Procedures of Statistics. New York: McGraw-Hill, 1960.
15. R. F. Nau, Forecasting, Feb. 2005. [Online]. Available: http://www.duke.edu/~rnau/rsquared.htm
16. *IEEE Standard for Synchrophasors for Power Systems*, IEEE Std. C37.118-2005, 2005.

Automatic Discovery and Application of Significant Relationships Between Steady-state Operation Data and Transient Stability Level in Electric Power Industry

Youping Fan*, Songkai Liu, Cheng Tian

School of Electrical Engineering, Wuhan University, Wuhan 430072, China

Abstract
The assessment of transient stability level is important for the automation of production processes in electric power industry. The connotative relationships of steady-state operation data and general critical clearing time $\overline{CCT}$ are explored in a large data set for power system. A novel online transient security assessment method is presented based on relationships exploration. Each relationship is given scores by the maximal information coefficient and Pearson correlation coefficient. Some highly ranked linear and nonlinear relationships are detected out and shown. Meanwhile, the generalized nonlinear relationships exploration coefficient is presented to discover connotative nonlinear relationships directly. Curve fitting is used for the explored linear relationships and functional nonlinear relationships to estimate $\overline{CCT}$ of new operation states. Weibull distribution and generalized extreme value distribution are adopted for distribution fitting of $\overline{CCT}$, and cumulative probability curve is used to determine the value range of $\overline{CCT}$ for each transient security level. The method is tested on a 21-bus system and various test results indicate it is accurate and effective. It can give accurate estimation results of $\overline{CCT}$, relative degree of transient stability and security level of transient stability. The applicability will not be influenced by the change of structure and scale since the selection of input features is based on data statistics and mining, and the way of selection is more intelligent than the current techniques. The automatic identification of transient stability level is meaningful for uninterrupted production in power industry.
Keywords: POWER SYSTEM AUTOMATION, TRANSIENT SECURITY ASSESSMENT, AUTOMATIC IDENTIFICATION, SIGNIFICANT RELATIONSHIPS, LARGE DATA SETS

Nomenclature

P-V	Active power-voltage amplitude.
PGi	Active power of generators at bus i.
QGi	Reactive power of generators at bus i.
Vi	Voltage amplitude of bus i.
θi	Voltage phase angle of bus i.
Pi_j	Active power from bus i to bus j.
Qi_j	Reactive power from bus i to bus j.
Si_j	Apparent power from bus i to bus j.
Pi_X	Active power from bus i to equipment X.
Qi_X	Reactive power from bus i to equipment X.
Si_X	Apparent power from bus i to equipment X.
I%i_X	Load rate percent of the equipment X at bus i.
SSE	Sum of squares due to error.
RMSE	Root mean squared error.
R-square	Coefficient of determination.

1. Introduction

For the uninterrupted production in power industry, it is necessary to implement the automatic identify of the transient stability level. There has been a continually increasing interest and investigation into assessment of transient stability and operation security [1]-[4]. Transient stability or large disturbance rotor angle stability is concerned with the ability of maintaining synchronism when the system is subjected to a severe disturbance, such as a short circuit on a transmission line. At present security scanning and assessment for power system mainly rely on a large number of fault simulations. Considering the large scale of current power system, different types of equipment, the real-time changing load and changing generator output, the enumerated probabilistic fault simulation analysis method is not able to provide real-time assessment results or effective control measures information of improving the security level. In general, a single simulation method cannot meet the demand of intelligent decision for the power system control. Therefore, there is a pressing need to develop a fast online transient security assessment method that could analyze security level [5], [6] and forewarn the system operators to take necessary preventive actions in case need arises.

Transient security assessment is a problem with inherent complexity, non-linearity, uncertainty and the need for online monitoring. Exploring the possible connotative relationships of operation data and transient security based on knowledge engineering technology and data mining [7]-[9] is a very attractive idea, which is based on a large number of accumulated samples. The samples rely on fault simulation scan of transient stability. It is designed to discover the connotative relationships in the fault scan results and power flow information, which may be useful to assess security level. This security level assessment idea makes up for the deficiency of simulation scan: 1) once the security level assessment rules have been established by offline data mining in a large set, the computation speed of online security assessment will be fast; 2) considering that power flow of power system is easy to observe, adjust and control, it will be convenient to improve the security level based on the explored relationships and it plays a role of aid decision making. However, it still faces many challenges to achieve direct security assessment based on operating information and data mining. It is needed to figure out how the steady-state operating information of the system affects the transient security under a certain fault. Another challenge is: the WAMS of a large power system may collect a huge amount of operating data, which not only contains the features of high correlation for security level, but also the ones of weak correlation. How to effectively select the dimension of input features, extract the high correlation features, and eliminate redundant features is a key step in the security level assessment based on artificial intelligence theory [10]-[12]. The purpose of feature selection is to select and classify high correlation features from a large number of original features, which requires reducing the dimension and minimum information loss of representing the research object. By selecting the features highly associating with the research object, the purpose that a d-dimensional set of features is extracted from a D-dimensional set of features ($d << D$) can be achieved [13], [14].

Lots of methods have been applied to the transient stability assessment, such as artificial neural networks (ANN), pattern recognition techniques [15], decision trees [16], and fuzzy neural networks [17]. Some optimization algorithms such as simulated annealing algorithm and ant colony algorithm are also applied to transient stability assessment [18], [19], and the purpose is to choose better input features. In [19], the number of variables in the optimal variable group given by optimization algorithms is small, which reduce the dimension well. However, the final accuracy is not particularly high, and the new additional variables cannot be directly given by optimization algorithms when the assessment accuracy needs to be improved. In [23], [24], CCT in the case of an assumption fault is calculated precisely with a short time. In [16], [17], transient stability prediction accuracy for an occurred fault is relatively high. But Transient security assessment for a steady-state operation state is not given to the operators in [16], [17], [21], [22]. In [23], Numerical simulation is used for stochastic transient stability assessment. The combined

effect of multiple system faults and the load uncertainty on transient stability is taken into account, and the transient stability is measured by a stability index from the statistical point of view. But the connotative association of steady-state operation data and transient stability security is not given.

The novel online assessment method for transient security presented in this paper is based on relationships exploration in a large data set, which is created on offline power flow reports and fault simulations. General critical clearing time $\overline{CCT}$ is defined based on the statistics of historical faults. A relationship between each operation variable and $\overline{CCT}$ is gi- ven two scores by the maximal information coefficient (MIC) [24] and the Pearson correlation coefficient (PCC) [25]. The input features are selected based on the connotative relationships explored in the data set, which are the highly scored ones. The explored relationships of operation variables and transient security are presented and explained, including linear and nonlinear ones. Generalized nonlinear relationships exploration coefficient and two kinds of models for it are presented in this paper since MIC cannot rank nonlinear relationships directly and $MIC-\rho^2$ provided by [24] does not do well .The linear and functional nonlinear relationships are applied to online assessment method for transient security based on curve fitting. The way to select features is based on the rank of relationships, which is different from the conventional optimization algorithms. The rank of relationships can explain why optimization algorithms give a better chance to some variables, and the number of input features needed is less than 2% of the total number of operating variables. Distribution fitting and cumulative probability of $\overline{CCT}$ are presented to determine the value range of $\overline{CCT}$ for each level. The method can give accurate estimation results of $\overline{CCT}$ and assess the transient security very well. It can overcome the curse of dimensionality of large-scale power systems. The applicability will not be influenced by the change of the structure and scale since input features selected are based on data statistics and mining. The method has a certain requirement for the number of previous offline simulation samples. The automatic assessment processes for transient security are shown in Fig. (1). The data of input features can be obtained from phase measurement units (PMUs) in power industry. The method is economical since it requires a relatively small number of PMUs.

