IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received December 30, 2021, accepted February 7, 2022, date of publication February 21, 2022, date of current version April 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3151240

Bearing Fault Diagnosis Under Small Data Set Condition: A Bayesian Network Method With Transfer Learning for Parameter Estimation

YONGYAN HOU[1], (Member, IEEE), AO YANG[1], WENQIANG GUO[2], (Member, IEEE), ENRANG ZHENG[1], QINKUN XIAO[3], (Member, IEEE), ZHIGAO GUO[4], AND ZIXUAN HUANG[2]

[1]School of Electrical and Control Engineering, Shaanxi University of Science and Technology, Xi'an, Shaanxi 710021, China
[2]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi 710021, China
[3]College of Electronic Information Engineering, Xi'an Technological University, Xi'an 710021, China
[4]School of Electronic Engineering and Computer Science, Queen Mary University, London E1 4NS, U.K.

Corresponding author: Wenqiang Guo (guowenqiang@sust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62071366 and Grant 61271363, in part by the Xi'an Science and Technology Bureau Plan Project under Gant 2019216514GXRC001CG002GXYD1.1, and in part by the General Project of Science and Technology Department of Shaanxi Province of China under Grant 2020SF-286.

ABSTRACT Bearings are broadly applied in various types of industrial systems. Fault diagnosis, as a promising way for reliability of modern industrial internet of thing applications, has attracted increasing attention from both academia and industry fields. Being ideal modeling and inference tool in uncertainty situations, Bayesian network (BN) is becoming increasingly popular in many systems. However, in practical uncertain and complicated engineering surroundings, it's difficult or expensive to collect massive labeled fault data for the sake of fault diagnosis model learning. To address the issue of BN parameter learning under small data set conditions, this paper proposes a Varying Coefficient Transfer Learning (VCTL) algorithm based on aggregation and transfer learning, that considers both knowledge from the resource domain and the resource relevance contributions. The balancing weight function is designed to determine whether the learning task in the resource domain is activated. Relevance weight factors are proposed to measure the relevance of resource and target parameters quantitatively, by combing parameter information from resource domains with those obtained from the target domain, using maximum a posterior (MAP) or maximum likelihood estimation (MLE). Finally, the target parameters are aggregated with both the target initial parameters and the parameter knowledge from the resource domain. Based on VCTL, a bearing fault diagnosis approach is proposed and verified. The experimental results show that, under the condition of the small data set, learning accuracy of VCTL algorithm with varying coefficient aggregation is better than MLE algorithm, MAP algorithm or state-of-the-art parameter transfer method, local linear pooling transfer learning (LoLP) algorithm. Under the condition of sufficient data set, learning accuracy of VCTL algorithm approaches the classical MLE or MAP, and the correctness of the proposed algorithm is verified. Moreover, we illustrate the successful application to real-world bearing fault diagnosis case with VCTL, where we had access to expert-provided resource knowledge and real fault diagnosis data.

INDEX TERMS Bayesian network, fault diagnosis, small data set, parameter learning, transfer learning.

I. INTRODUCTION

Bearing has been broadly used in rotating machinery and plays a significant role in the mechanical system in under complex and variable surroundings. Fault diagnosis (FD) of bearing is essential to reduce the incidence of catastrophic failures and heavy economic losses, and bearing FD has attracted increasing attention from both academia and industry fields [1]–[3].

With the rapid development of industrial Internet of Things (IIoTs), many FD techniques have been successfully applied in the modern industries, such as the vibration-based

The associate editor coordinating the review of this manuscript and approving it for publication was Laura Celentano.

method, current-based method, acoustic emission-based method, sound-based method, torque-based method, and rotating encoder-based method, etc [4]–[6]. The paper [5] shows a good survey for fault diagnosis methods and their applications in rotating machinery. In the past decades, Artificial Intelligence (AI)-based fault diagnosis methods have made remarkable achievements in data-driven FD solutions, by using machine learning techniques. The AI-based fault diagnosis methods often consist of data acquisition, fault feature extraction and fault pattern identification steps. In AI-based fault diagnosis schemes, the most commonly used classifiers are K nearest neighbor (KNN), neural network (NN), and support vector machine (SVM) [7].

In recent years, more and more machine learning methods have been introduced to bearing fault diagnosis. Hereinto, deep learning methods have been introduced to bearing fault diagnosis. Based on the deep learning framework, to address the problem of bearing fault diagnosis, while Yang *et al.* verified long-short term memory recurrent neural network (LSTM) method [8], Wu *et al.* also used LSTM scheme which generated auxiliary datasets from Case Western Reserve University bearing dataset, and then grey wolf optimization algorithm was introduced to learn parameters of joint distribution adaptation for locomotive bearing model [9]. While an adaptive deep belief network (DBN) method is presented with good classification report by a large number of comparative experiments for rolling bearing fault diagnosis in [10], a stacking auto-encoder (SAE) method is also investigated [11]. Based on Convolutional neural network (CNN), fault diagnosis under noisy environment and different working load is also advanced [12]. These studies verified that machine learning methods could overcome the human being methods' imitation effectively by data driven techniques.

However, deep networks frequently face the dilemma of how to construct the deep models with the lack of explainable and theoretical support. And deep networks are used often under a common assumption of a large amount of labeled training data is available and either the training data or the testing data is drawn from the same feature space [9]. In practical engineering applications, uncertainties and complexities often arise owing to noise, varying load conditions, and abnormal signal acquisition ways, etc. Therefore, how to use a small number of labeled data to build a reliable FD model in uncertain surroundings becomes particularly important. These problems restrict the deep learning methods' application.

In the field of uncertain knowledge representation and reasoning, Bayesian networks (BNs) are one of the most effective probabilistic graphical models [13], [14]. They have been broadly applied in various fields such as medical sciences, brain sciences, ecology, and manufacturing fields, etc [15]–[18]. Under the condition of sufficient modeling sample data, Guo *et al.* proposed a BN structure model to implement bearing fault diagnosis [19], with maximum likelihood estimation (MLE) parameter learning and existed inference algorithms to handle the uncertainty of reasoning with incomplete evidence.

When the BN model structure has been defined and the size of the data is relatively small, the maximum a posterior (MAP) probability estimation [20] is the most commonly used learning method. Since in real-world decision support problems that we wish to model as BNs, there are typically limited or small observation data, how to improve the accuracy of parameter learning under the condition of small data set has been one of the hot topics.

When the data size for target modeling is small, transfer learning (TL) has brought impressive progress to the state-of-the-art across a variety of machine learning tasks, including image classification, natural language processing, object recognition and so on [21]. To solve the abnormal condition identification modeling problem for the magnesia smelting process, the paper [22] shows a good survey for transfer learning methods using computational intelligence. TL is an AI technique, which can improve learning in the new task by transferring knowledge from the relevant learned task. In order to effectively transfer the classifier model between different domains, many methods have been investigated for transfer learning. The learned model knowledge is shared with the target model to optimize the learning effect of the model. Fundamental challenges in TL include computing how much multiple resources contribute their relevance for target modeling and how to fuse resource and target information. Combining resource domain network and a transferring learning, many methods are proposed to improve the performance of the decision support systems [23]. Ref. [1] used CNN and Multi-layer Perceptron (MLP) to train several base models with a mount of source data, and the models are transferred to target data with different level of variations. Ref. [9] proposed an adaptive deep transfer learning method using a long-short term memory (LSTM) recurrent neural network model. Then instance-transfer learning is used to generate some auxiliary data sets. The joint distribution between a generated auxiliary data set and target domain data set are estimated to construct the fault diagnosis model. But both the proposed LSTM model structure and parameters are too complicated and subjective to use in practice.

In the context of transfer learning in BNs, focusing on structure learning, Ref. [24] proposed the BN estimation algorithm from the related tasks based on the score-based method. To construct the target BN parameters, Chen *et al.* [25] proposed a linear aggregation method by weighted average initial parameters from resource domain parameters. But, such averaging method fails to exploit the contribution of the different resource domain data to the target domain parameter learning. Ref. [26] presented a Bayesian model averaging framework to estimate the structure and parameter. However, the parameter Dirichlet distribution assumption without closed form solution restricts the method's application. Luis *et al.* presented local linear pooling (LoLP) aggregation methods by considering the conditional probabilistic table (CPT) confidence [27].

However, without considering the relevance between the resource samples and the target samples, this method is too simplistic, which only relies on the CPT entry size and dataset size. Moreover LoLP fusion assumes every resource is equally related to the target. By integrating the expert knowledge, Ref. [22] and [28] presented a BN parameters transfer learning method regarding the varying balance between target and resource model. However, experimental comparisons of varying balance method and fixed (or Freeze) balance method are not investigated. Moreover, the similarities of alternative resources and weights for fusion function purely depend on the expert knowledge, and resource sample data potentials are not exploited for learning.

To address the issue of the bearing fault diagnosis under small data set condition, this paper proposed a fault diagnosis method based on varying coefficient transfer learning (VCTL) when the data amount for BN parameter learning is small.

Major contributions of this paper include the following:

(i) A transfer learning parameter aggregation model is proposed, which can exploit sample potentials from multiple resource networks for target BN parameter estimation. In the view of the presented aggregation model, when the target samples size for learning is small, the target BN parameter estimation can be improved with the help of transfer knowledge from resource models; with the increasing amount of target samples, the learning of target BN parameters depends more and more on the target samples itself automatically.
(ii) A relevance weight factor function is designed, which can quantify resource relevance with target samples.
(iii) The robustness of presented varying coefficient transfer learning (VCTL) is verified in irrelevant resources with better noise tolerance than classical LoLP approach by experimental studies.
(iv) Under small data set condition, a bearing fault diagnosis approach is proposed and verified by real case studies based on VCTL.

This remainder of this paper is organized as follows. Section II exposes the previous knowledge. In Section III, to address the issue of BN modeling under small data set condition, we propose a varying coefficient transfer learning for BN parameter estimation and bearing fault diagnosis algorithm based on BN modeling in Section IV. We experimentally validate the proposed approach by benchmark networks and real bearing fault diagnosis case studies in Section V. Finally, conclusion is described in Section VI.

