A Data Analytics Framework for Regional Voltage Control

Duotong Yang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Electrical Engineering

Virgilio A.Centeno (Chair)

James S. Thorp

Jaime De La Reelopez

Pratap Tokekar

Steve C. Southward

July 11th 2017

Blacksburg, Virginia U.S.A

Keywords: Regional Voltage Control, Data Analytics, Decision Tree, Online Boosting. Wide Area Measurements, Data Mining, Network Partition, Machine Learning

A Data Analytics Framework for Regional Voltage Control

Duotong Yang

ABSTRACT

Modern power grids are some of the largest and most complex engineered systems. Due to economic competition and deregulation, the power systems are operated closer their security limit. When the system is operating under a heavy loading condition, the unstable voltage condition may cause a cascading outage. The voltage fluctuations are presently being further aggravated by the increasing integration of utility-scale renewable energy sources. In this regard, a fast response and reliable voltage control approach is indispensable.

The continuing success of synchrophasor has ushered in new subdomains of power system applications for real-time situational awareness, online decision support, and offline system dianostics. The primary objective of this dissertation is to develop a data analytic based framework for regional voltage control utilizing high-speed data streams delivered from synchronized phasor measurement units. The dissertation focuses on the following three studies: The first one is centered on the development of decision-tree based voltage security assessment and control. The second one proposes an adaptive decision tree scheme using online ensemble learning to update decision model in real time. A system network partition approach is introduced in the last study. The aim of this approach is to reduce the size of training sample

database and the number of control candidates for each regional voltage controller. The methodologies proposed in this dissertation are evaluated based on an open source software framework.

A Data Analytics Framework for Regional Voltage Control

Duotong Yang

GENERAL AUDIENCE ABSTRACT

Modern power grids are some of the largest and most complex engineered systems. When the system is heavily loaded, a small contingency may cause a large system blackout. In this regard, a fast response and reliable control approach is indispensable. Voltage is one of the most important metrics to indicate the system condition. This dissertation develops a cost-effective control method to secure the power system based on the real-time voltage measurements. The proposed method is developed based on an open source framework.

Acknowledgements

This dissertation is one of the milestones in my research career. I have been fortunate to learn theories and concepts which would have been impossible if I had not extensively carried out the needed research.

I would like to express my deepest gratitude to my advisor, Dr. Virgilio Centeno, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I would like to extend my sincere appreciation to Dr. James Thorp. He has been a mentor and an invaluable source of inspiration throughout my PhD research. I would not have been able to complete my PhD without his support.

I would like to say thank you to Dr. Jaime De La Ree, who throughout my educational career have supported and encouraged me to believe in my abilities. My gratitude also goes to Dr. Pratap Tokekar and Dr. Steve Southward, for their genuine interest and support. I would like to knowledge my friends and lab mates at the power lab and personals at Dominion Virginia Power. It is an honor to work with these great minds.

Last but not least, I would like to thank my parents, Binggui and Lishan, my brother Duojiong, my sister Weiyang, and my grandma Jingluan. In spite of the distance, they are always with me in all my endeavors.

I dedicate this dissertation to
my father Binggui Yang and my mother Lishan Zhao
for their constant support and unconditional love

Contents

List of Figures

List of Tables

Chapter 1
Introduction & Background

The continuously growing demand for electricity has forced modern power systems to operate closer to their secure operating margin. One of the great challenges for power electric utilities and regional transmission organization is being able to meet system-wide voltage security. Therefore, it is critically important for operators and coordinators to have a wide area visualization tool for situation awareness and decisions making. With continued proliferation of Phasor Measurements Units (PMUs) and machine learning techniques, controllers driven by PMU measurements and synchrophasor-based state estimator (SE) would be the key stones to achieve this goal [1] [2] [3]. A recent DOE funded project is also aiming to design a regional voltage controller that providing open-loop control decisions within a control region with dense PMU installations. It is also expected that this new voltage controller be able to minimize the control decision per device per day. This dissertation investigates a new intelligent framework for developing this PMU data-driven controller.

1.1 Motivation & Objective

The study of voltage control for transmission grids has a long history, different methodologies had been developed in the past century. The hierarchical voltage control based on pilot buses selection was successfully applied in French grid back to 1980s [4] [5]. Researchers proposed a zone-based voltage control method [6] [3] and successfully implemented in China's national grid [7]. In North America, PJM as one of the largest RTO, is providing coordinated control by solving optimal power flow (OPF) in constraints of transmission losses and voltage profiles, reactive power reserves, etc. [8].

Computing single OPF for pre-contingency and post-contingency leads to an unimaginably high number of dimension variables as well as constraints. It would be difficult to find an acceptable solution method that meets convergence, robustness requirements, and more importantly allows for practical real-time implementation [9].

Instead of solving one single constrained OPF, researchers from Denmark and United States proposed a machine learning based control method [10]. This proposed methodology includes two decision trees (DT), one for security assessment and another one for online decision support for generator rescheduling. In the beginning, the insecure operating condition (OC) is detected by the first DT, then the final control trajectory is searched out from numerous new OCs based on different generator output with minimum economic cost.

While this approach provides faster online situational awareness and controllability, its reliability and accuracy are highly depending on the methodology used to generate the learning data set of OCs. In addition, there are two primary issues inherent in DT-based methodology should be considered. First, if the

framework is applied to discrete control devices across the whole system, the number of control combinations and their corresponding OCs will be tremendous. Second, the topology changes due to forced outages of network component could result in difference between the offline learning database and the real-time operating conditions, and ultimately give rise to inaccurate classification decisions. It is necessary to establish a timely update scheme for the trained model in order to adapt to the modified system configurations and new operating conditions.

To provide a cost-effective and robust decision for regional voltage control, a control methodology based on parallel DTs is studied in this dissertation. In this methodology, each DT provides voltage security assessment (VSA) for each control decision. In regarding the frequent change of system configurations and high-speed data streams, a tree update approach for trained system model is developed [10]. This approach employs adaptive ensemble decision-tree learning method to update decision tree models in real time [11]. Finally, a network partition technique is introduced to break down the power system network into non-overlapping set of coherent bus groups with unique voltage stability problem [12]. The goal of this partition method is to reduce system model as well as number of control candidates to be considered, and eventually reduce the size of training database for each DT.

1.2 Enabling Wide-Area-Monitoring

Since the development of the first phasor measurement unit (PMU) at Virginia Polytechnic Institute & State University back in early 1980s [13], entities across industries and academia have being conducting research and development to explore the applications of this technique for more than 30 years. In the recent decade especially since the 2003 Northeast blackout [14], utilities in United States started to actively deploy PMUs throughout the power grid. One of the key applications of

PMU is wide area measurement system (WAMS) which consists of PMUs placed in power system substations.

Consider the dynamic nature of power system: random load variation [15], frequent topology changes, any possible system contingencies, fast switching [16] or disconnection of electric components etc., lack of wide-area visibility prevents early identification of any possible cascading events. Without WAMS, utilities rely on current SCADA system which provides system conditions every 4-6 seconds leaving them with no information and sequence of the dynamic events on the grid during a short period of time.

In contrast, synchrophasor data provided by PMU allow the system to collect and share high speed, real time, and time-synchronized grid conditions across the entire system. If multiple PMUs are placed in power system, their outputs are streamed to phasor data concentrators (PDC) within substation level and transferred to PDCs in regional level that matching their time stamps and provide snapshot of the system states. Different applications that providing system monitoring, protection, and control can be implemented at each PDC.

One of these key applications is the commercialized linear state estimator (LSE) developed in Virginia Tech in 2013 [17]. This application exclusively employs PMU data to provide a three-phase linear tracking state estimator. This application is implemented in Dominion Virginia Power's 500 kV network.

1.3 Local Voltage Control

Local voltage controller runs on the EMS system. It provides automatic voltage control regulation and reactivate power support for transmission network. The control is accomplished by controlling Load Tap Changers (LTC) in transformers, capacitor banks, static VAR compensator, etc. The control strategies use locally

available information, such as bus voltage magnitude measurements, reactive power flow through local transformer, or reactive power output from local generator.

Since the controller is designed to perform control locally, its impact to the other substations or regions is limited. Therefore, the number of taps of control transformers, size of capacitor bank, and the delay parameter for each control are highly system dependent.

1.4 Co-ordinated Regional Voltage Control

Voltage control and reactive power support are long recognized as a critical problem in power system operation. Traditional voltage control strategies rely on independent substation level voltage/reactive power set point that is commissioned since the infrastructures installation. That means the system upgrades with configuration changes imposed challenges to guarantee the electric power system reliability. Therefore, it becomes necessary to manage the generation, transmission, and consumption of electric power in a regional manner.

"Primary" voltage control which relies on automatic voltage regulators (AVR) of synchronous machines, load tap changing transformers, and automatic capacitor banks, acts to compensate for small voltage and reactive power demand disturbance. Since primary voltage control is inadequate to deal with large changes in system state, operators can control the reactive power support devices which are available for manual operation. Such control is called "secondary" voltage control [18]. It is noted that this type control heavily relies on operating personnel. What is more, the control scheme assumes negligible interactions with the neighboring regions. Therefore, an intuitive reduced information structure voltage control should be employed to co-ordinate multiple "secondary" voltage regulations in a system level.

In this dissertation, the proposed system wide voltage control remains at utilities level.

1.5 Power System Analysis using Decision Tree

Thanks to massive investments on wide area measurement system, power grid monitoring and control is experiencing a transition from human centric to a data centric decision making process. Data mining was identified as a key approach for development of data centric power grid as early as 1968 [19]. However, the investigation of DTs for voltage security assessment sparked interest on in the late eighties [1] [13]. In the early nighties, Rovnyak and Thorp [5] demonstrated how DTs can be constructed off-line and then utilized on-line for prediction transient stability in real time. Due to the wide deployment of PMUs in the recent decade, real time system security assessment combining DT and synchrophasor measurements became possible. For example, in 2007, Kai [21] proposed an PMU-based online dynamic-security-assessment scheme for large-scale interconnected power systems using DT. Since then, this technique was also demonstrated for islanding detection [22], system adaptive protection [2], and system preventive control [7].

Though many different data-mining based scheme such as neural networks [23] and support vector machine [24] have been tried on these problems, DT still remains the prominent machine learning tool for system security assessment. This is due to the “white box” nature of DT, which satisfies the principles for security assessment. Model users, including policy makers and system operators, tend to favor this type of interpretable model not only for reliable prediction but also for the after-the fact auditing and forensic purposes.

The recent studies of ensemble learning demonstrated more robust and accurate mining power for power system study [25] and raised a discussion between

the transparency and accuracy trade-off. Innocent Kamwa [4] suggested that for tasks requiring high performance model (load forecasting), ensemble learning methodology will be more acceptable. For tasks requiring clearly identifiable decision making (protection), models that provide better transparency are necessary.

1.6 Organization of the Dissertation

The dissertation is organized as follows:

Chapter 1: Introduction & Background - This chapter explains the background of technologies involved in the proposed methodology. Detail introduction about wide area monitoring, coordinated voltage controller, and decision-tree based controller are presented. Finally, this chapter concludes with the organization of the dissertation.

Chapter 2: Voltage Security Assessment – The concept of voltage security assessment is introduced. The methodologies of how to determine the system security boundary and provide security assessment are presented. A simulation is conducted based on proposed methodology on the PSSE test system.

Chapter 3: Parallel-Decision-Trees based Voltage Security Assessment and Control – Presented in this chapter is the development of decision tree based voltage security assessment. Based on the security assessment approach, the parallel-decision-trees based voltage control scheme is presented. The proposed voltage controller is evaluated using IEEE 118 bench mark model.

Chapter 4: Adaptive Decision Tree for Regional Voltage Control –A review of ensemble learning, offline AdaBoost, and online boosting is presented. This is followed by an introduction of adaptive decision tree framework for voltage security assessment and control. A more robust weak learner induction using Hoeffding tree

is presented. The efficacy of this adaptive decision tree framework is evaluated by introducing topology change scenario in the IEEE benchmark model.

Chapter 5: Power System Network Partition for Regional Voltage Control – This chapter firstly illustrates the concept of voltage control areas identification and the algorithm based on structural weakness. Then it introduces a new data analytic based approach to properly select the parameters for determining the size and number of voltage control areas. The simulation is conduced based on the IEEE 118 bus system. The result shows that the model is partitioned into two sub-systems. The buses in each of the sub-systems are sharing the similar voltage security pattern.

Chapter 6: Structure of Voltage Controller Software Framework – This chapter explains the architecture of a generic software framework for both local voltage and regional voltage control. Then it presents a customized design for a regional voltage controller. Finally, the software testing methodology is presented.

Chapter 7: Summary – The contributions of the dissertation are summarized. The vision of improving this framework and a concept of real time simulation platform are shown.

Chapter 2
Voltage Security Assessment

Voltage collapse is a critical problem that threatens power system secure operations. Usually, it is associated with: 1. a continuous load increase; 2. a critical contingency. During a voltage collapse event, the system will suffer a low voltage problem in a localized area where it has deficient reactive power support. Without timely operation actions, the low voltage problem may cause a large-scale blackout instead of a localized outage.

One key point in the voltage security assessment (VSA) is the identification of loadability limits. A very commonly used method is the PV curve method, the knee point of a PV curve indicates the maximum power transfer without causing voltage instability. Further increase in the load over the knee point will cause the progressive voltage decline. These knee points are bifurcation points of the nonlinear power system model which can be seen in [1].

Translating the loadability limit into mathematical terms, they correspond to a scalar function ζ of P, where P represents the load demand variables, e.g. $\zeta = aP + Q$. We can consider it as an optimization problem:

$$\max_{\mathbf{p},\boldsymbol{\mu}} \ \zeta(\mathbf{P}) \tag{2. 1}$$

Subject to $\boldsymbol{\varphi}(\boldsymbol{\mu}, \mathbf{P}) = 0$

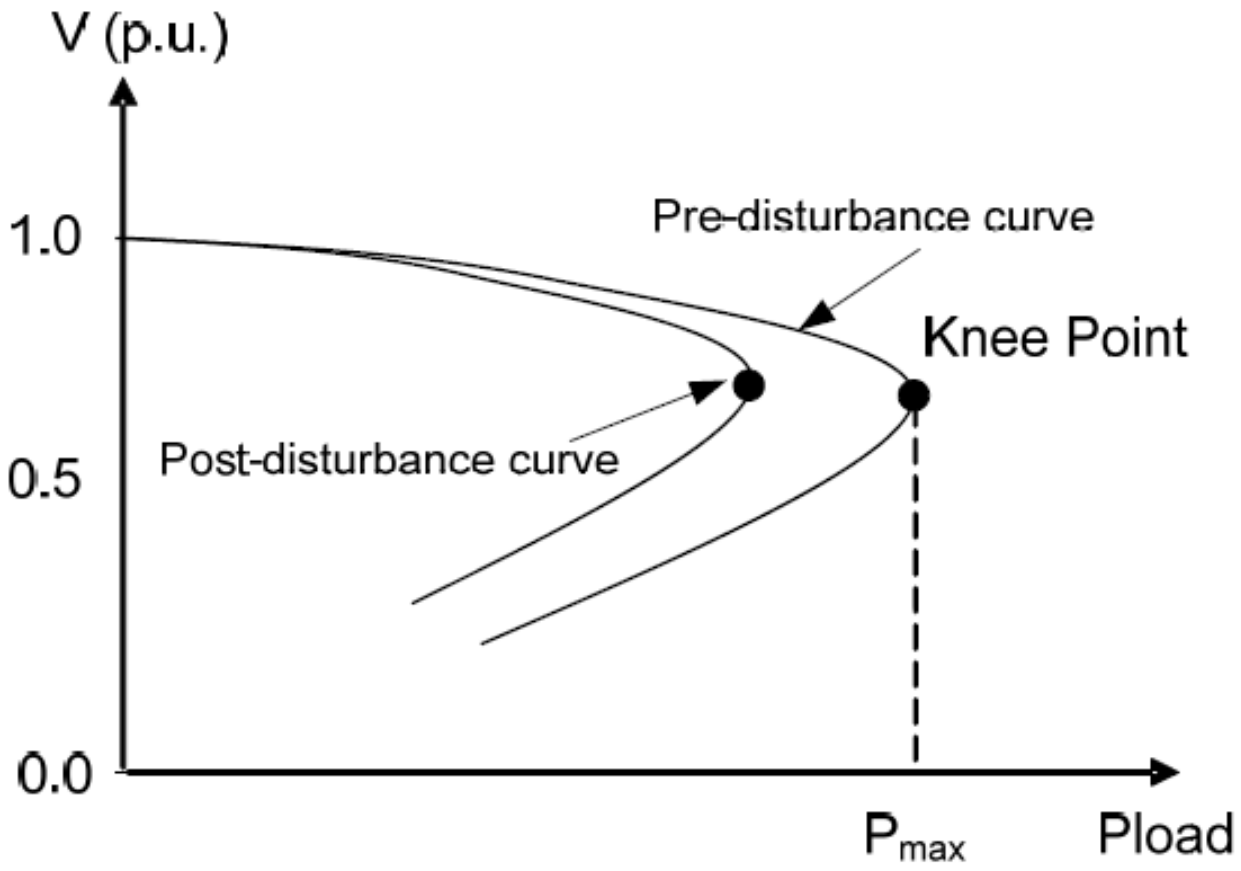


Figure 2-1 PV Curve Indicates Loadability Limit

where $\boldsymbol{\varphi}(\boldsymbol{\mu}, \mathbf{P}) = 0$ represents the a set of N algebraic equations in N algebraic variables. $\boldsymbol{\mu}$ denotes a set of state variables, for instance, $\boldsymbol{V}$ and $\boldsymbol{\theta}$.

To solve this problem, we define the Lagrangian:

$$\mathcal{L} = \varphi(\boldsymbol{\mu}, \mathbf{P}) + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\varphi}(\boldsymbol{\mu}, \mathbf{P}) = \varphi(\boldsymbol{\mu}, \mathbf{P}) + \sum_{\mathrm{i}} w_i \varphi_i(\boldsymbol{\mu}, \mathbf{P}) \quad \textbf{(2. 2)}$$

where $\boldsymbol{w}$ is a vector or Lagrange multipliers. As shown in [1], as a loadability limit, the Jacobian of $\boldsymbol{\nabla_{\mu}\varphi}$ of the steady state equations $\boldsymbol{\varphi}(\boldsymbol{\mu}, \mathbf{P}) = 0$ is singular which means that the system has reach its loadability limit. The system is voltage stable before it reaches its loadability limit.

2.1 Determine the Secure Operation Limit

Even when the system state is voltage secure for a given OC, it is desirable to know how far the system is away from its loadability limit. Therefore, it is necessary for us to determine a security margin. The security margin rely on the large deviations of parameters, e.g. system load increase or generation re-dispatch. The large deviations, also called "stress", are characterized by a "direction" in the parameter space. According to reference [1], the margin can be called either (1) post-contingency limit, or (2) secure operation limit. The first type of margin is determined by applying a given contingency and then stress the system along the specified direction. For the second type of margin, the system is stressed as far as it can while it is able to withstand some specified contingencies.

For our case, since the low voltage and voltage insecure problems are due to the deficiency of reactive power support and heavy summer load, the security margin should be determined based on the second type. Besides some specified contingencies, the OCs will be considered as insecure if they cannot satisfy other aspects of power system security: e.g. line's thermal limit, angle difference of

between buses, high voltage violation, low voltage violation, reactive power reserves, etc.

2.2 Determination of the Security Boundary

Power flow divergence can be considered as voltage instability. However, in some special cases, power flow divergence may result from purely numerical problems that do not associate to a physical voltage instability [1]. In addition, if it is a truly unstable case, we are left without any information regarding the nature and location of the problem. Therefore, it would be reasonable to create a new boundary called security boundary that shows the system getting closer to collapse. The concepts of security boundary, secure operation limit, and loadability limit are shown in Figure 2-2 (assuming the system has only two PQ buses).

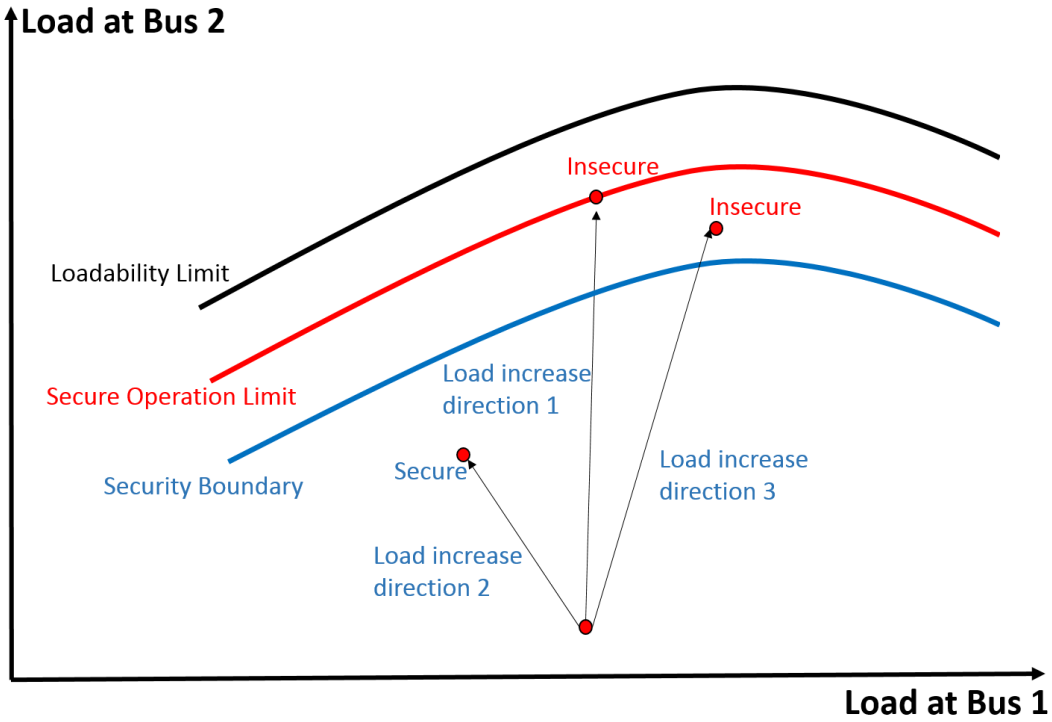


Figure 2-2 Boundary indicating voltage instability

Numerous researches have conducted studies to find he security boundary. L-index proposed in reference [26] is able to give us a sign that indicates if the system is becoming voltage unstable. Reference [27] computes a parameter vector which is able to indicate how far away the system is from the voltage instability. In this work, the boundary is determined by computing the Euclidean distance between each single OC and all the insecure OCs. The idea is illustrated as follows:

Assuming the secure operation limit and loadability limit are overlapping, the power system can be modeled by the following equation:

$$\boldsymbol{\varphi}(\boldsymbol{\mu}, \mathbf{P}) = 0 \tag{2. 3}$$

$$\mathbf{P} = [\mathbf{P}_{\mathrm{bus}_1}, \mathbf{P}_{\mathrm{bus}_2} \dots \mathbf{P}_{\mathrm{bus}_m}]$$

where $\mathbf{P}$ is a parameter vector of real and reactive load powers at all the PQ buses. It is assumed that the current operating power $\mathbf{P}_0$ at which the corresponding equilibrium $\boldsymbol{\mu}_\mathbf{0}$ is stable. As the load powers vary, the equilibrium $\boldsymbol{\mu}$ will change in the state space. Based on our assumption, when the load consumption reaches a critical value $\mathbf{P}_*$, the system can lose its stability by $\boldsymbol{\mu}$ disappearing in a saddle node bifurcation point. Such set of critical load power $\mathbf{P}_*$ are denoted as a voltage stability margin $\sum$ which can be seen in Figure 2-3 (assuming the system has only two PQ buses).

