Transfer Learning-Based Power System Online Dynamic Security Assessment: Using One Model to Assess Many Unlearned Faults

Chao Ren, *Student Member, IEEE*, and Yan Xu, *Senior Member, IEEE*

***Abstract*—This letter proposes a novel data-driven method for pre-fault dynamic security assessment (DSA) of power systems. To address the large number of potential faults, the proposed method aims to use one trained model to work for multiple faults. Firstly, a hybrid learning based DSA model is initially trained by one fault database. Then, based on transfer learning technique, the model is transferred to an unknown different but related fault by iteratively minimize the marginal and conditional distribution differences between the trained data and unknown data. Thus, the extensibility of the DSA model is greatly enhanced and the need for training a large number of models is eliminated. Test results have demonstrated the high accuracy of the proposed method.**

***Index Terms*—Data-driven, dynamic security assessment, hybrid ensemble model, transfer learning.**

I. Introduction

INTELLIGENT data-driven method is promising to achieve online DSA [1]. Its principle is to train a learning model (e.g., a classifier or predictor) from a DSA database. Once well trained, the model can be applied for online DSA with real-time measurements. Its advantages are much faster DSA speed, less data requirement, and stronger generalization capability.

For pre-fault DSA, different DSA models are trained for different potential faults, and each DSA model is applied for a specific fault [2]. It is generally assumed that the training data (i.e., DSA database) and the unknown data (i.e., online measurement) follow the same distribution, so that the model can achieve a satisfactory application accuracy. However, this assumption may not always hold, especially when the power system experiences significant unexpected changes and/or unlearned faults, e.g., the DSA training database only covers a certain number of faults, but at the online application stage, a new fault that is not in the database may appear. In this case, the existing DSA models should also be able to assess such unlearned faults. Since there are a huge number of potential faults in a practical power system, such extensibility is pressingly needed, yet not addressed in the literature.

Manuscript received May 12, 2019; revised August 15, 2019; accepted October 10, 2019. Date of publication October 16, 2019; date of current version January 7, 2020. This work was supported in part by the Ministry of Education (MOE), Republic of Singapore, under Grant 2019-T1-001-069 (RG75/19). The work of Y. Xu was supported by Nanyang Assistant Professorship from Nanyang Technological University, Singapore. Paper no. PESL-00110-2019. *(Corresponding author: Yan Xu.)*

The authors are with Nanyang Technological University, Singapore 639788 (e-mail: renc0003@e.ntu.edu.sg; eeyanxu@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2019.2947781

TABLE I
NOMENCLATURE

Notation	Description	Notation	Description
D_s	Source domain	$\mathbf{X}$	Input matrix
D_t	Target domain	$\mathbf{X_s}$	Source domain matrix
n_s	Number of source samples	$\mathbf{X_t}$	Target domain matrix
n_t	Number of target samples	$\mathbf{Y_s}$	Source domain label
x	Single sample	$\mathbf{K}$	Adaptation matrix
m	Number of features	$\mathbf{G}$	Common feature space matrix
tr(·)	Trace of a matrix	$\mathbf{H}$	Centering matrix
λ	Regularization parameter	$\mathbf{R}$	MMD matrix
c	Classes Condition	d	Dimensionality of the samples
u	Number of ELM or RVFL	J	Total MMD

This letter aims to utilize one trained DSA model to assess unlearned faults which are not trained before. Based on the transfer learning theory [3], the DSA model trained for one fault (source domain) is adapted to a different but related fault (target domain) by reducing the distribution differences between the trained data and unknown data. In this way, the extensibility of the DSA models can be greatly enhanced, since a small number of DSA models can cover a large number of fault scenarios. Moreover, based on the transfer performance, the correlation between different faults can be revealed, thus different faults can be aggregated as one.

