN. The predicted results on the testing samples are shown in Fig. 8.

In order to evaluate the performance of the above prediction methods, we use the root mean squared error (RMSE) and mean absolute percentage error (MAPE) as the evaluation criteria in this study. They are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2} \tag{73}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{74}$$

where y_t and $\hat{y}_t$ are actual and predicted values, respectively. The model with the smallest RMSE and MAPE is considered to be the best. The corresponding calculated results are listed in Table 5.

It is clear from Table 2 that the DER prediction model has the smallest RMSE and MAPE. The experimental results demonstrate that the DER prediction model is superior to LSSVM and RBFNN methods in term of prediction accuracy for this tank slow fault. On the other hand, one prominent advantage of the DER prediction method over LSSVM and RBENN is its interpretability of the trained parameters as mentioned in Section 5.3.

From the above numerical study, we can see that the initial DER prediction model is not accurate and the parameters optimal learning algorithm can be used to obtain the optimal parameters effectively if partial input and output data are available. However, the

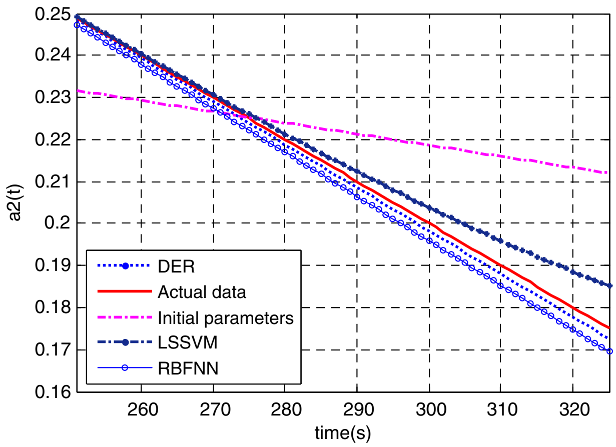

Fig. 8. Comparison of the two tanks fault prediction methods.

Table 5
Comparison of the prediction errors.

	Initial parameters	LSSVM	RBFNN	ER-based	DER
MAPE	0.2380	0.0259	0.0064	0.0044	0.0016
RMSE	0.0233	0.0052	0.0042	0.0032	0.0019

above comparative study is preliminary. It only shows that the DER prediction model is capable of generating credible results as other well-known common prediction methods in a case where all of them are applicable to numerical output. However, the DER prediction method has been developed not only to solve numerical output problems such as that represented above, but also to deal with prediction problems involving output values in the form of belief distribution structure as Eq. (21). A more complex problem is examined in the next section.

7. A practical case study—gyroscopic drift prediction

To demonstrate the detailed implementation process of the proposed approach under belief distribution outputs, we will examine a gyroscopic drift prediction problem in this section.

The dynamically tuned gyroscope (DTG) is used widely in the fields of missile, aviation, spaceflight, navigation, earth gauging, oil drilling, etc. As a key inertial device, gyroscope plays an important role in a inertial navigation system and its drift has a direct influence on navigation precision. Any fault in gyroscope is always the main source of faults in the control of a navigation system. So it is necessary and important to build a gyroscopic drift forecasting model, analyze the changing trend of any drift and predict gyroscopic faults.

7.1. Prediction model based on DER approach for DTG

In this study, the systems drift data are collected in a single axis dynamically tuned gyroscope performance reliability trial from leaving factory. The data sets include the time-to-drift data for 90 suits of gyroscope. Table B.1 in Appendix A.5 lists the original data. Table B.1 indicates that the gyroscopic drift is a time series. So it is reasonable to assume that the current gyroscopic drift value y_t is related to the most recent y_{t-1} or even extended into the past values $y_{t-2},\ldots,y_{t-p}$. This is because the next gyroscopic drift value is dependent on the current level of gyroscope state to a certain extent. Let $(y_{t-1}, y_{t-2},\ldots,y_{t-p})$ be the input vector of the DER prediction model, So Eq. (40) can be represented as follows:

$$\hat{y}(t) = \hat{y}_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \tag{75}$$

where $\hat{y}(t)$ is a scalar representing the predicted DTG drift at time t.

In this study, the drift data of gyroscope in Table B.1 were employed as the data set to train and test the DER model. This study is conducted with a relatively larger number 6 for the order of autoregressive terms. Therefore, we can transform 90 observation values to 84 input patterns. So Eq. (75) can be represented below:

$$\hat{y}(t) = \hat{y}_t = f(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}) \tag{76}$$

In Table B.1, the original input and output data were all provided as numerical numbers, as shown in columns 2, 4, 6. So there is a need to equivalently transform the numerical value into the belief structures.

7.2. Referential points of gyroscopic fault state

Note that the referential values of an attribute and the types of input information are problem specific, Thus, their definitions depend on the problems in hand.

For the DTG fault prediction problem, suppose gyroscopic fault state is classified into several categories like 'Small' (F_1), 'Medium' (F_2), and 'Large' (F_3). Define the following discernment framework of gyroscopic fault state:

$$F = \{F_j, j = 1, 2, 3\} = \{\text{Small}, \text{Medium}, \text{Large}\} \tag{77}$$

All related factors to gyroscopic fault may then be assessed with reference to this framework using the rule-based information transformation technique.

Similar to Section 5.3, it is assumed that equivalence rules can be acquired. Three equivalence rules could be acquired as follows:

1. If gyroscopic drift is 1°/h, then fault probability is small (or $F_1 = 1$).
2. If gyroscopic drift is 1.4°/h, then fault probability is medium (or $F_2 = 1.4$).
3. If gyroscopic drift is 2.4°/h, then fault probability is large (or $F_3 = 1$).

The quantified results above are listed in Table 6 as follows:

The given values for the input $(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6})$ can be transformed into the belief structures with respect to the defined three referential points using the rule based transformation technique. Take the twentieth input vector (1.36, 1.33, 1.4, 1.35, 1.26, 1.16) and the corresponding output value 1.26 for example. For 1.36, this can be equivalently transformed to belief distribution structure as follows:

$$S(1.36) = \{(F_1, 0.1), (F_2, 0.9), (F_3, 0)\}$$

Similarly, other values can be equivalently transformed to

$$\begin{aligned}
S(1.33) &= \{(F_1, 0.175), (F_2, 0.825), (F_3, 0)\} \\
S(1.4) &= \{(F_1, 0), (F_2, 1), (F_3, 0)\} \\
S(1.35) &= \{(F_1, 0.125), (F_2, 0.875), (F_3, 0)\} \\
S(1.26) &= \{(F_1, 0.35), (F_2, 0.65), (F_3, 0)\} \\
S(1.16) &= \{(F_1, 0.6), (F_2, 0.4), (F_3, 0)\} \\
S(1.26) &= \{(F_1, 0.35), (F_2, 0.65), (F_3, 0)\}
\end{aligned}$$

The other given input training data sets can be transformed in a similar way.

Table 6
The referential points of gyroscopic drift.

Linguistic terms	Small (F_1)	Medium (F_2)	Large (F_3)
Numerical values (°/h)	1.0	1.4	2.4

7.3. Illustration of the DER prediction approach under numerical outputs

For illustration, the first 75 sets of data are used as the training data for parameter estimation. In this case, the output valuey_t in Eq. (40) is a numerical value. After training, the remaining nine sets of data are used for testing the trained DER prediction model. The initial parameter vector P is given by experts. The initial attribute weights of $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}$ and y_{t-6} are all set to be equivalent. The initial parameter vector P is set to.

$P = [w_1, w_2, w_3, w_4, w_5, w_6, u(F_1), u(F_2), u(F_3), \gamma]$ = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6, 2.2, 2.4, 3, 0.2]. Similar to Section 5.3, FMINCON function in Matlab is used, the error tolerance is set to 0.000001, and the maximum iteration is set to 60. After training, the trained parameter vector P is obtained as follows:

$$P = [0.4835, 0.2332, 0.1235, 00.0863, 0.0496, 0.0239, 1.0756, 1.3660, 2.3425, 0.5751],$$

where $w_1 > w_2 > w_3 > w_4 > w_5 > w_6$ denotes that y_{t-1} has the most influence on the prediction results $\hat{y}_t$ and y_{t-6} has the least influence on $\hat{y}_t$, and the utilities $u(F_1) > u(F_2) > u(F_3)$ satisfy the constraint (58) denoting F_1 is preferred to F_2 and F_2 is preferred to F_3. The belief decaying coefficient γ is adjusted to 0.5751 after training.

The test results are illustrated in Fig. 9, where the comparisons are shown between the observed output and the simulated output that is generated using the trained DER prediction model. As shown in Fig. 9, it is obvious that the predicted outcomes generated by the trained DER prediction model match the actual observed ones closely. It further demonstrates that the DER prediction model can simulate a real system with a great accuracy.