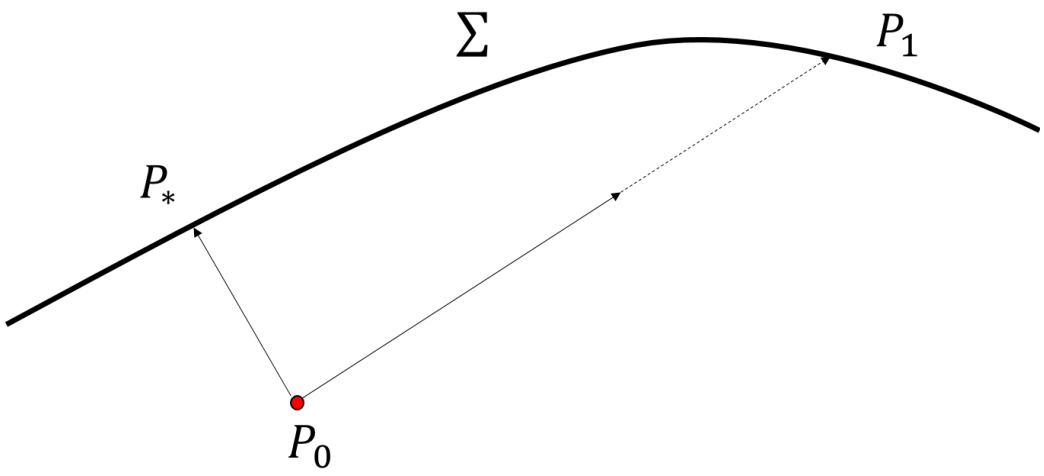


Figure 2-3 insecure point $\boldsymbol{P_0}$ and the voltage stability margin $\sum$

In the proposed methodology, an index is created to measure distance between current OC $\mathbf{P}_0$ to voltage stability margin $\sum$ by computing shortest Euclidean distance $\|\mathbf{P}_* - \mathbf{P}_0\|$. Note that the direction of the power change is not required, since the OCs will be marked as insecure once the shortest distance is less than a threshold value [8].

2.3 Pre-contingency Controls

Prior to any contingency, operators react to stress imposed on system when the OC enters security boundary. Usually, these control actions are used to keep the voltage profile within limits and to maximize the reactive reserves readily available to face incidents. Typical types of such actions include:

- Shunt/series capacitor switching
- Discrete operation of SVC and Statcom
- Adjustment of LTCs on transformers

As can be seen from Figure 2-4, it is beneficial to determine the needed actions if we know the exact security boundary. A list of control candidates is considered,

each of them corresponding to a known direction of control which is characterized by the distance between the insecure OC and the secure OCs bounded by the security boundary.

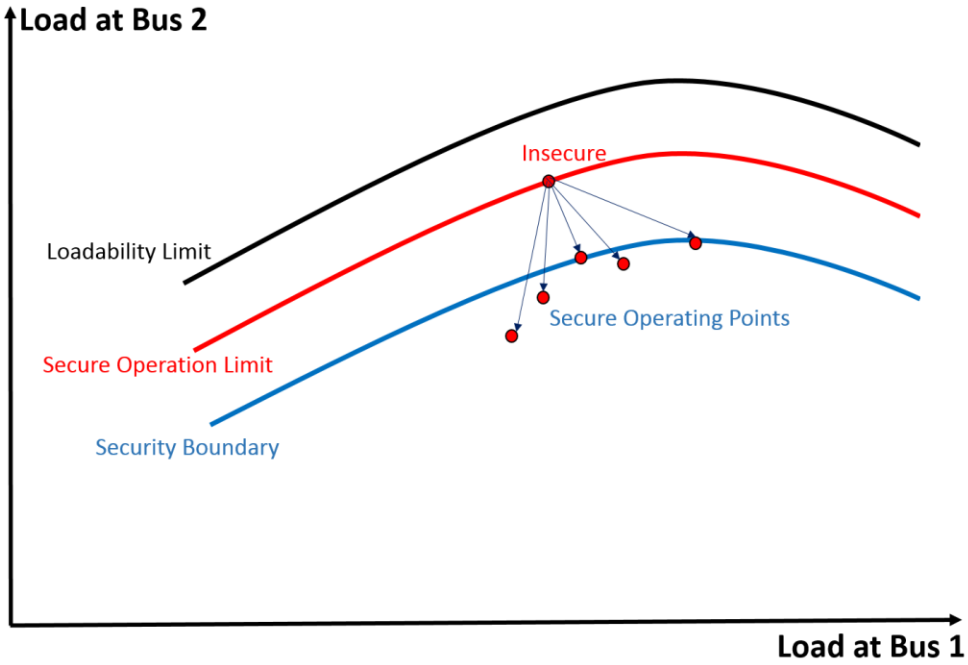


Figure 2-4 OCs after different control

2.4 VSA for PSSE test system

In this simulation, we made the following assumptions:

- The secure operation limit and loadability limit are overlapping, which means that the system becomes insecure when it reaches the bifurcation point.
- Contingencies are not considered in the simulation.
- The model is divided into 3 regions which can be seen in Figure 2-5. PQ buses in the same region will have the same percentage load change.
- Shunt capbank at bus 203 is switched off in the base case
- 10,000 OCs are generated based on the load increase in different regions.

- Bus voltage less than 0.9 pu is assumed as low voltage
- The generation re-dispatch is not considered and thus the swing bus generates power to meet the load demand.

The OCs that close to security boundary are determined based on the following

$$d_i = \|\mathbf{P_{insecure}} - \mathbf{P_i}\| \quad \textbf{(2. 4)}$$

$$d_{minimum} = \min(\boldsymbol{d}) \quad \textbf{(2. 5)}$$

where $\boldsymbol{d} = [d_1, d_2 \ldots, d_i]$

$if\ d_{minimum} < \delta$, the OC will be considered as insecure OCs. In this case, $\delta = 1$.

A three dimensions view of system OCs is shown in Figure 2-6 and Figure 2-7. The OC is considered as unstable when the power flow computation diverges or the power flow iteration exceeds a certain limit. Different colors show different OC status: e.g., “blue” indicates that the system is secure; “red” shows that the system is stable but it has low voltage; “magenta” means the system is getting closer to unstable limit; “black” represents the unstable system condition. Figure 2-6 shows the OCs of test system with no control operated (all capbanks are in off status). Figure 2-7 shows all of the scenarios of the system when shunt capbank at bus 203 (50 MVAR) is switched on. It can be seen that when the capbank is switched on, more reactive power is supplied to the system and thus less unstable OCs are shown. Table 2.1 shows the numbers of all different OCs for both scenarios.

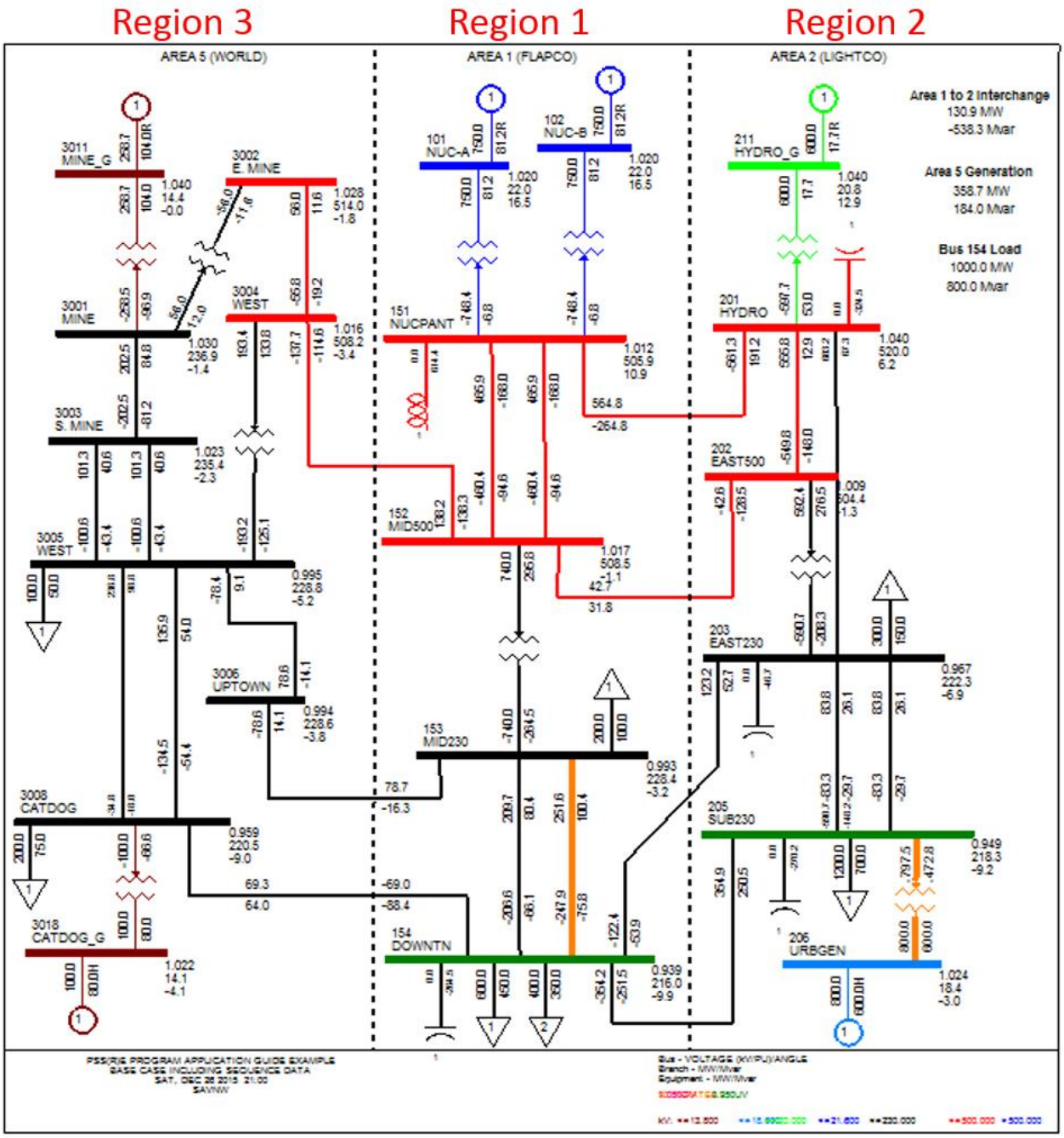


Figure 2-5 PSSE Test System

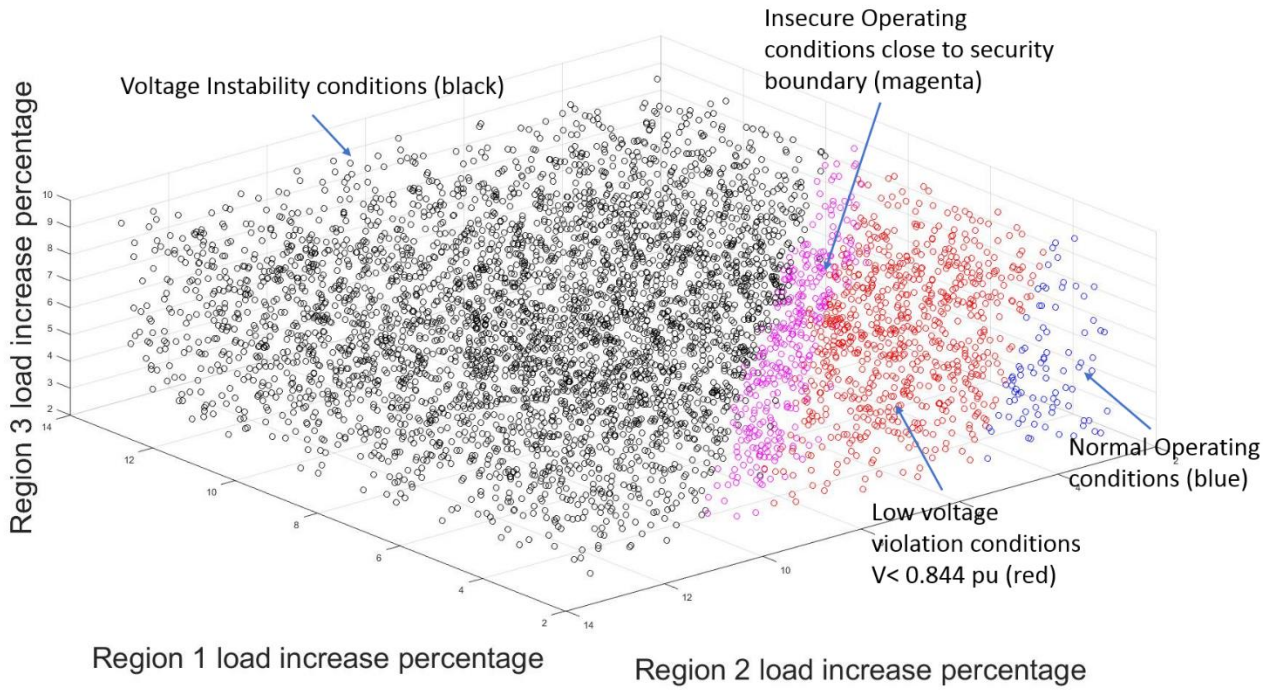


Figure 2-6 OCs before control PSSE Test System

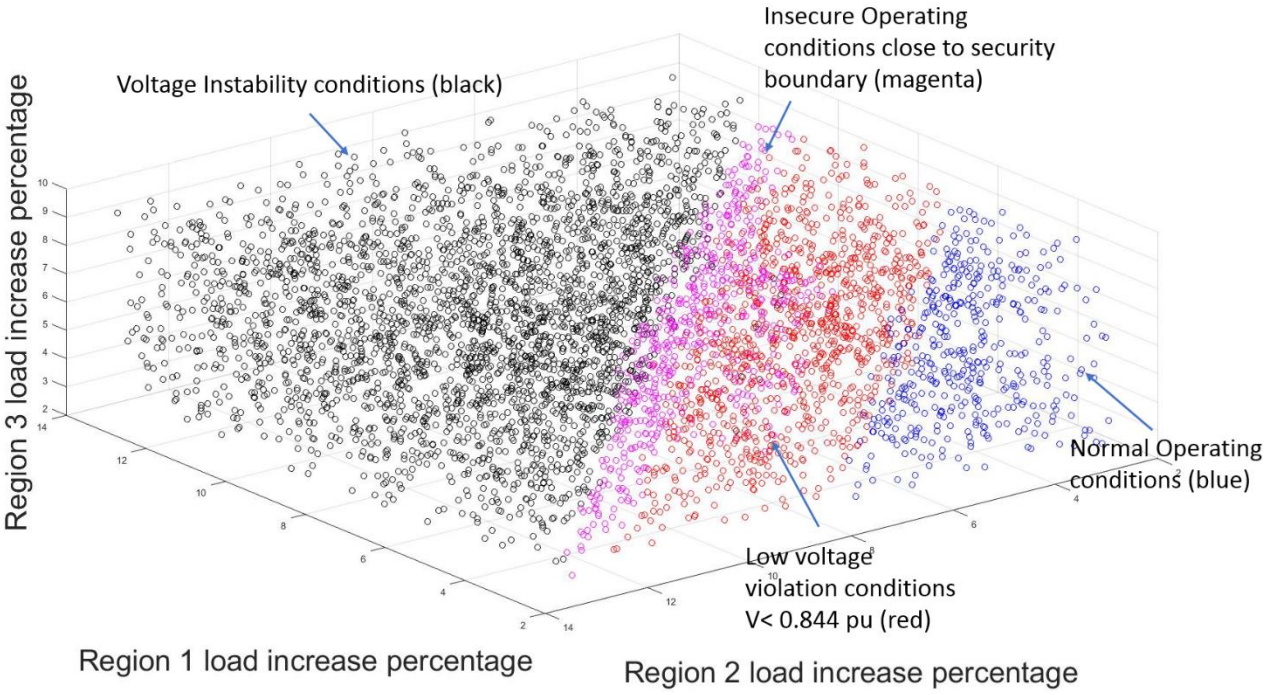


Figure 2-7 OCs after control PSSE Test System

Table 2.1 Simulation summary of PSSE test system

Conditions	Number (before control)	Number (after control)
Secure OCs	172	888
Low voltage OCs	1486	2076
Insecure OCs	686	1176
Voltage instability OCs	7656	5866

2.5 Summary

This chapter firstly introduced the concept of voltage security assessment and power system secure operation limit. Since when the system approaches the secure operation limit voltage collapse may occur, that leaves us without any information, by identifying power system security boundary we are able to provide situational awareness for system operators. In this work, the voltage security boundary is approximated by finding the insecure OCs whose Euclidean distance to the secure operation limits are less than a user-defined threshold value δ. The simulation is conducted based on PSSE test system. Two different scenarios for pre and post-control system operating conditions are presented. The voltage security boundary is clearly viewed using such method. To perform an accurate and timely evaluation of VSA for a certain OC, a supervised learning method is introduced and implemented in the next chapter.

Chapter 3
Parallel-Decision-Trees based Voltage Security Assessment and Control

Decision tree (DT) learning was first introduced as a predictive modeling approach which is commonly used in statistics, data mining, and machine learning. The goal of DT is to create a model that predicts the value of a target variable based on several input variables. A DT can be generated or "learned" by splitting the source set into subsets or a predefined set of classes based on recursive partitioning. The process of recursive partitioning is to partition the space into disjoint subsets by increasing the "purity" of each class. In order to compare each potential splitting attribute, several functions are introduced to measure the node's impurity: Gini index, Entropy, Towing, CHAID, etc.[28] The splitting is completed when the subset of node has all the same value of the target variable or a specific stopping criterion is met (e.g., the minimum size of DT).

DT technique is also widely implemented in the power system industry. Model users such as energy policy makers, power system operators, and utilities reliability engineers tend to favor its transparency and interpretability. As shown in reference [5, 6] DT is implemented online for predicting transient stability in real time. In reference [7], DT based voltage security assessment is introduced to characterize the vulnerability of the current operating condition. Reference [2] shows a DT based security/dependability adaptive protection scheme which aims to reduce the likelihood of manifestation of hidden failures. A two-DTs based systematic approach is adopted to provide preventive control for power system in Denmark [7].

One of the key goals of this dissertation is to develop a DT based regional voltage controller. This chapter develops that controller by first proposing a DT based voltage security assessment methodology. Several parallel DTs corresponding to a set of control combinations are trained. Each DT decides the security status of a post-control system operating condition. Since possible configuration change of system will create more new operating conditions, a more effective online DT training methodology is introduced in this dissertation.

3.1 Decision Tree based Voltage Security Assessment

As discussed in Chapter 1, increase in electric power demand forces the power system to operate with narrower margins of stability. To better understand the distance between stability margin and current operating condition, DT based methodology is introduced to perform voltage security assessment. Classification and regression trees (CART), the two main types of DT are effective data mining tools to handle prediction problem by capturing the mechanism hidden inside the data. In the past decades, they have been widely used as alternatives to other learning methods, such as linear regression, discriminant analysis, and other method used

based on algebraic models. As shown in Figure 3-1, trees are directed graphs that start with one tree node and branch to many. The root node contains the whole learning sample while each remaining node contains a subset of learning sample in the node directly connected above it. The classification decision is made at each

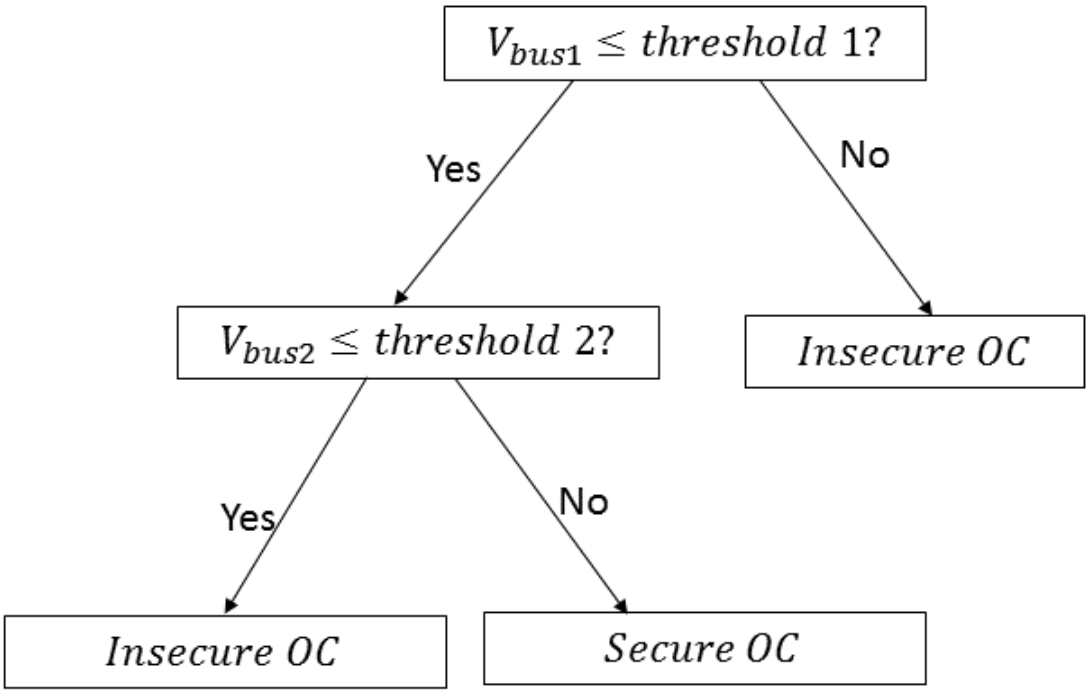


Figure 3-1 DT based VSA for two load busses system

terminal node that holding a class label. For DT based VSA, the terminal node are labeled with OC security status: Secure/Insecure. The number of objects below each terminal node indicates the number of misclassified cases. It is noted that the tree is binary since each node is split into only two subsets.

DT is trained based on exhaustive searches on all numeric input attributes. The process begins at the root node which encloses the learning sample. The splitting is conducted by measuring the "purity" or entropy of each subset in a recursive manner which is known as recursive binary partitioning. A split is said to be optimal when its child nodes have minimal entropy:

$$Entropy(\boldsymbol{S}) = -p_s \log_2 p_s - p_i \log_2 p_i \quad (3.1)$$

where $\boldsymbol{S}$ represents the learning database for the DT while p_s and p_i indicates the secure and insecure proportions of $\boldsymbol{S}$. This process is repeated on each derived subset. The process is stopped when the subset at a node has all the sample value of the target variable. However, such exhaustive DT can over fit to the training data. To avoid overfitting, different pruning and early stopping methods for producing a smaller tree are introduced [29].

Complicated decision trees might have a limited generalization capabilities, i.e., though it classify all training instances, it fails to do so in a new and unseen data set. [28]. To estimate how the structured DT can generalize to unknown dataset or a real problem, one set of sample data are used to build the tree, a disjoint set from the sample data is seperated for evaluation. k fold cross-validation is implemented in this study to evaluate the decision tree result. In k fold cross-validation, the original learning sample is partitioned into k equal size sub-data-sets. For each validation, a single subset of the data is used for DT training while the remaining $k - 1$ subsets are used for testing. The concept of k fold cross-validation is shown in Figure 3-2.