II. Proposed Methodology

The proposed method is based on transfer learning theory. The domain data consists of a feature space of input $\mathbf{X}$ and a marginal probability distribution $P(\mathbf{X})$. We denote the known fault data in the source domain as $D_s = \{(\mathbf{x}_{s1}, y_{s1}), \ldots, (\mathbf{x}_{sn}, y_{sn})\}$, where the input $\mathbf{X}_s$ are operating variables of the system (e.g., power generation/load, and bus voltage magnitude.) and y is the corresponding class label (stable or unstable). Similarly, the unlearned fault data in the target domain is denoted as $D_t = \{\mathbf{x_{t1}}, \ldots, \mathbf{x_{tn}},\}$, where $\mathbf{X}_t$ is the input without label. The training data D_s are labeled, while the testing data D_t are unlabeled. The nomenclature is given in Table I.

In practice, for an unlearned fault, its data space, the marginal distribution and conditional distribution are different from the trained fault, that is, $\mathbf{X}_\mathrm{s} \neq \mathbf{X}_\mathrm{t}$, $P(\mathbf{X}_\mathrm{s}) \neq P(\mathbf{X}_\mathrm{t})$ and $P(\mathbf{Y}_\mathrm{s}|\mathbf{X}_\mathrm{s}) \neq P(\mathbf{Y}_\mathrm{t}|\mathbf{X}_\mathrm{t})$. The purpose of proposed method is to utilize the DSA model trained by one fault for other faults that are not learned by the model before.

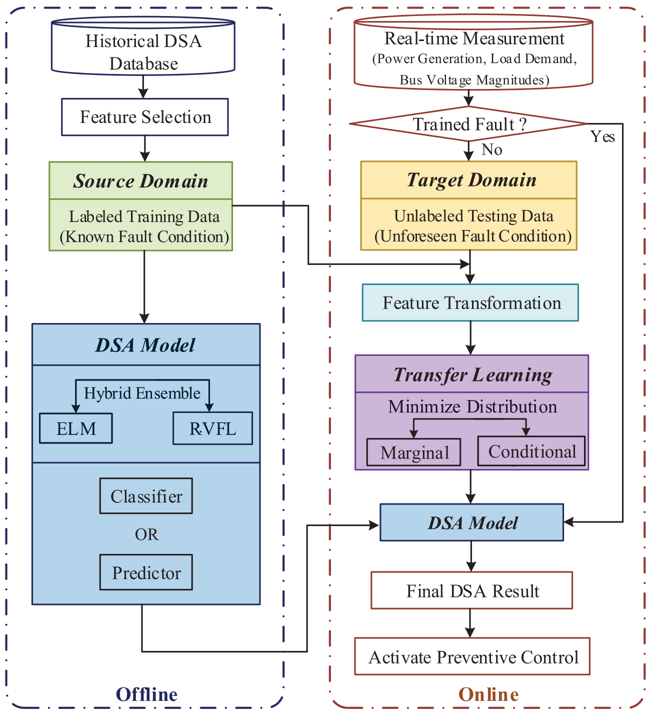


Fig. 1. Framework of the proposed method.

A. General Framework

The framework of the proposed method is illustrated in Fig. 1. At the offline training stage, a DSA model is trained by a fault database (source domain). The feature inputs to the model are P/Q power generation, load demand, and bus voltage magnitudes. The *RELIEF-F* algorithm is used to select the critical features which are mostly related to the stability status. Due to the page limit, such details are not given in the letter but referred to the reference [2]. At the online application stage, the feature inputs are real-time measured. If the fault to be assessed is already trained, the features will be directly sent to the DSA model; otherwise they will be considered as target domain, and then go to feature transformation and transfer learning. By minimizing marginal distributions and conditional distribution differences between the unknown features and the known features, an adaption matrix is obtained. After that, the unknown features will be adapted by this matrix so that they can be utilized by the DSA model. The key steps are elaborated in the sequel.