The EKF prediction method is used for the comparative study. The training samples and the testing samples are the same as the DER prediction model above. The predicted results on the testing samples are shown in Fig. 10.

The MAPE and RMSE are used to assess the prediction accuracy. The MAPE and RMSE for the DER prediction model is 1.35% and 0.0212, respectively. On the other hand, The MAPE and RMSE for the EKF prediction model is 3.16% and 0.0543, respectively. The experimental results demonstrate that the prediction performance of DER prediction model outperforms EKF prediction method in term of prediction accuracy for the DGT drift prediction.

7.4. Simulation results based on belief distribution output

For illustration purpose, the first 75 sets of data are used as the training data for parameter estimation. In this study, the output value y_t in Eq. (40) is a belief distribution structure through the rule based transformation similar to Section 7.2. After training, all the 84 sample data are used for testing the trained DER prediction model. The initial parameter vector P is given by experts. Suppose that the initial attribute weights are all set to be equivalent. For the belief distribution output, it is unnecessary to initialize the utilities of the fault assessment grades $u(F_i)$, $i = 1, 2, 3$, so the initial parameter vector P is set to

$$P = [w_1, w_2, w_3, w_4, w_5, w_6, \gamma] = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6, 0.2]$$

For the belief distribution output, there is a need to solve the multiple objective nonlinear optimization problems that are defined by Eq. (61). To solve the multiple objective optimization models given by Eq. (63), FMINIMAX function in Matlab is used. In this example, the error tolerance is set to 0.000001, and the maximum iteration is set to 60 to avoid dead loop in the optimal learning process. After training, the trained parameter vector P can be obtained as follows:

$$P = [0.0012, 0.7844, 0.0036, 0.0099, 0.0077, 0.1931, 3.8695],$$

where we can see that y_{t-2} has the most influence on the prediction result $O(\hat{y}(t))$, and the larger the attribute weights, the greater influence they have on the prediction results. After training, the belief decaying coefficient γ is adjusted to 3.8695.

The test results are illustrated in Figs. 11–13, where the comparisons are shown between the observed output and the simulated output that is generated using the trained DER prediction model. As shown in Figs. 11–13, it is obvious that the predicted

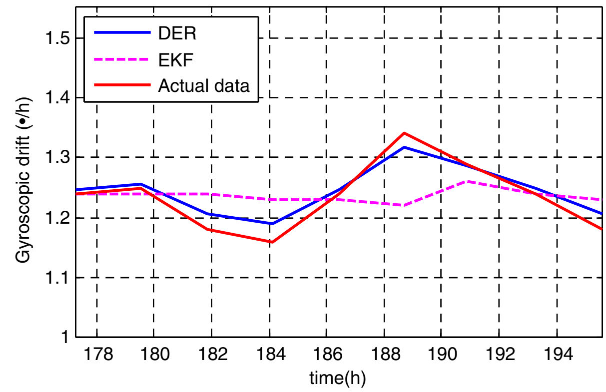

Fig. 10. Gyroscopic drift prediction results with three methods.

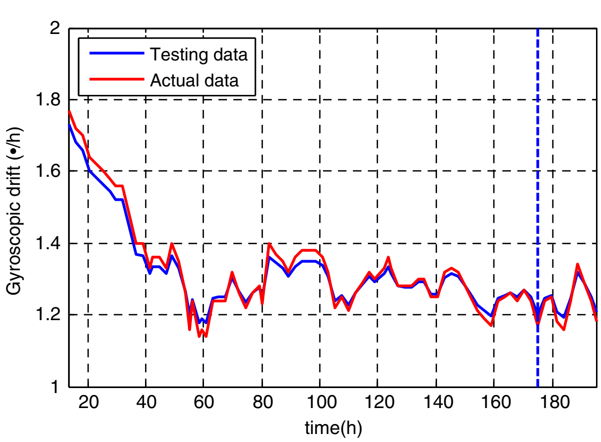

Fig. 9. Gyroscopic drift prediction results with DER.

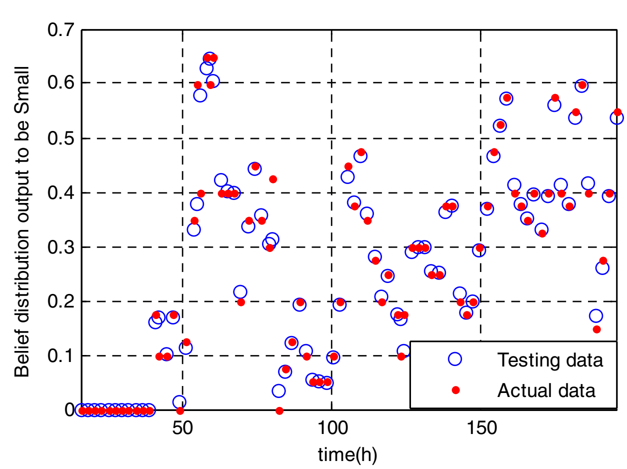

Fig. 11. Belief distribution output to be small.

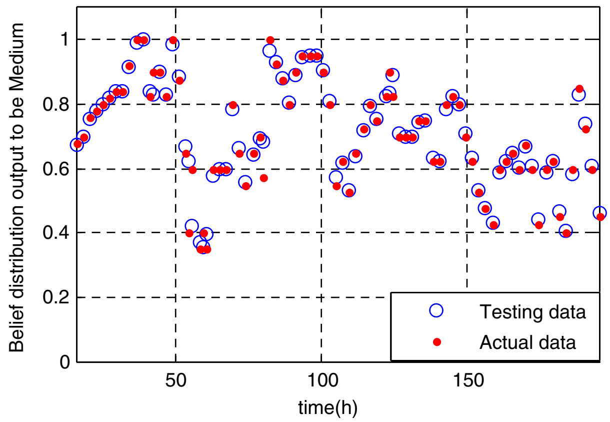

Fig. 12. Belief distribution output to be medium.

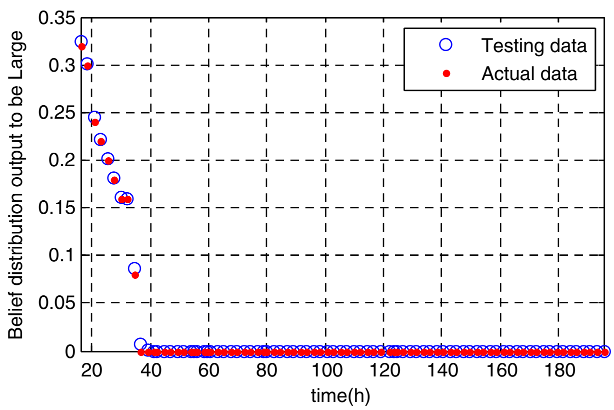

Fig. 13. Belief distribution output to be large.

Table 7
The prediction errors.

	Small	Medium	Large
MAPE	0.0387	0.0202	0.0019
RMSE	0.0132	0.0132	0.0011

outcomes generated by the trained DER prediction model match the actual observed ones closely.

In order to further demonstrate the proposed DER prediction method under belief distribution output, the MAPE and RMSE between the observed values and the predicted belief degrees to gyroscopic fault state "Small", "Medium" and "Large", respectively are used to evaluate the prediction accuracy. The calculated results are listed in Table 7 as follows.

Obviously, the predicted outcomes of DER prediction method match the observed ones closely.

The above results show that the DER prediction method can be used not only to solve numerical output problems such as that represented in Section 5, but also to effectively deal with prediction problems involving output values in the form of belief distribution structure through solving multiple objective optimization problems.

8. Concluding remarks

Due to inherent nonlinearity and uncertainty that widely exist in system models, most real world fault prediction problems involve both quantitative data and qualitative information as well as various types of uncertainties such as incompleteness and fuzziness, which significantly increase complexity and difficulty for applying the existing methods to fault prediction. Support to solution of such fault prediction problems requires powerful methodologies that are capable of dealing with uncertainty in a way that is rational, systematic, and flexible. The evidential reasoning approach provides a novel, flexible and systematic way to deal with such uncertainty and support the fault prediction.

In this paper, to conduct dynamic fusion, we extended the Dempster's combination rule to a dynamic combination style. Based on this, we developed the original evidential reasoning approach to the dynamic evidential reasoning approach, which took account of time effect by introducing the belief decaying factor. For fault prediction problems, we investigated a novel fault prediction method based on the dynamic evidential reasoning approach. The fault prediction model based on the DER approach was established and several optimization models for locally training a DER prediction model can be used to determine the related parameters. The optimization models were either single or multiple objective nonlinear optimization problems. The models can be used to fine tune the DER prediction model whose initial parameters are decided by expert's knowledge or common sense judgment.