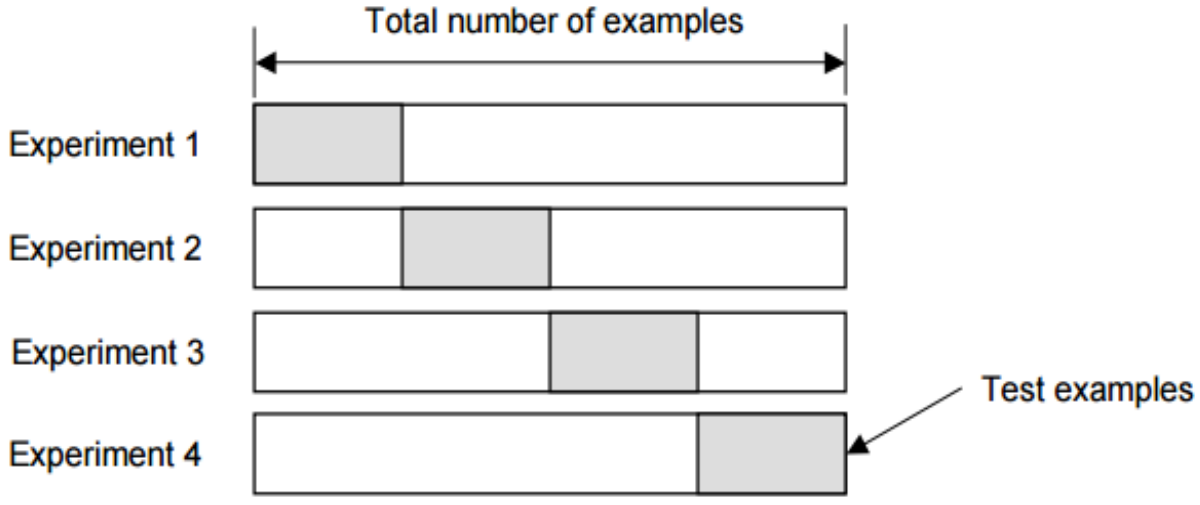


Figure 3-2 K fold cross validation

3.2 Initial Learning Database Preparation

The learning sample is set of data with m rows and $p + 1$ columns, where m denotes the number of OCs and p represents the number of measurement vectors $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_p}\}$. $\boldsymbol{v_1}$ represents a column of measurements for an attribute. The fist column contains the class of each measurement vector $Class = \{C_1, C_2, \dots, C_m\}$ known prior to the tree training process. Table 3.1 shows a typical setting of learning sample for the advocated method.

Table 3.1 Example learning database

OCs	**Class**	$\boldsymbol{V_{bus1}}$	$\boldsymbol{V_{bus2}}$	**…**	$\boldsymbol{V_{busp}}$
1	Secure	1.02	0.98	…	1.01
2	Insecure	0.91	0.96	…	1.03
3	Insecure	0.90	0.89	…	0.94
…	…	…	…	…	…
N	Secure	1.01	1.00	…	0.99

3.3 Parallel Trees for Post-Control Decision

DT can be also implemented as a predictive model which maps the measurements and post-operation decision. The advocated control methodology grows decision trees for post-control decision making.

All of the control options in the system are treated as independent control decisions. Assuming the system has M control options, there will be 2^M control combinations (including the one with no control operated). For a system with 6 controlled capacitor banks and with one capacitor bank operating at the beginning, the control combination is[1,0,0,0,0,0]. The total number of decision combinations would be 64. For each control combination, a learning database for both training and

testing is generated. The label for each OC is the post-control OC status, which is also determined by offline VSA, introduced in Chapter 2.

For each learning sample database, "Secure" indicates the control is acceptable for the current OC with state reflected by the measurements. A "Secure" OC should be able to satisfy multiple voltage security constraints, shown in Table 3.2. For 2^M combinations, there could be 2^M parallel Decision Trees (DTs).

Table 3.2 Constraints for Voltage Security Violation

	Constraints
1	Reaches the security boundary
2	Low Voltage Violation
3	High Voltage Violation
4	Line Thermal Limit
5	N-1 Contingency stable
6	Minimum Reactive Power Reserve
7	Power Flow Divergence

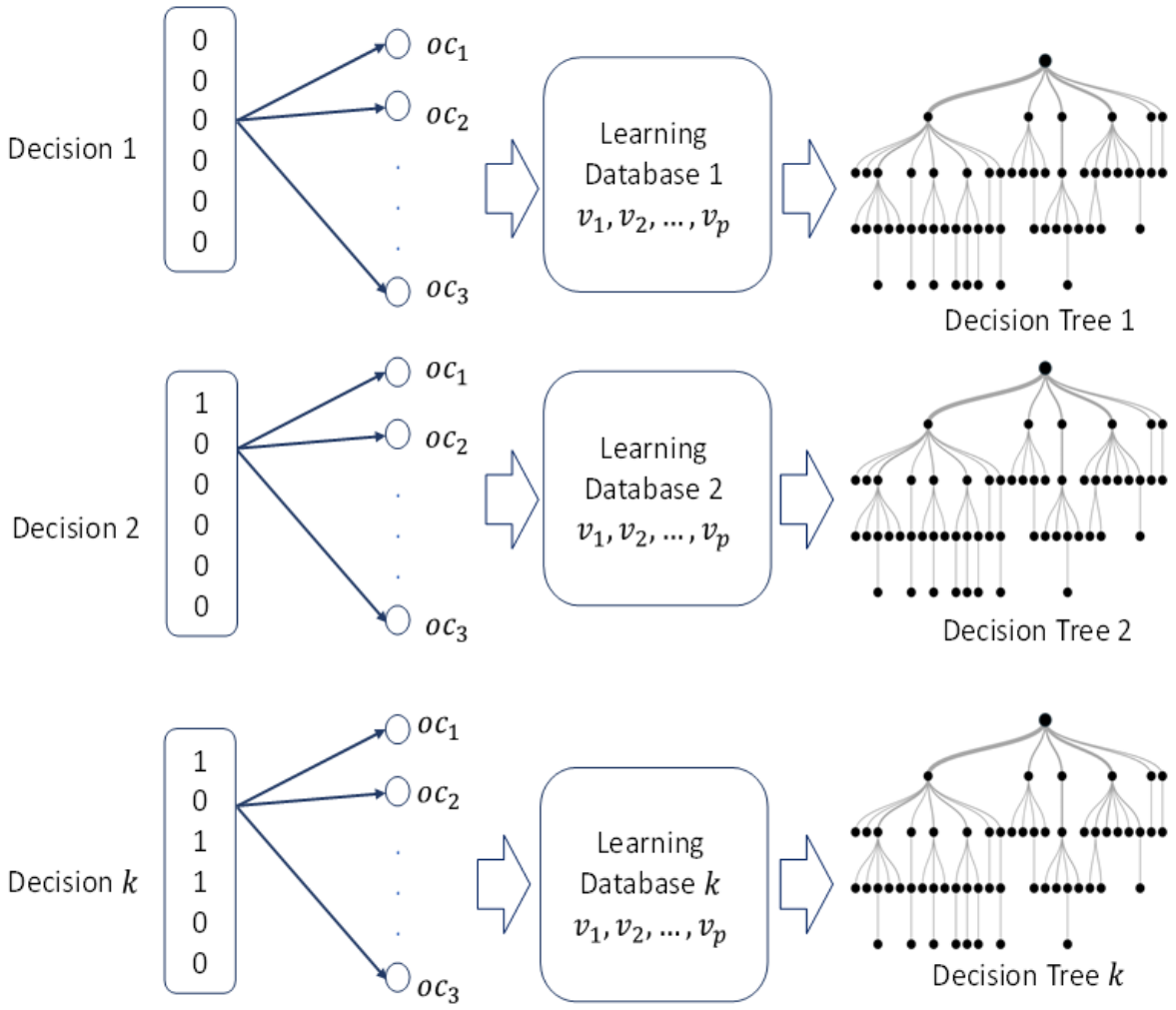


Figure 3-3 Learning Databases for Post-Control

The concept of parallel-tree-based control and the logics of the proposed controller are respectively shown in Figure 3-4 and Figure 3-4. Voltage measurements from PMUs and LSE collected in real-time continuously provide snapshots of the system at each time stamp. Collected measurements will fall into the first tree for VSA only, and will be compared with the critical splitting nodes inhere in the tree. If the terminal node indicates the current OC is "insecure", the rest of the trees will be activated simultaneously. Otherwise, if the tree provides "secure" control decision, the control decision with the tree is selected; if more than one tree provide "secure", the tree with fewer operations involved is selected.

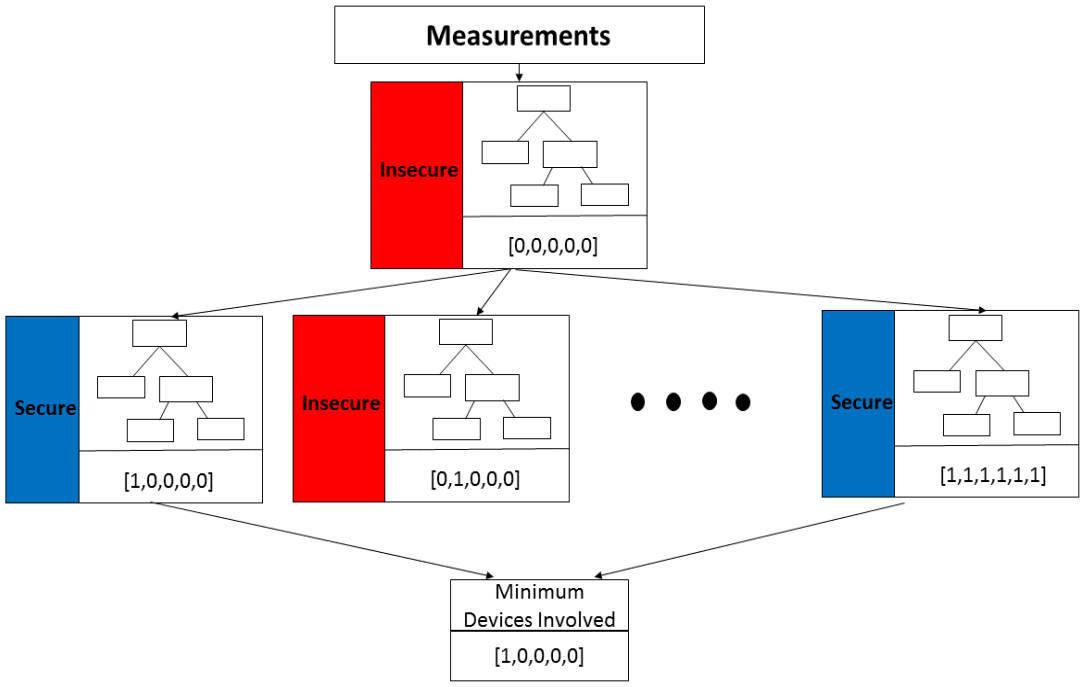


Figure 3-4 Logic of parallel tree based voltage control

3.4 Simulation based on IEEE 118 Bus System

The parallel-trees-based voltage control is implemented in IEEE 118 bus system which can be seen in Figure 3-5. In this simulation, we have the following assumptions:

1. The IEEE 118 bus system is used for case study. The system is divided into 3 different areas as suggested in [27]. Load buses within each area are assumed to have the same loading pattern that the load is scaled up and down in the same percentage.
2. Power factors are constant for all PQ buses
3. The generation re-dispatch is implemented by scaling the same amount of $Pgen$ as load for each region. (Since the generation changes depends on the independent load random variables via an economic dispatch operation, $Pgen$ is actually a dependent variable.)

4. Bus voltage less than 0.844 pu is assumed as low voltage. Noted that 0.844 pu is less than the requirement suggested by PJM (0.95 pu), it is utilized for simulation purpose only.
5. Voltage magnitudes at all buses are selected for learning database generation.
6. In this case, the control options are fixed capacitor banks only, located at buses 34, 44, 45, 48, 74, and 105. The capacity of each capacitor bank is shown in Table 3.3. In the initial condition, all selected fixed capacitor banks are switched off.

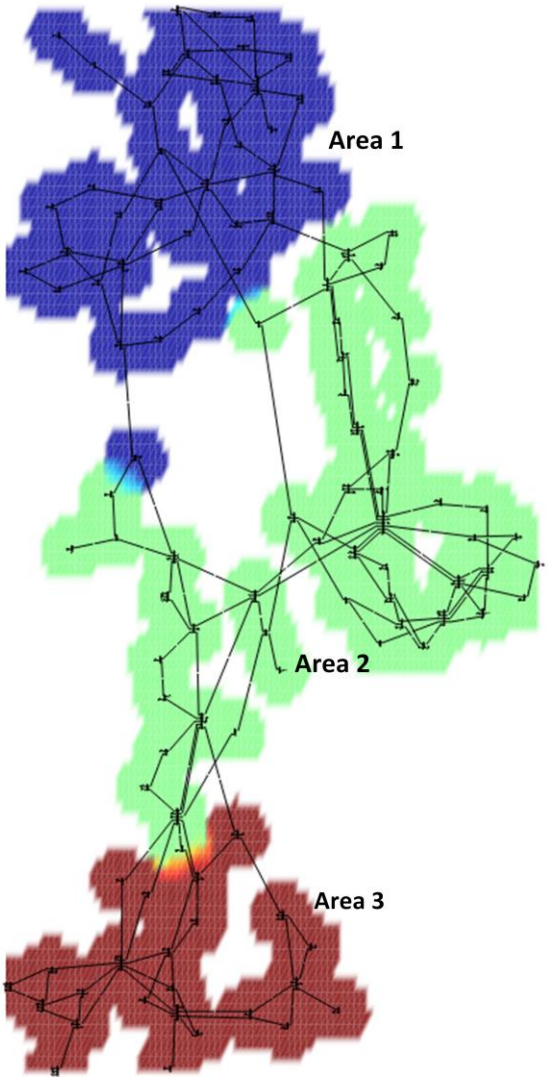


Figure 3-5 IEEE 118 bus system

Table 3.3 Capacitor banks available for control

Bus Number	Capacity of Capbanks (MVAR)
34	50
44	40
45	40
48	100
74	88
105	20

Overall, 25000 OCs are generated by scaling up the loads within 100% - 150% of their base case value for each area. The outputs of generators re-dispatch the same amount of load in the same area. VSA is implemented to determine the secure and insecure OCs. In this work, it is assumed that the load capacity limit and the secure operation limit mentioned in section 2 are overlapping. The unstable OCs are removed from the initial database, since they cannot provide useful information about the system condition. For all of these secure/insecure OCs, 60% of them are used for training while the rest are reserved for periodic update and testing. The initial trees are trained offline. For example, in the database for switching on cap bank at bus 44, the number of secure and insecure OCs are shown in Table 3.4.

Table 3.4 Number of secure/insecure OCs

OC	Training	Testing and Update
Secure	9948	6615
Insecure	2199	1483

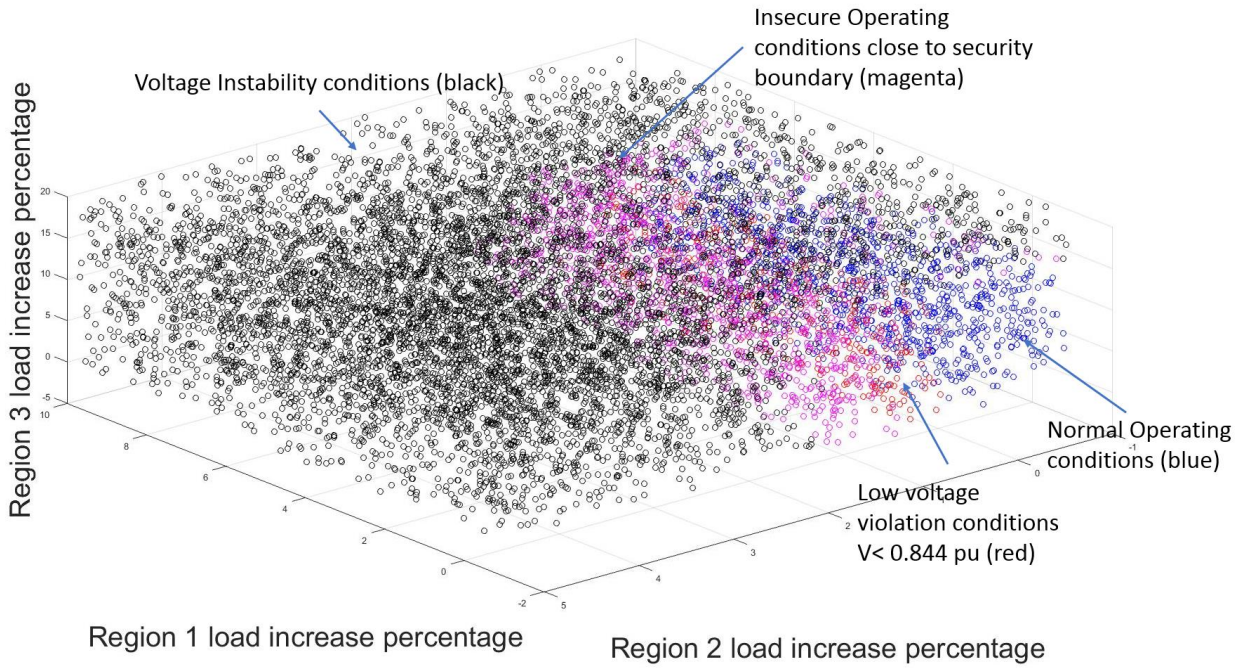


Figure 3-6 OCs before control IEEE 118 bus system

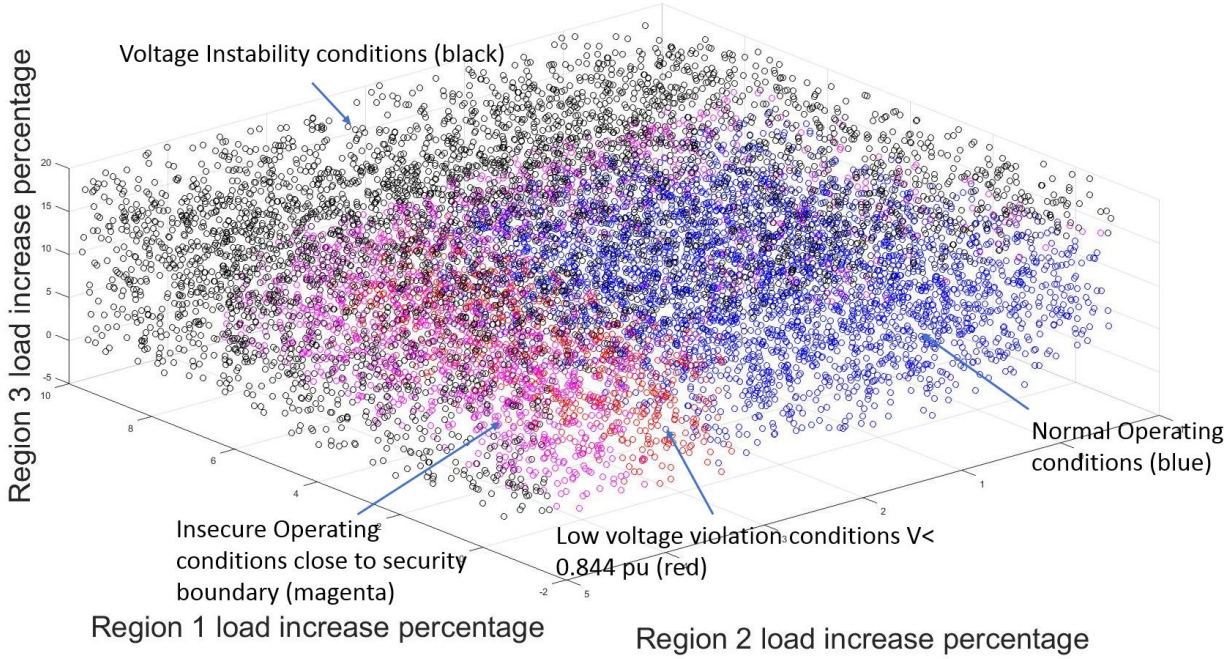


Figure 3-7 OCs after control IEEE 118 bus system

After the capbank at bus 117 is switched on, 10,000 OCs under different load condition shown in

Figure 3-7. As can be seen, the number of insecure OCs , low voltage conditions, and unstable OCs decreases.

The DT for post-control scenario which switching on capbank 44 is built by default classification and regression tree and validated using 10 folds cross validation. The cross-validation error for decision tree is computed as:

$$Cross\ Validation\ Error\ = 0.0147$$

$$Cross\ Validation\ Accuacy\ = 1 - Cross\ Validation\ Error\ = \ 0.9853$$

$$= 98.53\%$$

Build parallel DTs for all 64 controls, their cross validation accuracies are shown as Figure 3-8. The low error rates of cross validation indicate that the trained tree is able to provide accurate VSA for each control decision.

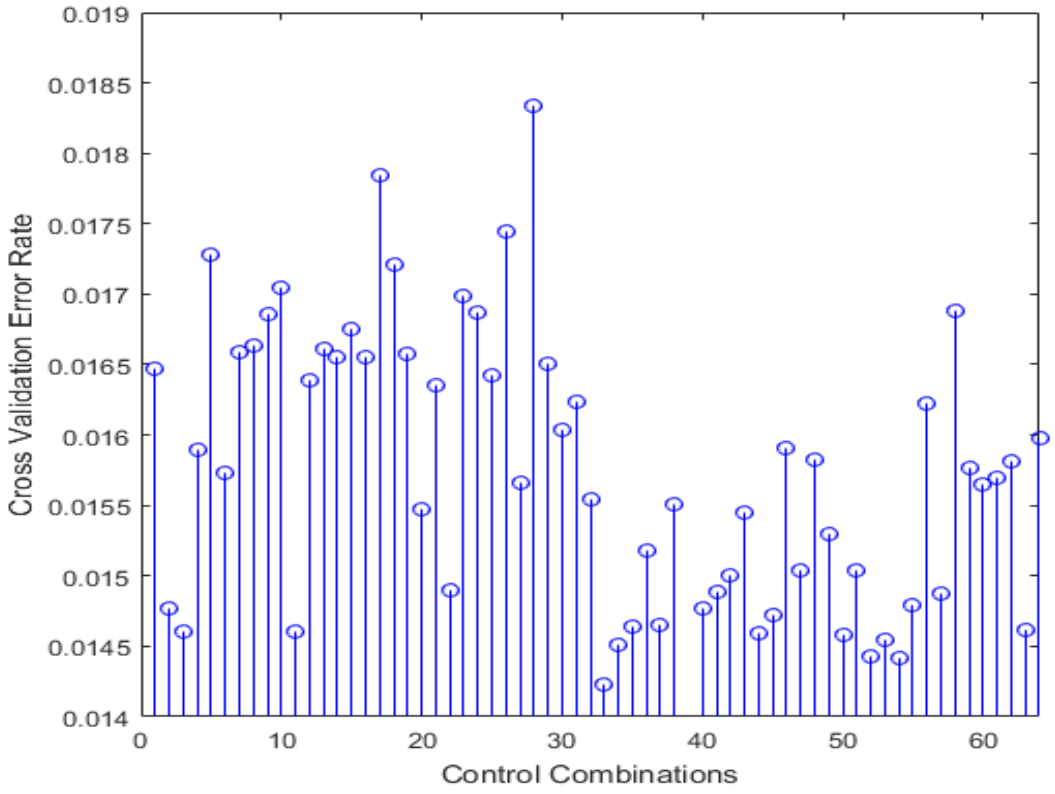


Figure 3-8 Accuracy of cross validation for 64 decision trees

A scenario is created to evaluate the performance of the proposed regional voltage controller. In this scenario, the initial system load is scaled up by 70% and kept increasing with a step size of 0.8% system-wise. The voltage measurements and capbank statuses at buses 34, 44, 45, 48, 74, and 105 are shown in Figure 3-9. At time instance 29, the VSA module detects an insecure status, then the rest of parallel DTs are activated to search out for a control strategy to bring the system back to secure status. The feasible control combinations, which can provide secure control decisions, are shown in Table 3.5. The first control decision [0, 0, 1, 0, 0, 0] with minimum control candidates involved and minimum reactive power consumption is selected.