II. BACKGROUND KNOWLEDGE

A. BAYESIAN NETWORK

A Bayesian network is made up of a set of probability distributions connected by a directed acyclic graph (DAG) and n variables. The following is the joint probability distribution for BN [29,30]:

$$P(X_1, X_2 \ldots X_n) = \prod_{i=1}^{n} P(X_i | P_a(X_i)) \tag{1}$$

where $P_a(X_i)$ represents the conditional probability distribution of the parent node set of X_i in G, $P(X_i | P_a(X_i))$ represents the probability of each value of a variable containing a given parent node value in G.(See [29] for more details.)

Let $\theta_{ijk} = P(X_i = k | P_a(X_i) = j)$ be a parameter of node X_i, given parent state j, the i^{th} node takes the value of the k^{th} state ($1 \le i \le n$, $1 \le j \le q_i$, $1 \le k \le r_i$). Obviously, node X_i has $r_i \times q_i$ parameters, which constitute a $r_i \times q_i$ dimensional matrix, which is called the conditional probability table (CPT) of node X_i.

With various types of inference algorithms based on complete or incomplete observation evidence, BNs can perform backward or diagnostic analyses. In fault diagnosis, exact inference algorithms were commonly used.

B. FAULT DIAGNOSIS FEATURES EXTRACTION FOR BN MODEL

Features extraction is one of the most significant components in the machine-learning approach. The quality of feature extraction will influence the diagnosis performance significantly.

The original time domain bearing vibration data acquired by transducers or dynamometers is frequently used for analysis. One of the most significant and extensively used techniques in signal processing disciplines is the fast Fourier transform (FFT), which can convert time domain data into frequency domain data. We utilize the FFT algorithm to process all vibration data after collecting vibration signals from K different types of rotor-bearing system working circumstances, including normal and abnormal types.

The energy of time domain signals is equivalent to the energy of frequency domain signals, according to Parseval's theory [19].

$$W = \int_{-\infty}^{\infty} f^2(t)dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\mathrm{j}w)|^2 dw \tag{2}$$

Multiple frequency energy (MFE) is the energy of the multiple frequency and its reflection in frequency domain. It is given by

$$W(I) = \frac{1}{2\pi} \int_{Iw_n - w_c}^{Iw_n + w_c} |F(\mathrm{j}w)|^2 dw \tag{3}$$

where $W(I)$ is the I^{th} MFE of a signal; ω_n is base frequency; ω_c is the multiple frequency surplus (MFS) used to extract MFE, and its value is $\eta\omega_n$, where η is a constants($0 < \eta < 1$); $[I\omega_n - \omega_c, I\omega_n + \omega_c]$ is defined as the interval in frequency domain whose total energy constitutes the MFE, abbreviated as IMFE.

When the bearing system is under varied situations, such as bearing inner race defects and rotor imbalances, MFE can

intuitively reflect the changes at distinct multiple frequencies. We use MFES as the fault signal's characteristics, and then use it to diagnose rotating machinery faults.

Based on acceleration sensors, the original sampled vibration signals data in time-domain are obtained, $s = \{tag\_sf_s(n) | s, K, m, N \in Z+; L = mN; s = 1, \ldots, K, n = 0, \ldots, L-1\}$. tag_s is the number of typical fault types, and the related acquisition signal set $f_s(n)$ is composed of m groups of N length data (N takes 1024 in this paper). Set fault type number K=4, types of bearing working conditions are summarized as follows: bearing normal fault (NF), inner race fault (IR), outer race fault (OR) and ball fault (BF) in this paper. Then s=1, 2, 3, 4 can represent NF, IR, OR and BF respectively. Take j as the group number, then $f_s(n)$ is given by

$$f_s(n) = \sum_{j=1}^{m} f_{s,j}(n'); \quad n' = 0, \cdots, N-1 \tag{4}$$

FFT results of the signals $f_{s,j}(n')$ are acquired as follows:

$$F_{s,j}(k) = \sum_{n'=0}^{N-1} f_{s,j}(n') e^{-j\frac{2\pi}{N}kn'} \tag{5}$$

where $k = 0, 1, \ldots,$ N -1; $j = 1, 2, \ldots,$ m.

With respect to the signal of discrete Fourier transform nature, the spectrum of FFT along the N/2 point is symmetry. $\left|F_{s,j}(k')\right|$ is divided into v segments using isometric segmentation ($k' = 0, 1, \ldots, \frac{N}{2} - 1$), and frequency signal $W_{s,j,u}$is obtained by piecewise summation (u $= 1, \ldots,$ v), where v is the number of fault feature component and takes 2$*$K in this paper. The u^{th} fault characteristic vector (FCV) in group j ($j = 1, \ldots,$ m), $W_{s,j,u}$, is computed by

$$W_{s,j,u} = \sum_{k'=0}^{u\frac{N}{2v}-1} \left|F_{s,j}(k')\right| \tag{6}$$

FCVs are frequently discretized as attribute values to speed up BN inference. We partition the variable range evenly into multiple r pieces, as proposed in [19], and obtain discrete-FCV attribute values between 1 and r. (r is a positive integer). In this article, r is set to 3 to represent the FCV component of bearing characteristic information in the frequency domain as low, moderate, or high energy. respectively.

C. BAYESIAN NETWORK PARAMETER LEARNING

Conditional probabilities are predicted to occur when other events are known to occur, according to the BN model. This information is frequently found in a conditional probability table (CPT). CPTs can be obtained in a variety of ways, but the two most common are knowledge elicitation from experts and data-driven parameterization using machine learning algorithms. CPTs are also constructed using machine learning approaches by learning BN parameters from historical data.

When the model structure has been defined but the parameters are unknown, MLE is a method for estimating the parameters of the model with supplied observation data. When there is sufficient data, the maximum likelihood estimation approach is typically employed to improve parameter learning accuracy. The MLE can be defined as following [29]:

$$\theta_{ijk}^{MLE} = \frac{N_{ijk}}{\sum_k N_{ijk}} \tag{7}$$

where N_{ijk} represents the j^{th} state of the parent node in the sample data, and the i^{th} node takes the statistical value of the kth state.

The Dirichlet distribution can be used to model the prior distribution of parameters in BN in practice. To provide point estimation of the BN parameter, MAP estimation is relied on empirical data [31]. MAP is estimated as:

$$\theta_{ijk}^{MAP} = \frac{N_{ijk} + \alpha_{ijk}}{\sum_k (N_{ijk} + \alpha_{ijk})} \tag{8}$$

where α_{ijk} is a hyper-parameter, which is a constant. When there is no expert judgment, the K2 ($\alpha_{ijk} = 1$) or BDeu ($\alpha_{ijk} = \frac{1}{r_i \times q_j}, \forall i, j, k$) priors are commonly used [32], [33].

Let G be a network structure, the sample complexity bound for parameter learning with a fixed BN structure can be determined as follows [15]:

$$C \geq \frac{1}{2\lambda^{2(d+1)}} \frac{1+\varepsilon^2}{\varepsilon^2} \log \frac{NK^{(d+1)}}{\delta} \tag{9}$$

where C is the instance number required for modeling to obtain a PAC-bound in the error ε, d is the maximum number, ε is Kullback-Leibler distance (KLD) error with confidence δ; N is the total node number in a BN and K is the maximal variable cardinality.

D. TRANSFER LEARNING FOR BN PARAMETER ESTIMATION

To complete the BN model, the parameters or CPTs for each variable must be calculated given a Bayesian network structure. Large historical data sets can yield accurate estimations, whereas small data sets typically yield poor estimation results. In practice, however, obtaining sufficient and available historical data is difficult, if not impossible. The goal of this research is to use aggregation from many resources to fill all of the conditional probability tables in the context of employing transfer learning from auxiliary parameter knowledge tasks.

Transfer learning is a data acquisition learning strategy that addresses the problem of insufficient target data. The target network is a set of parameters that must be estimated; any network in M resource domains that is comparable to the target network is referred to as a resource network, which can give information for learning the target network. Figure 1 depicts the methodology for estimating target network parameters via transfer learning.

A target network is represented as $\Delta^T = \{D^T, G^T, V^T\}$, where D^T stands for the data of the target network, G^T represents the structure of the target network, and V^T represents the dimension of the target network [26].

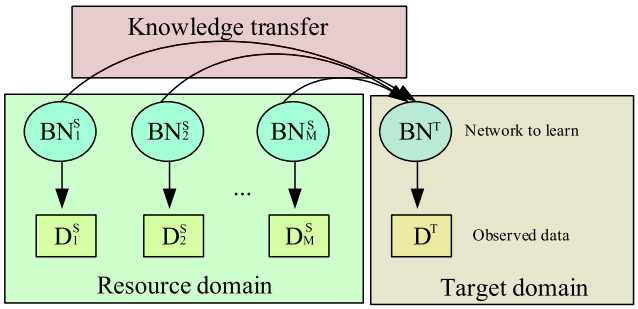


FIGURE 1. Target bayesian network learning with transfer learning.

Meanwhile, the resource network is represented as $\Delta^S = \{D^S, G^S, V^S\}$, where D^S stands for the data of the resource network, G^S represents the structure of the resource network, and V^S represents the dimension of the resource network.

Once we have a set of CPTs in terms of the target and resource networks, involving the same variables, we can proceed to combine them. There are several aggregation functions or operations commonly used, such as sum, maximum, minimum, count, and average.

Luis presents a method for transferring the LoLP, commonly known as a weighted linear average [27], based on satisfying the structural consistency and parameter dimension consistency (CPT entry) requirements. The probability P(X) is the weighted sum of the probabilities from the target and resource networks, which is written as:

$$P(X) = k_M \times \sum_{L=1}^{M} \omega_L P_L(X) \tag{10}$$

where $P_L(X)$ represents the conditional probability of the L^{th} model involving variable X in M models, ω_L is the weight associated with the probability, and k_M is the normalized factor (see more details in [27]).

E. DIRICHLET COMPOUND MULTINOMIAL SIMILARITY

In many applications, resource networks and target network are known to be structurally similar. They share the same variable entries. The sole ambiguity in this scenario is determining which of several potential resource networks is most relevant to a target.