The main difference between the DER prediction method and other current prediction approaches in handling uncertainties was that the new method was based on the dynamic evidential reasoning approach, which was developed on the basis of the D–S theory of evidence. On the other hand, the distinctive feature of the new model is that only partial input and output information is required, which can be either incomplete or vague, either numerical or judgmental, or mixed. Another advantage of the new method is its interpretability of the training parameters over other nonlinear models such as LSSVM and RBFNN.

The detailed implementation procedure of the new fault prediction techniques based on the dynamic evidential reasoning approach was demonstrated by two numerical examples. Compared with other prediction methods such as LSSVM, RBFNN, ER-based and EKF, the results show that the DER prediction method developed is not only capable of generating credible results as other well-known common prediction methods (LSSVM and RBFNN) in a case where all of them are applicable to numerical output, but also effective to deal with prediction problems involving output values in the form of belief distribution structure through solving multiple objective optimization problems. The latter feature is desirable in situations where expert judgments and qualitative information are available for supporting fault prediction.

Although the promising results obtained in this work revealed the potential of the proposed approach for predicting system fault, the validation of the proposed method needs to be more extensively investigated for more complex prediction problems. On the other hand, note that in this paper the parameters in the DER prediction models are determined by optimization models, which, nevertheless, are local optimization models in some sense. Therefore, different initial parameters might lead to different results, and the optimization models proposed in this paper might converge to local optimization solutions. So another direction for future research is to apply more efficient global optimization algorithms to train the parameters of the DER prediction model to improve prediction performance, such as genetic algorithm or simulated annealing.

Acknowledgements

The work was supported by the National Science Foundation of China (NSFC) under Grant No. 60736026 and the National Natural

Science Fund for Distinguished Young Scholars of China under Grant No. 61025014, National High Technology Research Development Program (863 program) of China under Grant No. 2008AAJ112.

Appendix A. The proof of the synthesis theorems

In this section, we prove the conclusion that we took for granted to develop the new DER approach in Section 2. These include the basic synthesis theorem, the consensus synthesis theorem, the complete synthesis theorem and the incomplete synthesis theorem.

A.1. Basic synthesis theorem

In the DER approach, the combined degrees of belief $\beta_n(t)$ ($n = 1,\ldots,L$) and $\beta_F(t)$ are generated using (27)–(33). These belief degrees are between zero and one and are summed to one as proved in the following Theorem 1.

Proof of Theorem 1. Similar to Yang's proof of the ER algorithm, we can directly conclude that (28) is held for any $i \in \{1,\ldots,L\}$. As such, (27)–(31) is then the straightforward implementation of the evidence combination of the D–S theory and thus ensure that

$$\sum_{n=1}^{N} m_{n,I(L)}(t) + m_{F,I(L)}(t) = \sum_{n=1}^{N} m_{n,I(L)}(t) + \bar{m}_{F,I(L)}(t) + \tilde{m}_{F,I(L)}(t) = 1 \tag{A.1}$$

Therefore, from (32) and (33), we have

$$\sum_{n=1}^{N} \beta_n(t) + \beta_F(t) = \sum_{n=1}^{N} \frac{m_{n,I(L)}(t)}{1-\bar{m}_{F,I(L)}(t)} + \frac{\tilde{m}_{F,I(L)}(t)}{1-\bar{m}_{F,I(L)}(t)} = \frac{1}{1-\bar{m}_{F,I(L)}(t)}\left(\sum_{n=1}^{N} m_{n,I(L)}(t) + \tilde{m}_{F,I(L)}(t)\right) = 1 \tag{A.2}$$

As $K_{I(i+1)}(t)$ calculated by (31) is always positive, we must have

$$m_{n,I(L)}(t) \geqslant 0 \text{ for all } n = 1,\ldots,N, \quad \bar{m}_{F,I(L)}(t) \geqslant 0 \text{ and } \tilde{m}_{F,I(L)}(t) \geqslant 0 \tag{A.3}$$

With attributes weights normalized using (19) and the reliability factor $\alpha(t) = e^{-\gamma t}$, $\gamma \geqslant 0$, we can obtain $0 < \alpha(t) \leqslant 1$, and from (A.1) and (A.3), there must be $\bar{m}_{F,I(L)}(t) < 1$. From (32), (33) and (A.3), we then have

$$\beta_n(t) \geqslant 0, \quad n = 1,\ldots,N \text{ and } \beta_F(t) \geqslant 0 \tag{A.4}$$

From (A.3) and (A.4), we therefore conclude

$$\beta_n(t) \leqslant 1, \quad n = 1,\ldots,N \text{ and } \beta_F(t) \leqslant 1 \qquad \square \tag{A.5}$$

A.2. Complete synthesis theorem

First of all, by a simple induction, we easily see that the followings hold.

Lemma 2. *With the quantity $\bar{m}_{F,I(L)}(t)$ inductively defined by (30), we have*

$$\bar{m}_{F,I(i)}(t) = K_{I(i)}(t)\prod_{i=1}^{i}(1 - w_i\alpha(dt_i)) \tag{A.6}$$

where $K_{I(i)}(t)$ is inductively defined by (31).

Lemma 3. *If all assessments $S(e_i(t_i))$, $i = 1,\ldots,L$ are complete, we have*

$$m_F(t) = \bar{m}_{F,I(L)}(t) = K_{I(L)}(t)\prod_{i=1}^{L}(1 - w_i\alpha(dt_i)) \tag{A.7}$$

i.e., $\tilde{m}_{F,I(L)}(t) = 0$, and consequently, the final assessment by (32) and (33) is also complete.

As such, we give the Proof of Theorem 4 in the following.

Proof of Theorem 4. From (22), (24) and (25), the basic probability assignments are given by

$$\begin{aligned}
m_{n,i}(t) &= w_i\alpha(dt_i)\beta_{n,i}(t_i) \quad \text{for } n \in I^+,\ i = 1,\ldots,L\\
m_{k,i}(t) &= w_i\alpha(dt_i)\beta_{k,i}(t_i) \quad \text{for } k \in I^-,\ i = 1,\ldots,L\\
\bar{m}_{F,i}(t) &= 1 - w_i\alpha(dt_i)\\
m_{F,i}(t) &= w_i\alpha(dt_i)\left(1 - \sum_{n=1}^{N}\beta_{n,i}(t_i)\right) \quad \text{for } i = 1,\ldots,L
\end{aligned}$$

According to Lemma 3, we can obviously obtain

$$\tilde{m}_{F,I(i)}(t) = 0 \text{ for all } i = 1,\ldots,L$$

Therefore,

$$\beta_F(t) = \frac{\tilde{m}_{F,I(L)}(t)}{1-\bar{m}_{F,I(L)}(t)} = 0$$

Since $m_{k,I(1)}(t) = m_{k,1}(t) = 0$ for $k \in I^-$, using (22)–(26), we then have $m_{k,I(i)}(t) = 0$ for $k \in I^-$ at $i = 2$. Suppose, $m_{k,I(r)}(t) = 0$ for $k \in I^-$ at $r = 2$, using (22)–(26), we obtain

$$\begin{aligned}
m_{k,I(r+1)}(t) &= K_{I(r+1)}(t)\big(m_{k,I(r)}(t)m_{k,r+1}(t) + m_{k,I(r)}(t)m_{F,r+1}(t) + m_{F,I(r)}(t)m_{k,r+1}(t)\big)\\
&= K_{I(r+1)}(t)\big(0\times 0 + 0\times m_{F,r+1}(t) + m_{F,I(r)}(t)\times 0\big)\\
&= 0 \text{ for } k \in I^-.
\end{aligned}$$

Thus, we deduce $m_{k,I(i)}(t) = 0$ for $k \in I^-$ at any $i = 1,\ldots,L$. We therefore conclude that

$$\beta_k(t) = \frac{m_{k,I(L)}(t)}{1-\bar{m}_{F,I(L)}(t)} = 0 \text{ for } k \in I^-.$$

That is $\sum_{k\in I^-}\beta_k(t) = 0$, thus from (A.2), we finally conclude

$$\begin{aligned}
1 &= \sum_{n=1}^{N}\beta_n(t) + \beta_F(t) = \sum_{n=1}^{N}\beta_n(t) + \beta_F(t)\\
&= \sum_{n\in I^+}\beta_n(t) + \sum_{k\in I^-}\beta_k(t) + \beta_F(t) = \sum_{n\in I^+}\beta_n(t) + 0 + 0\\
&= \sum_{n\in I^+}\beta_n(t) \qquad \square
\end{aligned} \tag{A.8}$$

A.3. Consensus synthesis theorem

In this section, we prove that the combined degrees of belief generated using (22)–(33) satisfy the synthesis theorem.