Table 3.5 Control Combination

Instance	**Cap 34**	**Cap 44**	**Cap 45**	**Cap 48**	**Cap 74**	**Cap 105**
1	0	0	1	0	0	0
2	0	1	0	0	0	0
3	1	0	0	0	0	0
4	0	0	0	0	1	0

Since the control is executed, the system condition is changed. All 63 DTs are updated with new simulated OCs. As the load demand keeps rising, at the time instance 53, the VSA module detects another insecure OC. This time, the controller indicates that the control combination [0, 0, 1, 0, 1, 0] can secure the system.

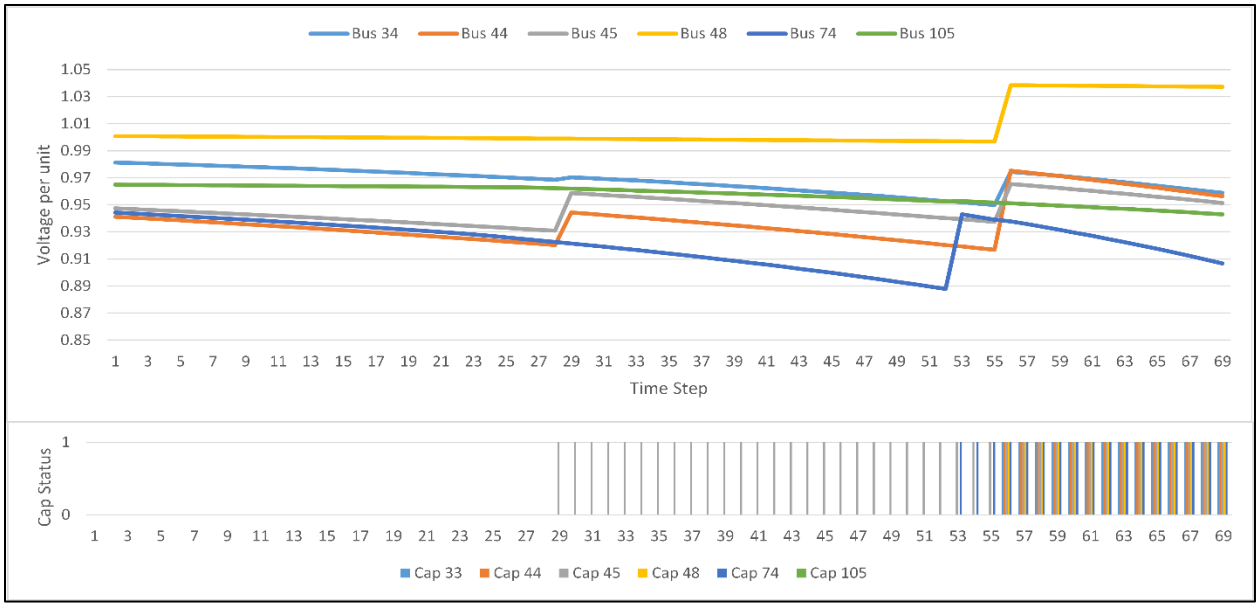


Figure 3-9 Voltage measurement and capbank status

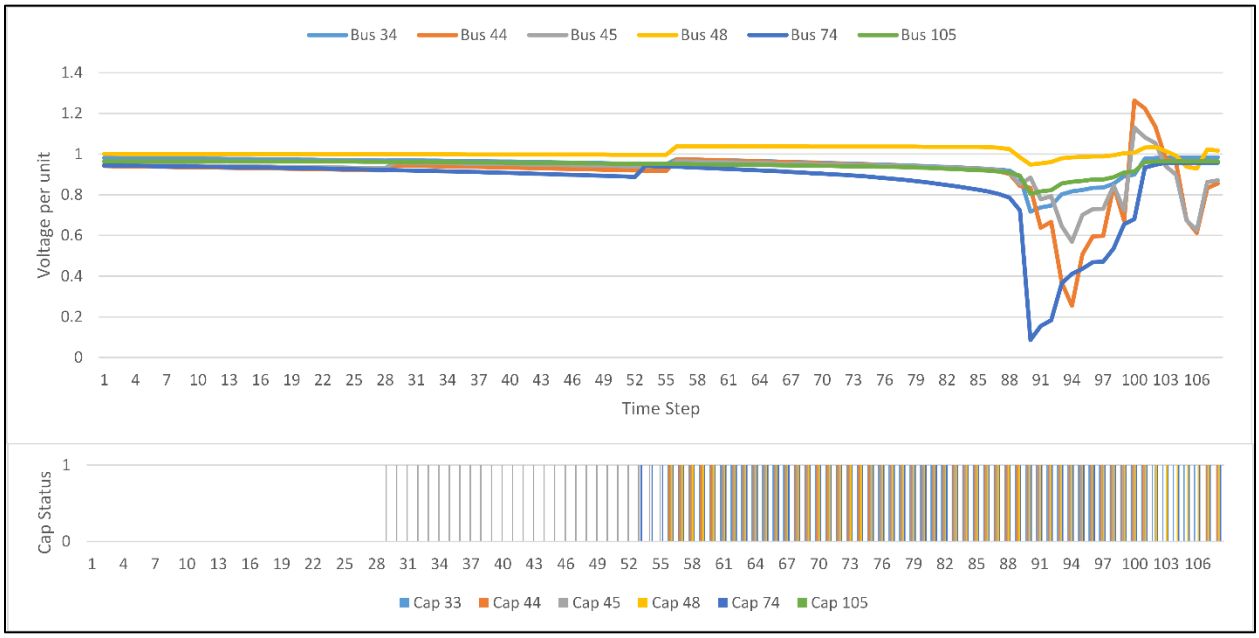


Figure 3-10 Voltage measurement and capbank status when system becomes unstable

At instance 57, the system is still able to provide secure control by switching on capbank at buses 34, 44, and 46. After this instance, the system starts detecting insecure OC while it is not able to provide secure control decision. At instance 89, a

significant voltage drop is observed and the whole system voltage starts to collapse. From then, the system shows huge distortion of voltage profile. It is noticed that during the whole simulation, capbank 105 did not participate at any control.

3.5 Summary

This chapter has presented and demonstrated a DT based control methodology for regional voltage. At the beginning, it introduced the basic concept of DT, DT based VSA, and k fold cross validation for tree validation. Then this chapter explained the learning sample database preparation required for DT training as well as the parallel trees based voltage control method. Different from DT based VSA, DT based control is conducted by providing VSA for post-control OC. Each control combination is associated with one learning sample database. Simulation is conducted based on IEEE 118 bus system with 6 control capbanks located at buses 34,44,45,48,74, and 105. A total of 64 DTs are trained for all 64 control different combinations. The accuracy of decision tree is validated by using 10 fold cross-validation. The validation result shows that the misclassification rates for all DTs are less than 0.0185 which demonstrates the effectiveness of the advocated method. The trained DTs are uploaded to controller written in C# and a scenario is created to test the whole control system by scaling up the load at each time instance. The results show that the proposed controller is able to secure the system while minimizing the number of control candidates as well as the reactive power injection.

Chapter 4
Adaptive Decision Tree for Regional Voltage Control

In the previous chapter, DT-based regional voltage control technique is evaluated and shows promising result when system's OC is approaching system security boundary. However, as discussed in [3], modern power system networks are actually experiencing frequent topology change during daily operation. Notably, the scheduled outage of transmission line and transformer significantly change the direction of power flow and thus modify the OC of system. Frequent topology changes in power system can result in the difference between the actual system operating conditions and the initial learning sample database. To guarantee the reliability of trained decision model, it is necessary to incorporate new available training cases and update the trained DTs.

Re-training the whole DT from scratch might not be a cost-effective way since the OCs including the new available training cases could be tremendous. In this advocated method, an online boosting technique [30] widely used for computer vision is implemented to update the classifier in an online manner. This chapter starts with the basic illustration of the offline boosting algorithm called AdaBoost. An online version of this boosting algorithm is also introduced. The algorithm is tested using the IEEE 118 Bus system by introducing scheduled line outage. Finally, the performance of the algorithm is evaluated by compared with the MATLAB default single DT training.

4.1 Naive Bayes Classifier, Logistic Regression, and Addictive Logistic Regression

Naive Bayes classifier is a classifier that applies Bayes' theorem with naive independence assumptions between the different features. Given a problem instance to be classified: $\boldsymbol{x} = \{x_1, \dots, x_n\}$ where x_i represents each feature, the instance probability can be represented by:

$$p(y_k|x_1, \dots x_n) = (p(y_k)p(\boldsymbol{x}|y_k))/p(\boldsymbol{x}) \quad \textbf{(4.1)}$$

Where y_k represents the k^{th} class. It can be expressed using Bayesian probability:

$$Posterior = \frac{Prior \times Likelihood}{evidence} \quad \textbf{(4.2)}$$

It can be also written using the chain rule as:

$$p(y_k, x_1, \dots x_n)$$
$$= p(x_1|x_1, \dots, x_n, y_k)p(x_2|x_3, \dots, x_n, y_k) \dots p(x_{n-1}|x_n, y_k)p(x_n|y_k)p(y_k)$$

Based on the naïve conditional independence assumptions that each feature is conditionally independent and thus, the posterior can be written as:

$$p(y_k|x_1, \dots, x_n) \; is \; approximated \; to \; p(y_k) \prod_{i=1}^{n} p(x_i|y_k)$$

$$pp(y_k|x_1, \dots, x_n) = \frac{1}{Z} p(y_k) \prod_{i=1}^{n} p(x_i|y_k) \tag{4.3}$$

where $Z = p(\boldsymbol{x})$ is a scaling factor.

Therefore, a Bayes classifier can be determined by using a decision rule to combine each independent model. One common rule is to pick the most probable hypothesis. This is also called maximum a posteriori (MAP) decision rule.

$$\hat{y} = argmax_{k=\{1,\dots,K\}} \, p(y_k) \prod_{i=1}^{n} p(x_i|y_k) \tag{4.4}$$

This is a generative approach. The key factor is to estimate likelihood and prior and finally use the Bayes rule to predict the estimate. There is another discriminative approach that estimate posteriori directly which is called logistic regression.

Using Bayes rule, the conditional log likelihood can be expressed as:

$$l(\boldsymbol{w}) = \sum_{j} \ln p(y_k|\boldsymbol{x}_k, \boldsymbol{w}) \tag{4.5}$$

where $\boldsymbol{w}$ is a set of parameters for sigmoid:

$$\sigma\left(w_0 + \sum_i w_i x_i\right) = \frac{1}{1 + e^{-w_0 - \sum_i w_i x_i}}$$

Therefore, the conditional log likelihood can be expressed as:

$$l(\boldsymbol{w}) = \sum_{j} y_k \ln p(y_k = 1|\boldsymbol{x}_k, \boldsymbol{w}) + (1 - y_k) \ln p(y_k = 0|\boldsymbol{x}_k, \boldsymbol{w}) \tag{4.6}$$

This is a concave function of $\boldsymbol{w}$, it can be solved by gradient descent optimization method. This indicates that Naive Bayes' parameters can be converted to Logistic

Regression parameters. However, Logistic Regression makes no assumptions about the likelihood $p(\boldsymbol{x}|y_k)$.

4.2 Offline Boosting

The concept of boosting emerged from the area of machine learning technique. The basic concept is to boost the accuracy of weak learners by combining the outputs from all weak learners and then average the result using a weighted vector. Particularly, the introduction of AdaBoost [31] gained much attention in the field of machine learning in the past decades. The output of the other learning algorithms (weak learners) is combined into a weighted sum that significantly improve their performance. Two key terms for AdaBoost are introduced as follows:

Weak Learner: A weak learner is a classifier that provides classification result slightly better than random guessing (For a single layer tree stump, the error rate should be less than 50%). A weak learner h^{weak} can be generated by using any given learning algorithm.

Strong Classifier: A stronger classifier H^{strong} is yielded by computing the linear combination of N weak learners.

$$H^{strong} = sign(\sum_{n=1}^{N} \alpha_n h_n^{weak}(x)) \tag{4.7}$$

The basic work flow of offline AdaBoost can be concluded as follows: Given a training set $\boldsymbol{X} = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_L, y_L)\}$, $\boldsymbol{x_n} \in \mathbb{R}^m$, $y_i \in \{-1, +1\}$ with uniform distributed weight $w_i(x_i) = \frac{1}{L}$, a weak classifier can be trained based on $\boldsymbol{X}$ and weights. Based on the training error e_n of the h_n^{weak}, h_n^{weak} will be assigned with a voting factor which is computed by

$$\alpha_n = \frac{1}{2}\ln((1 - e_n)/e_n) \tag{4.8}$$

The weight w_i increases when the corresponding sample is misclassified, and vice versa. At each boosting iteration, a new weak learner is added in to the strong classifier sequentially until a certain number of weak learners is met. [32] shows that for binary classification problems, if the weak learners consistently have error slightly better than random guessing, the final strong classifier drops to zero exponentially fast. The training and generalization error of AdaBoost is also proved as strong bounded based on the AdaBoost theoretical property introduced in [31].

4.2.1 Ensemble Learning

The goal in for a learning algorithm is not to find an exact representative for the training data, but to identify the data generation process by creating a statistical model. Therefore, the critical idea of the learning algorithm is to provide generalization performance. It is very common to see that "models with too few parameters performs badly" and "models with too many parameters performs badly". This is because a model with too few parameters is so simple that it will have a larger bias, while a model with too many parameters has too much flexibility that it will have higher variance. Intuitively, high bias means a large deviation and high variance indicates a weak match. Typically, the increase of decision tree depth can increase the variance while reduce the bias. To solve the problems associated with bias and variance, ensemble learning methods are applied. The existing methods for ensemble learning include Boosting, Bagging, Bayesian parameter averaging, etc. Bagging is derived from the name "Bootstrap Algorithm with Aggregation". Bootstrap draws random sample from give dataset with replacement, then aggregate all the model outputs to reduce the variance. Boosting is designed primarily to reduce

bias and also variance. Compared between boosting and bagging, Bagging is purely a variance-reduction technique and since trees tend to have high variance, bagging is able to produce good results. Early versions of AdaBoost have a connection with bagging and thus its major success comes with the variance reduction. However, boosting out performs bagging when:

1. Rather than weighted resampling where each training case is assigned with weight w_i, a weighted tree-growing algorithm is implemented.
2. "Tree stumps" are used as weak learners which is represented by a binary tree with only one layer. "Tree stump" has lower variance but high bias.
3. Instead of using squared-error loss to fit an additive model, AdaBoost is using a better loss function that similar to, but not the same as, the binomial log-likelihood.

Reference [33] illustrates the ensemble learning algorithm in a view of statistics that AdaBoost can interpreted as estimation procedure for fitting an additive logistic regression model. To better understand the learning procedure statistically, it is important to determine two aspects: the learning algorithm's structural model and its error model. The structural model is most important since it gives the function space of the approximator. The error model determines the criterion to be optimized in the estimation of the structural model. Assume that the classification problem is a pattern recognition completion, boosting is able to select the candidates (weak learners) to put together a "dream team" of experts. The main idea of boosting is to weight the data set based on how well the performance of the model so far: if data points are classified accurately, the data point is assigned with a lower weight; if data points are misclassified, a higher weight will be assigned to them. The purpose of this process is to let the algorithm focus on the portion of data which are not predicated well.

4.2.2 Weak Learner

The weak learner utilized in this study is “tree stump” which is a decision tree with one single node that immediately connected to the terminal nodes. A “tree stump” is produced by searching all the features and the best splitting point until it finds a decision stump that has the minimum error rate. It is accomplished in a brute force manner. Compared with a fully trained strong learner, a weak learner is less prone to over fit and thus the tree stump in the advocate method is not pruned. For AdaBoosting, the weak leaner is assumed to have accuracy just a little bit better than random guessing; For a binary class problems, this means each of them is having error below 0.5, i.e, each ε should be at most γ for $\gamma > 0$. Given this assumption and sufficient data, a strong classifier with arbitrarily small generalization error can be provided.

4.3 Online AdaBoost

The proposed online boosting algorithm is designed to correspond to its offline Adaboost method. This type of algorithm was first introduced in [31]. The given algorithm has also proved that under certain conditions, it could converge to the popular AdaBoost method as the number of weak learners and training examples approaches infinity. Online boosting starts gaining attentions in the recent decades, especially in the field on computer vision tracking due to its efficiency and simplicity.

In the online boosting algorithm’s pseudocode which is shown in Figure 4-1, $\boldsymbol{h_M}$ represents the set of weak learners trained so far, $(\boldsymbol{x}, y)$ denotes the latest arriving case, and $update$ is the algorithm that returns an updated weak classifier based on training sample and current hypothesis. In this case, the weak classifier is updated using methodology suggested in [3]. The new coming example’s weight is

set as λ. λ_n^{corr} denotes the sum of correctly classified example while λ_n^{wrong} represents sum of wrongly classified examples seen so far at stage n. h_n is serving as a selector that picking the h_m from the weak classifier pool based on the misclassification rate. The final strong classifier is a linear combination of N selectors.

Algorithm 1 Algorithm for Online Boosting

Input new arriving training case $\{\mathbf{x}, y\}, y \in \{-1, +1\}$
Output strong classifier $H(x)$
Initialize $\lambda = \lambda_{n,m}^{corr} = \lambda_{n,m}^{wrong} = 1; H(x) = null$
for all $n = 1, 2, ..., N$ **do**
    for all $m = 1, 2, ..., M$ **do**
        $h_{n,m} = update(h_{n,m}, \{\mathbf{x}, y\})$
        if $h_{n,m}(\mathbf{x}) = y$ **then**
            $\lambda_{n,m}^{corr} = \lambda_{n,m}^{corr} + \lambda$
        else
            $\lambda_{n,m}^{wrong} = \lambda_{n,m}^{wrong} + \lambda$
        end if
        $e_{n,m} = \frac{\lambda_{n,m}^{wrong}}{\lambda_{n,m}^{corr}}$
    end for
    $m^* = argmin_m(e_{n,m})$
    $e_n = e_{n,m^*}$
    $h_n = h_{n,m^*}$
    $\alpha_n = \frac{1}{2} \cdot \ln(\frac{1-e_n}{e_n})$
    if $h_n = y$ **then**
        $\lambda = \lambda \cdot \frac{1}{2(1-e_n)}$
    else
        $\lambda = \lambda \cdot \frac{1}{2e_n}$
    end if
    $H(x) = H(x) + \alpha_n \cdot h_n$
end for
return $H(x)$

Figure 4-1 Algorithm for Online Boosting

4.3.1 Periodic Update

Voltage measurements from PMUs and LSE continuously captures snapshots of system for each time stamp in real time. New post-control OCs associated with potential control combination, predicted loading changes/generator dispatch, and possible topology changes are generated. Figure 4-2 shows the flow chart of proposed scheme using online boosting. The three stages for this scheme are as follows: Stage 1, an ensemble classifier model is initialized using AdaBoost algorithm with offline training data set. Each weak learner in this ensemble model is a short tree stump. The voting weights for each weak learner is generated to achieve a weighted sum of the ensemble model. Stage 2, instead of re-training the ensemble model with the whole data set, the model is updated with one new OC at one time. In this stage, each tree stump is further induced incrementally. Then, the voting weights of all tree stumps in the ensemble model are re-computed according to the training errors. Stage 3, the application of VSA utilizes the updated ensemble model to assess each control option and current OC security status according to the voltage measurements captured with PMUs.

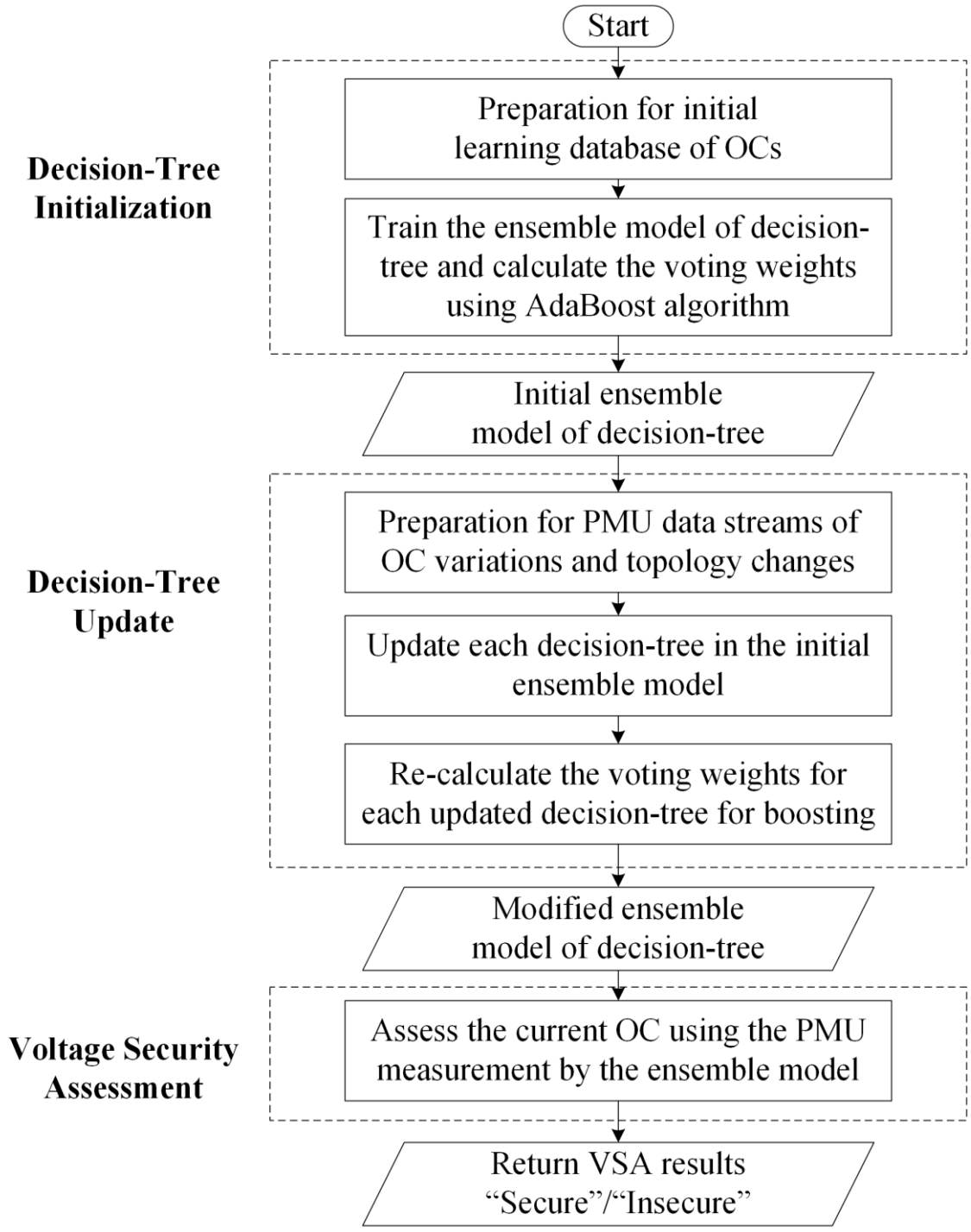


Figure 4-2 Flow Chart of Proposed Scheme for Online Application

4.4 Numerical Studies using Online Boosting

The control decision by switching on capacitor bank at bus 44 is selected for classifier performance evaluation. The initial tree is trained based on the offline AdaBoost method incorporated with 30 weak learners, and the number of selectors is also 30.