To measure the relevance between target and resource data, authors in [30] adopt the Bayesian model comparison for two hypotheses as: H_1 is the relevance hypothesis that the resource and target data share a common CPT, and H_0 is the independent hypothesis that the resource and target data have distinct CPTs. If there are M resource networks, for discrete data, the likelihood of H_1 and the L^{th} resource network ($L \in [1, M]$) to the target is referred as $R_L\left(D^T, D^{S_L}\right)$ in the Dirichlet compound multinomial (DCM) distribution [30]:

$$\begin{aligned} R_L(D^T, D^{S_L}) &= p(D^T | D^{S_L}, H_1^S) \\ &= \frac{\Gamma(A^{X^{S_L}})}{\Gamma(N^{X^T} + A^{X^{S_L}})} \prod_{i=1}^{n} \frac{\Gamma(N_i^{X^T} + \alpha_i^{X^{S_L}})}{\Gamma(\alpha_i^{X^{S_L}})} \end{aligned} \tag{11}$$

where $i = 1, \ldots, n$ is the variable number, N^{X^T} is the observation counts of the target parameter in data D^T, and $N^{X^T} = \sum_i N_i^{X^T}$; $\alpha_i^{X^{S_L}}$ indicates the observation counts from the L^{th} resource domain, and $A^{X^{S_L}} = \sum_i \alpha_i^{X^{S_L}}$.

III. VARYING COEFFICIENT TRANSFER LEARNING FOR BN PARAMETER ESTIMATION

A. TRANSFER LEARNING PARAMETER AGGREGATION MODEL

The goal of transfer learning is to use experience or knowledge gained from multiple tasks to improve performance on a similar but distinct task.

The target BN parameter estimation exploiting transfer learning aggregation is calculated as following:

$$\theta_{ijk}^{VCTL} = \beta\theta_{ijk}^{T} + (1-\beta)\theta_{ijk}^{S} \tag{12}$$

where θ_{ijk}^{T} is the initial BN parameter of the target network, and θ_{ijk}^{S} stands for the fused BN parameter of the resource networks; β ($0 \le \beta \le 1$) represents the balancing weight between the parameters in the resource domain and target domain.

The size of target domain sample as N^T is readily derived by statistics. In addition, according to Eq. (9), the small data set sample threshold C indicates that the target domain samples are insufficient or sufficient for parameter estimation. The sample size factor(SSF) is calculated in the following way:

$$\text{SSF} = \frac{N^T}{C} \tag{13}$$

Then, the balancing weight in Eq. (12) is designed as following:

$$\beta = \begin{cases} 1 - \dfrac{1}{e^{\text{SSF}}}, & if\ \text{SSF} \le \text{Tr} \\ 1, & \text{else} \end{cases} \tag{14}$$

where Tr is a constant for deciding whether an auxiliary parameter learn task in resource domain is triggered or not. In practice, Tr often takes 3 or 5.

Note that:

(1) Since SSF in Eq.(13) is non-negative, the balancing weight in Eq.(12) and (14) is always between 0 and 1;

(2) When SSF > Tr in Eq. (14), this hints that there is sufficient data can be used to estimating parameter by classical learning approach, such as MLE or MAP;

(3) β is varying and when it tends to 1, the final target aggregation parameters prefer target samples to resource samples for learning in the proposed transfer aggregation model.

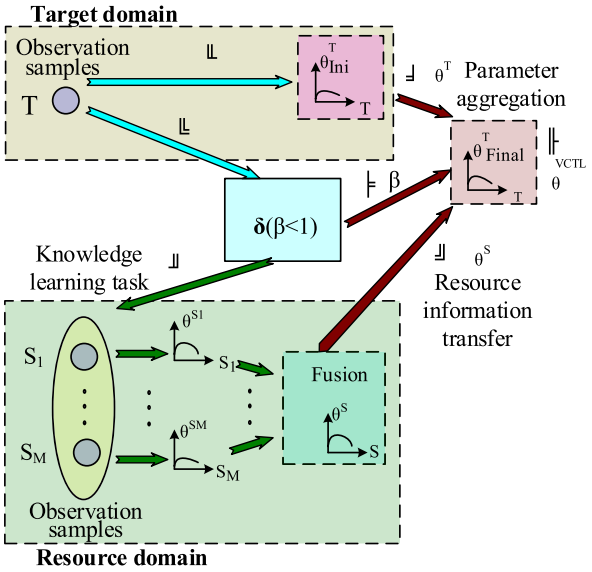


FIGURE 2. An architecture of transfer learning parameter aggregation model for the target BN.

This study proposes a target parameter aggregation model based on the transfer learning mechanism, as shown in Fig. 2.

As shown in Fig. 2, when the target data are obtained, the initial BN parameter can be estimated by MAP, and this procedure is indicated as "①". At the same time, as indicated as "②", sample threshold C can also be determined by Eq. (9) in an auxiliary task.

If balancing weight β for modeling equals to 1, we can notice that the observation samples are sufficient from Eq. (13) and (14). And the knowledge learning task in resource domain will not start, and the parameter aggregation result θ^{VCTL} is determined by target domain learning result θ^{T}.

If β is smaller than 1, the learning task in the resource domain is triggered as indicated as "④", which can provide the fused parameter θ^{S} by parameter learning in the resource domain.

In Fig. 2, the indicator function $\delta(\beta < 1)$ output "1" or "0". While output is 1, it means $\beta < 1$. Then in the resource domain, auxiliary parameter learning and fusion task will be activated, and outputs of "⑤" and "⑥" are provided for computing parameter aggregation according to Eq. (12); otherwise, the final parameter estimation in "⑦" will only depend on target samples, i.e., the sample data size is sufficient for modeling.

B. VARYING COEFFICIENT TRANSFER LEARNING FOR BN PARAMETER ALGORITHM ESTIMATION

If there are M resource networks, and let $\theta_{ijk}^{S_L}$ represent the BN parameter of the L^{th} resource network ($L \in [1, M]$) in the VCTL model, and ω_{S_L} is the relevance weight factor of the L^{th} resource network. To measure the relatedness between target and resource samples, the relevance weight factor is proposed as Eq.(15).

An overview of our Varying Coefficient Transfer Learning for BN parameter estimation is given in Algorithm 1.

$$\omega_{S_L} = \begin{cases} \dfrac{R_L\left(D^T, D^{S_L}\right)}{\sum_{L=1}^{M} R_L\left(D^T, D^{S_L}\right)}, & if\ size\ of\ (D) \neq 0 \\ \dfrac{1}{M}, & else \end{cases} \tag{15}$$

where $R_L\left(\mathrm{D^T}, \mathrm{D}^{S_L}\right)$ represents the DCM similarity between target sample data set and the sample data set from L^{th} resource network. Note thatω_{S_L} varies regarding the similarity between the data set in resource samples and the target samples. If and only if no target sample is obtained, ω_{S_L} is unified with the same weight 1/M for further parameter estimation. The sum of all ω_{S_L} always equals to 1.

The fused BN parameter of the M resource networks is calculated as follows:

$$\theta_{ijk}^{S} = \sum_{L=1}^{M} \omega_{S_L} \theta_{ijk}^{S_L} \tag{16}$$

where $\theta_{ijk}^{S_L}$ is the learned parameters from L^{th} resource network samples by MLE.

Algorithm 1: Varying Coefficient Transfer Learning Algorithm for BN Parameter

input : Target network $\mathrm{G^T}$ and sample set $\mathrm{D^T}$
Resource network $\mathrm{G^S}$ and sample set $\mathrm{D^S}$
Threshold Tr for learning in the resource domain
output: BN parameter $\theta_{\mathrm{ijk}}^{\mathrm{VCTL}}$
Count dataset size N^{T} regarding to $\mathrm{D^T}$
Calculate sample threshold C by Eq (6)
for *all i, j, k* **do**
  Calculate the initial target network parameter $\theta_{\mathrm{ijk}}^{\mathrm{T}}$ by MAP **if** $N^T \geq Tr * C$ **then**
    $\theta_{\mathrm{ijk}}^{\mathrm{VCTL}} \leftarrow \theta_{\mathrm{ijk}}^{\mathrm{T}}$;
  else
    %auxiliary learning task
  end
  Compute balancing weight by Eq. (14)
  Count the number of source networks M
  for *L=1:M* **do**
    Compute resource network S L parameter by MLE;
    Calculate relevance weight factor by Eq. (15)
  end
  Compute fused resource parameter by Eq. (16);
  Calculate the final target network parameter by Eq. (12);
end
end
return $\theta_{\mathrm{ijk}}^{\mathrm{VCTL}}$

Note that:

(1) Varying relevance weight factor can reflect the contribution of the data from different resource samples regarding their similarity with the target domain samples quantitatively; each resource contribution is compared to target net by similarity evaluation. Obviously, by Eq.(15) and (11), the most relevant weight factor is assigned to the biggest weight value;

(2) By introducing the proposed varying coefficients, i.e., balancing weight β and relevance weight factors ω_{S_L}, how much parameter knowledge transferring from the resource domain to target domain, can change along with relevance weight factors in parameter estimation automatically.

In general, when the target sample data set size is limited or small, the target domain parameter estimation can be improved by the assistance of the learned parameter knowledge transfer from the resource networks.

C. COMPUTATIONAL COMPLEXITY

The overall number of relevance estimation and MAP/MLE operations in the VCTL method determines its computational complexity. Because it can be as fast as linear programming solvers, we treat the calculation of MAP or MLE as an elementary operation O(1). If M resources are available, the time complexity of similarity and MLE calculation in VCTL is M∗O. (1). As a result, the computational complexity of VCTL is O(1)+M∗O(1), with M always being a constant. This shows that the computational complexity of VCTL is nearly identical to that of MAP.

For example, processing the classical Weather (also known as Sprinkler) Bayesian network took 1.31 milliseconds in VCTL and 1.03 milliseconds in MAP (see Table 5, row 5 column 1) on our computer (Intel core i7 CPU @2.6 GHz).

IV. BEARING FAULT DIAGNOSIS ALGORITHM BASED ON BAYESIAN NETWORK MODELING

The bearing fault diagnosis system based on Bayesian network modeling can be described and depicted in Fig. 3. Five sub-modules that make up the diagnosis system are data collection, inference and learning modules, knowledge module, output module, and man-machine interface.

Sensors in the data acquisition module keep track of environmental data and bearing status. After being processed in features samples and stored in a database, the acquired data could be available to the fault diagnostic inference engine for the purpose of BN learning or inference. When the target characteristic signal appears, the effecter will instantly start the information processor in the reasoning and learning module to construct a new fault diagnosis task.

There are two types of knowledge in knowledge module: one is BN modeling information, such as methods for fault diagnosis BN structure, and parameter learning algorithm which uses VCTL in this paper; the other is the inference engine model knowledge which exploits Junction tree inference algorithm to conduct BN reasoning.