Proof of Theorem 3. It can clearly see that the assumption of Theorem 3 is the same as that given in Theorem 4 with $|I^+| = 1$.

Therefore, Theorem 3 immediately follows from Lemma 3 and Theorem 4.

A.4. Incomplete synthesis theorem

The incomplete synthesis theorem is to say that if any basic assessment is not complete, then the assessment of the associated general attribute will not be complete either. We will give the proof of the incomplete synthesis theorem as follows.

Proof of Theorem 5. As if $w_i = 0$, then the attribute e_i will play no role in the aggregation. Thus, without loss of generality, we assume that $0 < w_i < 1$ for all $i = 1,\ldots,L$ in the following.

From (17), (28)–(30), it is easy seen that

$$m_{F,I(L)}(t) = K_{I(L)}(t)\prod_{i=1}^{L}\left[w_i\alpha(dt_i)m_{F,i}(t) + 1 - w_i\alpha(dt_i)\right] \tag{A.9}$$

and in addition, if there is an incomplete assessment, say $S(e_j(t_j))$, then $w_j\alpha(dt_j)m_{F,j}(t) > 0$, resulting in

$$w_j\alpha(dt_j)m_{F,j}(t)\prod_{i=1,i\neq j}^{L}(1 - w_i\alpha(dt_i)) > 0 \tag{A.10}$$

Then from Lemma 2, (28) and (A.3), we have

$$\begin{aligned}\tilde{m}_{F,I(L)}(t) &= m_{F,I(L)}(t) - \bar{m}_{F,I(L)}(t)\\ &= K_{I(L)}(t)\prod_{i=1}^{L}\left[w_i\alpha(dt_i)m_{F,i}(t) + 1 - w_i\alpha(dt_i)\right]\\ &\quad - K_{I(L)}(t)\prod_{i=1}^{L}(1 - w_i\alpha(dt_i))\\ &= w_j\alpha(dt_j)m_{F,j}(t)K_{I(L)}(t)\prod_{i=1,i\neq j}^{L}(1 - w_i\alpha(dt_i)) > 0 \end{aligned} \tag{A.11}$$

From (33), we have

$$\beta_F(t) = \frac{\tilde{m}_{F,I(L)}(t)}{1 - \bar{m}_{F,I(L)}(t)} > 0$$

which we desired. This completely concludes the Proof of Theorem 5. □

A.5. The proof of the equivalence between the recursive and analytical DER approach

In the following, we apply the mathematic induction principle to prove the equivalence between the recursive DER approach and the analytical DER approach, in which the combination and the normalization of evidence are integrated. Then, we derive Eqs. (24) and (25),

In the following, for simplicity, we use $K_l(t)$, $l = 1,\ldots,L$ to denote the normalization factor for combining l pieces of evidence in the analytical ER algorithm. Therefore, $K_L(t)$ is equal to $k(t)$ in Eq. (11).

First of all, let us combine two factors with normalization. The combined probability masses generated by aggregating the two factors are given as follows. (see Table B.1)

Table B.1
Gyroscopic drift data (°/h).

2.12	2.04	1.92	1.88	1.83	1.77	1.72	1.7	1.64
1.62	1.6	1.58	1.56	1.56	1.48	1.4	1.4	1.33
1.36	1.36	1.33	1.4	1.35	1.26	1.16	1.24	1.14
1.16	1.14	1.24	1.24	1.24	1.32	1.26	1.22	1.26
1.28	1.23	1.4	1.37	1.35	1.32	1.36	1.38	1.38
1.38	1.36	1.32	1.22	1.25	1.21	1.26	1.29	1.32
1.30	1.33	1.36	1.33	1.28	1.28	1.28	1.30	1.30
1.25	1.25	1.32	1.33	1.32	1.28	1.25	1.21	1.19
1.17	1.24	1.25	1.26	1.24	1.27	1.24	1.17	1.24
1.25	1.18	1.16	1.24	1.34	1.29	1.24	1.18	1.17

$$\begin{aligned}K_{I(2)}(t) &= \left[1 - \sum_{n=1}^{N}\sum_{q=1,t\neq n}^{N} m_{n,I(1)}(t)m_{q,2}(t)\right]^{-1} = \left[1 - \sum_{n=1}^{N}\sum_{q=1,q\neq n}^{N} m_{n,1}(t)m_{q,2}(t)\right]^{-1}\\ &= \left[1 - \sum_{n=1}^{N}\left(\sum_{q=1}^{N} m_{n,1}(t)m_{q,2}(t) - m_{n,1}(t)m_{n,2}(t)\right)\right]^{-1}\\ &= \left\{1 - \sum_{n=1}^{N}\left[m_{n,1}(t)\left(\sum_{q=1}^{N} m_{q,2}(t) - m_{n,2}(t)\right)\right]\right\}^{-1}\\ &= \left\{1 - \sum_{n=1}^{N}[m_{n,1}(t)(1 - m_{F,2}(t) - m_{n,2}(t))]\right\}^{-1}\\ &= \left(1 - \sum_{n=1}^{N} m_{n,1}(t) + m_{F,2}(t)\sum_{n=1}^{N} m_{n,1}(t) + \sum_{n=1}^{N} m_{n,1}(t)m_{n,2}(t)\right)^{-1}\\ &= \left[m_{F,1}(t) + \left(m_{F,1}(t)\sum_{n=1}^{N} m_{n,2}(t) + m_{F,2}(t)\sum_{n=1}^{N} m_{n,1}(t) + \sum_{n=1}^{N} m_{n,1}(t)m_{n,2}(t)\right.\right.\\ &\quad \left.\left. + \sum_{n=1}^{N} m_{F,1}(t)m_{F,2}(t)\right) - m_{F,1}(t)\sum_{n=1}^{N} m_{n,2}(t) - \sum_{n=1}^{N} m_{F,1}(t)m_{F,2}(t)\right]^{-1}\\ &= \left[m_{F,1}(t)\left(1 - \sum_{n=1}^{N} m_{n,2}(t)\right) + \sum_{n=1}^{N}\prod_{i=1}^{2}(m_{n,i}(t) + m_{F,i}(t)) - Nm_{F,1}(t)m_{F,2}(t)\right]^{-1}\\ &= \left[m_{F,1}(t)m_{F,2}(t) + \sum_{n=1}^{N}\prod_{i=1}^{2}(m_{n,i}(t) + m_{F,i}(t)) - Nm_{F,1}(t)m_{F,2}(t)\right]^{-1}\\ &= \left[\sum_{n=1}^{N}\prod_{i=1}^{2}(m_{n,i}(t) + m_{F,i(t)}) - (N-1)\prod_{i=1}^{2} m_{F,i}(t)\right]^{-1}\\ &= \left[\sum_{n=1}^{N}\prod_{i=1}^{2}(m_{n,i}(t) + \tilde{m}_{F,i}(t) + \bar{m}_{F,i}(t)) - (N-1)\prod_{i=1}^{2}(\tilde{m}_{F,i}(t) + \bar{m}_{F,i}(t))\right]^{-1}\\ &= K_2(t).\end{aligned}$$

Then we have

$$\begin{aligned}m_{n,I(2)}(t) &= K_{I(2)}(t)(m_{n,1}(t)m_{n,2}(t) + m_{n,1}(t)m_{F,2}(t) + m_{F,1}(t)m_{n,2}(t))\\ &= K_{I(2)}(t)[m_{n,1}(t)(m_{n,2}(t) + m_{F,2}(t)) + m_{F,1}(t)m_{n,2}(t)]\\ &= K_{I(2)}(t)[m_{n,1}(t)(m_{n,2}(t) + m_{F,2}(t)) + m_{F,1}(t)(m_{n,2}(t) + m_{F,2}(t)) - m_{F,1}(t)m_{F,2}(t)]\\ &= K_{I(2)}(t)[(m_{n,1}(t) + m_{F,1}(t))(m_{n,2}(t) + m_{F,2}(t)) - m_{F,1}(t)m_{F,2}(t)]\\ &= K_{I(2)}(t)\left[\prod_{i=1}^{2}(m_{n,i}(t) + m_{F,i}(t)) - \prod_{i=1}^{2} m_{F,i}(t)\right]\\ &= K_{I(2)}(t)\left[\prod_{i=1}^{2}(m_{n,i}(t) + \bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)) - \prod_{i=1}^{2}(\bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t))\right]\\ &= K_2(t)\left[\prod_{i=1}^{2}(m_{n,i}(t) + \bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)) - \prod_{i=1}^{2}(\bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t))\right], \quad n = 1,\ldots,N\end{aligned}$$