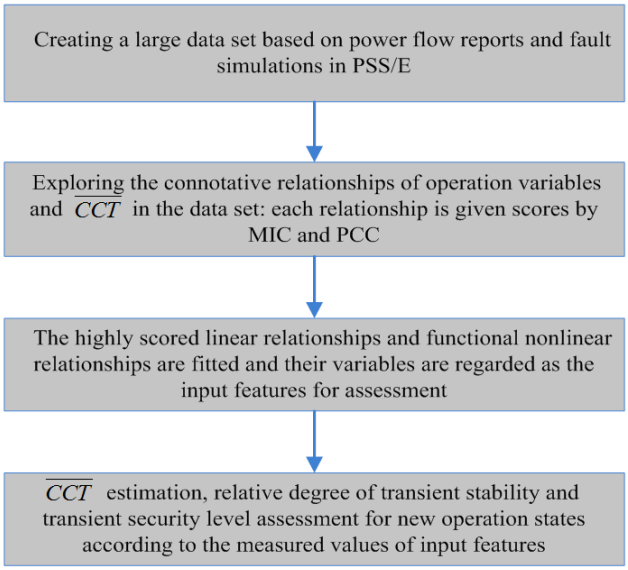


Figure 1. Automatic assessment processes for transient security

2. Problem Statement and Supporting Mathematical Methods

2.1. General Critical Clearing Time ($\overline{CCT}$)

For a power system, the transient stability is usually described with the critical clearing time (CCT) when a certain fault occurs. CCT can be used as the index for transient stability assessment. Compared with other indices, CCT is easier for operators to understand how security a system is during the operation. The system with a longer CCT is considered to be operated at a higher transient security level. The delay in fault clearing from the CCT means loss of synchronous operation of the generators in a power system. For an operation state, the CCT of a single fault position cannot represent the transient stability of the whole system well and a better index is ne- cessary. Therefore, fault tests of various positions are taken into consideration, which are based on the statistics data of historical faults. A general critical clearing time ($\overline{CCT}$) of the whole system is presented in this paper, which is given by Equation (1).

$$\overline{CCT} = \sum_{i=1}^{n} \mu_i CCT_i \qquad (1)$$

where n is the total number of fault test position, μ_i is the percentage of historical statistics for fault i, and CCT_i is CCT of fault i.

It is should be noted that if the statistics data of historical faults is not provided, the transmission lines with heavy power flow or some other ones considered to be key can be selected as fault test positions.

2.2. Maximal Information Coefficient (MIC)

MIC is a measure of dependence for two-variable relationships and it can capture relationships both functional and not in large data sets [24]. In this pa-

per, MIC is introduced into the field of power system transient stability and it is used to explore the connotative relationships of $\overline{CCT}$ and operation variables in power system. MIC is based on the idea if there is a relationship between two variables, a grid can be drawn on the scatter plot of the two variables that partitions the data to encapsulate the relationship. MIC gives a score to measure the relationship between two variables based on the data pairs of variables.

Given a finite set D of ordered pairs, the x-values of D are partitioned into x bins and the y-values of D are partitioned into y bins, allowing empty bins. Such a pair of partitions can be called an x-by-y grid. Given a grid G, let $D|_G$ be the distribution induced by the points in D on the cells of G. The distribution on the cells of G is obtained by letting the probability mass in each cell be the fraction of points in D falling in that cell. For a fixed D, different grids G result in different distributions $D|_G$. For a data set D of two-variable, the MIC of their relationship is given by (2), (3).

For a finite set $D \subset R^2$ and positive integers x, y,

$$I^*(D,x,y) = \max I(D|_G) \tag{2}$$

where the maximum is over all grids G with x columns and y rows, and $I(D|_G)$ denotes the mutual information of $D|_G$.

The MIC of two-variable data with sample size n and grid size less than $B(n)$ is given by

$$MIC(D) = \max_{xy<B(x,y)} \left\{ \frac{I^*(D,x,y)}{\log \min\{x,y\}} \right\} \tag{3}$$

where $\omega(1) < B(n) \le O(n^{1-\varepsilon})$ for some $0 < \varepsilon < 1$.

$B(n) = n^{0.6}$ is used because it's found to work well in the research [24]. MIC falls between 0 and 1. Some properties of MIC are as follows. (1) MIC assigns scores that tend to 1 to all never-constant noiseless functional relationships; (2) MIC assigns scores that tend to 1 for a larger class of noiseless relationships; (3) MIC assigns scores that tend to 0 to statistically independent variables.

3. Creating a Large Data Set Of $\overline{CCT}$ and System Operation Variables

The accidents of transient stability failure are unusual in the practice of power system operation, which leads to the lack of instability data samples and cannot satisfy the requirements of data mining. Therefore, power system fault simulation is usually used to obtain samples. In this paper, software PEE/S and Python programming is used to obtain the original operating data and analyze transient stability. The flow chart is illustrated in Fig. (2).

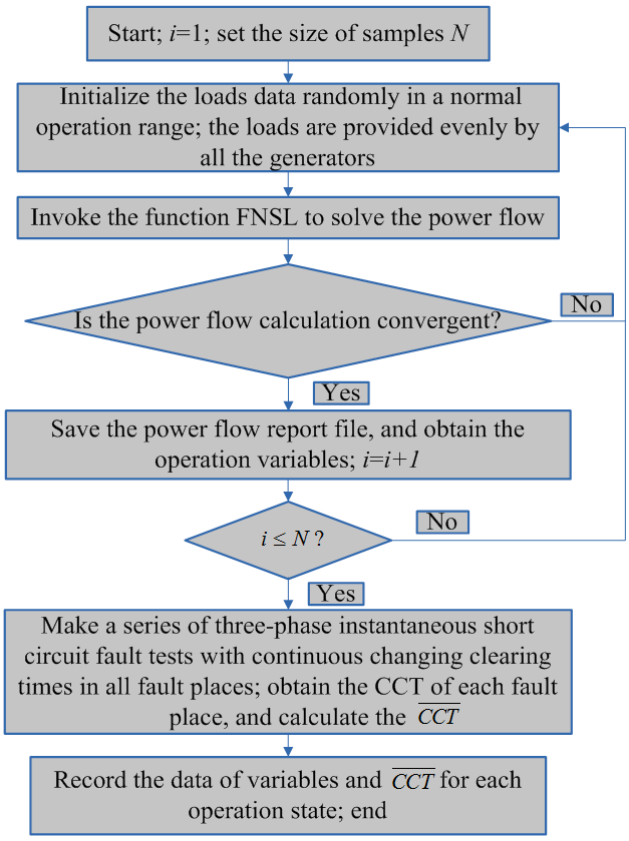


Fig. (2). Flow chart for creating a large data set of operation variables and $\overline{CCT}$ based on PSS/E

The test case used in the paper is a 21-bus test system provided by PSS/E, which is shown in Fig. (3). Three-phase instantaneous short circuit faults are tested at the head of transmission line in the case. The position and percentage of historical statistics for each fault are shown in Table 1. In the power flow reports, 470 operation variables are finally selected after eliminating some constant ones. Then a large data set of operation variables and $\overline{CCT}$ is created, which is a matrix with 471 rows (470 physical variables and $\overline{CCT}$) and 150 columns.