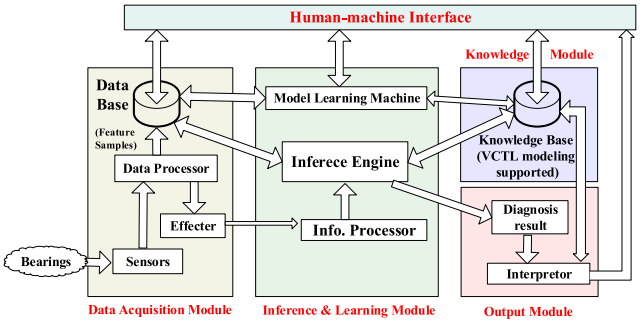


FIGURE 3. The intelligent fault diagnosis system using time-varied Bayesian network modeling.

The model learning machine in the reasoning and learning module plays a significant role, as it may access historical data in the database and create or change fault diagnosis inference engine models using a number of existing Bayesian network (structure and parameters) learning techniques. The new fault diagnostic task is created by the information processor, and the evidence inference engine for fault diagnosis inference is coordinated by the information processor.

The output module logically and intuitively interprets the inferred fault diagnosis results to give decision support via the man-machine interface. And the following are the functions of the man-machine interface: presenting the fault diagnosis system's status and findings; updating each module via commands, such as database maintenance, learning algorithms, inference algorithms, and expert knowledge.

The steps of the bearing fault algorithm based on BN and transfer learning aggregation are presented as follows:

Step 1: Set the number of sample datasets m and the fault diagnosis belief threshold parameter θ^*. Set the number q for the fault type "Bearing," the initial fault diagnosis type parameter $s = \{1, \ldots, q\}$ and initial type tag $tag\_s = \{1, \ldots, q\}$. Generally, θ^* is between 0.7 and 0.85; q is often between 3 and 5.

Step 2: Acquire the vibration signals based on acceleration or other sensors and obtain the sampled data; $data\_s = \{tag\_sf_s(n) \,|m, N \in Z+; L = mN; n = 0, \ldots, L-1\}$; N is often 1024. Meanwhile $f_s(n)$ is m groups of N length data. Take$j$ as the set number, then $f_s(n)$ is given by Eq. (4);

Step 3: Process the vibration signals $f_{s,j}(n')$ acquired by acceleration sensors with FFT by (5). The reason of FFT is used lies that vibration fault signals in time domain are not intuitively clear in various frequency components for the characteristic retrieval.

Step 4: Calculate the fault characteristic vector. According to the signal of discrete Fourier transform nature, the spectrum along the N/2 point is symmetry. Divide $\left|F_{s,j}\left(k'\right)\right|$ into v segments using isometric segmentation ($k' = 0, 1, \ldots, \frac{N}{2} - 1$), and acquire the $W_{s,j,u}$ by piecewise summation ($u = 1, \ldots, v$). Take frequency signal $W_{s,j,u}$ as u^{th} the fault characteristic vector in group j which is given by (6).

Step 5: To boost the inference speed for BN, discretize the characteristic vector by isometric segmentation. The continuous variable $W_{s,j,u}$ is represented by numerical attribute value by variable range divided evenly into r parts. The attribute values are between 1 and r. And r often takes 3 or 4. The attributes reflect various frequency components, such as low, middle and high frequency components.

Step 6: Construct the fault diagnosis BN structure. Take the fault type as the parent node "Bearing" and v target characteristics data as the children nodes, connect the parent node and each child node with directed arcs from parent node to children nodes, where the arc arrows point to children nodes from the same parent node. Thus, the structure of BN can be established.

Step 7: Determine the BN's conditional probability parameter. Sample threshold C for parameter modeling is calculated using Eq. (9).

If m $\geq$ 5C, id., the sample size for parameter is sufficient, the BN's conditional probability parameters can be learned using q types of discrete fault characteristic data via EM algorithm, then go to Step 9 for evidence inference; otherwise, the sample size for parameter is small, resource domain information for parameter learning using VCTL is started.

Step 8: The target BN parameter estimation exploiting transfer learning aggregation is calculated using Eq. (12). The methods to derive samples in resource domains are same in the description from Step 2 to Step 5. The fault diagnosis BN model is implemented.

Step 9: Obtain target domain observation characteristics evidence in order to diagnose the fault. Consider the variables m=1, s=, and tag s=. Repeat Steps 2 through 5, the fault diagnosis inference based on observation evidence can be ready.

Step 10: Enter the target characteristics observations which may be complete or incomplete ones for single BN inference engine. The diagnosis belief values are updated using the junction tree inference algorithm.

Step 11: If $\theta > \theta*$, compute the fault diagnosis according to Eq. (17), then output diagnosis results and the diagnosis process stops; otherwise, go back to Step 9 to acquire further target observations through the sensor system.

The output of type node x_S in fault diagnosis BN can be achieved by

$$x_S^* = \arg\max_l (P(x_s^l | ev = \{W_1, W_2, \cdots, W_v\})) \qquad (17)$$

where W_v is the v^{th} target characteristic observation); x_s^l is the l^{th} event of node "Bearing", $1 \leq l \leq$ q, q is often 3 or 4.

V. EXPERIMENT EXPERIMENTAL EVALUATION

We first assess our VCTL parameter learning methodology on three benchmark networks from the BN repository (http://www.bnlearn.com/bnrepository/) before moving on to real bearing defect diagnosis case studies utilizing our proposed transfer learning aggregation method. The learning performances of BN parameters are compared using the classic MLE algorithm, MAP algorithm, Freeze method (transfer learning without target domain samples), LoLP method (transfer learning with fixed coefficient for aggregation), and VCTL (transfer learning with varying coefficient for aggregation) algorithms under various data conditions. Functions and subroutines are used to implement algorithms based on the BNT toolbox [35] in MATLAB R2014a. All the experiments were performed on an Intel core i7 CPU running at 2.6 GHz and 8 GB RAM.

TABLE 1. Descriptions of weather, Asia, and Alarm BNs.

Name	Paras #	Nodes	Arcs
Weather	9	4	4
Asia	18	8	8
Alarm	509	37	46

Paras #: Total number of parameters in each BN.

A. SYNTHETIC SAMPLE EXPERIMENT SETTINGS

Firstly, we utilize BN models with known structure and parameters, which we can use as a gold standard for comparing alternative learning strategies. We ran experiments on three benchmark BN networks (Weather, Asia, and Alarm), with parameter sizes ranging from small to medium to high. Table 1 contains the details and descriptions of these BNs. We employ the "sample bnet" function in the BNT package [35] as the sampling generator to create varied sample sizes for parameter learning among a series of original BN models.

The Weather model structure is shown in Fig. 4, and the ground truth parameters come from BN repository. The threshold value of small data set C is 121 by Eq. (9), when λ takes 1, d is 2, N takes 4, δ is 0.05 and ε takes 0.1. This indicates that, in the target domain, the modeling sample set size for learning smaller than C=121, for example, 100 or less, the sample set is referred as insufficient or small. Without changes of the network structure and the parameter or CPT dimension, three resource network models are obtained by the soft-noise simulation [26], with different CPTs. These three models can provide the knowledge as resource domain information for parameter estimation in the target domain.

B. PARAMETER LEARNING WITH DIFFERENT TRANSFER STRATEGIES

In this section, we look into VCTL and compare it to other BN parameter learning algorithms. In general, these learning strategies can be classified into two types: no information transfer methods (MLE and MAP algorithm) and information transfer methods. The latter are further divided into two types: freeze methods (which use the estimated resource parameter

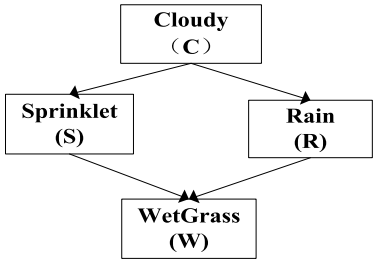


FIGURE 4. Weather BN model.

TABLE 2. Parameter index and ground truth value for node 'W'.

CPT Index	Index Event	Ground Truth
P1	P(W = 1 | S = 1, R = 1)	1.00
P2	P(W = 1 | S = 1, R = 2)	0.10
P3	P(W = 1 | S = 2, R = 1)	0.10
P4	P(W = 1 | S = 2, R = 2)	0.01
P5	P(W = 2 | S = 1, R = 1)	0.00
P6	P(W = 2 | S = 1, R = 2)	0.90
P7	P(W = 2 | S = 2, R = 1)	0.90
P8	P(W = 2 | S = 2, R = 2)	0.99

directly without relying on target samples) and fine-tune methods (which rely on target samples) (using both target and resource samples for modeling). Clearly, LoLP and the suggested VCTL are two strategies for fine-tuning learning.

We compare our technique to the state-of-the-art LoLP [27] using typical BNs (Weather, Asia, and Alarm). The setting is the same as Ref. [27]: only transfers between target and resource nodes with the same node index are allowed. Sample sets are drawn for each reference BN by sampling with the ''sample_bnet'' function [35]. The sample sets are utilized to learn the BN parameters.

The most complicate CPT instance for the Weather BN model is node ''W,'' as depicted in Fig. 4. Table 2 shows the ground truth values for node ''W.'' In the example network, we assign an index to each parameter or CPT for simplicity. The number ''1'' in index event indicates that the event is ''false,'' whereas ''2'' indicates that the event is ''true.''

The simulation technique is as follows: samples are taken from 1 to 1500 instances for each target BN. The first 100 samples are used as target samples for small data set learning in the Weather model, as illustrated in Table 3. Similarly, three groups of 500 different samples can be obtained from various resource networks using the same structures but slightly different CPTs. For parameter learning, resource samples are used to transfer information.

This experiment deals with a target domain data size of 100. This implies that we are dealing with small data modeling because the target data set size is smaller than the threshold C. Eq. (9). For parameter estimation, the MLE and MAP approaches are utilized. In the meantime, LoLP, Freeze, and VCTL are being used to aggregate target and resource

TABLE 3. Samples for 'weather' BN model.

Set #	C	S	R	W
1	1	2	1	2
2	2	1	1	1
3	2	1	2	1
...	...	...	...	...
35	2	1	2	1
...	...	...	...	...
1498	2	1	2	2
1499	2	1	1	1
1500	1	2	1	1

TABLE 4. KLD performance of MLE, MAP, Freeze, LoLP and VCTL. (target sample size: 100).