The transmission line between bus 15 and 33 is tripped on the test system. New training cases and testing cases are created using the proposed approach in the previous sections but with a different system topology. 4000 of these new cases are used for the periodic update, and another 4000 of them are reserved for online validation. Among these new cases, 82% of them are secure OCs while the rest of them are insecure OCs. The performance of online boosting approach is evaluated by comparing it with single decision tree training using default MATLAB tree training. The computation time and misclassification error rate are recorded and illustrated in Figure 4-3. The online boosting scheme turns out to be more accurate than single DT training while the computation time spent by online boosting for tree update is much less than re-training the tree from scratch. The computation was run under the environment of MATLAB on a workstation with Intel Core i7-4790 3.6 GHz CPU and 32 GB memory.

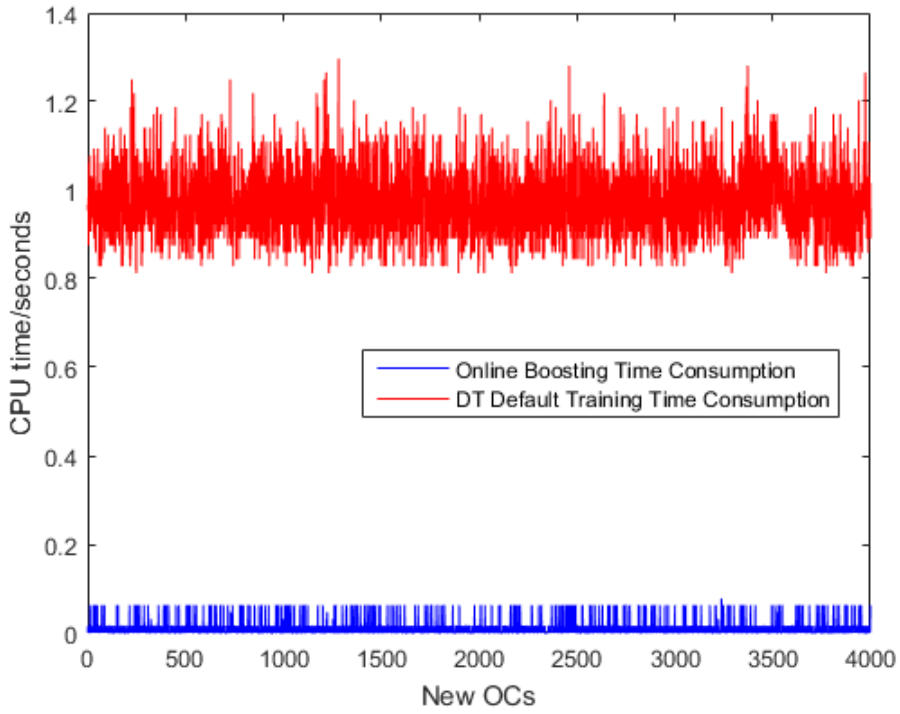


Figure 4-3 Computation time for tree update

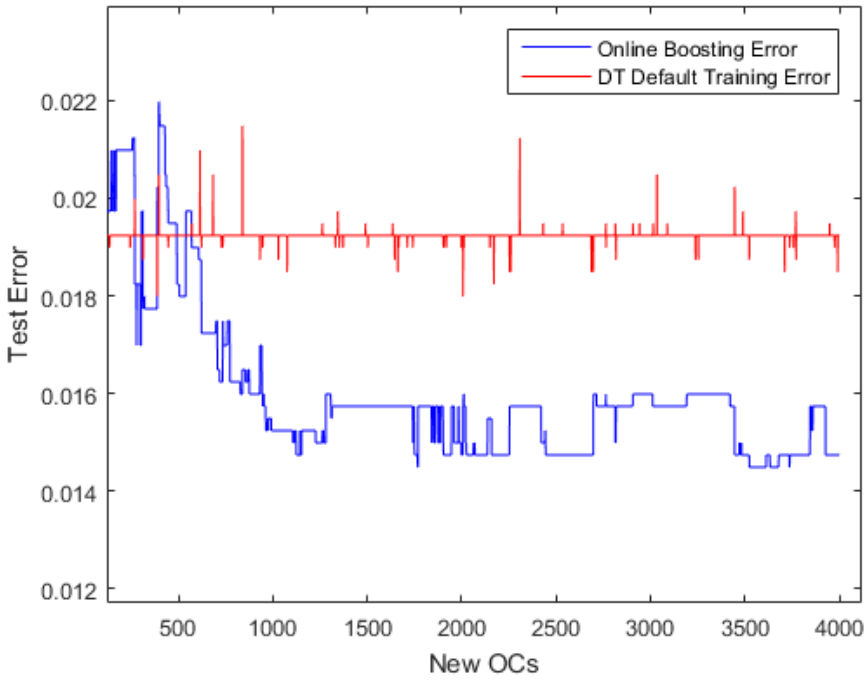


Figure 4-4 Test error rate for online boosting and single DT training

4.5 Hoeffding-Tree-based Learning

In the previous study, an adaptive decision-tree-based systematic method for open-loop regional voltage control was developed. The proposed scheme employs the VSA approach introduced in Chapter 2 to generate a DT training sample database for each control combination. The online-boosting method is implemented to adaptively update the trained decision trees. This approach is evaluated based on the IEEE 118 bus system. The cross validation scheme shows that the parallel trees trained for different control combinations can predict the post-control security status for OC in a low error rate. Finally, a topology change scenario is created to test the online boosting scheme. Simulation result shows that the proposed method is able to reduce the computation burden and have a lower misclassification error compared with the default decision tree training method.

The online boosting approach proposed in [7] which only modifies the voting weights of ensemble DTs based on the prediction error, and flips the decision with an error rate higher than 50% during each periodic update. This method also lacks theoretical supports, even though this method reduces the computational complexity while retaining the tree structure, the feature nodes are no longer the best variables to conduct a prediction and might lead to a higher prediction error rates in the long run. Incremental Tree Induction (ITI) is a lossless online updating algorithm that provides the same DT as the one built from scratch after restructuring the trained model with the new coming data set. However, this method has a memory management problem and it is time consuming since it involves recursive transpositions that requires the storage of all training and updating data within the decision nodes of the DT. In regards to reducing the time required for mining high speed data streams, Hoeffding trees (HTs) is an alternative incremental decision tree method that proposed and be able to be trained in constant time per sample.

In this section, in order to guarantee the robustness of the proposed technique after possible topology changes or any unpredictable system conditions captured by PMU data streams, an adaptive ensemble decision-tree-based VSA methodology using Very Fast Decision Tree (VFDT) system for updates is proposed. Initially, the ensemble DT model consists of multiple DTs with voting weights generated by an adaptive boosting method discussed in the previous section. Instead of retraining the whole set of DTs from scratch, an adaptive DT updating scheme using VFDT system is introduced to update the ensemble trees. The proposed framework provides more benefits to memory management for high-speed data stream mining which requires the basic statistical analysis and the new coming data instead of storing the whole data set for updating.

The Very Fast Decision Tree (VFDT) system based on Hoeffding trees is designed and implemented in Reference [34]. VFDT is capable of applying various kinds of heuristic measure methods, such as Information Gain and Gini Index to find out the best feature. The system is implemented in a distributed wireless sensor network application [35].

4.5.1 Hoeffding Bound

During the tree induction process, whether a leaf node should be split into a new attribute node or not is decided based on a criterion of heuristic measurement. Alternatively, In [34], Domingos and Hulten have utilized Hoeffding Bound, which is also known as additive Chernoff bound, to serve as a criterial of the statistical condition about how many new samples should be observed before executing a new split in a high-speed data streams framework. Assuming $Z_1, Z_2, \dots, Z_n$ are independent bounded random variables in the range of [a,b], according to the Hoeffding Inequality, $\forall t > 0$,

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})\right|\right\} \leq 2\exp\left(-\frac{2nt^2}{B^2}\right) \quad \textbf{(4.9)}$$

where $\bar{Z} = (Z_1 + Z_2 + \cdots + Z_n)/n$. And B represents the range of existing random variables $B = b - a$. With a confidence of $1 - \delta$, the true mean μ is at least $\bar{Z} - \epsilon, i.e.$

$$\Pr(\mu \leq \bar{Z} - \epsilon_H) \geq 1 - \delta \quad \textbf{(4.10)}$$

By the equations, the Hoeffding bound is defined as:

$$\epsilon_H = \sqrt{\frac{B^2 ln\left(\frac{1}{\delta}\right)}{2n}} \quad \textbf{(4.11)}$$

4.5.2 Hoeffding Tree

With confidence $1 - \delta$ as shown in Equation (5.3), the tree have received enough information to split a leaf node into branches as long as the difference of the highest two heuristic measurements $\Delta\bar{G} = G(\bar{X}_a) - G(\bar{X}_b)$ is larger than ϵ_H. This is the core principle for Hoeffding tree induction. The detail of the algorithm is shown in Figure 4-5. The inputs for the Hoeffding induction process are the trained DT using AdaBoost, new operating condition data set D, split confidence δ, and grace period n_{min}. For each single training instance, the DT model sorts it into a leaf node by comparing the attribute values from the instance and the leaf node. Then, the training instance is stored at the leaf node holding sufficient statistics needed to make decisions about further growth. According to the grace period $n_{\min}$, the code from 7-18 is executed periodically. Line 13 involves the "null" attribute X_ϕ which considers the benefit of not splitting. The "null" attribute prevents splitting unless an attribute is better than X_ϕ based on the Hoeffding bound and user defined

threshold τ. τ is also called "tie-breaking" which considers the situation of two or more attributes causing low Hoeffding bound and the setting of τ allows continuing the induction.

Algorithm 1 Hoeffding Tree Induction Algorithm

procedure HOEFFDINGTREE($DT, \mathcal{D}, \boldsymbol{X}, G, \delta, n_{\min}$)
initialize DT with a root **if** the decision tree is empty
for all training instances $(\boldsymbol{x}_k, y_k) \in \mathcal{D}$ **do**
    Sort the instance into the leaf node l using DT
    Update sufficient statistics in l
    Increment n_l, the number of instances seen by l
    if $n_l \bmod n_{\min} = 0$ **and** instances seen by l are not the same class **then**
        Compute $\bar{G}_l(X_i)$ for each attribute $X_i \in \boldsymbol{X}$
        Let X_a be the attribute with highest $\bar{G}_l$
        Let X_b be the attribute with second-highest $\bar{G}_l$
        Compute the Hoeffding bound in equation (10)
    end if
    if $X_a \neq X_\emptyset$ **and** $(\bar{G}_l(X_a) - \bar{G}_l(X_b) > \epsilon_H$ **or** $\epsilon_H < \tau)$ **then**
        Replace the leaf node l with an attribute node splitting on X_a
        for all branches under the new attribute node **do**
            Add a new leaf with initialized sufficient statistics
        end for
    end if
    return DT
end for
end procedure

Figure 4-5 Hoeffding Induction Tree Algorithm

The induction process of HT is independent of the distribution from training data. This is the primary property of HT that benefits the application of online learning algorithm. Compared with the batch learning, the HT induction could be more vulnerable to the overfitting when concept drifting data is injected [36]. The problem can be solved by introducing active and adaptive post-pruning to adjust for changes in concept or a new training method based on Bayesian approach. This topic requires further study.

4.5.3 Numerical Studies with Hoeffding Tree Algorithm

The proposed scheme is evaluated using the IEEE 118 bus system and the same data sets created in section 4.3.2. Sixty percent of the data set are used for initial training, and the rest are implemented for testing the initial ensemble model. Figures show the computation time and test error during the updating process for a DT in the ensemble model. It is observed that, with more and more instances incorporated, the database grows increasingly and thus the default methodology takes more time to rebuild the DT from scratch. By contrast, the Hoeffding tree algorithm saved much computation time by inducing at the leaf nodes with sufficient data according to the computed Hoeffding bound. It should be noticed that the computation time for each update using Hoeffding tree is only related to the size of update data set for each instance. The step differences of red line in Figure 4-7 reveals the changes in the DT model using Hoeffding tree. Despite that the Hoeffding tree algorithm is relatively conservative when the leaf nodes splitting is applied, the testing error is still able to reach the same level as the default DT training method.

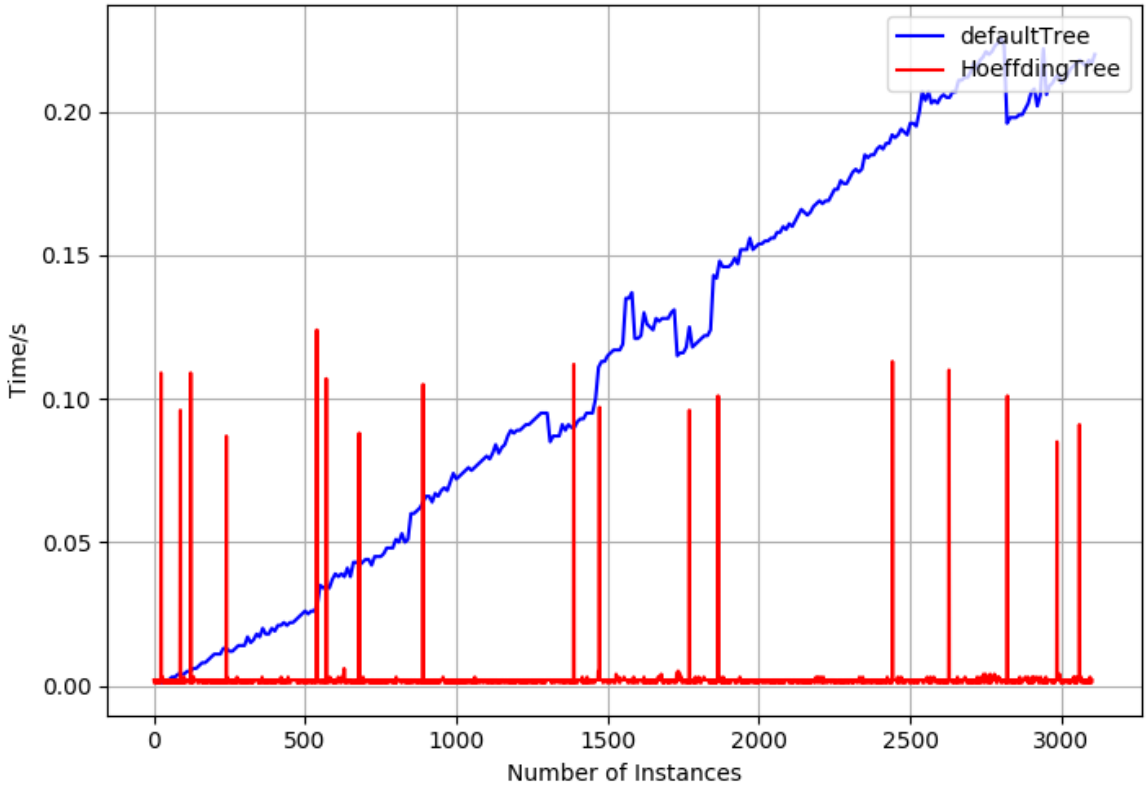


Figure 4-6 Computation time for modifying DT model

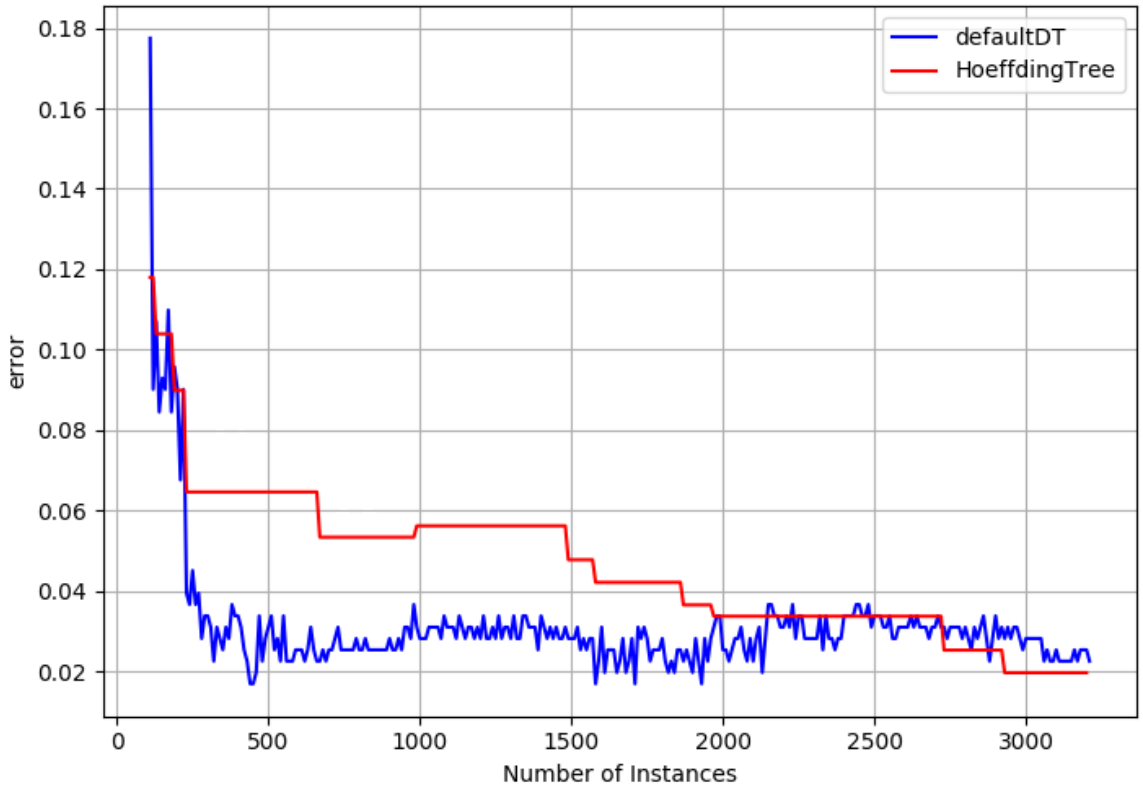


Figure 4-7 Testing error for modified DT model

4.6 Summary

This chapter introduced the concept of ensemble learning and online boosting. Then an adaptive decision-tree-based systematic method corresponding to the offline AdaBoost method was proposed for open-loop regional voltage security control. This method was implemented to adaptively update the trained DTs. This approach was evaluated based on IEEE 118 bus system. A topology change scenario was created to test the online boosting scheme. Simulation results show that the proposed method is able to reduce the computation burden and have a lower misclassification error compared with the default decision tree training method. To provide a more efficient update for the weak learner, the Hoeffding tree based induction algorithm was introduced. The weak learner's model structure was further induced incrementally based on the Hoeffding bound, which guarantees that the inductions are necessary with sufficient statistics to improve the performance. The simulation was conducted with the same data set generated for Online-boosting. The results show that the new induction algorithm provides better performance than the original DT training method by reducing the time consumption while retaining the accuracy.

Chapter 5
Power System Network Partition Methodology for Regional Voltage Control

Voltage collapse is shown to occur due to the following issues: 1: weakness of the boundaries of groups of PQ and PV buses. 2: not enough reactive power support within these group of buses. According to [15], due to the weakness of the boundary lines connecting between the group of buses and the rest of the system, the voltage and angle within such a group of buses are changing coherently for any disturbance. Therefore, this group of buses are called Voltage Control Area (VCA). The control within these buses should be coordinated and the control devices within this area would be independent of the control devices outside of the area. That also means that if the VCA is identified, most of the control devices outside of the area would not be considered as the voltage control candidates. Based on the discussion in Chapter 4, if all the control devices are considered as control candidates then the number of control combinations would be tremendous. Therefore, finding the VCAs can significantly reduce the number of control combination and thus minimize the computation burden for the advocate methodology.

The zoning algorithm for power system network division is an optimization method that requires the definition of the proximity measure and a clustering criterion which describe the rule to form a number of zones utilizing the proximity measure. For example, studies conducted in [37] [38] use the concept of "electrical distance" to evaluate the degree of similarity for any two buses. Reference [39] proposed a new zoning methodology based on the concept of electrical distance. This reference defined a "VAR control space" which is considered as an r dimensional Euclidean space where each PQ bus node can be represented by a coordination vector $\{x_{i1}, x_{i2}, \dots, x_{ir}\}$. r represents the number of reactive power sources. S_{ij} indicates the sensitivity of the ith node's voltage with respect to the jth reactive power support. x_{ij} is defined as:

$$x_{ij} = -\log(|S_{ij}|) \tag{5.1}$$

Each component of a bus's coordination vector represents the level of coupling between the node and a specified reactive power source. For two load buses $bus_m = \{x_{m1}, x_{m2}, \dots, x_{mr}\}$ and $bus_n = \{x_{n1}, x_{n2}, \dots, x_{nr}\}$, the electrical distance between them is defined by:

$$D_{mn} = \sqrt{|x_{m1} - x_{n1}|^2 + \cdots + |x_{mr} - x_{nr}|^2} \tag{5.2}$$

The singletons are merged in an iterative manner to construct the dendrogram. Finally the nodes strongly coupled with the same set of reactive supports would be clustered into the same group based on the defined average linkage criterion L:

$$L = \frac{1}{\sum(G_I) * \sum(G_J)} \sum_{m \in G_I} \sum_{n \in G_J} D_{mn} \tag{5.3}$$

where G_I represents the I^{th} cluster.

5.1 Voltage Control Area Identification based on Structural Weakness

A power system can break down into a non-overlapping set of coherent bus groups called "Voltage Control Area" (VCA) with unique voltage instability problems [19]. The algorithm developed in [40] identifies the VCAs by finding the weakest transmission elements connected to each bus. These elements would be eliminated, and finally the isolated buses would be grouped as the same area. The algorithm is implemented by using the reduced Jacobian matrix which are the reactive power/voltage Jacobian for all buses in the network.

For a given operating condition, the network model can be expressed in the following linearized model:

$$\begin{bmatrix}\Delta P\\ \Delta Q\end{bmatrix} = \boldsymbol{J}\begin{bmatrix}\Delta\theta\\ \Delta V\end{bmatrix} \tag{5.4}$$

$$\boldsymbol{J} = \begin{bmatrix}J_{P\theta} & J_{PV}\\ J_{Q\theta} & J_{QV}\end{bmatrix} \tag{5.5}$$

where ΔP and ΔQ indicate the incremental change for bus real power and reactive power respectively. $\Delta\theta$ represents the incremental change angle while ΔV shows the incremental change in bus voltage magnitude. $\boldsymbol{J}$ represents the full Jacobian matrix.