$$\begin{aligned}\tilde{m}_{F,I(2)}(t) &= K_{I(2)}(t)(\tilde{m}_{F,1}(t)\tilde{m}_{F,2}(t) + \tilde{m}_{F,1}(t)\bar{m}_{F,2}(t) + \bar{m}_{F,1}(t)\tilde{m}_{F,2}(t))\\ &= K_2(t)[\tilde{m}_{F,1}(t)(\bar{m}_{F,2}(t) + \tilde{m}_{F,2}(t)) + \bar{m}_{F,1}(t)\tilde{m}_{F,2}(t)]\\ &= K_2(t)[\tilde{m}_{F,1}(t)(\bar{m}_{F,2}(t) + \tilde{m}_{F,2}(t)) + \bar{m}_{F,1}(t)(\tilde{m}_{F,2}(t) + \bar{m}_{F,2}(t)) - \bar{m}_{F,1}(t)\bar{m}_{F,2}(t)]\\ &= K_2(t)[(\bar{m}_{F,1}(t) + \tilde{m}_{F,1}(t))(\bar{m}_{F,2}(t) + \tilde{m}_{F,2}(t)) - \bar{m}_{F,1}(t)\bar{m}_{F,2}(t)]\\ &= K_2(t)\left[\prod_{i=1}^{2}(\tilde{m}_{F,i}(t) + \bar{m}_{F,i}(t)) - \prod_{i=1}^{2}\bar{m}_{F,i}(t)\right]\end{aligned}$$

$$\bar{m}_{F,I(2)}(t) = K_{I(2)}(t)\bar{m}_{F,1}(t)\bar{m}_{F,2}(t) = K_2(t)\prod_{i=1}^{2}\bar{m}_{F,i}(t)$$

Suppose the following equations are true while combining the first $(l-1)$ evidence, $3 \leqslant l \leqslant L$,

$$m_{n,I(l-1)}(t) = K_{l-1}\left[\prod_{i=1}^{l-1}(m_{n,i}(t) + \bar{m}_{F,i}(t) + \tilde{m}_{F,i}(t)) - \prod_{i=1}^{l-1}(\bar{m}_{F,i} + \tilde{m}_{F,i})\right], \quad n = 1,\ldots,N$$

$$\tilde{m}_{F,I(l-1)}(t) = K_{l-1}(t)\left[\prod_{i=1}^{l-1}(\tilde{m}_{F,i}(t) + \bar{m}_{F,i}(t)) - \prod_{i=1}^{l-1}\bar{m}_{F,i}(t)\right]$$

$$\bar{m}_{F,I(l-1)}(t) = K_{l-1}(t)\prod_{i=1}^{l-1}\bar{m}_{F,i}(t)$$

where $K_{l-1}(t)$ denotes the normalization factor for combining $l-1$ pieces of evidence at time t in the analytical ER algorithm.

Then the results combining $m_{I(l-1)}(t)$ with the lth evidence $m_l(t)$ can be formulated as follows:

$$
\begin{aligned}
K_{I(l)}(t) &= \left[1-\sum_{n=1}^{N}\sum_{q=1,q\neq n}^{N} m_{n,I(l-1)}(t)m_{q,l}(t)\right]^{-1} = \left[1-\sum_{n=1}^{N}\left(m_{n,I(l-1)}(t)\sum_{q=1,q\neq n}^{N} m_{q,l}(t)\right)\right]^{-1} \\
&= \left\{1-\sum_{n=1}^{N}\left[m_{n,I(l-1)}(t)\left(\sum_{q=1}^{N} m_{q,l}(t)-m_{n,l}(t)\right)\right]\right\}^{-1} \\
&= \left\{1-\sum_{n=1}^{N}\left[m_{n,I(l-1)}(t)\left(1-m_{F,l}(t)-m_{n,l}(t)\right)\right]\right\}^{-1} \\
&= \left[1-\sum_{n=1}^{N} m_{n,I(l-1)}(t)+m_{F,l}(t)\sum_{n=1}^{N} m_{n,I(l-1)}(t)+\sum_{n=1}^{N} m_{n,I(l-1)}(t)m_{n,l}(t)\right]^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\left(\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l-1} m_{F,i}(t)\right)\right.\right. \\
&\quad -m_{F,l}(t)\sum_{n=1}^{N}\left(\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l-1} m_{F,i}(t)\right) \\
&\quad \left.\left.-\sum_{n=1}^{N} m_{n,l}\left(\prod_{i=1}^{l-1}(m_{n,i}+m_{F,i})-\prod_{i=1}^{l-1} m_{F,i}\right)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\left(\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l-1} m_{F,i}(t)\right)\right.\right. \\
&\quad -m_{F,l}(t)\sum_{n=1}^{N}\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))+m_{F,l}(t)\sum_{n=1}^{N}\prod_{i=1}^{l-1} m_{F,i}(t) \\
&\quad \left.\left.-\sum_{n=1}^{N} m_{n,l}(t)\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))+\sum_{n=1}^{N} m_{n,l}(t)\prod_{i=1}^{l-1} m_{F,i}(t))\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\left(\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l-1} m_{F,i}(t)\right)\right.\right. \\
&\quad -\sum_{n=1}^{N}\left((m_{F,l}(t)+m_{n,l}(t))\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))\right)+m_{F,l}(t)\sum_{n=1}^{N}\prod_{i=1}^{l-1} m_{F,i}(t) \\
&\quad \left.\left.+\sum_{n=1}^{N} m_{n,l}(t)\prod_{i=1}^{l-1} m_{F,i}(t)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\left(\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l-1} m_{F,i}(t)\right)-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))\right.\right. \\
&\quad \left.\left.+\sum_{n=1}^{N}\prod_{i=1}^{l} m_{F,i}(t)+\prod_{i=1}^{l-1} m_{F,i}(t)\sum_{n=1}^{N} m_{n,l}(t)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\left(\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l-1} m_{F,i}(t)\right)-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))\right.\right. \\
&\quad \left.\left.+N\prod_{i=1}^{l} m_{F,i}(t)+\prod_{i=1}^{l-1} m_{F,i}(t)\sum_{n=1}^{N} m_{n,l}(t)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-\sum_{n=1}^{N}\prod_{i=1}^{l-1} m_{F,i}(t)-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))\right.\right. \\
&\quad \left.\left.+N\prod_{i=1}^{l} m_{F,i}(t)+\prod_{i=1}^{l-1} m_{F,i}(t)\sum_{n=1}^{N} m_{n,l}(t)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\sum_{n=1}^{N}\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-N\prod_{i=1}^{l-1} m_{F,i}(t)-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))\right.\right. \\
&\quad \left.\left.+N\prod_{i=1}^{l} m_{F,i}(t)+\prod_{i=1}^{l-1} m_{F,i}(t)\sum_{n=1}^{N} m_{n,l}(t)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\frac{1}{K_{l-1}(t)}-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))+N\prod_{i=1}^{l} m_{F,i}(t)\right.\right. \\
&\quad \left.\left.+od_{i=1}^{l-1} m_{F,i}(t)\sum_{n=1}^{N} m_{n,l}(t)-\prod_{i=1}^{l-1} m_{F,i}(t)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\frac{1}{K_{l-1}(t)}-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))+N\prod_{i=1}^{l} m_{F,i}(t)-\prod_{i=1}^{l-1} m_{F,i}(t)\left(1-\sum_{n=1}^{N} m_{n,l}(t)\right)\right]\right\}^{-1} \\
&= \left\{1-K_{l-1}(t)\left[\frac{1}{K_{l-1}(t)}-\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))+N\prod_{i=1}^{l} m_{F,i}(t)-\prod_{i=1}^{l} m_{F,i}(t)\right]\right\}^{-1} \\
&= \left\{K_{l-1}(t)\left[\sum_{n=1}^{N}\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))-(N-1)\prod_{i=1}^{l} m_{F,i}(t)\right]\right\}^{-1} \\
&= \left(\frac{K_{l-1}(t)}{K_l(t)}\right)^{-1} = \frac{K_l(t)}{K_{l-1}(t)}
\end{aligned}
$$