Method	Weather	Asia	Alarm
MLE	0.21 ± 0.01	0.34 ± 0.23	0.51 ± 0.28
MAP	0.11 ± 0.08	0.14 ± 0.08	0.22 ± 0.11
Freeze	0.05 ± 0.02	0.05 ± 0.01	0.05 ± 0.02
LoLP	0.03 ± 0.01*	0.04 ± 0.02	0.04 ± 0.03
VCTL	0.03 ± 0.01*	0.03 ± 0.01*	0.03 ± 0.02*

TABLE 5. Average running time cost of different algorithms under different models (Ms) (target sample size: 100).

Method	Weather	Asia	Alarm
MLE	0.99 ± 0.17	1.01 ± 0.22	1.11 ± 0.12
MAP	1.03 ± 0.15	1.33 ± 0.11	1.54 ± 0.18
Freeze	1.05 ± 0.15	1.54 ± 0.12	1.57 ± 0.15
LoLP	1.11 ± 0.18	1.62 ± 0.20	1.64 ± 0.23
VCTL	1.31 ± 0.25	1.95 ± 0.24	2.13 ± 0.33

samples in order to estimate parameters. The average KLD between learned CPTs and ground truth CPTs is used to calculate the results.

The average KLD between learned CPTs and ground truth CPTs is used to calculate the results quantitatively. We execute 15 trials using a random data sample in each experiment and present the mean and standard deviation results.

The learning results of different algorithms for different BN models are presented in Table 4 regarding KLDs, with statistically significant improvements for the best result over competitors indicated with asterisks $*$ ($p \leq 0.05$). And the average running time costs (milli second, ms) are listed in Table 5. (The experimental data are available upon request.)

Experimental results indicate that:

(i) When the target sample size for learning is small, as shown in Table 4, compared with methods without information transfer (MLE or MAP), parameter learning using transfer method (Freeze, LoLP or VCTL) provides better reduction of KLD compared to the ground truth. The results demonstrate information transfer can help to improve parameter estimation results.

(ii) Comparison results between freeze method (without target sample information) and fine-tune method (LoLP or VCTL) demonstrate that fine-tune methods do improve performance in target BN parameter learning. In general, VCTL achieves the best learning results with the smallest KLD in almost each experiment by exploiting information potentials from both resource and target data.

(iii) With increasing BN model complexity in node and parameter number as shown in Table 5, the time consumption of nearly every algorithm increases but did not change significantly. The computational complexity of the VCTL algorithm is similar to that of the MAP algorithm without notable time consumption cost.

C. PARAMETER LEARNING WITH DIFFERENT TARGET SAMPLE SIZES

The objective of following experiments is to evaluate how different target sample sizes affect the learning accuracy in transfer learning framework.

Experiment settings:

- Number of resource domain networks M takes 3;
- Resource network data sizes are 500 sets in this experiment;
- Tr takes 5;
- Size of target domain samples in the experiment is set to a small data set (from 1 to 50 groups) at first, and then the sample size of the target domain is gradually increased to 1500.

Given the Weather model, Fig. 5 shows the KL divergence comparisons of different techniques with a sample size of 1∼1500 groups. The logarithm of KL divergence, decibel value (dB), is utilized for display simplicity. The average of all outcome values is calculated over 15 repetitions. (Data can also be obtained upon request.)

On the x-axis, we progressively selected a larger number of target sample sizes from left-to-right. Different methods, each using independent curves, are used to investigate the KL divergence comparison of parameter learning.

When the target domain sample size is from 1 to 50, Fig. 6 shows box-plots of MLE and MAP parameter learning algorithms without information transfer on KLD. Fig. 7 shows box-plots of Freeze, VoLP and VCTL parameter learning algorithms with resource information transfer on KLD.

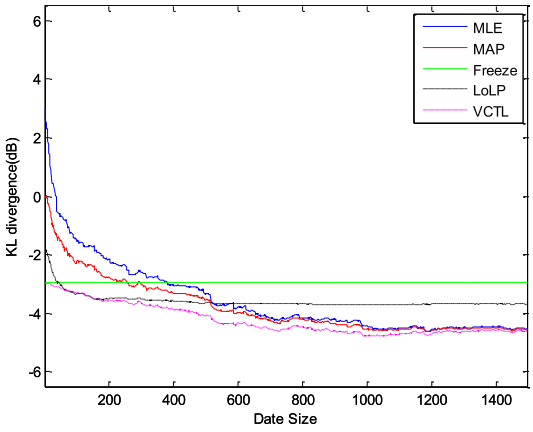
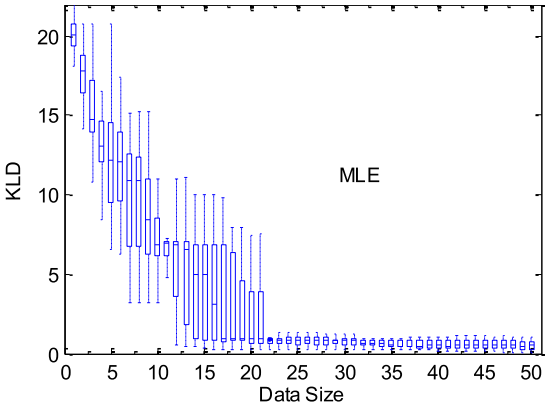
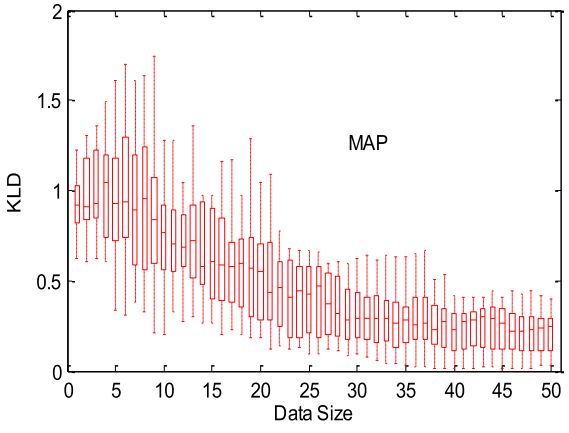


FIGURE 5. KL divergence comparisons of different methods of node 'W' under the sample size of 1∼1500 group.



(a) MLE learning results

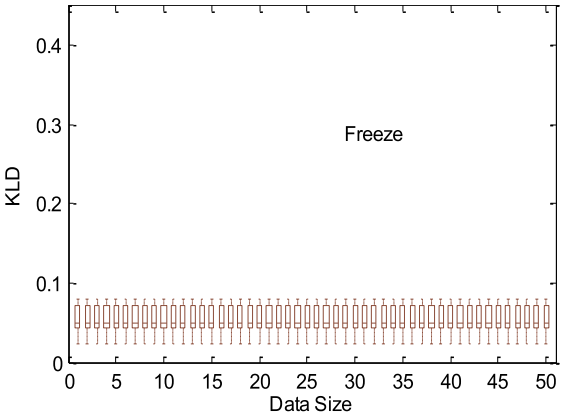


(b) MAP learning results

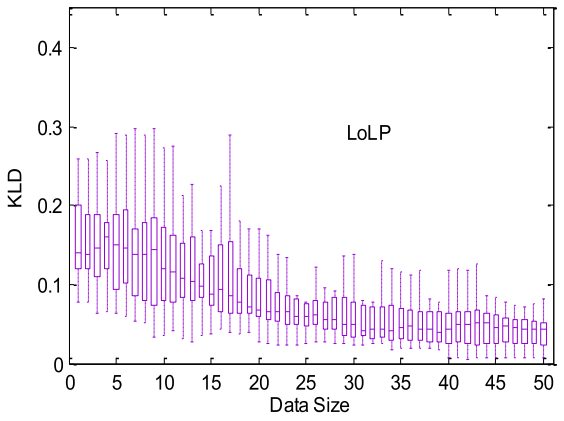
FIGURE 6. Comparison of learning results without information transfer learning methods under sample size from 1 to 50.

Experimental results indicate that:

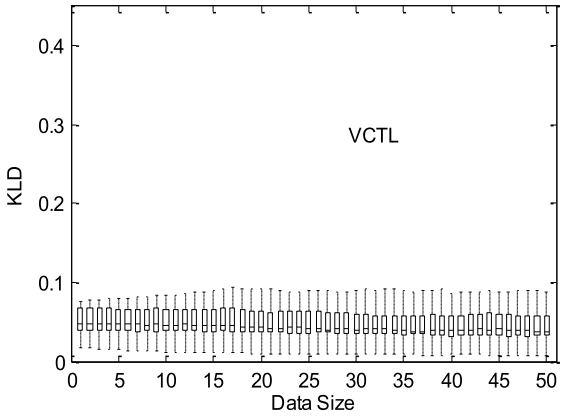
(i) In general, as indicated by Fig. 5, the more sample data sizes (on the x-axis) we can obtain, the better parameter estimation accuracy (on the y-axis) we can derive using these BN parameter learning methods;



(a) Freeze learning results



(b) LoLP learning results



(c) VCTL learning results

FIGURE 7. Comparison of learning results with information transfer learning methods under sample size from 1 to 50.

(ii) When the sample size is small (e.g., less than 44 in Fig. 6 and Fig. 7), thanks to the information transfer, Freeze or LoLP is superior to MLE or MAP. And VCTL provides the best KLD performance;

(iii) When the target sample size for modeling is sufficient, e.g., under the data size of 1500 in Fig. 5, the VCTL method tends to perform the well as the classical MLE or MAP method does in KLD. However, as shown in Fig. 5 the Freeze or LoLP method didn't provide the dramatic improvement in KLD with increasing target data. This suggests that VCTL can perform well as the classical MLE or MAP approach in this situation.

The reason is that: Freeze method results strongly depend on the prior knowledge learned from the resource domain only; LoLP performs better than Freeze when sample data size is bigger and bigger (for instance, 44 in this example). VCTL provides the best KLD results benefit from both varying balance coefficients regarding the varying contribution of target samples and the available information among the resource sample set. And as the target data size of instances increases, the VCTL learning accuracy improves by using more and more available observed data.

In general, under the condition of small data size, every transfer learning method is better than using the small target data only. Our VCTL significantly almost outperforms the alternatives on parameter estimation accuracy in each case.