B. Feature Transformation

Dimension reduction method can learn the feature representation of transformation by minimizing the reconstruction error of input data. The purpose of generating common feature space is to integrate low-dimensional robust feature representation for training data and testing data, which maintains the inherent attributes of two domains after adaptation. Denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathrm{R}^{mxn}$ the input matrix, and $\mathbf{H} = \mathbf{I} - \frac{1}{n_s+n_t} ll^T$, where $\mathbf{I}$ is regarded as l as the ones vector, then the covariance matrix can be formulated as $\mathbf{XHX}^{\mathrm{T}}$. In this paper, the proposed method integrates input matrix via Principal Component Analysis to make the common feature space variance maximum as follow:

$$\max_{\mathbf{K^TK=I}} \operatorname{tr}(\mathbf{K^TXHX^TK}) \tag{1}$$

An adaptation matrix $\mathbf{K}$ can be obtained, hence, the common feature space $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_{\mathrm{n}}] = \mathbf{K}^{\mathrm{T}}\mathbf{X}$ can be generated.

C. Transfer Learning

Maximum Mean Discrepancy (MMD) is a criterion to measure the difference between different data distributions [4]. Compared to the parametric criteria, MMD can measure a nonparametric distance and void expensive distribution calculation caused by the parametric criteria. Taking datasets D_s and D_t, for example, the MMD calculates the empirical estimate of distances between source domain and target domain in the d-dimensional embedding.

In the common feature space, even though the d-dimensional feature transformation representation has been integrated, the distribution divergence between source domain and target domain is still very large. In order to reduce the differences between marginal distribution $P(\mathbf{X}_{\mathrm{s}})$ and $P(\mathbf{X}_{\mathrm{t}})$ of different faults, the robust transferable features of the source domain and target domain can be extracted by explicitly minimizing a proper distance measure. The MMD of marginal distributions is used in this paper as the distance measures as follows:

$$\begin{aligned}\mathrm{MMD}^2(\mathbf{X_s},\mathbf{X_t}) &= \left\| \frac{1}{n_s}\sum_{a=1}^{n_s} \mathbf{K^T} x_a - \frac{1}{n_t}\sum_{b=n_s+1}^{n_s+n_t} \mathbf{K^T} x_b \right\|^2 \\ &= \operatorname{tr}(\mathbf{K^TXR}_0\mathbf{X^TK})\end{aligned} \tag{2}$$

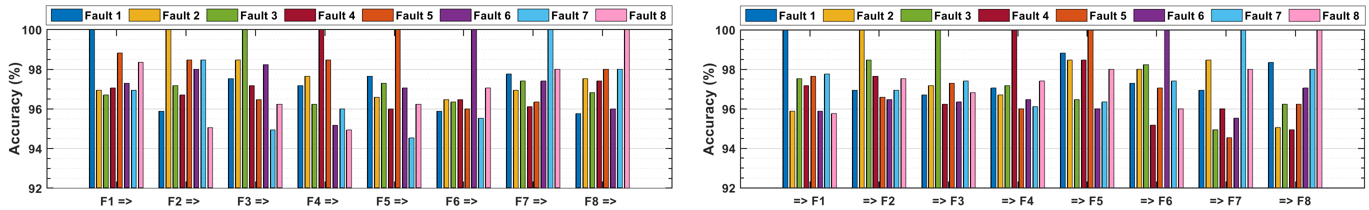
where $\mathbf{X} = \{\mathbf{X_s},\ \mathbf{X_t}\}$, $\mathbf{R_0}$ is the MMD matrix.

However, reducing the marginal distributions cannot ensure that the conditional distributions would be drawn close at the same time. The difference between conditional distributions $P(\mathbf{Y}_{\mathrm{s}}|\mathbf{X}_{\mathrm{s}})$ and $P(\mathbf{Y}_{\mathrm{t}}|\mathbf{X}_{\mathrm{t}})$ are reduced by minimizing the class-conditional distribution. Here, MMD is also used to measure the class-conditional distributions as follows:

$$\begin{aligned}&\mathrm{MMD}^2(\mathbf{X_s}^{(c)},\mathbf{X_t}^{(c)}) \\ &= \left\| \frac{1}{n_s^{(c)}}\sum_{x_a\in D_s^{(c)}}^{n_s} \mathbf{K^T} x_a - \frac{1}{n_t^{(c)}}\sum_{x_b\in D_t^{(c)}}^{n_s^{(c)}+n_t^{(c)}} \mathbf{K^T} x_b \right\|^2 \\ &= \operatorname{tr}(\mathbf{K^TXR}_c\mathbf{X^TK})\end{aligned} \tag{3}$$

where $\mathbf{R}_c$ is the MMD matrix involving class label c.