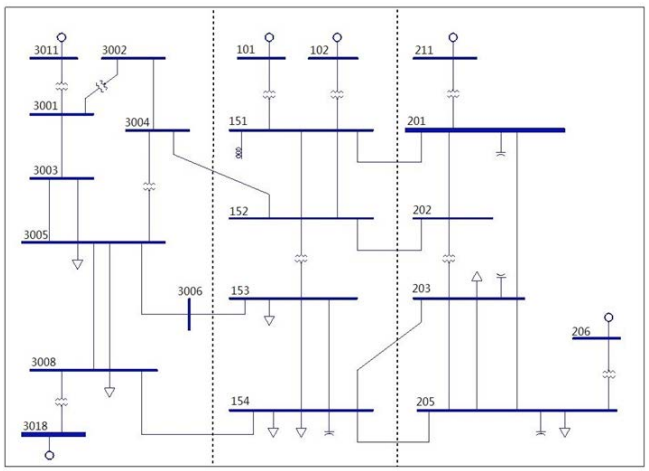


Figure 3. 21-bus test system provided by PSS/E

Table 1. Statistics of Historical Faults

No.	Transmission Line	Percentage of Historical Statistics
1	Bus 205 to bus 203	40%
2	Bus 206 to bus 205	40%
3	Bus 3005 to bus 3003	5%
4	Bus 3008 to bus 154	5%
5	Bus 3018 to bus 3008	10%

4. Exploring the Connotative Relationships of Operation Variables and $\overline{CCT}$

4.1. Top Relationships Explored by MIC and PCC

MIC and PCC are applied to detect connotative relationships of operation variables and $\overline{CCT}$ in the created data set. Table 2 shows the top 2% of relationships by MIC and Table 3 shows the top 2% of relationships by PCC. Specially, the relationship ranked 1st in Table 2 and the one ranked 1st in Table 3 are shown successively in Fig. (4A) and Fig. (4B). A relationship highly ranked by PCC is with high degree of linearity. A relationship highly ranked by MIC shows a certain relationship between the two variables, but maybe not linear.

Table 2. Top 2% of Relationships by MIC

Var. 1	Var. 2	MIC	MIC Rank	PCC	PCC Rank
$\overline{CCT}$	*S152_151_1*	0.854	1	-0.834	118
$\overline{CCT}$	*S152_151_2*	0.854	2	-0.834	119
$\overline{CCT}$	*Q3005_3006*	0.832	3	-0.924	1
$\overline{CCT}$	*S202_201*	0.829	4	-0.872	49
$\overline{CCT}$	*Q153_3006*	0.828	5	0.912	8
$\overline{CCT}$	*Q3004_152*	0.825	6	-0.917	4
$\overline{CCT}$	*Q151_102*	0.824	7	0.844	92
$\overline{CCT}$	*Q151_101*	0.824	8	0.844	93
$\overline{CCT}$	*V151*	0.814	9	0.845	84
$\overline{CCT}$	*QG102*	0.814	10	-0.844	87

Table 3. Top 2% of Relationships by PCC

Var. 1	Var. 2	PCC	PCC Rank	MIC	MIC Rank
$\overline{CCT}$	*Q3005_3006*	-0.924	1	0.894	3
$\overline{CCT}$	*Q3006_3005*	0.917	2	0.847	22
$\overline{CCT}$	*Q3006_153*	-0.917	3	0.847	23
$\overline{CCT}$	*Q3004_152*	-0.917	4	0.831	7
$\overline{CCT}$	*Q3004_3002*	0.916	5	0.877	35
$\overline{CCT}$	*Q3005_3003_2*	0.916	6	0.871	38
$\overline{CCT}$	*Q3005_3003_1*	0.916	7	0.871	39
$\overline{CCT}$	*Q153_3006*	0.912	8	0.874	5
$\overline{CCT}$	*Q3002_3001*	0.902	9	0.865	56
$\overline{CCT}$	*Q3002_3004*	-0.902	10	0.865	68

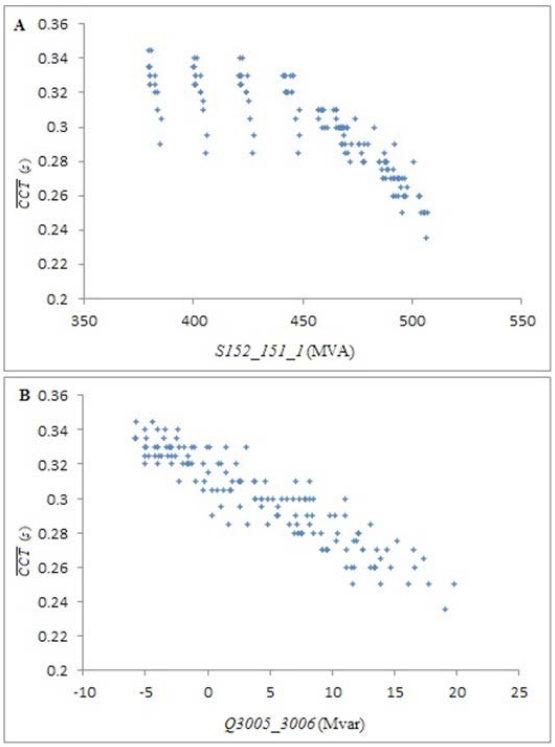


Figure 4. Scatter plots. (A) S152_151_1 and $\overline{CCT}$. (B) Q3005_3006 and $\overline{CCT}$

These relationships are not easy to be found directly since they are connotative in mass data. Generally reasonable explanation can be given for these relationships from the perspective of power system. The variable S152_151_1 is selected as an example in Fig. (4A), which represents the apparent power of the first line from bus 152 to bus 151. A scenario is given to explain the relationship conveniently: with the increase of power system load, the power on the transmission line will increase; when load rating of the system is low, the increasing load doesn't make $\overline{CCT}$ decrease significantly; when load rating of the system is high, a slight increase of load will make $\overline{CCT}$ decrease significantly. By the way, a conclusion can be got: the sensitivity of $\overline{CCT}$ to the variable S152_151_1 increases with the increase of load.