D. LEARNING WITH DIFFERENT RESOURCE PARAMETER NOISE

In practice, parameter learning transfer from the irrelevant domain knowledge is sometimes inaccurate. Therefore, we considered CPT noises (bigger noise means less relevant) in the learning experiment to test the robustness of the estimators.

In this case, we use "soft noise" to simulate continuously varying relatedness among a set of resources [26]. The specific soft noise simulation procedure is as follows:

(1) the sizes of the data sets for target networks were set to be 15, 55 and 75;

(2) to simulate the irrelevant samples, the sample data were generated from the known true parameters of the networks. Because domain knowledge with significant noise could not be taken into account, only small noises for resource CPTs, varying from 5%, 10%, to 15%, were incorporated into the parameter learning experiments;

(3) because the resource networks are learned from varying samples, they will vary in degree of relatedness to the target, with noise from 5% to 15% under 500 samples in each resource network;

(4) in each experiment, we run 15 trials with random data samples and report the mean and standard deviation of the KLD.

The learned results under different noise conditions are summarized in Table 6, Table 7 and Table 8. The best results are highlighted with asterisks.

Note that:

(1) for the algorithms introduced transfer, such as Freeze, LoTL, and VCTL, resource noises negatively affected performance on parameter estimation. VCTL performs

TABLE 6. Average KLD under different resource CPT noises (target sample size: 15).

Resource CPT Noises	Freeze	LoLP	VCTL
(a) Weather Network			
5%	0.04 ± 0.02	0.12 ± 0.05	0.04 ± 0.01*
10%	0.07 ± 0.03	0.12 ± 0.05	0.06 ± 0.03*
15%	0.10 ± 0.04	0.15 ± 0.05	0.10 ± 0.04*
(b) Asia Network			
5%	0.04 ± 0.02	0.12 ± 0.04	0.04 ± 0.02*
10%	0.07 ± 0.02	0.13 ± 0.05	0.06 ± 0.03*
15%	0.10 ± 0.04	0.15 ± 0.05	0.10 ± 0.04*
(c) Alarm Network			
5%	0.05 ± 0.02	0.15 ± 0.05	0.04 ± 0.02*
10%	0.07 ± 0.03	0.15 ± 0.06	0.06 ± 0.03*
15%	0.12 ± 0.04	0.17 ± 0.07	0.10 ± 0.04*

TABLE 7. Average KLD under different resource CPT noises (target sample size: 55).

Resource CPT noises	Freeze	LoLP	VCTL
(a) Weather Network			
5%	0.04 ± 0.02	0.11 ± 0.04	0.04 ± 0.02*
10%	0.06 ± 0.01*	0.11 ± 0.05	0.05 ± 0.01*
15%	0.10 ± 0.04	0.08 ± 0.04	0.08 ± 0.03*
(b) Asia Network			
5%	0.04 ± 0.02	0.11 ± 0.04	0.04 ± 0.02*
10%	0.07 ± 0.01*	0.12 ± 0.04	0.05 ± 0.01*
15%	0.11 ± 0.04	0.09 ± 0.05	0.08 ± 0.03*
(c) Alarm Network			
5%	0.04 ± 0.01*	0.13 ± 0.04	0.04 ± 0.02
10%	0.08 ± 0.02*	0.17 ± 0.05	0.06 ± 0.02*
15%	0.12 ± 0.05	0.10 ± 0.04	0.09 ± 0.04*

the best with the smallest KLD in almost each noise experiment.

(2) as to the noises and the same sample size, the smaller networks (such as Weather network) was more robust than that of bigger networks (such as the Asia and Alarm) with smaller KLDs. The reason lies in that the smaller networks need fewer samples for modeling.

(3) when resource parameter noise increases, the target learning KLD values increase. Under the condition of target sample size is scarce (e.g., 15 in the experiments in Table 6), when the resource CPT noises is 15%, learned KLD errors are over 0.1 in almost each case, which may not be acceptable for domain experts. Therefore, before applying the transfer learning method for parameter estimation, it is advisable to verify and validate the parameter noise or domain knowledge.

TABLE 8. Averaged KLD under different resource CPT noises (target sample size: 75).

Resource CPT Noises	Freeze	LoLP	VCTL
(a) Weather Network			
5%	0.04 ± 0.01	0.04 ± 0.02	0.03 ± 0.01*
10%	0.06 ± 0.03	0.06 ± 0.02	0.04 ± 0.02*
15%	0.09 ± 0.03	0.07 ± 0.05	0.05 ± 0.02*
(b) Asia Network			
5%	0.04 ± 0.01	0.04 ± 0.03	0.03 ± 0.02*
10%	0.06 ± 0.04	0.06 ± 0.04	0.05 ± 0.02*
15%	0.11 ± 0.04	0.07 ± 0.05	0.07 ± 0.03*
(c) Alarm Network			
5%	0.04 ± 0.02	0.04 ± 0.04	0.03 ± 0.02*
10%	0.07 ± 0.03	0.06 ± 0.03	0.05 ± 0.02*
15%	0.09 ± 0.04	0.08 ± 0.04	0.08 ± 0.03*

In general, given samples with noise in resource domain, VCTL can exploit the most relevant resources by varying coefficients in the aggregation model. VCTL algorithm makes up the defection that classical transfer learning algorithm without considering the relevance fitness and the varying contribution of the sample data. The varying coefficients in VCTL improves the learning accuracy for BN parameters dynamically and automatically.

VI. FAULT DIAGNOSIS FOR BEARING USING TRANSFER LEARNING AGGREGATION

In this section, we explore the proposed bearing fault algorithm based on BN and transfer learning aggregation.

A. EXPERIMENTAL SETTINGS

The data for the experiments came from the Bearing Data Center at Case Western Reserve University [6]. Electro-discharge machining was used to seed defects in motor bearings. At the inner raceway, rolling element (i.e. ball), and outer raceway, faults ranging in diameter from 7 mils (1 mil=0.001 inches) to 40 mils were introduced separately. The test motor's faulty bearings were reinstalled, and vibration data was taken for motor loads ranging from 0 to 3 horsepower. To acquire the data sets, destructive tests were used to provide working bearing data for s=1,2,3,4 typical fault types (normal-NF, inner fault-IR, outer fault-OR, and rolling fault-BF).

At 3 o'clock, 6 o'clock, and 12 o'clock on the driving end of the motor shell, acceleration sensors are mounted. Table 9 shows the 12k driving end bearing fault data used in this experiment. Types of working bearing data (normal, inner fault, and rolling fault) are all complete. However, for unknown reasons, there are missing data in the case of 14 mil for Outer 3 or 12 o'clock data, which is highlighted in bold in Table 9. To the best of the authors' knowledge, no transfer

TABLE 9. 12k drive end bearing fault data (Load:2 HP; motor speed:1750 RRM).

Fault Diameter (mil)	Outer 6 O 'clock	Outer 3(12) O 'clock	Other
7	Complete	Complete	Complete
14	Complete	**Missing**	Complete
21	Complete	Complete	Complete

TABLE 10. Bearing details (6205-2RS-JEM SKF).

Parameters	Values (mm)
Inner Race Diameter	25.001
Outer Race Diameter	51.998
Thickness	15.001
Ball Diameter	7.940
Pitch Diameter	39.039

BN learning for the 14 million example has been documented. For fault diagnosis, we aim to model the parameters for the 14 mil case using VCTL and BN knowledge transfer from the 7 mil and 21 mil examples. The fault diagnosis belief threshold parameter $\theta*$ takes 0.75.

The following experiments used the same equipment and the same motor speed and load to obtain fault data. As a result, the BN model structure for bearing fault diagnosis with a fault diameter of 7mil or 21mil is the same as the BN model structure for bearing fault diagnosis with a fault diameter of 14mil. Using the 14 mil instance as the target domain, we can provide resource models of 7 mil and 21 mil.

The 6205-2RS-JEM SKF deep-groove ball bearing (BB) from CWRU Bearing Data Center is listed in Table 10.

For drive end bearing issues, vibration data is collected using accelerometers under four bearing working conditions: NF, IR, OR, and BF, at 12,000 and 48,000 samples per second.

When a bearing fault occurs, vibration at a specific frequency are created. These faults characteristic frequencies are calculated as

$$\text{OR fault f}_0 = \frac{Nb}{2}F_r\left(1-\frac{d}{Nb}cos\alpha\right) \tag{18}$$

$$\text{IR fault } f_i = \frac{Nb}{2}F_r\left(1+\frac{\text{d}}{\text{D}}\text{cos}\alpha\right) \tag{19}$$

$$\text{BF fault } f_b = \frac{D}{2d}F_r\left(1-\left[\frac{\text{d}}{\text{D}}\right]^2\cos^2\alpha\right) \tag{20}$$

where Nb is the number of balls, Fr is the rotational speed of the rotor in hertz, d is the diameter of the ball, D is the pitch diameter, and α is the contact angle [19].

B. FAULT DIAGNOSIS FEATURES EXTRACTION FOR BN MODEL

Based on 12k Drive End Bearing Fault Data, Fig. 8 illustrates the bearing vibration signals in time domain for the normal, OR, IR, and BF under the condition of Table 10 for 7mil faults at 12,000 samples per second.

Set s={1,2,3,4}for the fault type "Bearing," N takes 1024, and r takes 3. By the presented method in Section IV, we get characteristic data. Given 1024 sampling data in time domain,

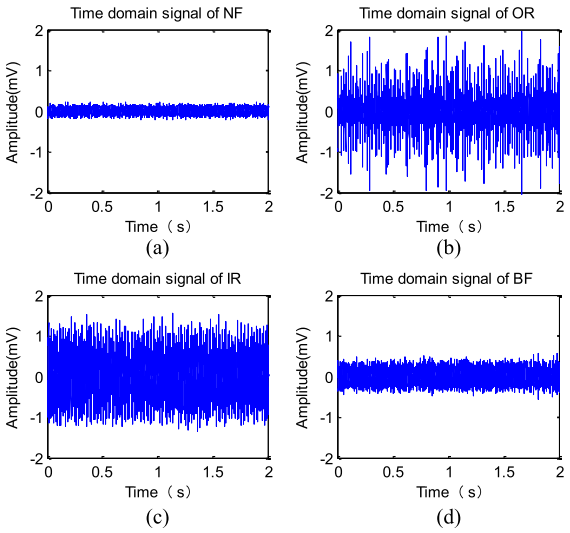


FIGURE 8. Bearing vibration signals in time domain for normal, OR, IR, and BF faults.