A simplified sensitivity model can be derived assuming the loadflow equations are decoupled. By letting $\Delta P = 0$, Equation 5.4 becomes:

$$\Delta \mathrm{Q} = \boldsymbol{J}_{\mathrm{r}} * \Delta V \tag{5.6}$$

where $\boldsymbol{J_r}$ represents the reduced Q-V Jacobian sub-matrix:

$$\boldsymbol{J}_r = \left[J_{QV} - J_{Q\theta}J_{P\theta}^{-1}J_{PV}\right] \tag{5.7}$$

To identify the VCAs, the weakest transmission elements connected to each bus in $\boldsymbol{J_r}$ can be located. To begin with, this algorithm will normalize the reduced Jacobian matrix $\boldsymbol{J_r}$ with the largest diagonal element of $\boldsymbol{J_r}$.

$$\boldsymbol{J}_{\mathrm{r,Norm}} = \boldsymbol{J}_r / \boldsymbol{max}(\boldsymbol{J}_r)_{\boldsymbol{ii}} \quad \textbf{(5.8)}$$

An α value is defined as a threshold to determine which element should be reduced. This threshold value can be initialized as:

$$\alpha = \min(\boldsymbol{J}_r) + (\mathrm{mid}(\boldsymbol{J}_r) - \min(\boldsymbol{J}_r)) \quad \textbf{(5.9)}$$

For each row of the normalized reduced Jacobian matrix, the absolute values of the off diagonal elements are sorted from smallest to largest. The smallest absolute values in J_r will be reduced from each row until the sum of the eliminated normalized Jacobian matrix is larger or equal to α. Cluster the buses that are still interconnected after the weakest branches elimination. Each cluster forms a single VCA.

The selection of α is able to determine the group of buses that experience the same voltage problem. When α is selected very small, very few clusters would be grouped since the voltage change within each cluster becomes larger. If a larger α is selected, the number of clusters would increase. The difference in voltage changes for buses within each cluster would become smaller. An algorithm is proposed in [41] to determine an appropriate α by introducing disturbances to the system. As suggested, the proper α value would be the smallest α value that identify the coherent pattern of the largest voltage changes caused by disturbance.

5.2 Determine Variable $\boldsymbol{\alpha}$ for VCA Identification

A new α selection algorithm using identical V-Q curve minima and the set of generators hitting their Q-limit while reaching the V-Q curve minima is presented in

[42] [41]. Figure 5-1 reveals the proposed approach using Kmeans to determine parameter α for modal analysis based VCA identification.

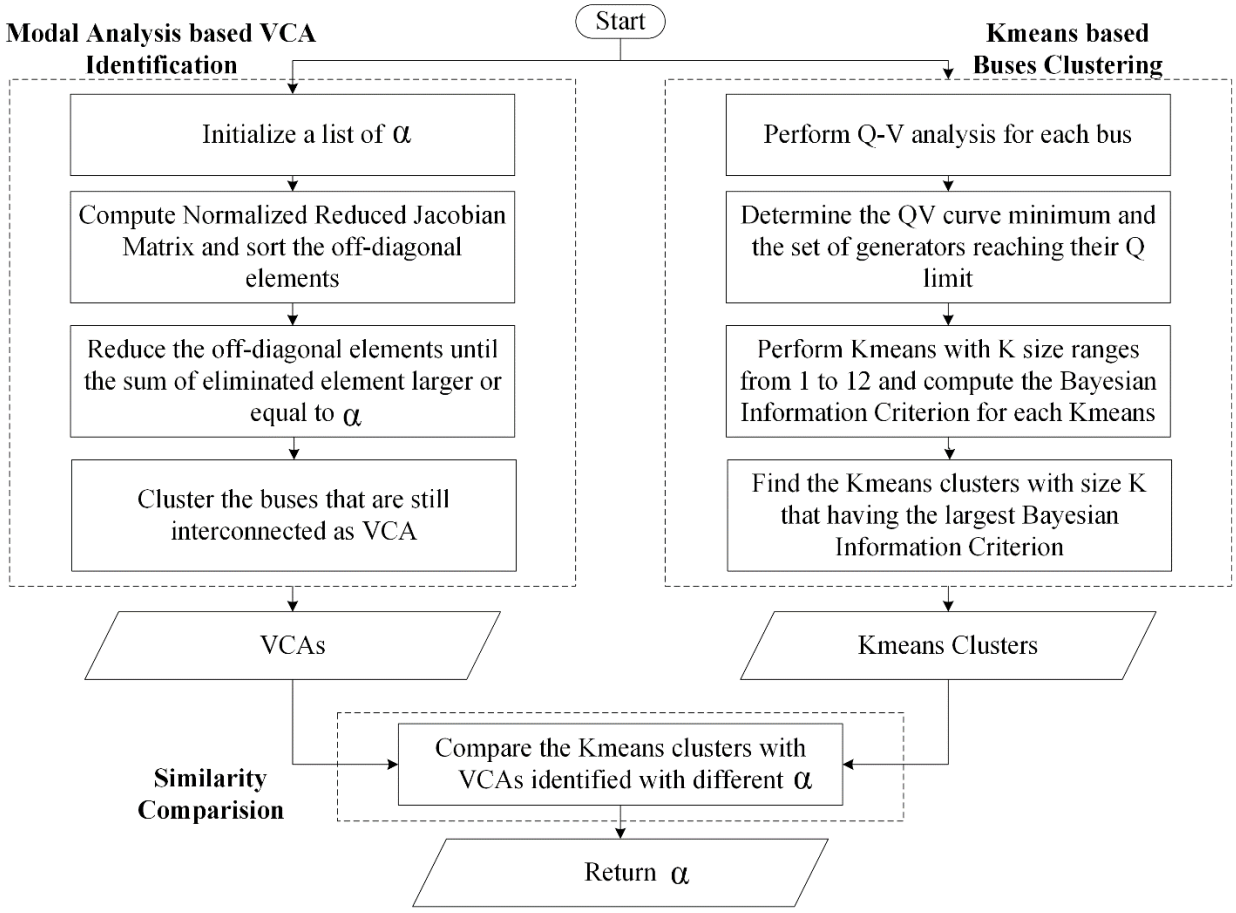


Figure 5-1 Proposed Scheme for VCA Identification

The major steps for this algorithms are introduced as follows:

1. Run a repeated load flow to plot a V-Q curve at each PQ bus in the system. Observe the set of generators that reach their Q-limit when plotting the V-Q curve. Locate the minimum point of the V-Q curve : Q_{min} and V_{min}. For example, The second and third columns of Table 5.1 show a set of normalized minimum. The attributes start from the third column show the status of neighboring generators' reactive power reserve basin. "1" indicates the generator has reached its reactive power limit.

Table 5.1 Normalized V_{min} , Q_{min} and generator hitting their Q limit

$QV\ Analysis$	V_{min}	Q_{min}	Gen_1	Gen_2	Gen_3	Gen_4	…	Gen_i
1	0.921907	0.535	1	0	0	1	…	0
2	0.909724	0.52	0	0	0	1	…	1
3	0.698074	0.685	1	0	0	1	…	1
…	…	…	…	…	…	…	…	…
K	0.808074	0.632	0	0	0	1	…	0

2. Cluster the buses that having the same reactive reserve basin (same set of generators that hit their Q-limit) and identical V_{min} and Q_{min} points. For instance, if bus 1 and bus 2 has the largest similarity compared with any other bus, then bus 1 and bus 2 will be considered in the same voltage control area [42].
3. Implement Kmeans algorithm on the database to find out the clusters and use Bayesian Information Gains to determine the best cluster size.
4. Use the Modal based method introduced in Section 5.1 to identify VCAs with different α values.
5. Compare the clustered areas obtained in step 3 with the VCAs identified in step 4. The optimum α should provide VCAs closest to the clusters obtained in step 3.

It has to be mentioned, that the cluster yielded in step 3 only defines the set of identical nose point of the V-Q curve but does not provide a general "voltage control area". It is only utilized to determine the optimum value of α.

5.2.1 Q Metric for Generator's Reactive Reserve Basin using Cardinal Sine Function

Using binary variable to indicate the generator's reactive power reserve basin status (1 means the generator reaches its limit while 0 means it is still within its range)

actually fails to describe the generator status when the generator's reactive power output is close to but not hitting its limit. In addition, it will be difficult to implement unsupervised learning technique such as Kmeans on database mixed with continues and binary variables. Euclidean distance used by Kmeans is not defined for categorical data such as binary status. It will fail to count the number of variable on which two cases disagree when the attributes are binary. In reference [43], Huang developed a simple method where Euclidean distance is computed for finding similarity between numerical data and Hamming distance for similarity between categorical data and finally combined both of them together with weights as one cost function. Motivated by Huang, Amir and Lipika [44] also proposed a mixed data clustering using K-means type algorithm by re-designing the distance measure to take care of the significance of an attribute while computing the distance between two data elements.

In this study, instead of using Kmeans type algorithm to cluster mixed database with continues and categorical data, a new metric is introduced using cardinal sine function (sinc) to measure the significance of each attribute. The sinc function is widely used in various signal process applications, including anti-aliasing, Lanczos resampling filter construction, and bandlimited interpolation of discrete-time signals. The sinc function is commonly defined by:

$$\begin{cases} sinc(x) = \dfrac{\sin(\pi x)}{\pi x} & x! = 0 \\ 1 & x = 0 \end{cases} \tag{5.10}$$

The function's plot shown in Figure 5-2. As long as input x getting close to zero, the function's output is close to 1. In this study, the input variable x is expressed as follows:

$$x = \frac{Q_{max,gen_i} - Q_{gen_i}}{Q_{max,gen_i}} \tag{5.11}$$

where Q_{max,gen_i} represents the i^{th} generator's maximum reactive power limit and Q_{gen_i} indicates the i^{th} reactive power output from generator.

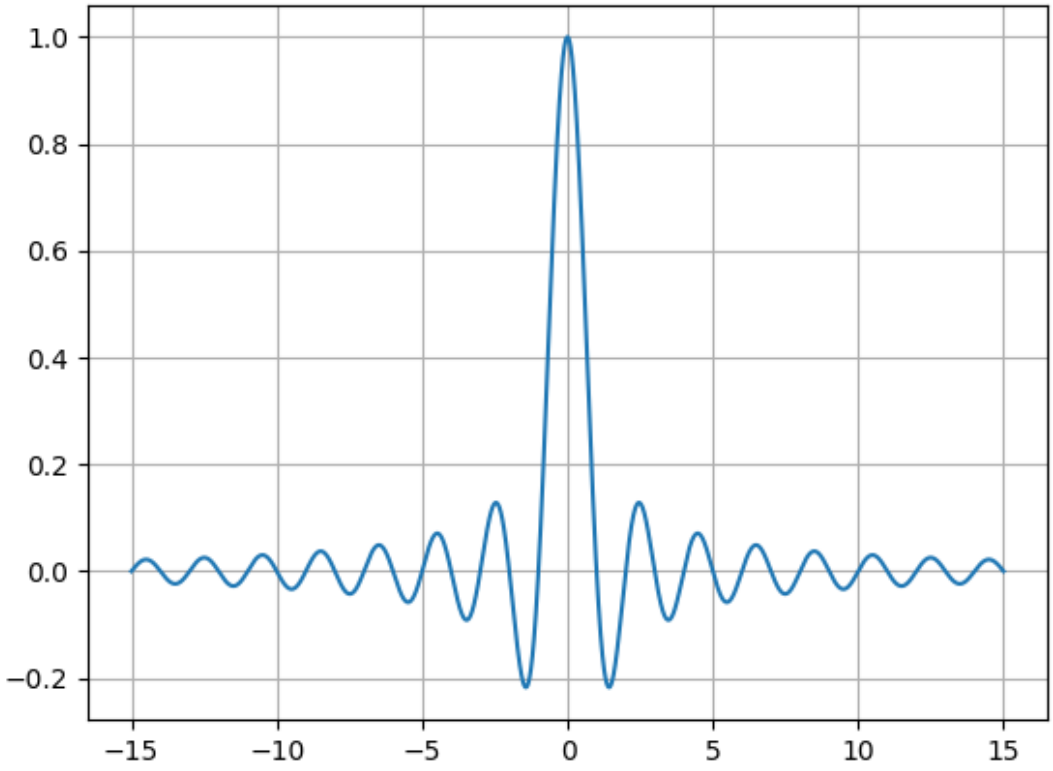


Figure 5-2 Sinc function

Compute the sinc function output based on x for each generators for the k^{th} QV analysis:

$$y_{gen_i}^{(k)} = sinc(x_{gen_i}^{(k)}) \quad \textbf{(5.12)}$$

$$\mathrm{Y^{(k)}} = [y_{gen_1}^{(k)}, y_{gen_2}^{(k)}, \cdots, y_{gen_i}^{(k)}] \quad \textbf{(5.13)}$$

These sinc function output is called the Q metric which is able to indicate how far away the generator is going to reach its limit. The Q metric as well as the minimum QV point are saved to the database for Kmeans clustering. An example is shown as follows:

Table 5.2 Normalized V_{min} , Q_{min} and Q metric for each generator

$QV\ Analysis$	V_{min}	Q_{min}	Gen_1	Gen_2	Gen_3	Gen_4	…	Gen_i
1	0.921907	0.535	1	0.0433	0.4907	1	…	0.98
2	0.909724	0.52	0.99	0.0121	0.0135	1	…	1
3	0.698074	0.685	1	0.00231	0.02	1	…	1
…	…	…	…	…	…	…	…	…
K	0.808074	0.632	0.031	0.017	0.23	1	…	0.87

Compared with Table 5.1, instead of using binary variable to indicate generator's status. Table 5.2 provides more information. For example, during the k^{th} QV analysis, the metric shows 0.98 which indicates that that Gen_i's reactive power output is close to its limit, while this information is not included in Table 5.1.

5.2.2 K-means clustering

K-means is one of the unsupervised learning algorithms that aims to partition numbers of observations into k clusters in which each observation belongs to the cluster with the nearest mean. The main idea is to define k centroids for each cluster. The next step is to take each point in the given data set and connect it to the nearest centroid. After this step, k new "means"(centroids) will be re-computed and a new binding has to be done between the same data set sets and the nearest new centroid. After several iterations, the K centroids would not change any more and thus the algorithm finally minimize an objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - \mu_j \right\|^2$$

where $\left\| x_i^{(j)} - \mu_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centroid μ_j. Assuming there are n sample feature vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n}$ all from the same class, and we know that they fall into k clusters, where k < n. m_i is

the mean of the vectors in cluster i. If the clusters are well separated, a minimum-distance classifier can be used to separate them. The procedure for finding the means are shown as follows:

- Make initial guesses for the means/centroids $c_1, c_2, \ldots, c_k$
- While there are changes in any centroids
    - Use the estimated centroids to classify the samples into clusters
    - For j rom 1 to k:
        - Replace c_j with the mean of all of the samples for cluster i
    - End for
- End while

5.2.3 Bayesian Information Criterion Scoring

To avoid constantly adding new centroids, Bayesian Information Criterion (BIC) scoring is introduced to provide an early stop. According to [45], the BIC is expressed as follows:

$$BIC(\emptyset) = \hat{l}_{\emptyset}(D) - \frac{\rho_{\emptyset}}{2} \cdot logR \quad \textbf{(5.14)}$$

where $\hat{l}_{\phi}(D)$ represents the maximum log-likelihood of the data set D according to hypothesis ϕ, and ρ_{ϕ} is the number of parameters in ϕ. BIC penalizes the increase of likelihood as the complexity of the model grow. Given two hypothesis ϕ_1 and ϕ_2, each one of them is a function of number of centroids K, element number R_j within the j^{th} cluster, and variance σ. Therefore, when ϕ_2 has more cluster than ϕ_1 and $BIC(\phi_2) > BIC(\phi_1)$, hypothesis ϕ_2 could be a better model than hypothesis

ϕ_1. In reference [46], Pelleg and Moore implemented the BIC coring for X-means clustering algorithm to determine the optimum number of cluster. The clustering algorithm is developed with "identical spherical assumption" that the data is modeled based on n Gaussians, each with identical variance σ and differing position μ_n. The probability of a data point i located at x_i can be expressed as:

$$P(x_i) = \sum_{j=1}^{K} P\left(x_i \in D_j\right) \cdot P(x_i | x_i \in D_j) \quad \textbf{(5.15)}$$

where $P\left(x_i \in D_j\right)$ indicates the probability of data point i is an element of cluster D_j. It can be computed under the maximum likelihood:

$$P\left(x_i \in D_j\right) = \frac{R_j}{R} \quad \textbf{(5.16)}$$

where R_j is the number of elements in the j^{th} cluster while R represents the number of elements in the whole given data set D. The second term $P(x_i | x_i \in D_j)$ represents the probability of element i is positioned at x_i given x_i is within in D_j. It is a multivariable Gaussian distribution with centroid located at μ_j and variance equal to σ^2.

$$P\left(x_i \middle| x_i \in D_j\right) = \frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}} \exp\left(-\frac{1}{2\sigma^2} \left\| x_i - \mu_j \right\|^2\right) \quad \textbf{(5.17)}$$

Substituting equations (5.15) and (5.16) in to (5.17) yields:

$$P(x_i) = \frac{R_j}{R}\frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}}\exp(-\frac{1}{2\sigma^2}\left\|x_i - \mu_{(i)}\right\|^2) \quad \textbf{(5.18)}$$

Convert the whole data set D into log-likelihoods:

$$l(D) = \log\prod_i P(x_i) \; = \; \sum_i \log P(x_i)$$

$$= \sum_{j=1}^{K}\sum_{x_i \in D_j}\left[\log(\frac{R_n}{R}) + \log\left(\frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}}\right) - \frac{1}{2\sigma^2}\left\|x_i - \mu_{(i)}\right\|^2\right]$$

$$= \sum_{j=1}^{K}\left[R_j\left(\log\frac{R_j}{R} - \frac{M}{2}\log(2\pi\sigma^2)\right) - \frac{1}{2\sigma^2}\sum_{x_i \in D_j}\left\|x_i - \mu_{(i)}\right\|^2\right] \quad \textbf{(5.19)}$$

By setting $\frac{\partial l(D)}{\partial \sigma} = 0$, the maximum likelihood estimate can be computed:

$$\hat{\sigma_j}^2 = \frac{1}{MR}\sum_i\left\|x_i - \mu_{(i)}\right\|^2 \quad \textbf{(5.20)}$$

According to reference [47], the estimate for the variance is proportional to the sum of the distances from each data point to the nearest centroid μ_j and therefore this sum can be separated with several sum by all clusters.

$$\sum_i\left\|x_i - \mu_{(i)}\right\|^2 = \sum_j\sum_{i \in D_j}\|x_i - \mu(i)\|^2 \quad \textbf{(5.21)}$$

Since each individual cluster is a spherical Gaussian, the unbiased estimator or the variance can be shown as:

$$\hat{\sigma}_j^2 = \frac{1}{M(R_j - 1)} \sum_{i \in D_n} \|x_i - \mu_j\|^2 \quad \textbf{(5.22)}$$

Combining equations (5.21) and (5.16), the sum can be changed to:

$$\sum_i \|\mathrm{x_i} - \mu_{(i)}\|^2 = M \sum_j (R_j - 1)\hat{\sigma}_j^2 \quad \textbf{(5.23)}$$

$$= M\left(\sum_j R_j - K\right)\hat{\sigma}_j{}^2 = M(R - K)\hat{\sigma}_j{}^2 \quad \textbf{(5.24)}$$

$$\hat{\sigma}^2 = \frac{1}{M(R - K)} \sum_i \|x_i - \mu_{(i)}\|^2 \quad \textbf{(5.25)}$$

Implement this maximum likelihood assumption from equation (5.23) and using $\sum_{j=1}^{K} R_j = R$, it yields:

$$\hat{l}(D) = \sum_{n=1}^{K} R_n log R_n - R log R - \frac{RM}{2}\log(2\pi\hat{\sigma}^2) - \frac{M}{2}(R - K) \quad \textbf{(5.26)}$$

This maximum likelihood can be broken into two parts, the first part is model dependent while the second part is model independent.

$$\hat{l}(D) = \left[\sum_{j=1}^{K} R_j \log(R_j) - \frac{MK}{2} - \frac{RM}{2}\log(\hat{\sigma}(\phi)^2)\right] - \left[\frac{MR}{2} + R log(R) + \frac{RM}{2}\log(2\pi)\right] \quad \textbf{(5.27)}$$

$$\hat{l}_{model-dependent}(D) = \left[\sum_{j=1}^{K} R_j \log(R_j) - \frac{MK}{2} - \frac{RM}{2}\log(\hat{\sigma}^2)\right] \quad \textbf{(5.28)}$$

$$\hat{l}_{model-independent}(D) = \left[\frac{MR}{2} + Rlog(R) + \frac{RM}{2}\log(2\pi)\right] \quad \textbf{(5.29)}$$

According to the BIC equation (5.14) and the model dependent part, the inequality equation holds when hypothesis $\emptyset_2$ is better than the first hypothesis. Therefore, $BIC(\emptyset_2) > BIC(\emptyset_1)$ can be re-wrote as

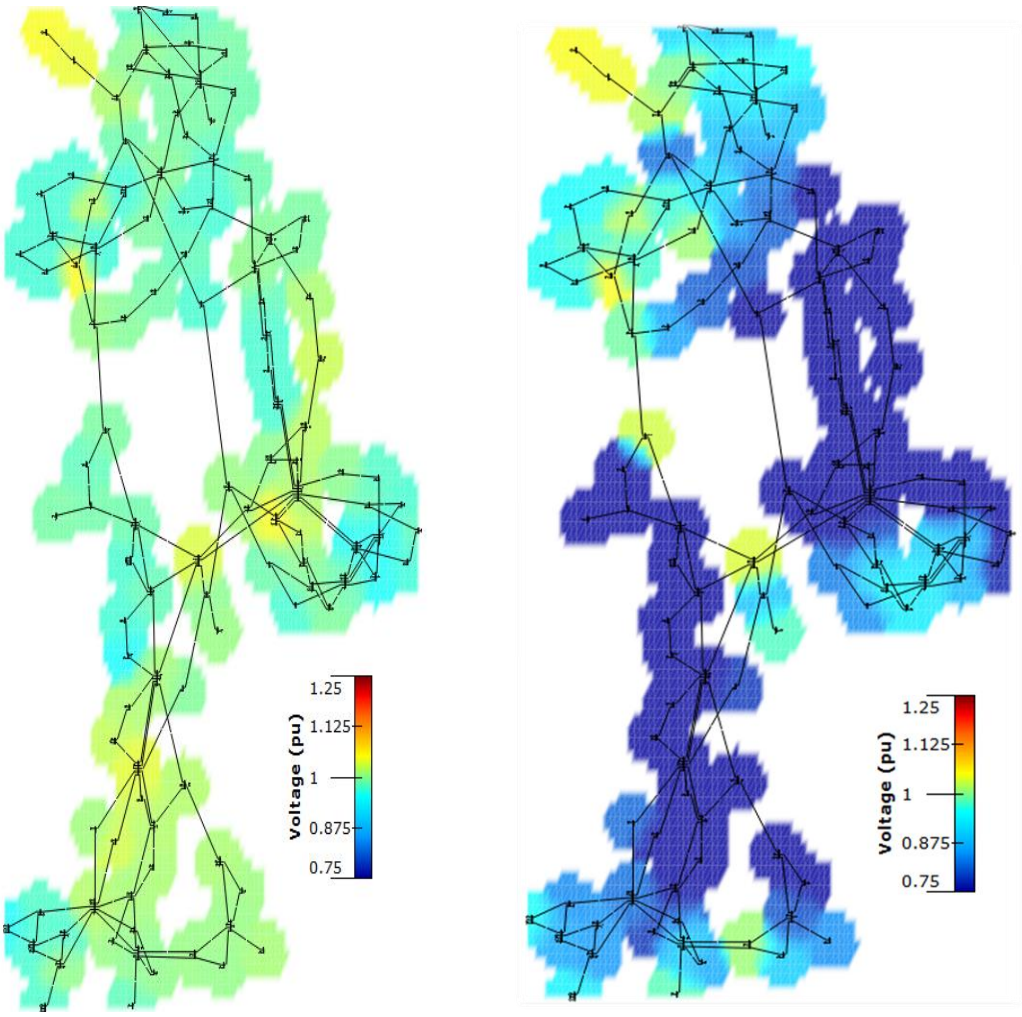
$$\hat{l}_{model-dependent}(D, \phi_1) - \frac{P_{\phi_2}}{2}\log(R) \; > \; \hat{l}_{model-dependent}(D, \phi_2) - \frac{P_{\phi_2}}{2}\log(R) \quad \textbf{(5.30)}$$

$$\left[\sum_{j=1}^{K} R_j(\phi_2)\log\big(R_j(\phi_2)\big) - \frac{MK(\phi_2)}{2} - \frac{RM}{2}\log(\hat{\sigma}(\phi_2)^2)\right] - \frac{P_{\phi_2}}{2}\log(R)$$

$$> \left[\sum_{j=1}^{K} R_j(\phi_1)\log\Big(R_j(\phi_1)\Big) - \frac{MK(\phi_1)}{2} - \frac{RM}{2}\log(\hat{\sigma}(\phi_1)^2)\right] - \frac{P_1}{2}\log(R) \quad \textbf{(5.31)}$$

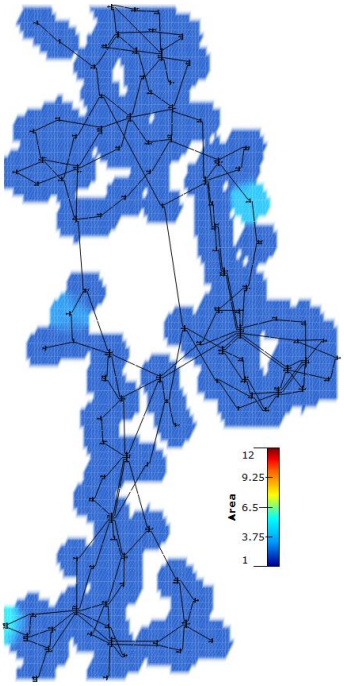
5.3 Simulation on IEEE 118 Bus System

The proposed approach for VCA identification is evaluated using IEEE 118 bus standard system network. The network is running in PSSE software and visualized based on its GPS information. The original system model is shown in Figure 5-3. (a) shows the voltage measurements when the system is under normal loading condition. (b) shows the system under heavy loading condition when the system load increases 150% evenly.