4.2. MIC versus PPMCC

Fig. (5A) shows that MIC versus PCC for all pairwise relationships in the data set. In different areas of Fig. (5A), different kinds of relationships can be found. Not every operation variable has a specific re-

lationship with $\overline{CCT}$. In the relationships, some are functional while some are not. In the functional relationships, some are linear while some are nonlinear. Some examples are as follows.

1. Fig. (5B): Unassociated variables are given low scores both by MIC and PCC. It indicates no specific relationship exists between the variable θ3002 and $\overline{CCT}$;

2. Fig. (5C): Ordinary linear relationships get high score under both MIC and PCC tests. It indicates an obvious linear relationship between the variable Q153_3006 and $\overline{CCT}$;

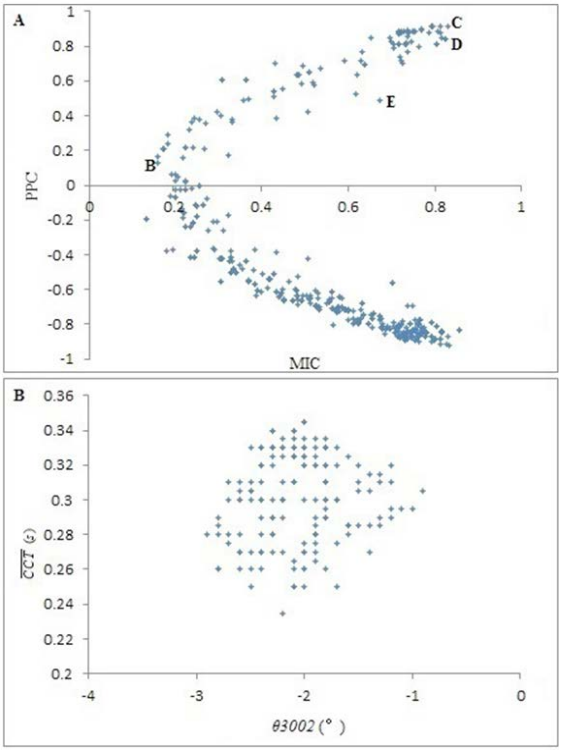
3. Fig. (5D) and Fig. (5E): Relationships can be detected by MIC but not by PCC if they are nonlinear. It indicates there exists a kind of nonlinear relationship between the variable and $\overline{CCT}$. Fig. (5D) shows a functional relationship between Q151_101 and $\overline{CCT}$, while Fig. (5E) shows a kind of specific relationship between Q205_206 and $\overline{CCT}$.

4.3. Generalized Nonlinear Relationships Exploration Coefficient

The direct exploration method for nonlinear relationships is specially discussed since linear and nonlinear relationships are all given highly scores by MIC. The generalized nonlinear relationships exploration coefficient is presented to explore nonlinear relationships directly. Two kinds of models for it are recorded as β_1^γ and β_2^α in the paper, which are given by Equations (4) and (5).

$$\beta_1^\gamma = MIC - |\rho|^\gamma (\gamma > 0) \quad (4)$$

$$\beta_2^\alpha = MIC - \alpha|\rho| (0 < \alpha \le 1) \quad (5)$$



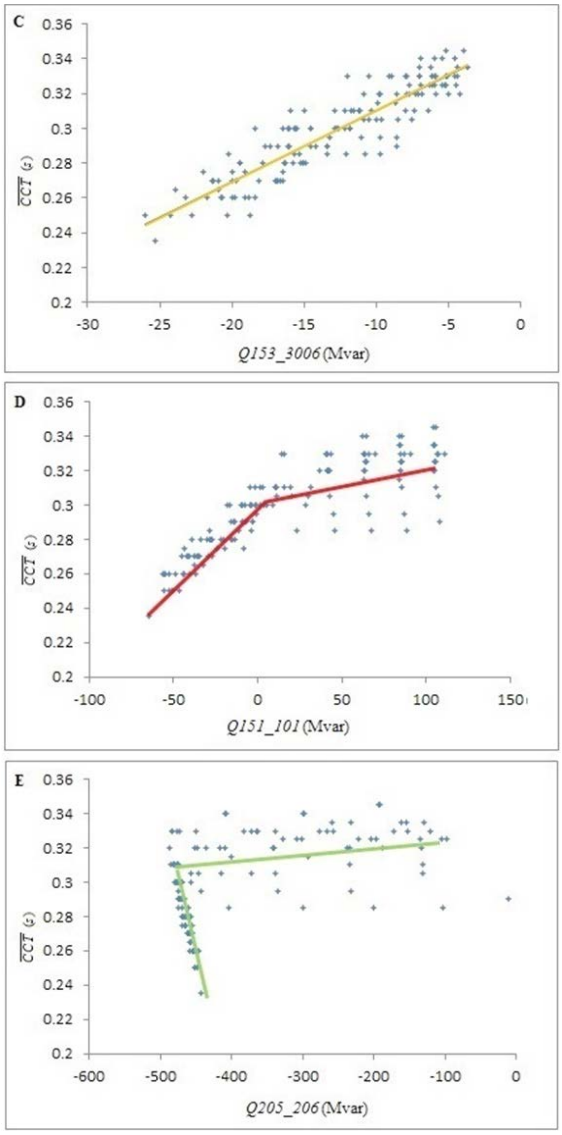


Figure 5. Application of MIC and PCC to the created data set. (A) MIC versus PCC for all pair-wise relationships in the data set. (B)-(E) Examples of relationships from (A)

β_1^γ: each integer in $(0, 20]$ is tested for γ. (1) For $\gamma = 1$, the top 2% of β_1^γ rank doesn't give obvious relationships. (2) For $2 \le \gamma \le 8$, a few of the relationships in the top 2% of β_1^γ rank are obvious. The increase of γ nearly doesn't affect the top 2% of β_1^γ rank. (3) For $9 \le \gamma \le 20$, the top 2% of β_1^γ rank keeps almost the same with that of MIC when γ increases, which still includes some linear relationships. Therefore, β_1^2 can be selected as an excellent one to present the ability of β_1^γ in the direct exploration for nonlinear relationships. The top 2% of β_1^2 rank is in Table 4 and the scatter plots of the relationships are shown in Fig. (6).

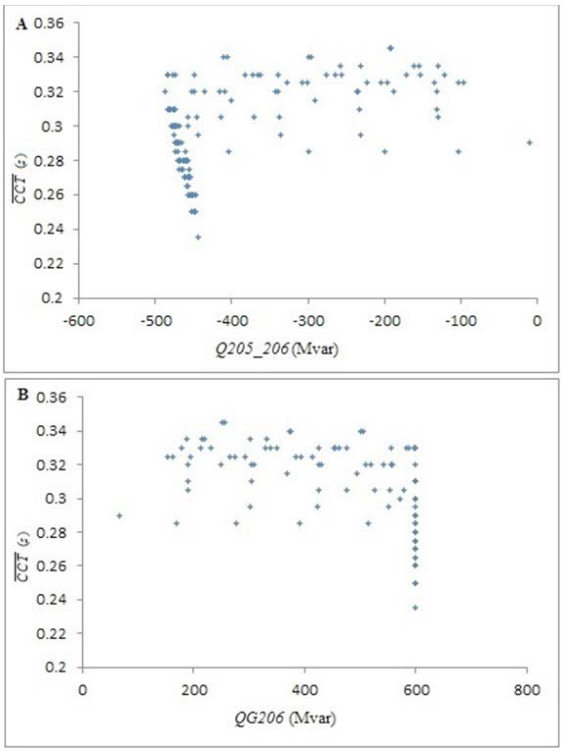


Figure 6. Scatter plots for top 2% relationships by $MIC-\rho^2$.