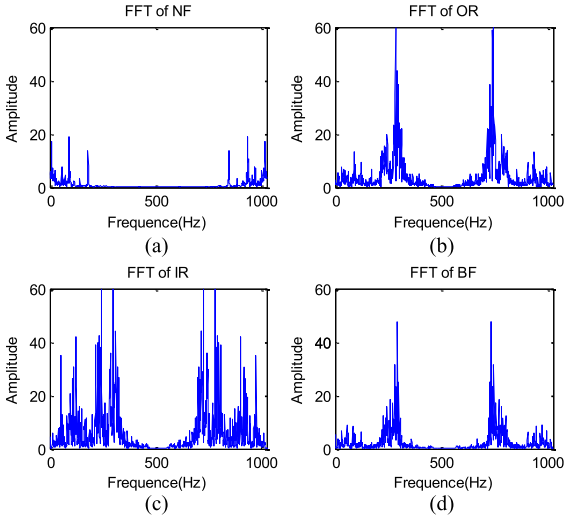


FIGURE 9. Instances of FFT results for normal, OR, IR, and BF faults (12k drive end bearing fault data).

Fig. 9 shows FCV instances by FFT in case of normal(s=1), OR(s=2), IR(s=3), and BF(s=4) respectively for 7mil (Drive End). Notice that both of the values of time domain and frequency domain data are continuous originally. To speed up BN modeling and inference, FCVs are discretized as attribute values. Fig. 10 indicates instances of bearing DFCV features for normal, OR, IR, and BF faults for $W_{s,80,u}$, where x-axis displays the value of characteristic vector component u from 1 to 8. Meanwhile, y-axis indicates discrete energy value from 1 to 3 which represents low, medium or high energy respectively.

In Table 11, the continuous characteristic data, FCVs, are listed, and the discrete characteristic data, DFCV, are summarized in Table 12 for 7mil bearing data cases.

C. MODELING FAULT DIAGNOSIS BN

Using the method of Step 6 in the proposed bearing fault algorithm based on BN and transfer learning aggregation,

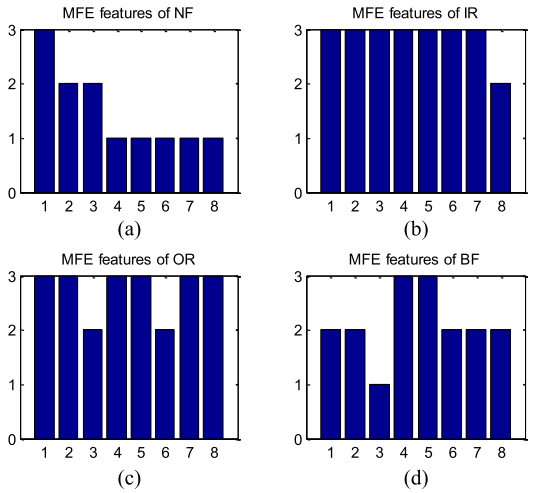


FIGURE 10. Instances of bearing DFCV features for normal, OR, IR, and BF faults.

TABLE 11. Continuous characteristic data (7Mil).

Fault Type	FCV u	Data Set Number (mv^2) 1	2	...	299	300
	1	336.3951	340.4934	...	330.3798	337.2891
	2	427.9363	424.9339	...	414.2499	424.2730
NF	3	440.2578	441. 0432	...	449.1633	422.7192
	...	...	...	...	...	...
	8	453.6609	449.8030	...	439.0764	453.9073
	1	653.2046	640.1032	...	639.3412	672.5022
	2	857.0047	968.9828	...	962.4355	971.4146
IR	3	1997.1693	1938. 8332	...	2015.7291	2079.7180
	...	...	...	...	...	...
	8	2098.2113	2044.9178	...	2134.7636	2179.3767
	1	358.9238	372.9187	...	351.6619	327. 0448
	2	1168.6855	1198.1725	...	1225.9335	1611.3843
OR	3	3605.8638	3418.5993	...	3872.5855	3999.5472
	...	...	...	...	...	...
	8	3941.5447	3854.0236	...	4258.5053	4298.1662
	1	238.7874	230.1268	...	218.6458	226.8484
	2	322.6112	298.6425	...	323.9650	323.4575
BB	3	906.7253	833.9853	...	965.0360	908.3716
	...	...	...	...	...	...
	8	959.4030	869. 6977	...	988. 4934	944.4238

Fig. 11 shows the topology of a Bayesian network fault diagnosis model with 7 mil fault sizes (resource BN 1). Similarly, data for the fault characteristic vector with fault diameters of 21 mil (resource BN 2) and 14 mil (target BN) can be derived, and bearing fault diagnosis BN model structures with fault diameters of 21 mil and 14 mil are shown in Fig. 12 and Fig. 13, respectively. The child nodes are eight feature vectors, and the parent node "Bearing" is separated into four states (normal state, inner fault, outer fault, and rolling fault). The 8 eigenvectors S10∼S17 are the child nodes of parent node Bearing_07 (7 mil). Meanwhile, T0∼T7 are eigenvectors in the child nodes of node Bearing_14 (14 mil). Each eigenvector uses 1, 2 and 3 to represent low, medium or high energy respectively.

Clearly, the resource networks and the target network meet the structural consistency (same BN structure) and parameter dimension consistency (same CPT entry)constraints for transferring learning. As a result, 7 mil and 21 mil data can be used as resource domain data, and transfer learning can be

TABLE 12. Discrete characteristic data (7 Mil).

Fault Type	FCV u	Data Set Number Set 1	2	...	299	300
	1	1	1	...	1	1
	2	2	2	...	2	2
NF	3	1	1	...	1	1
	...	...	...	...	...	...
	8	2	1	...	2	2
	1	3	3	...	3	3
	2	2	2	...	2	2
IR	3	2	2	...	2	2
	...	...	...	...	...	...
	8	2	2	...	2	2
	1	2	2	...	2	2
	2	2	3	...	3	3
OR	3	3	3	...	3	3
	...	...	...	...	...	...
	8	3	3	...	3	3
	1	1	1	...	1	1
	2	2	2	...	1	2
BB	3	1	1	...	1	1
	...	...	...	...	...	...
	8	1	1	...	1	1

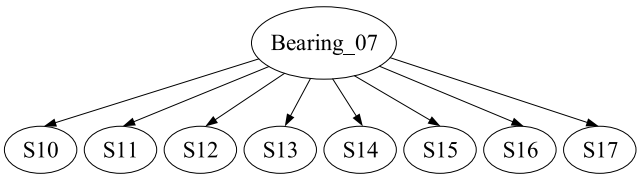


FIGURE 11. Bearing fault diagnosis model for 7 mil (resource BN 1).

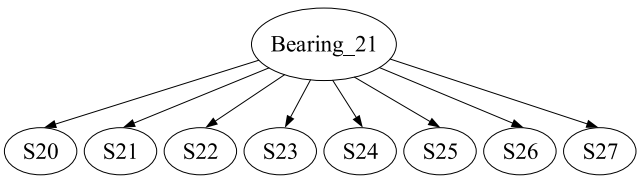


FIGURE 12. Bearing fault diagnosis model for 21 mil (resource BN 2).

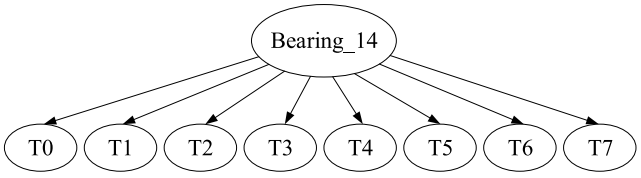


FIGURE 13. Bearing fault diagnosis model for 14 mil (target BN).

used to create a 14 mil diagnosis model in the target domain with missing data.

In this experiment, λ takes 1, d is 1, N is 9, δ is 0.05 and ε takes 0.1. Then, (16) determines the threshold value of small data set C is 209. The data size of 7 mil is 300, and for 21 mil, the data size is also 300. Let the data size of 14 mil takes 60. Since 60<C, the modeling for 14 mil fault diagnosis is under the condition of a small target sample data set.

As to the target domain (14 mil cases), we learn the target network in two scenarios: one is learning using

TABLE 13. Fault diagnosis results under modeling of small target sample data sets by VCTL (size: 60).

Method	Normal	Inner	Outer	Rolling	Average
LoLP	100%	78.72%	82.33%	73.31%	83.59%
VCTL	100%	93.96%	86.41%	80.81%	90.30%

TABLE 14. Fault diagnosis results under modeling of sufficient target sample data set (size: 300).

Normal	Inner	Outer	Rolling	Average
100%	95.78%	88.73%	83.32%	91.96%

first 60 group (small) data sets, another is learning using 300 group (sufficient) data sets for each fault type respectively. In the context of two 300 groups of resource network data sets (7 mil and 21 mil cases), this study employs the VCTL technique to learn target BN parameters (14 mil cases) from resource models (7 mil and 21 mil cases). Then, we conduct inference and diagnosis tests in the target domain employing distinct 115 groups of feature data sets to verify the validity of our presented bearing fault diagnosis BN approach with transfer learning for parameter estimation.

D. FAULT DIAGNOSIS RESULT

The target model for 14 mils is constructed using the provided VCTL, and the BN model with 115 target sample groups is used for inference verification. The diagnosis accuracy is reported in Table 13 utilizing the suggested VCTL aggregation method and the classic Junction Tree inference algorithm [29]. When 300 target samples are utilized for parameter learning, the results are shown in Table 14.

Table 13 demonstrates that when the target network sample data set size is small, the VCTL algorithm improves average diagnosis accuracy by 6.71% when compared to the standard LoLP approach. The reasons for the lowest diagnostic accuracy of the fault type ''Rolling'' lie that the evidence for rolling is quite close to evidence for inner fault, when the given samples for modeling are small. Generally, experimental results suggest that VCTL-based inference diagnosis is superior than the LoLP technique. With more target sample data utilized in BN modeling, the VCTL algorithm can slightly enhance fault diagnostic accuracy, as shown in Table 14.

The experimental results show that, in the case of a small data set, employing transfer learning aggregation approach, the BN parameter estimation can be modified from resource networks by VCTL. The proposed bearing fault diagnosis based on VCTL can yield reasonable results by transferring information from the resource domain.