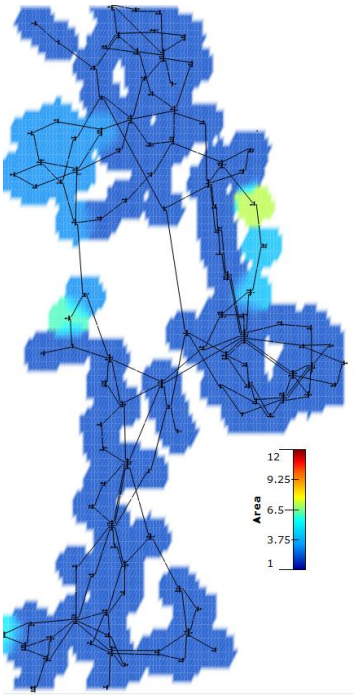


(a) (b)

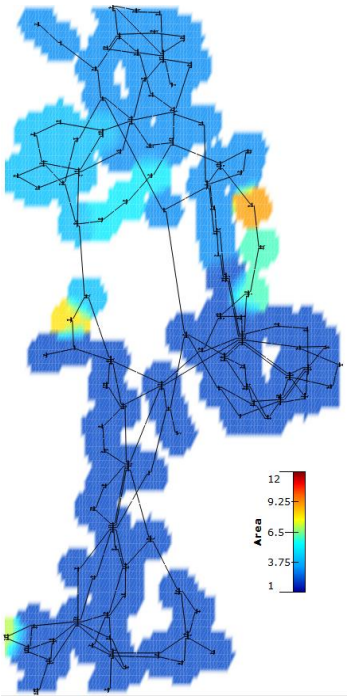
Figure 5-3 (a) Voltage measurements when system is under original loading. (b) Voltage measurements when system is under heavy loading (150%)



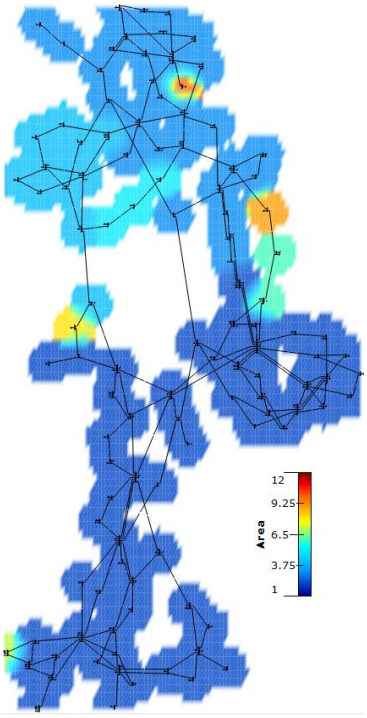
(a) $\alpha = 11$



(b) $\alpha = 11.6$



(c) $\alpha = 12$



(d) $\alpha = 12.9$

Figure 5-4 Identified VCA based on different α

To implement the proposed scheme, the α threshold value is initialized as 1 and increase with a step size 0.1 until α reaches 15. Figure 5-4 shows 4 different α values and their corresponding identified VCAs. As shown in the simulation, when the α value increases, more VCAs are identified since more weak transmission elements are removed. To determine which α should be the one indicating the VCAs, the methodology introduced in section 5.2 is implemented.

Figure 5-5 shows the plotted QV curve at bus 2 in IEEE 118 bus system. The red dot at the elbow of QV curve shows the minimum QV condition [Qmin = -338.68 MVAR, Vmin = 0.535 p.u].

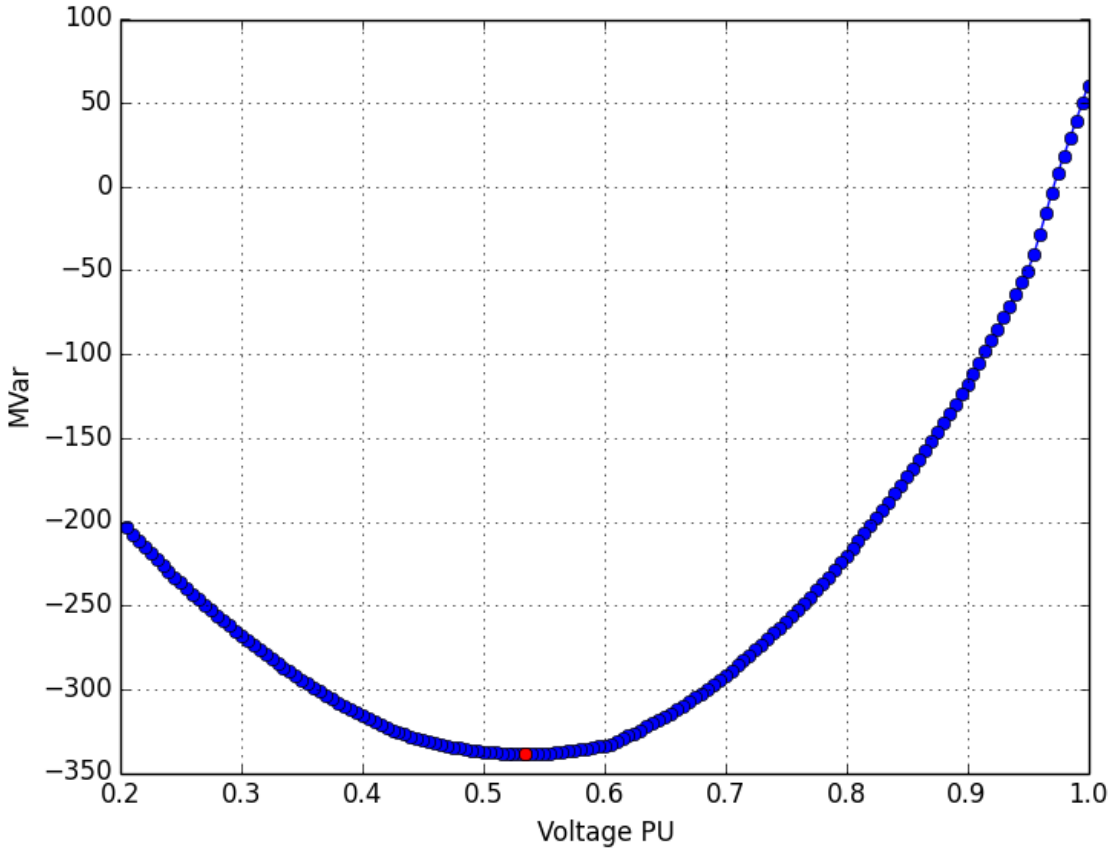


Figure 5-5 QV curve at bus 2

QV analysis is applied to the system for each PQ bus, the database saves all minimum QV points' reactive power output and voltage measurements as well as the Q metrics computed based on methodology introduced in section 5.2.1. Figure 5-6 shows the computed Q metric and the Binary status at bus 24 when the QV analysis is applied to all 64 PQ buses. It is observed that when bus 14 reaches its minimum QV point, the generator at bus 24 also reaches it reactive power limit. However, when the QV analysis is applied to bus 41, the binary status shows 0 while the Q metric equals to 0.932. The Q metric reveals that the generator is actually close to its limit. This information cannot be observed if only binary status is applied. Same situations are observed when QV analysis is applied to generator located at bus 21. It brings more information when the Q metric is introduced.

Figure 5-7 shows the computed Q metric at bus 1, 6, and 15. It can be seen the Q metric at bus 1 and 6 are sharing a similar pattern. Compared with the other two buses, the Q metric computed at bus 15 has a shape. Therefore, based on the observation, the attribute showing the Q metric at bus 1 and 6 had higher similarity then the Q metric computed at bus 15.

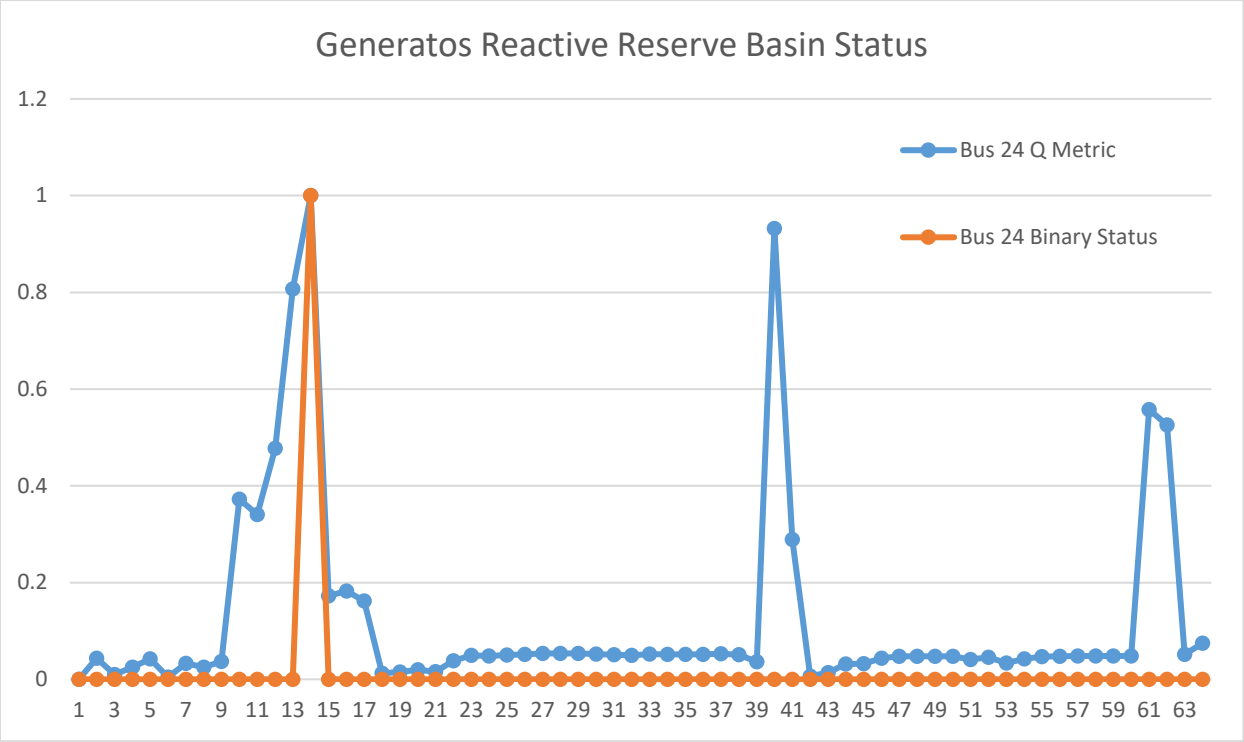


Figure 5-6 Q metric and binary status of the same bus

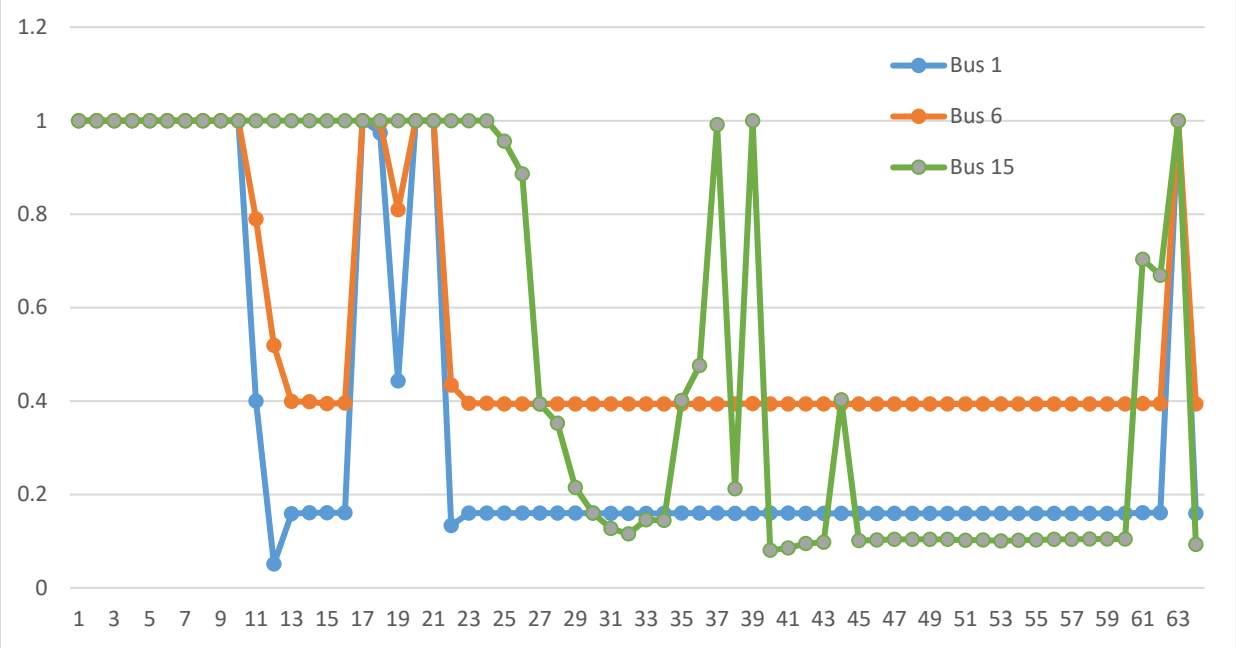


Figure 5-7 Q metric at bus 1, 6, and 15

The Kmeans clustering is implemented while the BIC is computed based on different number of clusters. Figure 5-8 shows the best BIC is the one associated with a cluster size equal to 3. Since Kmeans is initialized with different centroids, the results is greatly impacted by the centroid values. Therefore, the BIC computation runs 1000 times and the distribution of the identified K associated with the elbow BIC is shown in Figure 5-9.

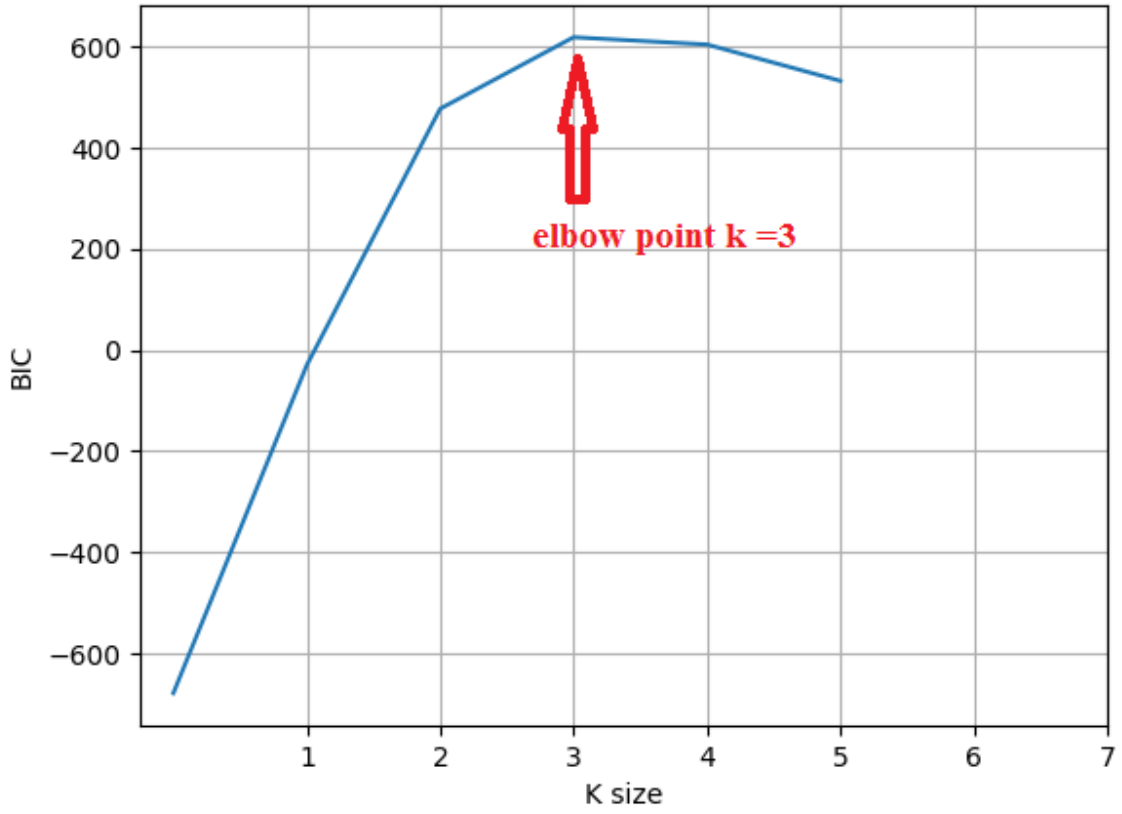


Figure 5-8 BIC corresponding to number of cluster

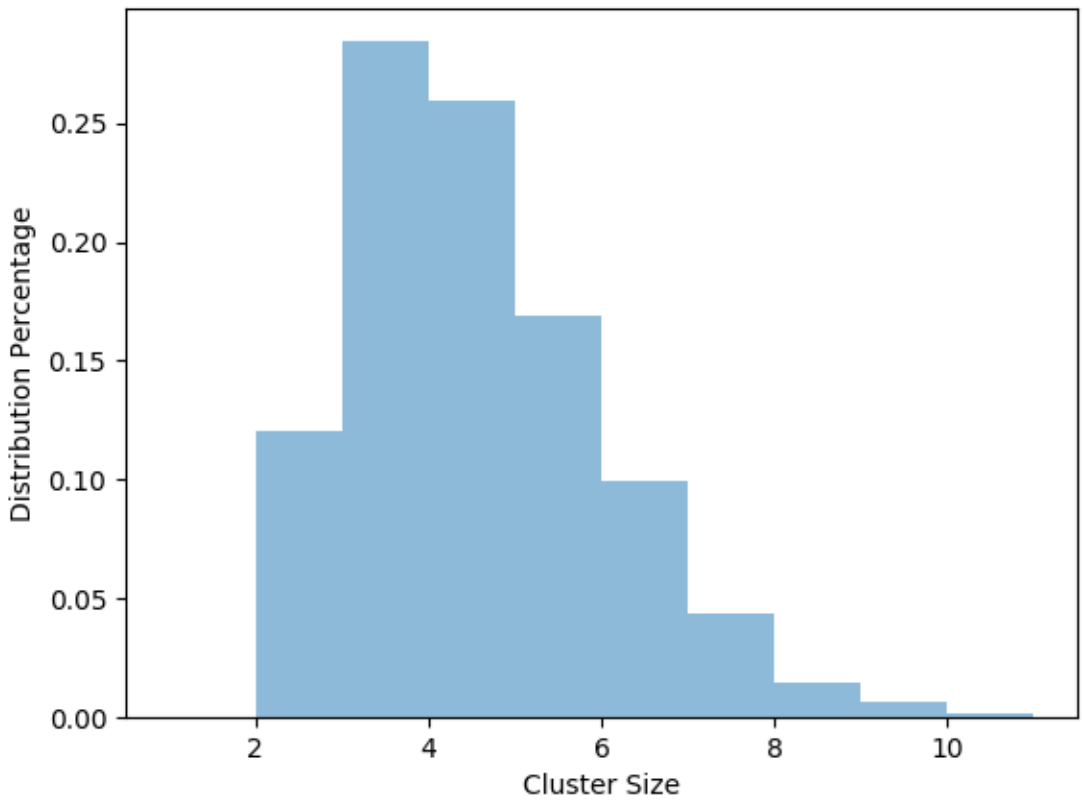


Figure 5-9 Normalized distribution of number of clusters

According to Figure 5-9, the peak of the distribution occurs when cluster size equals 3 and thus, K = 3 is selected as the number of cluster for Kmeans clustering. The Kmeans algorithm is implemented in the saved QV analysis database and final clustered system is shown in Figure 5-10. It can be seen that the whole system is divided into three major areas. Area 1 is completely separated from area 3. Some parts of area 3 are surrounded by area 2.