The marginal distributions and conditional distributions between source domain and target domain are drawn close under the new representation $\mathbf{G} = \mathbf{K}^{\mathrm{T}}\mathbf{X}$ by minimizing Eq. (2) and Eq. (3), respectively. In order to obtain the effective transferable feature representation, the difference in both marginal distributions and conditional distributions across different fault domains should be minimized simultaneously by resorting the pseudo



(a) (b)

Fig. 2. Online testing results of proposed method. (a) Each of fault is transferred to the remaining 7 faults. (b) Each of 7 faults is transferred to the remaining 1 fault.

Algorithm 1: Transfer Learning.

Input: Input matrix $\mathbf{X}$, label $\mathbf{Y_s}$, dimensionality of the samples d, regularization parameter λ.
Output: Adaptation matrix $\mathbf{K}$, common feature space matrix $\mathbf{G}$.
Initialize: Training data D_s and unlabeled testing data D_t.
begin
  Feature transformation in the common feature space by Eq. (1)
  Construct MMD matrix $\mathbf{R_0}$ by Eq. (2)
  while not convergence **do**
    Solve the generalized Eigen decomposition by Eq. (6)
    Select the d smallest eigenvectors
    Construct adaptation matrix $\mathbf{K}$
    Construct common feature space matrix $\mathbf{G} = \mathbf{K^T X}$
    Train the classifier on $\{(\mathbf{K^T}x_a, y_a)\}$, a=1…n_s
    Update pseudo testing labels $\{y_b\}$, b= n_s+1…n_s+ n_j
    Construct MMD matrix $\mathbf{R_c}$ by Eq. (3)
  end while
  Return the Adaptation matrix $\mathbf{K}$
end

labels of testing data [5]. The optimization problem can be comprised from Eq. (2) and Eq. (3) as follows:

$$\min J = \mathrm{MMD}^2(\mathbf{X_s},\mathbf{X_t}) + \sum_{c=1}^{C} \mathrm{MMD}^2(\mathbf{X_s}^{(c)}, \mathbf{X_t}^{(c)}) \quad (4)$$

where Eq. (4) is equal to Eq. (5) as follows:

$$\min_{\mathbf{K^T X H X^T K = I}} \sum_{c=0}^{C} \mathrm{tr}(\mathbf{K^T X R}_c \mathbf{X^T K}) + \lambda \left\|\mathbf{K}\right\|_F^2 \quad (5)$$

where λ is the regularization parameter to ensure the optimization problem to be well-defined. According to the constrained optimization theory, after deriving the Lagrange function for Eq. (5), the generalized Eigen decomposition can be calculated as follows:

$$\left(\mathbf{X} \sum_{c=0}^{C} \mathbf{R}_c \mathbf{X^T} + \lambda \mathbf{I}\right) \mathbf{K} = \mathbf{X H X^T K} \Lambda \quad (6)$$

where Λ is the Lagrange multiplier. The adaptation matrix $\mathbf{K}$ can be calculated by solving Eq. (6) for the d smallest eigenvectors. A complete process of transfer learning is summarized in Algorithm 1.

D. Hybrid Ensemble Learning Model for DSA

The DSA model can be a classifier or predictor using any an effective machine learning algorithm. In this paper, a hybrid ensemble learning model developed in our recent work [6] is used for DSA. It combines two different randomized learning algorithms, extreme learning machine (ELM) and random vector functional link (RVFL), as the single leaning unit for achieving higher performance. ELM and RVFL both show much faster learning speed and requires much less computation memory. Besides, the randomness in the input weights and biases can statistically improve the accuracy. Supposing the hybrid ensemble model includes u ELMs and u RVFLs, the prediction outputs from the learning units are combining to improve classification reliability. The final result is classified as secure and insecure by rule for classification. Note that other advanced DSA model reported in the literature can also be used.