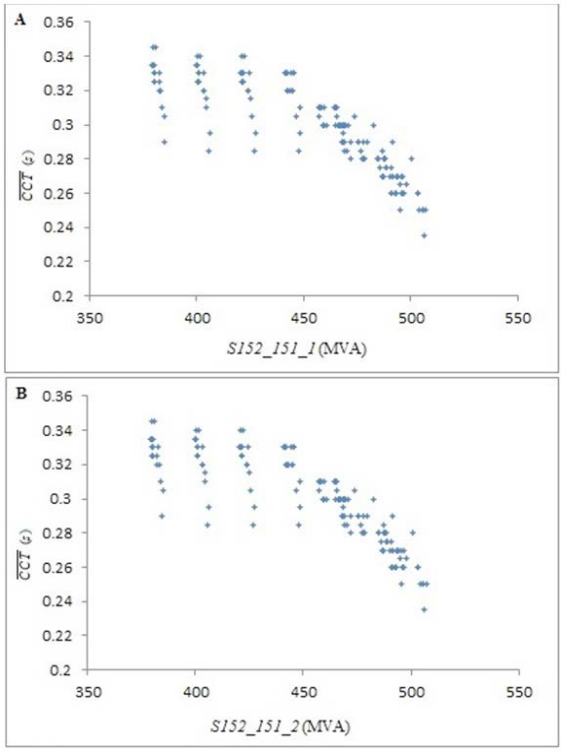


Figure 7. Scatter plots for top 2% relationships by $MIC-0.4|\rho|$

Table 4. Top 2% of Relationships by $MIC-\rho^2$

Var. 1	Var. 2	$MIC-\rho^2$	$MIC-\rho^2$ Rank	MIC	PCC
$\overline{CCT}$	*Q205_206*	0.433	1	0.672	0.489
$\overline{CCT}$	*QG206*	0.386	2	0.701	-0.561
$\overline{CCT}$	*Q206_205*	0.386	3	0.701	-0.561
$\overline{CCT}$	*Q205_201*	0.341	4	0.619	0.527
$\overline{CCT}$	*Q205_SHUNT*	0.329	5	0.509	0.422
$\overline{CCT}$	*S205_SHUNT*	0.329	6	0.509	-0.422
$\overline{CCT}$	*P202_152*	0.286	7	0.326	-0.175
$\overline{CCT}$	*P152_202*	0.296	8	0.326	0.175
$\overline{CCT}$	*S154_SHUNT*	0.285	9	0.434	-0.387
$\overline{CCT}$	*Q154_SHUNT*	0.285	10	0.434	0.387

Table 5. Top 2% of Relationships by $MIC-0.4|\rho|$

Var.1	Var. 2	$MIC-0.4\|\rho\|$	$MIC-0.4\|\rho\|$ Rank	MIC	PCC
$\overline{CCT}$	*S152_151_1*	0.520	1	0.854	-0.834
$\overline{CCT}$	*S152_151_2*	0.520	2	0.854	-0.834
$\overline{CCT}$	*Q151_102*	0.486	3	0.824	0.844
$\overline{CCT}$	*Q151_101*	0.486	4	0.824	0.844
$\overline{CCT}$	*S202_201*	0.480	5	0.829	-0.872
$\overline{CCT}$	*P151_152_1*	0.478	6	0.803	-0.813
$\overline{CCT}$	*P151_152_2*	0.477	7	0.803	-0.813
$\overline{CCT}$	*P152_151_2*	0.477	8	0.803	0.812
$\overline{CCT}$	*P152_151_1*	0.477	9	0.803	0.812
$\overline{CCT}$	*Q205_206*	0.285	10	0.673	0.489

β_2^α: α is tested from 0.1 to 1, and the step size is 0.1. (1) For $0.1 \le \alpha \le 0.3$, the top 2% of β_2^α rank is almost the same with that of MIC. (2) For $\alpha = 0.4$, many of the relationships in the top 2% of β_2^α rank are obvious. (3) For $0.5 \le \alpha \le 1$, most of the relationships in the top 2% of β_2^α rank are not obvious. Therefore, $\beta_2^{0.4}$ can be selected as an excellent one to represent β_2^α. The top 2% of $\beta_2^{0.4}$ rank is in Table 5 and the scatter plots of the relationships are shown in Fig. (7).

The explored nonlinear relationships in Fig. (7) are more clear and useful from the contrast of Fig. (6) and Fig. (7), which shows the model $\beta_2^{0.4}$ is more effective than β_1^2 in the direct exploration for nonlinear relationships. In fact, $MIC-\rho^2$ recommended by [24] is a narrow expression of Equation (5). On the face of it, a narrow expression such as $MIC-\rho^2$ or $MIC-|\rho|$ can be a measurement of nonlinearity. But it is considered in the paper that these expressions are not effective enough to explore nonlinear relationships in practice. As it is shown in Fig. (5) and

Fig. (7), an important or useful relationship is usually based on the basic linear relationships in engineering. In other words, a nonlinear relationship usually consists of two or more simple relationships. In fact, some nonlinear relationships such as Fig. (5D) are with linearity to a certain extent and each of them has a not small $|\rho|$. If a nonlinear relationship is imagined as a superstructure, $MIC-\rho^2$ or $MIC-|\rho|$ will break the infrastructure if it is used as a direct exploration tool for nonlinear relationships. Naturally, nonlinear relationships such as Fig. (5D) cannot be highly ranked or explored by $MIC-\rho^2$ or $MIC-|\rho|$.Therefore, the generalized nonlinear relationships exploration coefficient is with higher practical value in engineering than the narrow expression $MIC-\rho^2$, and the model β_2^{α} is believed to be effective and convenient than β_1^{γ} in the paper.

5. $\overline{CCT}$ Estimation and Online Transient Security Assessment

5.1. Input Features Selection

Since the connotative relationships are explored in a large data set of operation variables and $\overline{CCT}$, which can be used to estimate $\overline{CCT}$ and assess the transient security for a new operation state. Obviously, the accuracy of estimation is affected directly by the input features selection. The selected variables should be the ones who have obvious relationships with $\overline{CCT}$. Moreover, functional relationships are specially selected since curve fitting of them can be used for estimation conveniently. The total number M of the variables selected should be appropriate. Setting M too low can lead to inaccurate estimation, while setting M too low means the increase of economic cost in engineering application because of more measuring points. An appropriate selection is given based on estimation tests: the top 1% of functional nonlinear relationships in MIC rank and the top 1% of linear relationships by PCC. Variables finally selected from the ranking list by MIC are: *S152_151_1, S202_201, Q3004_152, Q151_101, QG102*, and Q3005_3006. Variables finally selected from the ranking list by PCC are: *Q3005_3006, Q3006_153, Q3006_3005, Q3004_3002 and Q3005_3003_1*.