Note that $K_{I(l)}(t)$ is not equal to $K_l(t)$. We then have

$$
\begin{aligned}
m_{n,I(l)}(t) &= K_{I(l)}(t)\left(m_{n,I(l-1)}(t)m_{n,l}(t)+m_{F,I(l-1)}(t)m_{n,l}(t)+m_{F,l}(t)m_{n,I(l-1)}(t)\right) \\
&= K_{I(l)}(t)\left[\left(m_{n,I(l-1)}(t)+m_{F,I(l-1)}(t)\right)m_{n,l}(t)+m_{F,l}(t)m_{n,I(l-1)}(t)\right] \\
&= K_{I(l)}(t)\left[\left(m_{n,I(l-1)}(t)+m_{F,I(l-1)}(t)\right)m_{n,l}(t)+m_{F,l}(t)\left(m_{n,I(l-1)}(t)+m_{F,I(l-1)}(t)\right)\right. \\
&\quad \left.-m_{F,I(l-1)}(t)m_{F,l}(t)\right] \\
&= K_{I(l)}(t)\left[\left(m_{n,I(l-1)}(t)+m_{F,I(l-1)}(t)\right)\left(m_{n,l}(t)+m_{F,l}(t)\right)-m_{F,I(l-1)}(t)m_{F,l}(t)\right] \\
&= K_{I(l)}(t)\left[(m_{n,l}(t)+m_{F,l}(t))\left(K_{l-1}(t)\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-K_{l-1}(t)\prod_{i=1}^{l-1} m_{F,i}\right.\right. \\
&\quad \left.\left.+K_{l-1}(t)\prod_{i=1}^{l-1} m_{F,i}\right)-K_{l-1}(t)m_{F,l}(t)\prod_{i=1}^{l-1} m_{F,i}(t)\right] \\
&= K_{I(l)}(t)K_{l-1}(t)\left[(m_{n,l}(t)+m_{F,l}(t))\prod_{i=1}^{l-1}(m_{n,i}(t)+m_{F,i}(t))-m_{F,l}(t)\prod_{i=1}^{l-1} m_{F,i}(t)\right] \\
&= K_{I(l)}(t)K_{l-1}(t)\left[\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l} m_{F,i}(t)\right] \\
&= \frac{K_l(t)}{K_{l-1}(t)}K_{l-1}(t)\left[\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l} m_{F,i}(t)\right] \\
&= K_l(t)\left[\prod_{i=1}^{l}(m_{n,i}(t)+m_{F,i}(t))-\prod_{i=1}^{l} m_{F,i}(t)\right] \\
&= K_l(t)\left[\prod_{i=1}^{l}\left(m_{n,i}(t)+\bar{m}_{F,i}(t)+\tilde{m}_{F,i}(t)\right)-\prod_{i=1}^{l}\left(\bar{m}_{F,i}(t)+\tilde{m}_{F,i}(t)\right)\right], \quad n=1,\ldots,N.
\end{aligned}
$$

$$
\begin{aligned}
\tilde{m}_{F,I(l)}(t) &= K_{I(l)}(t)\left(\tilde{m}_{F,l}(t)\tilde{m}_{F,I(l-1)}(t)+\tilde{m}_{F,l}(t)\bar{m}_{F,I(l-1)}(t)+\bar{m}_{F,l}(t)\tilde{m}_{F,I(l-1)}(t)\right) \\
&= K_{I(l)}(t)\left[\tilde{m}_{F,l}(t)\left(\bar{m}_{F,I(l-1)}(t)+\tilde{m}_{F,I(l-1)}(t)\right)+\bar{m}_{F,l}(t)\tilde{m}_{F,I(l-1)}(t)\right] \\
&= K_{I(l)}(t)\left[\tilde{m}_{F,l}(t)\left(\bar{m}_{F,I(l-1)}(t)+\tilde{m}_{F,I(l-1)}(t)\right)+\bar{m}_{F,l}(t)\left(\tilde{m}_{F,I(l-1)}(t)+\bar{m}_{F,I(l-1)}(t)\right)\right. \\
&\quad \left.-\bar{m}_{F,l}(t)\bar{m}_{F,I(l-1)}(t)\right] \\
&= K_{I(l)}(t)\left[\left(\tilde{m}_{F,I(l-1)}(t)+\bar{m}_{F,I(l-1)}(t)\right)\left(\tilde{m}_{F,l}(t)+\bar{m}_{F,l}(t)\right)-\bar{m}_{F,l}(t)\bar{m}_{F,I(l-1)}(t)\right] \\
&= K_{I(l)}(t)\left\{\left(\tilde{m}_{F,l}(t)+\bar{m}_{F,l}(t)\right)\left[K_{l-1}(t)\left(\prod_{i=1}^{l-1}(\tilde{m}_{F,i}(t)+\bar{m}_{F,i}(t))-\prod_{i=1}^{l-1}\bar{m}_{F,i}(t)\right)\right.\right. \\
&\quad \left.\left.+K_{l-1}(t)\prod_{i=1}^{l-1}\bar{m}_{F,i}(t)\right]-K_{l-1}(t)\prod_{i=1}^{l}\bar{m}_{F,i}(t)\right\} \\
&= K_{I(l)}(t)K_{l-1}(t)\left[\prod_{i=1}^{l}(\tilde{m}_{F,i}(t)+\bar{m}_{F,i}(t))-\prod_{i=1}^{l}\bar{m}_{F,i}(t)\right] \\
&= \frac{K_l(t)}{K_{l-1}(t)}K_{l-1}(t)\left[\prod_{i=1}^{l}(\tilde{m}_{F,i}(t)+\bar{m}_{F,i}(t))-\prod_{i=1}^{l}\bar{m}_{F,i}(t)\right] \\
&= K_l(t)\left[\prod_{i=1}^{l}(\tilde{m}_{F,i}(t)+\bar{m}_{F,i}(t))-\prod_{i=1}^{l}\bar{m}_{F,i}(t)\right]
\end{aligned}
$$

$$
\begin{aligned}
\bar{m}_{F,I(l)}(t) &= K_{I(l)}(t)\bar{m}_{F,l}(t)\bar{m}_{F,I(l-1)}(t) = K_{I(l)}(t)\bar{m}_{F,l}(t)K_{l-1}(t)\prod_{i=1}^{l-1}\bar{m}_{F,i}(t) \\
&= K_{I(l)}(t)K_{l-1}(t)\prod_{i=1}^{l}\bar{m}_{F,i}(t) \\
&= \frac{K_l(t)}{K_{l-1}(t)}K_{l-1}(t)\prod_{i=1}^{l}\bar{m}_{F,i}(t) = K_l(t)\prod_{i=1}^{l}\bar{m}_{F,i}(t)
\end{aligned}
$$

Therefore, according to mathematical induction principle, the above equations are true for any $l \in \{1,\ldots,L\}$. For $l = L$, we obtain the following normalized combined probability assignments generated by aggregating the L attributes:

$$
\begin{aligned}
m_{n,I(L)}(t) &= K_L(t)\left[\prod_{i=1}^{L}\left(m_{n,i}(t)+\bar{m}_{F,i}(t)+\tilde{m}_{F,i}(t)\right)-\prod_{i=1}^{L}\left(\bar{m}_{F,i}(t)+\tilde{m}_{F,i}(t)\right)\right], \quad n=1,\ldots,N, \\
\tilde{m}_{F,I(L)}(t) &= K_L(t)\left[\prod_{i=1}^{L}\left(\tilde{m}_{F,i}(t)+\bar{m}_{F,i}(t)\right)-\prod_{i=1}^{L}\bar{m}_{F,i}(t)\right], \\
\bar{m}_{F,I(L)}(t) &= K_L(t)\prod_{i=1}^{L}\bar{m}_{F,i}(t)
\end{aligned}
$$

According to the definition of the basic probability assignment function, we have

$$
\sum_{n=1}^{N} m_{n,I(L)}(t)+m_{F,I(L)}(t) = \sum_{n=1}^{N} m_{n,I(L)}(t)+\bar{m}_{F,I(L)}(t)+\tilde{m}_{F,I(L)}(t) = 1.
$$

from which we obtain

$$K_L = \left[\sum_{n=1}^{N}\prod_{i=1}^{L}(m_{n,i}(t)+\bar{m}_{F,i}(t)+\tilde{m}_{F,i}(t))-(N-1)\prod_{i=1}^{L}(\bar{m}_{F,i}(t)+\tilde{m}_{F,i}(t))\right]^{-1}.$$

Therefore, from (34)–(39), we obtain Eqs. (42)–(44) as follows:

$$\beta_n(t)=\frac{\prod_{i=1}^{L}(w_i\alpha(dt_i)\beta_{n,i}(t_i)+1-w_i\alpha(dt_i)+w_i\alpha(dt_i)\beta_{F,i}(t_i))-\prod_{i=1}^{L}(1-w_i\alpha_i(dt_i)+w_i\alpha(dt_i)\beta_{F,i}(t_i))}{D(t)}$$

$$\beta_F(t)=\frac{\prod_{i=1}^{L}(1-\alpha_i(dt_i)w_i+w_i\alpha(dt_i)\beta_{F,i}(t_i))-\prod_{i=1}^{L}(1-w_i\alpha_i(dt_i))}{D(t)}$$

$$D(t)=\sum_{q=1}^{N}\prod_{i=1}^{L}(w_i\alpha(dt_i)\beta_{q,i}(t_i)+1-\alpha_i(dt_i)w_i+w_i\alpha(dt_i)\beta_{F,i}(t_i))-(N-1)(1-w_i\alpha_i(dt_i)+w_i\alpha(dt_i)\beta_{F,i}(t_i))-\prod_{i=1}^{L}(1-w_i\alpha_i(dt_i))$$

References

Angeli, C., & Chatzinikolaou, A. (1999). Fault prediction and compensation functions in a diagnostic knowledge-based system for hydraulic systems. *Journal of Intelligent and Robotic Systems: Theory and Application, 26*(2), 153–165.