III. Simulation Results

Simulation test are performed on New England 10-machine 39-bus system to validate the proposed method. A large number of operating points are generated based on Monte-Carlo method, which randomly samples generation and load within a certain range. A total of 8 faults are studied which are the three-phase faults with inter-area corridor trip and cleared 0.25 s after their occurrences (see Table II). Transient stability criterion is used to label the instances.

Given the 8 fault databases, 8 DSA models are trained correspondingly. Then, each of them is transferred to the remaining 7 faults or each of 7 faults is transferred to the remaining 1 fault, and the testing results are shown in Fig. 2(a–b), where the symbol "$=>$" in x-axis denotes transfer from the source domain and the target domain, e.g., "F1 $=>$" means the DSA model trained by F1 is transferred to other faults; "$=>$F1" means the DSA model trained by other faults are transferred to F1; y-axis is the DSA accuracy of the transfer. It can be observed that the proposed method can accurately transfer the model to the unlearned faults. The average DSA accuracy of proposed method shown in Fig. 2 are compared with the original DSA model in Table III. It can be seen that the proposed method can achieve a high accuracy for the unlearned faults.

Besides, the mutual transfer accuracy (MTA), calculated as the average accuracy between any two faults, is used to study the relationship between different faults. The result is illustrated in Fig. 3. This MTA can be a measure to quantify the similarity or correlation of different faults. From Fig. 3, the accuracy between

TABLE II
SELECTED CONTINGENCY

Fault ID	F1	F2	F3	F4	F5	F6	F7	F8
Fault Setting	Fault bus 3 Trip 3-4	Fault bus 4 Trip 3-4	Fault bus 14 Trip 14-15	Fault bus 15 Trip 14-15	Fault bus 15 Trip 15-16	Fault bus 16 Trip 15-16	Fault bus 16 Trip 16-17	Fault bus 17 Trip 16-17

TABLE III
AVERAGE ACCURACY OF DIFFERENT METHODS

Method	**Average Accuracy**
Original DSA Model without Transfer Learning	82.25%
Proposed method	97.27%

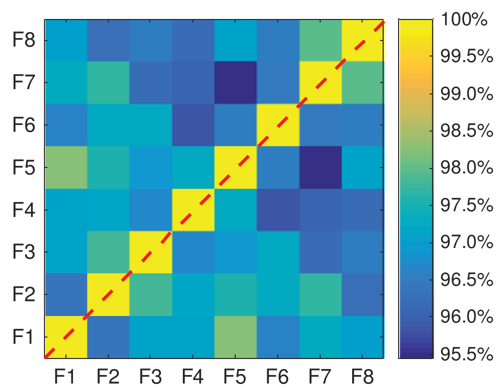


Fig. 3. MTA for eight faults.

different faults can be clearly observed, and the bright color (yellow) outperforms the dark color (blue) in term of MTA. The bright color means that these two faults are more similar (e.g., fault 1 and fault 5), and vice versa. In practice, these similar faults can be aggregated together for more efficient DSA purpose.

IV. CONCLUSION

A transfer learning method is proposed to use one trained pre-fault DSA model to assess multiple faults that have not been trained. Simulation results have validated its effectiveness and high accuracy. To the best of our knowledge, similar works have not been reported in the literature, and transfer learning can be a very promising method to solve other similar data-driven problems in power engineering.

REFERENCES

[1] Z. Y. Dong, Y. Xu, P. Zhang, and K. P. Wong, "Using IS to assess an electric power system's real-time stability," *IEEE Intell. Syst.*, vol. 28, no. 4, pp. 60–66, Jul./Aug. 2013.

[2] Y. Xu, Z. Y. Dong, J. H. Zhao, P. Zhang, and K. P. Wong, "A reliable intelligent system for real-time dynamic security assessment of power systems," *IEEE Trans. Power Syst.*, vol. 27, no. 3, pp. 1253–1263, Aug. 2012.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[4] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[5] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 2200–2207.

[6] C. Ren and Y. Xu., "A multiple randomized learning based ensemble model for power system dynamic security assessment," in *Proc. IEEE Power Energy Soc. General Meeting*, 2018, pp. 1–5: IEEE.