5.2. Clustering

Samples of $\overline{CCT}$ are clustered in 3-d space to test input selected variables. The value of $\overline{CCT}$ is in $(0.235,\ 0.345)$ and 0.290s is regarded as a boundary. In Fig. (8), a sample is green if the $\overline{CCT}$ is larger than 0.290s and is blue if the $\overline{CCT}$ is smaller than 0.290s. The variables of x, y, and z axis are S152_151_1, Q3005_300, andQ3006_153 respectively in Fig. (8A), which are the selected ones. The variables of x, y, and z axis are θ204, θ3018, and P3008_154 respectively in Fig. (8B), which are the stochastic ones. As it is clearly shown, the clustering in Fig. (8A) is much better than that in Fig. (8B).

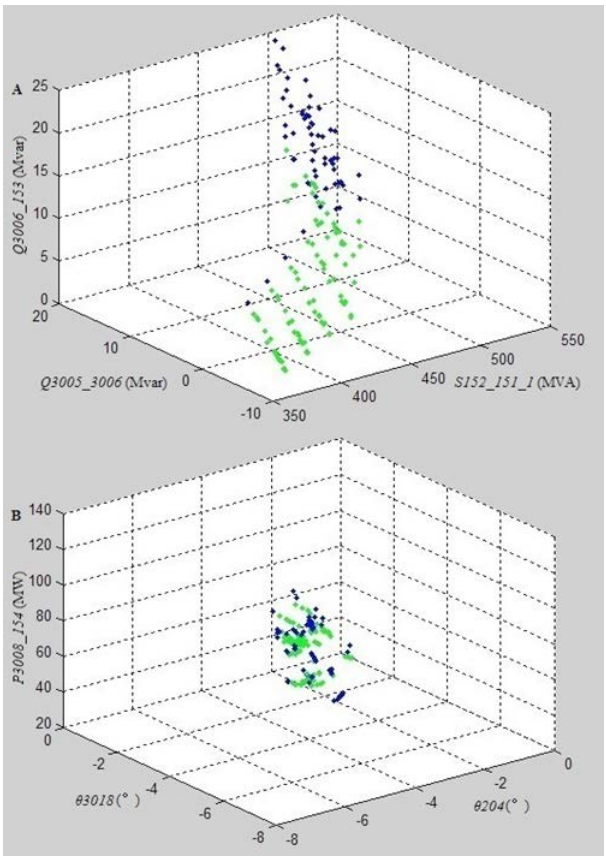


Figure 8. Clustering of $\overline{CCT}$. (A) Clustering based on 3 selected variables. (B) Clustering based on 3 stochastic variables

5.3. Curve Fitting

Polynomial fitting is better than other types of fitting after tests for the explored functional nonlinear relationships, and cubic polynomial fitting can satisfy the need of accuracy. For example, cubic polynomial fitting is used for the scatter plot of the variable S152_151_1 and $\overline{CCT}$ and the value of each index is: $SSE=0.017$, $RMSE=0.011$, $R-square=0.828$. The functional expression for nonlinear relationships is given by Equation (6) and the polynomial coefficients are shown in Table 6. The left 5 subfigures in Fig. (9) show the results of curve fitting for nonlinear relationships. The linear fitting is used for the explored relationships and the functional expression is given by Equation (7) and the coefficients are shown in Table 6. The right 5 subfigures in Fig. (9) are for them.

$$f_1(x)=P_1x^3+P_2x^2+P_3x+P_4 \quad (6)$$

$$f_2(x)=P_3x+P_4 \quad (7)$$

Table 6. Polynomial Coefficients

No.	Var.	P_1 (10^{-9})	P_2 (10^{-6})	P_3 (10^{-3})	P_4
1	*S152_151_1*	-42.97	49.96	-19.41	2.848
2	*S202_201*	3.982	-8.843	5.654	-0.7917
3	*Q3004_152*	-373.1	-161.6	-24.53	-0.9625
4	*Q151_101*	9.013	-4.319	-0.872	0.2988
5	*QG102*	-4.775	-1.790	-0.1388	0.3211
6	*Q3005_3006*	-	-	-3.605	0.3157
7	*Q3006_153*	-	-	-3.949	0.3397
8	*Q3006_3005*	-	-	3.949	0.3397
9	*Q3004_3002*	-	-	1.381	0.3140
10	*Q3005_3003_1*	-	-	1.417	0.3490

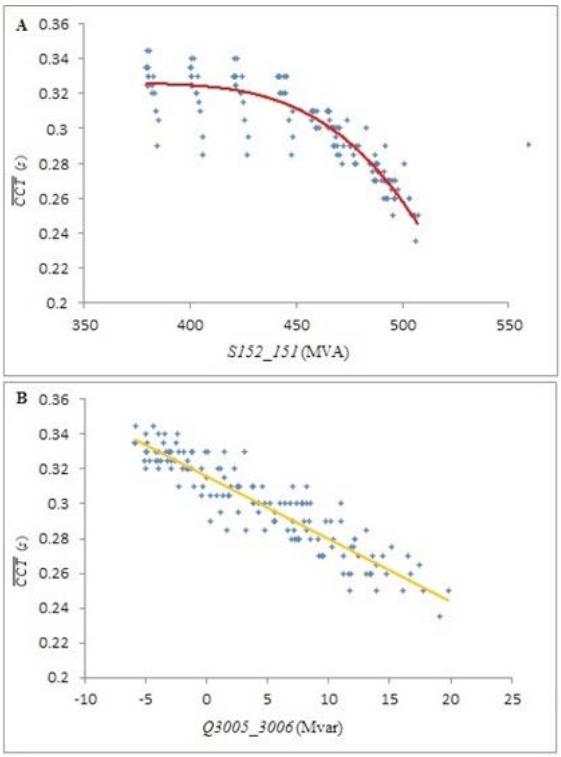


Figure 9. Curve fitting for selected relationships

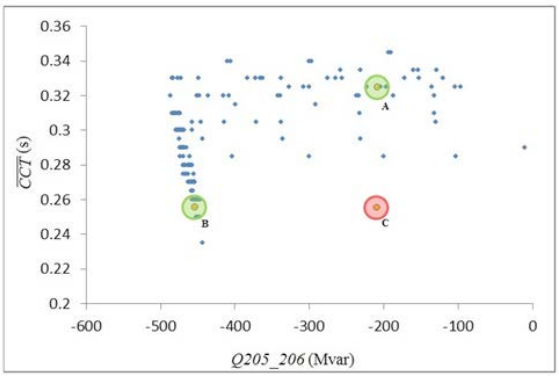


Figure 10. Curve fitting for selected relationships

It is should be noted, although the non-functional relationships cannot be used for estimation directly, they can check the estimation results to a certain extent. An example of using a non-functional relationship to check the estimation results is shown in Fig. (10). Usually some previous samples can be found in a small neighborhood of an accurate estimation result in the scatter plot (A and B), while not for an inaccurate one (C). Naturally, the using of multiple non-function relationships can give a more strict check for estimation.