Box, G. E., & Jenkins, G. M. (1970). *Time series analysis forecasting and control*. San Francisco: Holden-Day.

Biswas, G., Oliff, M., & Sen, A. (1988). An expert decision support system for production control. *Decision Support Systems, 4*, 235–248.

Bloch, I. (1996). Some aspects of Dempster–Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern Recognition Letters, 17*(8), 905–919.

Bauer, M. (1997). Approximation algorithms and decision making in the Dempster–Shafer theory of evidence – An empirical study. *International Journal of Approximate Reasoning, 17*, 217–237.

Binaghi, E., & Madella, P. (1999). Fuzzy Dempster–Shafer reasoning for rule-based classifiers. *International Journal of Intelligent Systems, 14*, 559–583.

Benferhat, S., SafHotti, A., & Smets, P. (2000). Belief functions and default reasoning. *Artificial Intelligence, 122*, 1–69.

Binaghi, E., Gallo, L., & Madella, P. (2000). A neural model for fuzzy Dempster–Shafer classifiers. *International Journal of Approximate Reasoning, 25*(2), 89–121.

Beynon, M., Cosker, D., & Marshall, D. (2001). An expert system for multi-criteria decision making using Dempster Shafer theory. *Expert Systems with Applications, 20*, 357–367.

Beynon, M. (2002a). DS/AHP method: A mathematical analysis, including an understanding of uncertainty. *European Journal of Operational Research, 140*(1), 148–164.

Beynon, M. (2002b). An investigation of the role of scale values in the DS/AHP method of multi-criteria decision making. *Journal of Multi-Criteria Decision Analysis, 11*, 327–343.

Connor, J. T., Martin, D., & Atlas, L. E. (1994). Recurrent neural networks and robust time series prediction. *IEEE Transaction on Neural Networks, 5*(2), 240–254.

Chen, L. H. (1997). An extended rule-based inference for general decision-making problems. *Information Sciences, 102*, 111–131.

Chen, M. Z., Zhou, D. H., & Liu, G. P. (2005). A new particle predictor for fault prediction of nonlinear time-varying systems. *Developments in Chemical Engineering and Mineral Processing, 13*(3–4), 379–388.

Chen, K. Y. (2007). Forecasting systems reliability based on support vector regression with genetic algorithms. *Reliability Engineering and System Safety, 92*, 423–432.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multi-valued mapping. *Annals Mathematical Statistics, 38*, 325–339.

Deng, J. L. (1982). Control problems and grey systems. *Systems Control Letters, 1*, 288–294.

Denoeux, T. (1997). Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition, 30*(7), 1095–1107.

Denoeux, T. (1999). Reasoning with imprecise belief structures. *International Journal of Approximate Reasoning, 20*, 79–111.

Dimitras, A. I., Slowinski, R., Susmaga, R., & Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research, 114*, 263–280.

Denoeux, T. (2000a). A neural network classifier based on Dempster–Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems and Humans, 30*(2), 131–150.

Denoeux, T. (2000b). Modelling vague belief using fuzzy-valued belief structures. *Fuzzy Sets and Systems, 116*, 167–199.

Denoeux, T., & Zouhal, L. M. (2001). Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems, 122*(3), 409–424.

Denoeux, T., & Masson, M. (2004). EVCLUS: Evidential clustering of proximity data. *IEEE Transaction on Systems, Man, and Cybernetics—Part B: Cybernetics, 34*(1), 95–109.

Flint, A. D., Ingleby, M., & Morton, D. (1992). A new generalization of the Hough transform in trend analysis. In *Proceeding of the 1992 IEEE international symposium on intelligent control* (pp. 261–267). Piscataway, NJ, USA: IEEE.

Fabre, S., Appriou, A., & Briottet, X. (2001). Presentation and description of two classification methods using data fusion based on sensor management. *Information Fusion, 2*, 49–71.

Fan, X. F., & Zuo, J. M. (2006a). Fault diagnosis of machines based on D–S evidence theory. Part 1: D–S evidence theory and its improvement. *Pattern Recognition Letters, 27*, 366–376.

Fan, X. F., & Zuo, J. M. (2006b). Fault diagnosis of machines based on D–S evidence theory. Part 2: Application of the improved D–S evidence theory in gearbox fault diagnosis. *Pattern Recognition Letters, 27*, 377–385.

George, T., & Pal, N. R. (1996). Quantification of conflict in Dempster–Shafer framework: A new approach. *International Journal of General Systems, 24*(4), 407–423.

Guo, M., Yang, J. B., & Chin, K. S. (in press). Evidential reasoning based preference programming for multiple attribute decision analysis under uncertainty. *European Journal of Operational Research*.doi:10./1016./j.ejor.2006.09.064.

Ho, S. L., & Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computer & Industrial Engineering, 35*(1–2), 213–216.

Hullermeier, E. (2001). Similarity-based inference as evidential reasoning. *International Journal of Approximate Reasoning, 26*, 67–100.

Huber, R. (2001). Scene classification of SAR images acquired from antiparallel tracks using evidential and rule-based fusion. *Image and Vision Computing, 19*(13), 1001–1010.

Hu, C. H., Cao, X. P., & Zhang, W. (2005). Fault prediction based on bayesian MLP neural networks. *GESTS Transaction Journal, 16*(1), 17–25.

Huynh, V. N., Nakamori, Y., Ho, T. B., & Murai, T. (2006). Multiple-attribute decision making under uncertainty: The evidential reasoning approach revisited. *IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems and Humans, 36*(4), 804–822.

Hu, X., Danil, V. P., & Donald, C. W. (2007). Time series prediction with a weighted bidirectional multi-stream extended Kalman filter. *Neurocomputing, 70*, 2392–2399.

Hu, C. H., Si, X. S., & Yang, J. B. (2010). Systems reliability prediction model based on evidential reasoning algorithm with nonlinear optimization. *Expert Systems with Applications, 37*(3), 2550–2562.

Ishizuka, M., Fu, K. S., & Yao, J. T. P. (1982). Inference procedures and uncertainty for the problem-reduction method. *Information Sciences, 28*, 179–206.

Jones, R. W., Lowe, A., & Harrison, M. J. (2002). A framework for intelligent medical diagnosis using the theory of evidence. *Knowledge-Based Systems, 15*, 77–84.

Jiang, Y. Q. (2004). Applying grey forecasting to predicting the operating energy performance of air cooled water chillers. *International Journal of Refrigeration, 27*(4), 385–392.

Krishnapuram, R. (1991). A belief maintenance scheme for hierarchical knowledge-based image analysis systems. *International Journal of Intelligent systems, 6*(7), 699–715.

Lu, K. S., & Saeks, R. (1979). Failure prediction for an on-line maintenance system in a poisson shock environment. *IEEE Transaction on Systems, Man, and Cybernetics, 9*(6), 356–362.

Lapedes, A., & Farber, R. (1987). *Nonlinear signal processing using neural networks prediction and system modeling*. USA: Los Alamos National Laboratory.

Liu, J., Yang, J. B., Wang, J., Sii, H. S., & Wang, Y. M. (2004). Fuzzy rule-based evidential reasoning approach for safety analysis. *International Journal of General Systems, 33*(2-3), 183–204.

Murphy, C. K. (2000). Combining belief functions when evidence conflicts. *Decision Support Systems, 29*, 1–9.

Monney, P. A. (2003). Analyzing linear regression models with hints and the Dempster–Shafer theory. *International Journal of Intelligent Systems, 18*(1), 5–29.

Khoshgoftaar, M. T. (2003). Fault prediction modeling for software quality estimation: Comparing commonly used techniques. *Empirical Software Engineering, 8*, 255–283.