VII. CONCLUSION

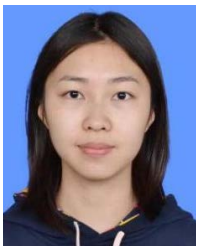
To improve the BN parameter estimation accuracy with small data size, this study proposes a BN parameter learning algorithm based on varying coefficient transfer learning. The varying contributions from resource networks and the target network are considered in a target BN parameter aggregation model. Based on the sample statistics information and the presented VCTL parameter estimation algorithm, the presented variable coefficients balance the initial parameter influence between the target domain and the resource domain by using the knowledge of the resource domain and the small data set threshold. Furthermore, a fault diagnosis for bearing utilizing transfer learning aggregation method is advanced. Experimental results show that, with small data set, VCTL method provides better learning accuracy than MLE, MAP and classical transfer learning methods. The application of real fault diagnosis modeling verifies the transfer learning aggregation method's reliability and effectiveness. The presented approach provides a novel way for intelligent system modeling, especially when the data size for BN parameter modeling is small.

REFERENCES

[1] X. Li, Y. Hu, M. Li, and J. Zheng, ''Fault diagnostics between different type of components: A transfer learning approach,'' *Appl. Soft Comput.*, vol. 86, pp. 105950–105961, Jan. 2020.
[2] C. Wang, Y. Xie, and D. Zhang, ''Deep learning for bearing fault diagnosis under different working loads and non-fault location point,'' *J. Low Freq. Noise, Vib. Act. Control*, vol. 40, no. 1, pp. 588–600, Mar. 2021.
[3] C. Che, H. Wang, Q. Fu, and X. Ni, ''Deep transfer learning for rolling bearing fault diagnosis under variable operating conditions,'' *Adv. Mech. Eng.*, vol. 11, no. 12, pp. 1–11, 2019.
[4] R. Liu, B. Yang, E. Zio, and X. Chen, ''Artificial intelligence for fault diagnosis of rotating machinery: A review,'' *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Feb. 2018.
[5] Y. Wei, Y. Li, M. Xu, and W. Huang, ''A review of early fault diagnosis approaches and their applications in rotating machinery,'' *Entropy*, vol. 21, no. 4, pp. 409–415, 2019.
[6] Y. Li, X. Wang, S. Si, and S. Huang, ''Entropy based fault classification using the case western reserve university data: A benchmark study,'' *IEEE Trans. Rel.*, vol. 69, no. 2, pp. 754–767, Jun. 2020.
[7] Y. Li, Y. Yang, X. Wang, B. Liu, and X. Liang, ''Early fault diagnosis of rolling bearings based on hierarchical symbol dynamic entropy and binary tree support vector machine,'' *J. Sound Vib.*, vol. 428, pp. 72–86, Aug. 2018.
[8] R. Yang, M. Huang, Q. Lu, and M. Zhong, ''Rotating machinery fault diagnosis using long-short-term memory recurrent neural network,'' *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 228–232, 2018.
[9] Z. Wu, H. Jiang, K. Zhao, and X. Li, ''An adaptive deep transfer learning method for bearing fault diagnosis,'' *Measurement*, vol. 151, pp. 107227–107241, Feb. 2020.
[10] H. Shao, H. Jiang, F. Wang, and Y. Wang, ''Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet,'' *ISA Trans.*, vol. 69, pp. 187–201, Jul. 2017.
[11] X. Zhou, X. Zhang, W. Zhang, and X. Xia, ''Fault diagnosis of rolling bearing under fluctuating speed and variable load based on TCO spectrum and stacking auto-encoder,'' *Measurement*, vol. 138, no. 2, pp. 163–174, 2019.
[12] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, ''A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load,'' *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.
[13] J. Li, P. Du, A. Y. Ye, Y. Zhang, C. Song, H. Zeng, and C. Chen, ''GPA: A microbial genetic polymorphisms assignments tool in metagenomic analysis by Bayesian estimation,'' *Genomics, Proteomics Bioinf.*, vol. 17, no. 1, pp. 106–117, Feb. 2019.
[14] S. Kabir, M. Walker, and Y. Papadopoulos, ''Dynamic system safety analysis in HiP-HOPS with Petri nets and Bayesian networks,'' *Saf. Sci.*, vol. 105, pp. 55–70, Jun. 2018.
[15] Y. Hou, E. Zheng, W. Guo, Q. Xiao, and Z. Xu, ''Learning Bayesian network parameters with small data set: A parameter extension under constraints method,'' *IEEE Access*, vol. 8, pp. 24979–24989, 2020.
[16] Z. Hao, Z. Xu, H. Zhao, and H. Fujita, ''A dynamic weight determination approach based on the intuitionistic fuzzy Bayesian network and its application to emergency decision making,'' *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 1893–1907, Aug. 2018.

[17] S. Mascaro, A. E. Nicholso, and K. B. Korb, "Anomaly detection in vessel tracks using Bayesian networks," *Int. J. Approx. Reasoning*, vol. 55, no. 1, pp. 84–98, Jan. 2014.
[18] N. A. Zaidi, G. I. Webb, M. J. Carman, F. Petitjean, W. Buntine, M. Hynes, and H. De Sterck, "Efficient parameter learning of Bayesian network classifiers," *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1289–1329, Oct. 2017.
[19] W. Guo, Q. Zhou, and Y. Hou, "Early classification of bearing faults based on MFES and Bayesian network," *Int. J. Digit. Content Technol. Appl.*, vol. 7, no. 11, pp. 116–126, Jul. 2013.
[20] Z. Zhou, E. Y. Lam, and C. Lee, "Nonlocal means filtering based speckle removal utilizing the maximum *a posteriori* estimation and the total variation image prior," *IEEE Access*, vol. 7, pp. 99231–99243, 2019.
[21] Y. Zhai, H. Cao, W. Deng, J. Gan, V. Piuri, and J. Zeng, "BeautyNet: Joint multiscale CNN and transfer learning method for unconstrained facial beauty prediction," *Comput. Intell. Neurosci.*, vol. 2019, no. 4, pp. 1–14, Jan. 2019.
[22] P. Yuan, Y. Sun, H. Li, F. Wang, and H. Li, "Abnormal condition identification modeling method based on Bayesian network parameters transfer learning for the electro-fused magnesia smelting process," *IEEE Access*, vol. 7, pp. 149764–149775, 2019.
[23] X. Liao, X. Wu, and L. Gui, "A cross-domain emotional classification that represents learning and migration learning," *J. Peking Univ.*, vol. 55, no. 1, pp. 37–46, 2019.
[24] D. Oyen and T. Lane, "Transfer learning for Bayesian discovery of multiple Bayesian networks," *Knowl. Inf. Syst.*, vol. 43, no. 1, pp. 1–28, Apr. 2015.
[25] A. L. P. Chen, J.-S. Chiu, and F. S. C. Tseng, "Evaluating aggregate operations over imprecise data," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 2, pp. 273–284, Apr. 1996.
[26] Y. Zhou, T. M. Hospedales, and N. Fenton, "When and where to transfer for Bayes net parameter learning," *Expert Syst. Appl.*, vol. 55, pp. 361–373, Aug. 2016.
[27] R. Luis, L. E. Sucar, and E. F. Morales, "Inductive transfer for learning Bayesian networks," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 227–255, May 2009.
[28] W. Guo, Z. Wen, Z. Guo, C. Xu, Q. Xiao, and L. Mao, "Varying balancing transfer learning for BN parameter estimation," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Hefei, China, Aug. 2020, pp. 2179–2184.
[29] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge, MA, USA: Cambridge Univ. Press, 2009.
[30] Y. Zhou, N. Fenton, and M. Neil, "Bayesian network approach to multinomial parameter learning using data and expert judgments," *Int. J. Approx. Reasoning*, vol. 55, no. 5, pp. 1252–1268, Jul. 2014.
[31] X.-G. Gao, Z.-G. Guo, H. Ren, Y. Yang, D.-Q. Chen, and C.-C. He, "Learning Bayesian network parameters via minimax algorithm," *Int. J. Approx. Reasoning*, vol. 108, pp. 62–75, May 2019.
[32] N. Dojer, "Learning Bayesian networks from datasets joining continuous and discrete variables," *Int. J. Approx. Reasoning*, vol. 78, pp. 116–124, Nov. 2016.
[33] D. Koller, *Probabilistic Graphical Models: Principles and Techniques-Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.
[34] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[35] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

YONGYAN HOU (Member, IEEE) is currently pursuing the Ph.D. degree. She is also an Associate Professor with the School of Electrical and Control Engineering, Shaanxi University of Science and Technology. Her research interests include Bayesian methods for data-driven modeling, prediction, risk management, and decision making. She applies these techniques to a wide range of real-world problems, for both academic research and industrial clients.



AO YANG received the B.E. degree from the School of Information Engineering and Artificial Intelligence, Shaanxi University of Science and Technology, in 2021. Her research interest includes probabilistic graphical models and applications.

WENQIANG GUO (Member, IEEE) received the Ph.D. degree in system engineering from Northwestern Polytechnical University, China, in 2011. He is currently a Professor with the School of Information Engineering and Artificial Intelligence, Shaanxi University of Science and Technology. His research interests include intelligent decision-making support systems, machine learning, and Bayesian network and its application. He has published more than 30 refereed articles in journals and conferences in these areas.

ENRANG ZHENG is currently a Professor and a Ph.D. Supervisor with the School of Electrical and Control Engineering. He is also currently a President Assistant with the Shaanxi University of Science and Technology. He has published two books and more than 100 refereed articles and has provided consulting to many major companies worldwide. His research interests include intelligent systems, machine learning, and data mining.

QINKUN XIAO (Member, IEEE) received the Ph.D. degree in system engineering from Northwestern Polytechnical University, in 2007. He is currently a Professor and a Ph.D. Supervisor with the College of Electronic Information Engineering, Xi'an Technological University. His research interests include probabilistic graphical models, machine learning, and data mining. He has published more than 80 refereed articles in journals and conferences in these areas.

ZHIGAO GUO received the Ph.D. degree in system engineering from Northwestern Polytechnical University, China, in 2019. He is currently a Postdoctoral Research Associate with Queen Mary University, London. His research interests include probabilistic graphical models and causal inference.

ZIXUAN HUANG received the B.Eng. degree from the School of Information Engineering and Artificial Intelligence, Shaanxi University of Science and Technology, in 2020. His research interests include Bayesian facial pain expression recognition, unsupervised learning, and reinforcement learning.