5.4. Distribution Fitting of $\overline{CCT}$, Relative Degree of Transient Stability and Security Level Division

The distribution fitting for samples of $\overline{CCT}$ is shown in Fig. (11A). Weibull distribution (WD) and generalized extreme value distribution (GEVD) both are better than other types of fitting after tests. Some distribution characteristics of $\overline{CCT}$ can be explored by distribution fitting: 1) $\overline{CCT}$ has a great probability to be around 0.31s. 2) $\overline{CCT}$ is hardly to be larger than 0.35s or smaller than 0.22s. The cumulative probability curves of WD and GEVD fitting can be used to assess the relative degree of transient stability for new operation states, and the one of GEVD is shown in Fig. (11B). Usually, $\overline{CCT}$ is not visual enough to describe security of transient stability for the power system operators when the system is steady, so the concepts relative degree of transient stability and security levels of transient stability are needed. In the paper, 5 security levels of transient stability are divided by the points with the y axis 0%, 20%, 40%, 60%, 80% and 100% respectively. The coordinates of the points are shown in Fig. (11B), and the 5 levels are as follows. Level 1: 0.323s-0.345s, level 2: 0.309-0.322, level 3: 0.296-0.308, level 4: 0.279-0.295, level 5: 0.245-0.278.

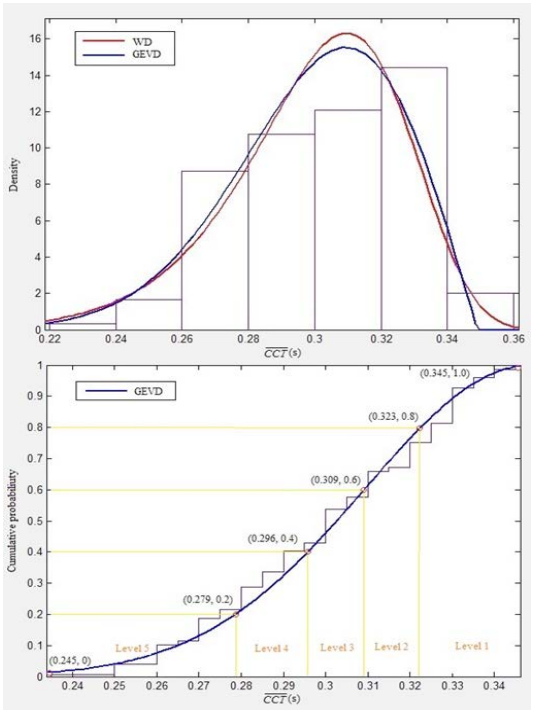


Figure 11. Statistical property of $\overline{CCT}$. (A) WD and GEVD for $\overline{CCT}$. (B) Cumulative probability of GEVD for $\overline{CCT}$

5.5. $\overline{CCT}$ Estimation and Transient Security Assessment Results

Each input variable can give an estimation result of $\overline{CCT}$, while multiple variables can make a more accurate result than a single one. For a group of functional nonlinear relationships, the estimation result is given by Equation (8). For a group of linear relationships, the estimation result is given by Equation (9). The final comprehensive estimation is given by Equation (10).

$$\overline{CCT_{MIC}} = \sum_{i=1}^{N_1} \frac{MIC_i}{\sum_{i=1}^{N_1} MIC_i} \overline{CCT_i} \tag{8}$$

$$\overline{CCT_{\rho}} = \sum_{j=1}^{N_2} \frac{|\rho_j|}{\sum_{j=1}^{N_2} |\rho_j|} \overline{CCT_j} \tag{9}$$

$$\overline{CCT} = \frac{N_1}{N_1+N_2} \overline{CCT_{MIC}} + \frac{N_2}{N_1+N_2} \overline{CCT_{\rho}} \tag{10}$$

where N_1 is the total number of selected functional nonlinear relationships, N_2 is the total number of selected linear relationships, MIC_i is the MIC score of relationship i, ρ_j is the PCC score of relationship j, $\overline{CCT_i}$ (or $\overline{CCT_j}$) is the single estimation result of $\overline{CCT}$ by relationship i (or j).

Table 7. $\overline{CCT}$ Estimation and Transient Security Assessment for New Operation States

No.	R (s)	R_1 (s)	R_2 (s)	R_r	R_l	S (s)
1	0.329	0.326	0.332	88.1%	1	0.330
2	0.321	0.321	0.320	77.9%	2	0.323
3	0.305	0.309	0.302	53.6%	3	0.310
4	0.286	0.288	0.283	27.7%	4	0.285
5	0.258	0.255	0.260	6.7%	5	0.255
6	0.328	0.324	0.333	87.0%	1	0.330
7	0.311	0.318	0.304	62.9%	2	0.314
8	0.289	0.294	0.285	31.1%	4	0.290
9	0.337	0.333	0.342	95.6%	1	0.340
10	0.335	0.330	0.339	94.1%	1	0.333
11	0.330	0.328	0.331	89.2%	1	0.329
12	0.321	0.324	0.317	77.9%	2	0.322
13	0.289	0.295	0.288	31.2%	4	0.292
14	0.260	0.256	0.264	7.5%	5	0.257
15	0.301	0.307	0.295	47.6%	3	0.309
16	0.292	0.298	0.285	34.9%	4	0.300
17	0.284	0.289	0.277	25.4%	4	0.285
18	0.274	0.279	0.269	16.0%	5	0.275
19	0.263	0.267	0.260	8.9%	5	0.266
20	0.251	0.252	0.251	4.3%	5	0.250

In the paper, $N_1 = 5$, $N_2 = 5$. The above method is applied to estimate the $\overline{CCT}$ for each new operation state. In the tests, 20 possible new operation states are set stochastically. The results are shown in Table 7. Some explanatory annotations for Table 7 are as follows.

1. R is the final estimation result, which is given by the program with the 10 selected variables as input features;
2. R_1 is given by the program with 5 variables as input features, which are corresponding to the selected top 1% of functional nonlinear relationships by MIC;
3. R_2 is given by the program with 5 variables as input features, which are corresponding to the selected top 1% of linear relationships by PCC;
4. R_r is the relative degree of transient stability;
5. R_l is the security level of transient stability;
6. S is the simulation value given by PSS/E, which is used to verify the accuracy of estimation results.