Mohandes, M. A., Halawani, T. O., & Reman, S. (2004). Support vector machines for wind speed prediction. *Renewable Energy, 29*(6), 939–947.

Pankratz, A. (1983). *Forecasting with univariate Box–Jenkins models: Concept and cases*. New York: Wiley.

Parikh, C. R., Pont, M. J., et al. (2001). Application of Dempster–Shafer theory in condition monitoring applications: A case study. *Pattern Recognition Letters, 22*(6–7), 777–785.

Pent-Renaud, S., & Denoeux, T. (2004). Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning, 35*(1), 1–28.

Pai, P. F., & Hong, W. C. (2006). Software reliability forecasting by support vector machines with simulated annealing algorithms. *The Journal of Systems and Software, 79*, 747–755.

Pai, P. F. (2006). Systems reliability forecasting based on support vector machines with genetic algorithms. *Mathematical and Computer Modelling, 43*, 247–262.

Rietman, E. A., & Beachy, M. (1998). A study on failure prediction in a plasma reactor. *IEEE Transaction on Semiconductor Manufacturing, 11*(4), 670–680.

Rakar, A., Juricic, D., & Ball, P. (1999). Transferable belief model in fault diagnosis. *Engineering Applications of Artificial Intelligence, 12*(5), 555–567.

Ruthven, I., & Lalmas, M. (2002). Using Dempster–Shafer's theory of evidence to combine aspects of information use. *Journal of Intelligent Information Systems, 19*(3), 267–301.

Rakar, A., & Juricic, D. (2002). Diagnostic reasoning under conflicting data: The application of the transferable belief model. *Journal of Process Control, 12*(1), 55–67.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letter, 9*(3), 293–300.

Siow, C. H. R., Yang, J. B., & Dale, B. G. (2001). A new modeling framework for organizational self-assessment: Development and application. *Quality Management Journal, 8*(4), 34–47.

Sonmez, M., Graham, G., Yang, J. B., & Holt, G. D. (2002). Applying the evidential reasoning approach to pre-qualifying construction contractors. *Journal of Management in Engineering, 18*(3), 11–119.

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Information Fusion, 8*, 387–412.

Tseng, F. M., Tseng, G. H., & Yuan, J. C. (2001). Fuzzy ARIMA model for forecasting the foreign exchange market. *Fuzzy Sets and Systems, 181*(1), 9–19.

Telmoudi, A., & Chakhar, S. (2004). Data fusion application from evidential databases as a support for decision making. *Information and Software Technology, 46*(8), 547–555.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Wallery, P. (1996). Measures of uncertainty in expert systems. *Artificial Intelligence, 83*, 1–58.

Wang, J., & Yang, J. B. (2001). A subjective safety based decision making approach for evaluation of safety requirements specifications in software development. *International Journal of Reliability, Quality and Safety Engineering, 8*(1), 35–57.

Wang, D., Zhou, D. H., & Jin, Y. H. (2004). A strong tracking predictor for nonlinear processes with input time delay. *Computers and Chemical Engineering, 28*(12), 2523–2540.

Wang, Y. M., Yang, J. B., & Xu, D. L. (2006). Environmental impact assessment using the evidential reasoning approach. *European Journal of Operational Research, 174*, 1885–1913.

Wang, Y. M., & Elhag, M. S. T. (2007). A comparison of neural network, evidential reasoning and multiple regression analysis in modeling bridge risks. *Expert Systems with Applications, 32*, 336–348.

Wang, W. (2007). An adaptive predictor for dynamic systems forecasting. *Mechanical Systems and Signal Processing, 21*, 809–823.

Wang, Y., & Guo, W. (2008). Local prediction of the chaotic fh-code based on LS-SVM. *Journal of Systems Engineering and Electronics, 19*(1), 65–70.

Wang, Y. M., & Elhag, M. S. T. (2008). Evidential reasoning approach for bridge condition assessment. *Expert Systems with Applications, 34*, 689–699.

Xu, K., Xie, M., Tang, L. C., & Ho, S. L. (2003). Application of neural networks in forecasting engine systems reliability. *Applied Soft Computing, 2*(4), 255–268.

Xu, D. L., & Yang, J. B. (2003). Intelligent decision system for self-assessment. *Journal of Multi-Criteria Decision Analysis, 12*, 43–60.

Xu, D. L., Yang, J. B., & Wang, Y. M. (2006). The evidential reasoning approach for multi-attribute decision analysis under interval uncertainty. *European Journal of Operational Research, 174*, 1914–1943.

Xu, D. L., Liu, J., Yang, J. B., Liu, G. P., Wang, J., Jemkinson, I., et al. (2007). Inference and learning methodology of belief-rule-based expert system for pipeline leak detection. *Expert Systems and Applications, 32*, 103–113.

Yager, R. R. (1992). Decision making under Dempster-Shafer uncertainties. *International Journal of General Systems, 20*, 233–245.

Yen, J. (1990). Generalizing the Dempster–Shafer theory to fuzzy sets. *IEEE Transaction on Systems, Man, and Cybernetics, 20*(3), 559–570.

Yang, J. B., & Singh, M. G. (1994). An evidential reasoning approach for multiple-attribute decision making with uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics, 24*(1), 1–18.

Yang, J. B., & Sen, P. (1994). A general multi-level evaluation process for hybrid MADM with uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics, 24*(10), 1458–1473.

Yang, S. K., & Liu, T. S. (1999). State estimation for predictive maintenance using Kalman filter. *Reliability Engineering and System Safety, 66*(1), 29–39.

Yang, J. B. (2001). Rule and utility based evidential reasoning approach for multi-attribute decision analysis under uncertainties. *European Journal of Operational Research, 131*, 31–61.

Yang, J. B., Dale, B. G., & Siow, C. H. R. (2001). Self-assessment of excellence: An application of the evidential reasoning approach. *International Journal of Production Research, 39*(16), 3789–3812.

Yang, S. K. (2002). An experiment of state estimation for predictive maintenance using Kalman filter on a DC motor. *Reliability Engineering and System Safety, 75*(1), 103–111.

Yang, J. B., & Xu, D. L. (2002a). On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 32*(2), 289–304.

Yang, J. B., & Xu, D. L. (2002b). Nonlinear information aggregation via evidential reasoning in multiattribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 32*(3), 376–393.

Yang, J. B., & Xu, D. L. (2004). Intelligent decision system for supplier assessment. In: DSS2004: The 2004 IFIP conference on decision support systems, Prato, Tuscany, Italy.

Yang, J. B., & Xu, D. L. (2005). An intelligent decision system based on evidential reasoning approach and its applications. *Journal of Telecommunications and Information Technology, 3*, 73–80.

Yang, J. B., Liu, J., Wang, J., Sii, H. S., & Wang, H. W. (2006a). Belief rule-base inference methodology using the evidential reasoning approach – RIMER. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30*(2), 266–285.

Yang, J. B., Wang, Y. M., Xu, D. L., & Chin, K. S. (2006b). The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties. *European Journal of Operational Research, 171*, 309–343.

Yang, B. S., & Kim, K. L. (2006). Application of Dempster–Shafer theory in fault diagnosis of induction motors using vibration and current signals. *Mechanical Systems and Signal Processing, 20*, 403–420.

Yang, J. B., Liu, J., Xu, D. L., Wang, J., & Wang, H. W. (2007). Optimization models for training belief-rile-based systems. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 37*(4), 569–585.

Zhou, D. H., & Frank, P. M. (1996). Strong tracking filtering of nonlinear time-varying stochastic systems with colored noise with application to parameter estimation and empirical robustness analysis. *International Journal of Control, 65*(2), 295–370.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35–62.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175.

Zhang, G. Z., Huang, D. S., & Quan, Z. H. (2005). Combining a binary input encoding scheme with RBFNN for globulin protein inter-residue contact map prediction. *Pattern Recognition Letters, 26*, 1543–1553.

Zhang, B., Chen, M. Y., Zhou, D. H., & Li, Z. X. (2006). Particle-filter-based estimation and prediction of chaotic states. *Chaos, Solitons & Fractals, 30*(5), 1273–1280.

Zhang, L. B., Wang, Z. H., & Zhao, S. X. (2007). Short-term fault prediction of mechanical rotating parts on the basis of fuzzy-grey optimizing method. *Mechanical Systems and Signal Processing, 21*, 856–865.

Zhou, Z. J., & Hu, C. H. (2008). An effective hybrid approach based on grey and ARMA for forecasting gyro drift. *Chaos Solitons & Fractals, 35*(3), 525–529.

Zhang, Z. D., & Hu, S. S. (2008). A new method for fault prediction of model-unknown nonlinear system. *Journal of the Franklin Institute, 345*, 136–153.