5. Conclusion

Transient security level has an effect on the continuous secure operation in electric power industry, which is essential for the automation of production. This paper has explored the connotative relationships of operation data and the general critical clearing time $\overline{CCT}$ in a large data set for power system, and then a novel online transient security assessment method is presented based on the explored relationships. $\overline{CCT}$ is defined based on the statistics data of historical faults. Power flow simulation for steady states and short circuit fault simulations of different positions are used to create a large data set of operation variables and $\overline{CCT}$. Each relationship of operation variable and $\overline{CCT}$ is given scores by MIC and PCC. The generalized nonlinear relationships exploration coefficient is presented to explore nonlinear relationships directly, including two kinds of models $MIC - |\rho|^{\gamma} (\gamma > 0)$ and $MIC - \alpha|\rho| (0 < \alpha \le 1)$. In the paper, $MIC - 0.4|\rho|$ is found to be more effective than other models and $MIC - \alpha|\rho| (0 < \alpha \le 1)$ is recommended in engineering since it can control the detracted linearity by parameter α. Some highly ranked linear and functional nonlinear relationships are detected out, shown and explained from the perspective of power system operation. Curve fitting is used for them to estimate $\overline{CCT}$ of new operation states. WD and GEVD are adopted for distribution fitting of $\overline{CCT}$. Transient security levels are divided based on the cumulative probability curve of GEVD. The $\overline{CCT}$ estimation results of new operation states are verified to be accurate, which can be used to assess the relative degree of transient stability and transient security level.

Different from conventional feature selection methods, the input features are selected from a great number of variables based on connotative relation-

ships exploration and data mining in this paper. It will be more intelligent and efficient than conventional optimization algorithms since each relationship is given a score and ranked clearly. The applicability of $\overline{CCT}$ estimation method will not be influenced by the change of the structure and scale because it is based on data statistics and mining. The estimation results are with high precision, which relies on the total number of previous operation samples to a certain extent. $\overline{CCT}$ estimation, relative degree of transient stability and transient security level assessment will be important bases for system operators of changing operation state to improve the security level in practice.

The online automatic identification of transient stability level can be implemented with the proposed automatic assessment processes and PMUs data, which is of great significance for the uninterrupted power production and supply in power industry. The proposed method is meaningful for power system operation and automation.

Conflict of Interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant (61074101, 50477018 and 51007093) and in part by the Specialized Research Fund for the Doctoral Program of Higher Education (20090141120062).

References

1. Z. Y. Dong, J. H. Zhao, and D. J. Hill, "Numerical Simulation for Stochastic Transient Stability Assessment," *IEEE Transactions on Power System,* vol. 27, no. 4, pp. 1741-1749, Nov. 2012.
2. D. Chatterjee and A. Ghosh, "Transient Stability Assessment of Power Systems Containing Series and Shunt Compensators," *IEEE Transactions on Power System*, vol.22, no. 3, pp. 1210-1220, Aug. 2007.
3. Y. Xu, et al, "Real-time transient stability assessment model using extreme learning machine," *IET Generation, Transmission & Distribution*, vol.5, no. 3, pp. 314-322, Mar. 2011.
4. A. M. Miah, "Study of a coherency-based simple dynamic equivalent for transient stability assessment," *IET Generation, Transmission & Distribution*, vol.5, no. 4, pp. 405-416, Apr. 2011.
5. K. Verma and K. R. Niazi, "Determination of vulnerable machines for online transient security assessment in smart grid using artificial neural network," *India Conference (INDICON), Annual IEEE*, 2011, pp. 1-5.
6. N. Amjady and S. F. Majedi, "Transient stability prediction by a hybrid intelligent system," *IEEE Transactions on Power System*, vol. 22, no. 3, pp. 1275-1283, Aug. 2007.
7. X. Tao, H. Renmu, W. Peng, and X. Dongjie, "Applications of data mining technique for power system transient stability prediction," *Proceeding of the 2004 IEEE International Conference on Electric Utility Deregulation, Restructuring and Power technology*, 2004, pp. 1389-392.
8. Z. H. Yu, X. X. Zhou, and Z. X. Wu, "Fast Transient Stability Assessment Based on Data Mining for Large-Scale Power System," *Proceeding of Transmission and Distribution Conference and Exhibition*, 2005, pp. 1-6.
9. T. W. Wang and L. Guan, "A data mining technique based on pattern discovery and k-nearest neighbor classifier for transient stability assessment," *Proceeding of Power Engineering Conference,* 2007, pp. 118-123.
10. R. Silipo and M. R. Berthold, "Input features' impact on fuzzy decision processes," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 30, no. 6, pp. 821-834, Dec. 2000.
11. T. I. Laine, et al, "Selection of input features across subjects for classifying crewmember workload using artificial neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 32, no. 6, pp. 691-704, Nov. 2002.
12. N. Kwak and C. H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143-159, Jan. 2002.
13. J.X. Chen and S. Wang, "Data visualization: parallel coordinates and dimension reduction," *Computing in Science & Engineering*, vol. 3, no. 5, pp. 110-113, Sep./Oct. 2001.
14. L. Shen and E. C. Tan, "Dimension reduction-based penalized logistic regression for cancer classification using microarray data," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 166-175, Apr. 2005.
15. T. W. Wang, L. Guan, and Y. Zhang, "A modified pattern recognition algorithm and its application in power system transient stability assessment," *Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*, *IEEE*,

2008, pp. 1-7.
16. I. Genc, R. Diao, and V. Vittal, "Computation of transient stability related security regions and generation rescheduling based on decision trees," *Proceeding of IEEE PES Summer Meeting*, 2010, pp.1-6.
17. N. Amjady, "Application of a new fuzzy neural network to transient stability prediction," *Proceeding of IEEE PES Summer Meeting*, 2005, pp. 69-76.
18. M. Cao and Y. J. Wang, "Application of binary particle swarm optimization in feature selection for transient stability assessment," *Proceeding of 2011 International Conference on Electric Information and Control Engineering (ICEICE)*, 2011, pp. 5719-5722.
19. X. Q. Zhang, L. Guan and T. W. Wang, "Kernel Feature Identification Based on Improved Ant Colony Optimization Algorithm for Transient Stability Assessment," *Transactions of China Electrotechnical Society*, vol. 25, no. 12, pp. 154-160, Dec. 2010.
20. K. R. Niazi, C. M. Arora and S. L. Surana, "Power system security evaluation using ANN: feature selection using divergence," *Electric Power Systems Research*, vol. 69, pp. 161-167, May 2004.
21. H. H. Al Marhoon, I. Leevongwat, and P. Rastgoufard, "A Practical Method for Power Systems Transient Stability and Security Analysis," *Proceeding of IEEE PES* Transmission and Distribution Conference and Exposition, 2012, pp. 1-6.
22. N. Yorino, A. Priyadi, H. Kakui, and M. Takeshita, "A New Method for Obtaining Critical Clearing Time for Transient Stability," *IEEE Transactions on Power System*, vol. 25, no. 3, pp. 1620-1626, Aug. 2010.
23. Z. Y. Dong, J. H. Zhao, and D. J. Hill. "Numerical Simulation for Stochastic Transient Stability Assessment," *IEEE Transactions on Power System*, vol. 27, no. 4, pp. 1741-1749, Nov. 2012.
24. N. R. David, A. R. Yakir, K. F. Hilary, R. G. Sharon, M. Gilean, J. T. Peter, S. L. Eric, M. Michael, and C. S. Pardis. "Detecting novel associations in large data sets," *Science*, vol. 334, pp. 1518-1524, Dec. 2011.
25. J. Benesty, J. Chen, and Y. Huang, "On the importance of the Pearson correlation coefficient in noise reduction," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 16, no. 4, pp. 757-765, May 2008.