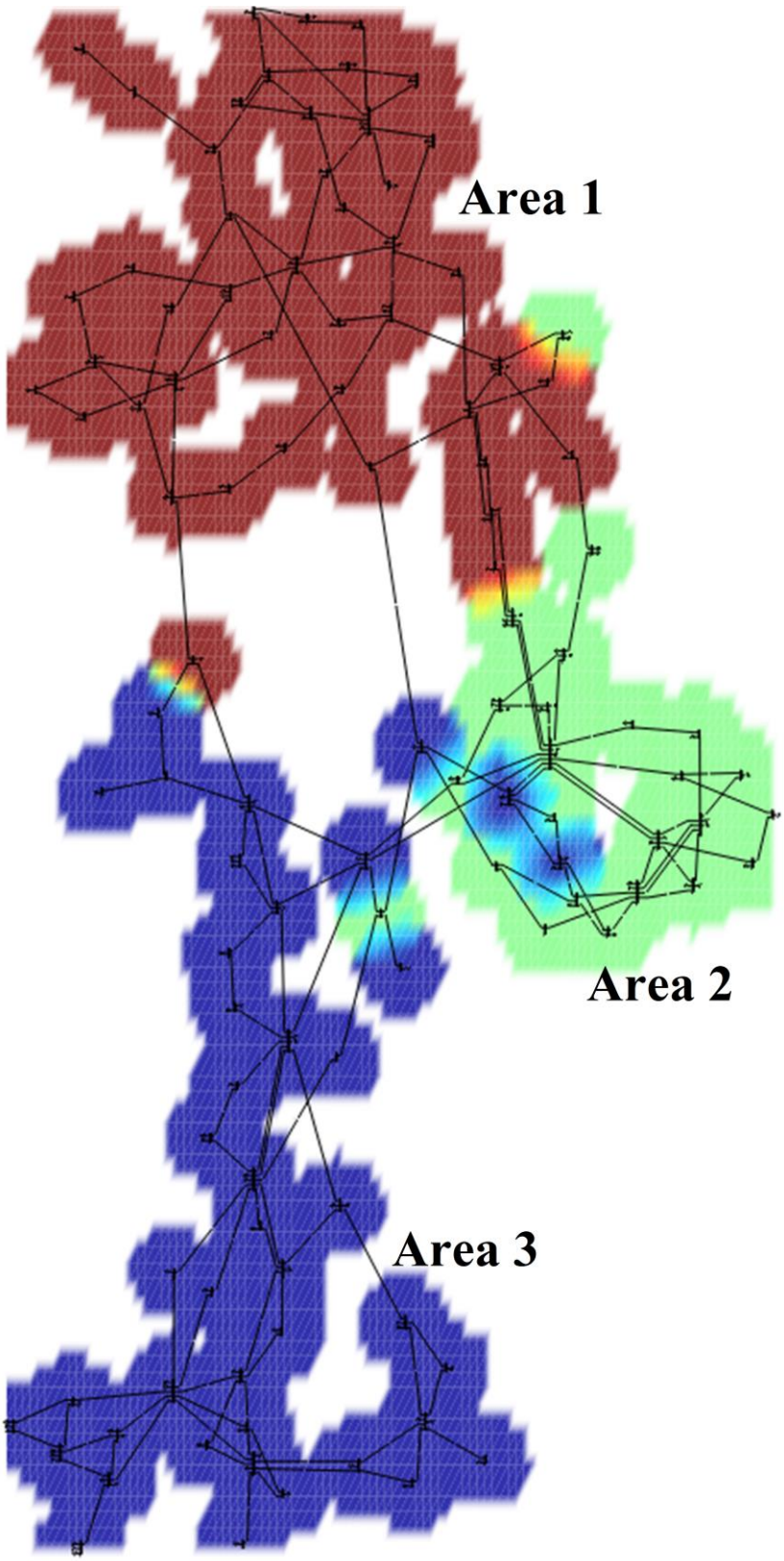


Figure 5-10 System divided using QV analysis

Comparing with the identified VCA showing in Figure 5-4, an α threshold of 12 provides the most similar clustering as the QV analysis shown in Figure 5-10. According to Figure 5-11 (a), the system can be decomposed into 2 major areas, and they are called the "North" and "South". As can be seen, the "Area 1" in Figure 5-11 (b) covers all the small sub-areas identified in the North found in Figure 5-11 (a).

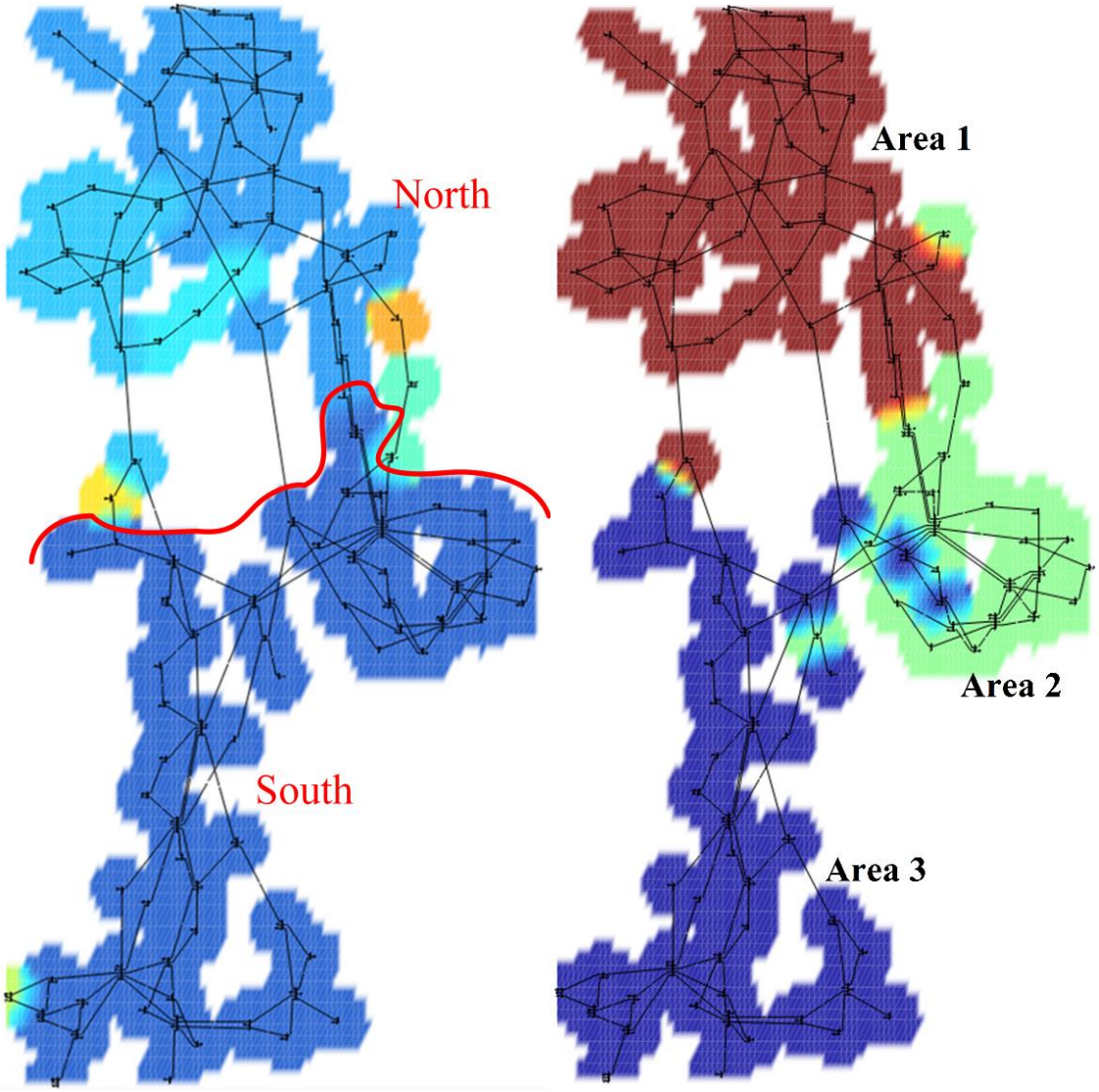


Figure 5-11 (a) Identified VCAs ($\alpha = 12$) , (b) Areas identified using QV analysis

The bus information of sub-system North and South are shown in Table 5.3.

Table 5.3 “North” and “South” of the system.

Sub-systems	**Buses**
North	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 43, 44, 45, 72, 87, 107, 113, 114, 115, 117
South	42, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 108, 109, 110, 111, 112, 116, 118

5.4 System Network Partition

While analyzing a large system’s voltage problem, power system engineers are usually interested in the behavior of a certain part of the system. Such part of the system is called the internal sub-system while the rest of the system is considered as the external sub-system. Based on the internal and external sub-system’s information, static equivalence or reduction can be implanted to reduce the complexity of the external system model while retaining the its impact on the study system. Through this process, a large power system can be reduced significantly while keeping its accuracy associated with some specific studies such as static or dynamic voltage control.

As discussed in the previous section, once the voltage control area (VCA) is determined, the internal and external sub-system can be identified and a reduced

model based on the VCAs information can be acquired. In this study, one of the classical methods for calculating static network equivalents called Radial Equivalents, Independent (REI) Network Reduction is implemented. Figure 5-12 shows the concept of REI, where f represents the node where the REI equivalent is developed. A is a set of nodes of internal sub-system while g_m and g_n are the nodes of the external network. f can directly or indirectly connect with external buses. $\bar{z}_{mf}$, $\bar{z}_{nf}$,and $\bar{y}_{f0}$ represented the impedances and shunt admittance after REI equivalence.

To conduct the REI, first, all the PV buses will be equivalent as a PQ loads. The load equivalent admittances will be added to the diagonal elements of the network admittance matrix. A relation between currents and voltages can have:

$$\begin{bmatrix} \bar{I}_r \\ 0 \end{bmatrix} = \begin{bmatrix} \bar{Y}_{rr} & \bar{Y}_{rc} \\ \bar{Y}_{cr} & \bar{Y}_{cc} \end{bmatrix} \begin{bmatrix} \bar{v}_r \\ \bar{v}_c \end{bmatrix} \quad \textbf{(5.32)}$$

where k represents the vector of indices of all generators contained in a given external network; f is the index of the neighboring bus at which the REI equivalence is computed; $\bar{v}_c$ is a vector of indices of the load buses in the external network; $\bar{v}_r$ is a vector yielded from the union of k and f.

Since the currents relative to loads are 0 and the admittance matrices has the load equivalent admittances already, $\bar{v}_c$ can be replaced as:

$$\bar{I}_r = (\bar{Y}_{rr} - \bar{Y}_{rc}\, \bar{Y}_{cc}^{-1} \bar{Y}_{cr}) \bar{v}_r = \bar{Y}_r \bar{v}_r \quad \textbf{(5.33)}$$

$$\begin{bmatrix} \bar{I}_k \\ \bar{I}_f \end{bmatrix} = \begin{bmatrix} \bar{Y}_{kk} & \bar{Y}_{kf} \\ \bar{Y}_{fk} & \bar{Y}_{ff} \end{bmatrix} \begin{bmatrix} \bar{v}_k \\ \bar{v}_f \end{bmatrix} \quad \textbf{(5.34)}$$

Where:

$$\bar{Y}_r = \begin{bmatrix} \bar{Y}_{kk} & \bar{Y}_{kf} \\ \bar{Y}_{fk} & \bar{Y}_{ff} \end{bmatrix}$$

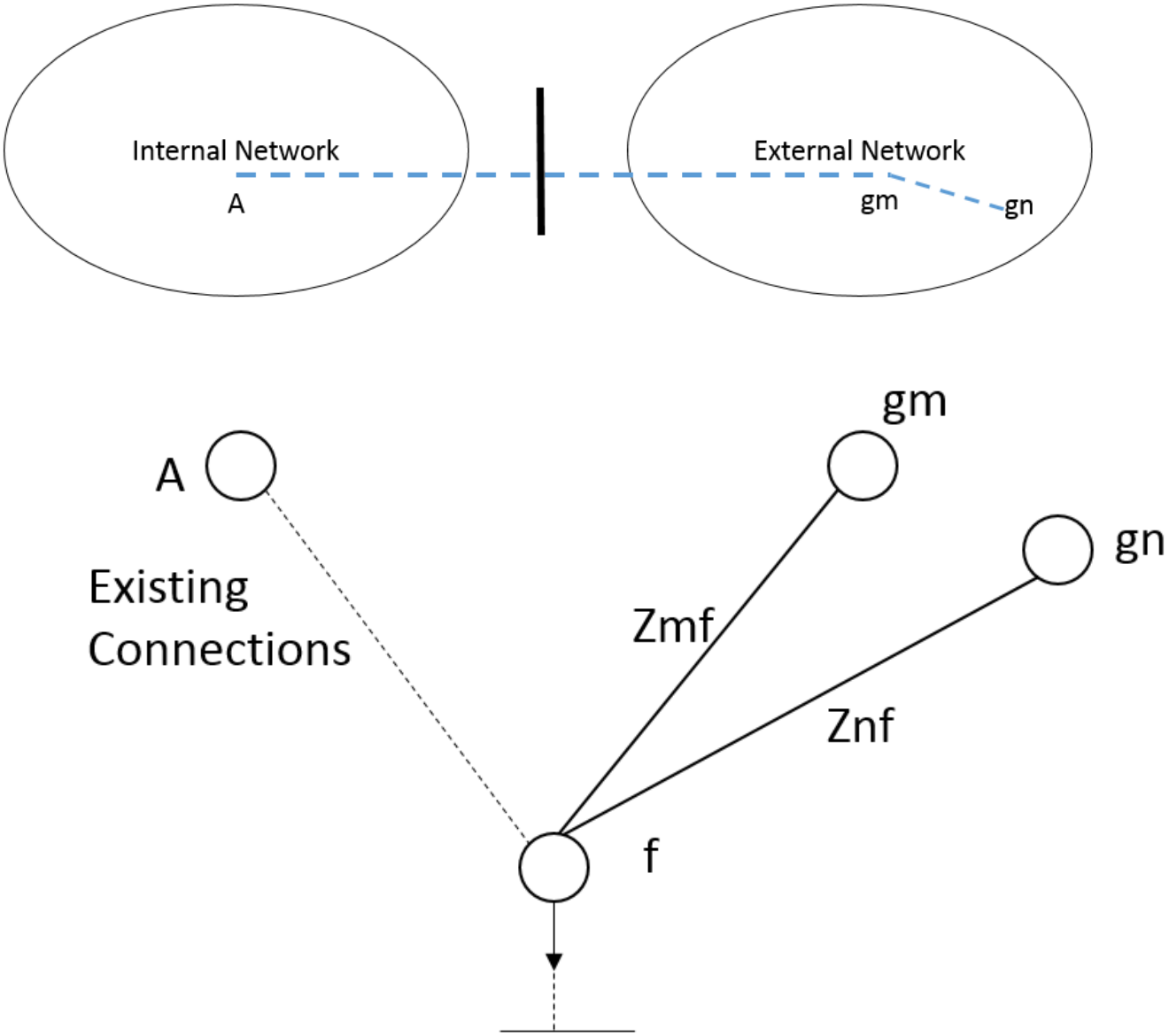


Figure 5-12 REI Equivalent

The elements of $\bar{Y}_{fk}$ represents the series admittances of the lines that between the frontier bus f and the retained generator at bus k. Therefore, we can have:

$$\bar{Y}_{fk} = \left[\frac{1}{\bar{z}_{f1}} \quad \frac{1}{\bar{z}_{f2}} \quad \dots \quad \frac{1}{\bar{z}_{fk}}\right] \tag{5.35}$$

Once the REI equivalent is completed, the node aggregation is straightforward:

$$\bar{Y}_{fg} = \frac{1}{\sum_k \bar{Y}_{fk}} \tag{5.36}$$

The final step is to aggregate the synchronous machine at fictitious bus g. The details of the generator aggregation is shown in [48].

The algorithm is implemented on IEEE 118 bus bench mark model using EEQV module in PSSE. The generators in the external sub-system are retained. The control variables locating in the internal and external areas are shown in Table 5.4

Table 5.4 Capbanks locations and Status in both sub-systems

Sub-systems	Location (Buses)	Before Close (MVAR)	After Close (MVAR)
North	34	0	50
	44	10	50
	45	10	50
South	48	0	100
	74	12	100
	105	0	20

The decomposed sub-systems are evaluated by introducing 1000 different loading conditions. The original system shows 47% of the OC are secure while 53% of them are insecure. Both of the sub-systems are compared with the original system network. Figure 5-13 and Figure 5-14 show the status difference between the original system and the sub-systems: 0 indicates that the sub-system shows the same security status as the original system, while 1 shows that security status of sub-system is changed. Both of the sub-systems show the same amount of OCs that have different security status from the original system. Since only 5 out of 1000 OCs showed different status, both of the sub-systems are said to be good estimator for each part of the original system.

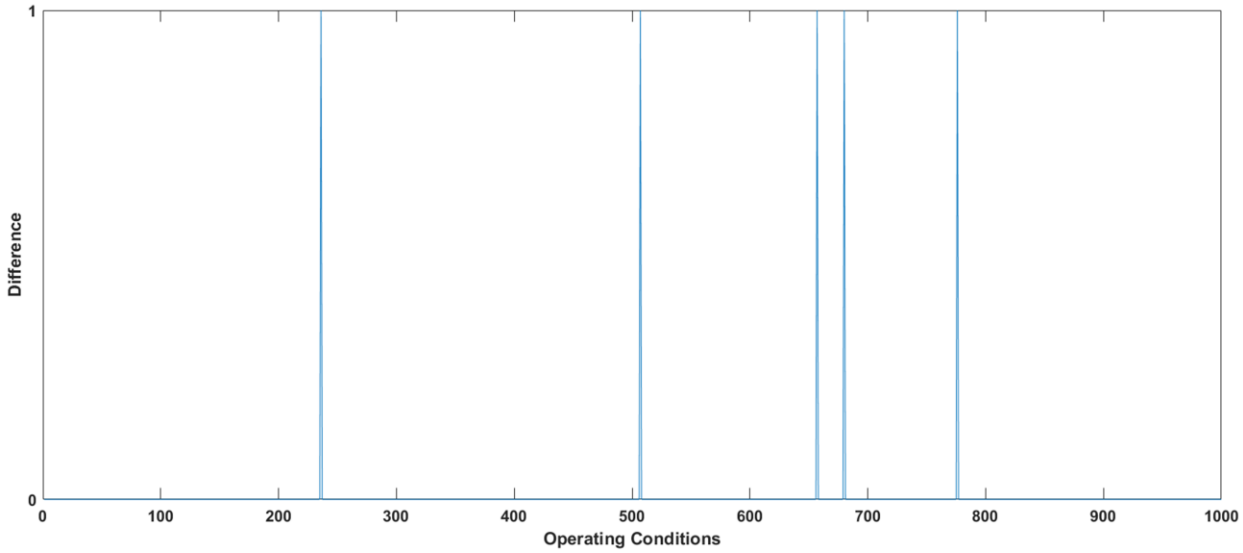


Figure 5-13 Status difference between “North” and the original system

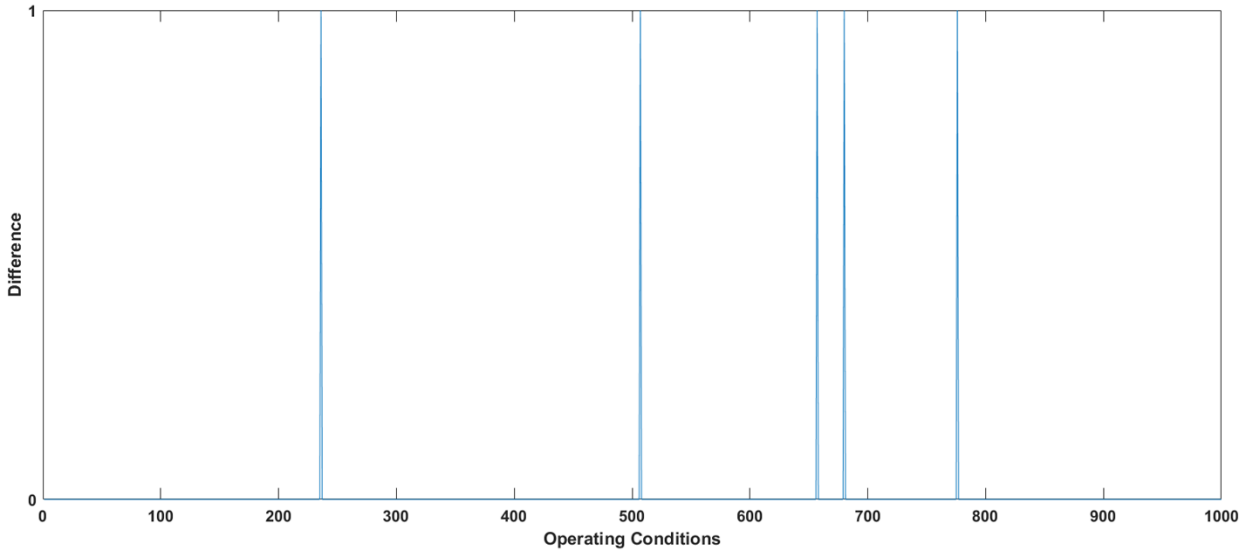


Figure 5-14 Status difference between “South” and the original system

5.5 Summary

This chapter introduces a system partition approach to decompose the system into sub-areas and thus reduce the computation burden for OC generation and control candidate selection. This chapter introduces a modal based method for reducing weak transmission stability boundaries and identify voltage control areas. This methodology requires a proper selection of function parameter α. To identify the optimum value of α, a new approach comparing modal based identified VCAs and areas clustered by QV analysis is proposed. The clustering procedure requires a database consisted of QV minimum (continue variable) for each target bus and the generator's status (binary status) during each QV analysis run. To avoid mining mixed learning sample with continue and binary data type, a new index called Q metric is introduced. This metric implements sinc function to compute the generator reactive basin reserve status for each generator when the target bus reaches its QV minimum.

The simulation is conducted based on IEEE 118 buses system. The simulation results show that the Q metric reveals more information for each generator than using binary variable. The clustered areas are compared with the VCAs identified using modal method. It is observed that the identified VCAs with $\alpha = 12$ show the highest similarity. Finally, the system is decomposed into two sub-systems according to the selected VCAs and each sub-system with smaller number of control candidates will have its own regional controller. These two sub-systems are compared with the original system model and the test results show that both of them are sharing the similar voltage security pattern as the original system.

Chapter 6
Structure of Voltage Controller Program

A software framework is an abstraction in which software providing a reusable design for a family of related problem domains. A well-defined software framework can be selectively changed by additional user-written code and thus providing application-specific software [49]. Some of the existing designs of local/region voltage control programs are still running perfectly and providing reliable control/protection for power system network. However, most of these programs are not designed as monolithic solution and it is difficult to identify their information flows. Based on these programs (OOP), it is almost impossible to use real high-speed data stream from PMU. This will also cause some unexpected hazards to data security. The work in this chapter addresses the system level design of a software framework for both local voltage controller and regional voltage controller. It also illustrated how to integrate the trained model using machine learning techniques into the controller program.

6.1 Architecture of Voltage Controller

The architecture of the voltage controller shows a “pipes and filter structure”. Three fundamental components of the voltage controller are shown in Figure 6-1:

1. The input adapter
2. The voltage controller adapter (i.e., the user defined application)
3. The Interface with EMS system

Each component of the controller is essentially a package, located at its own filter level in the architecture and has low coupling with other component. Each component may consist of several sub-packages, and a set of co-operating classes are included in each sub-package.

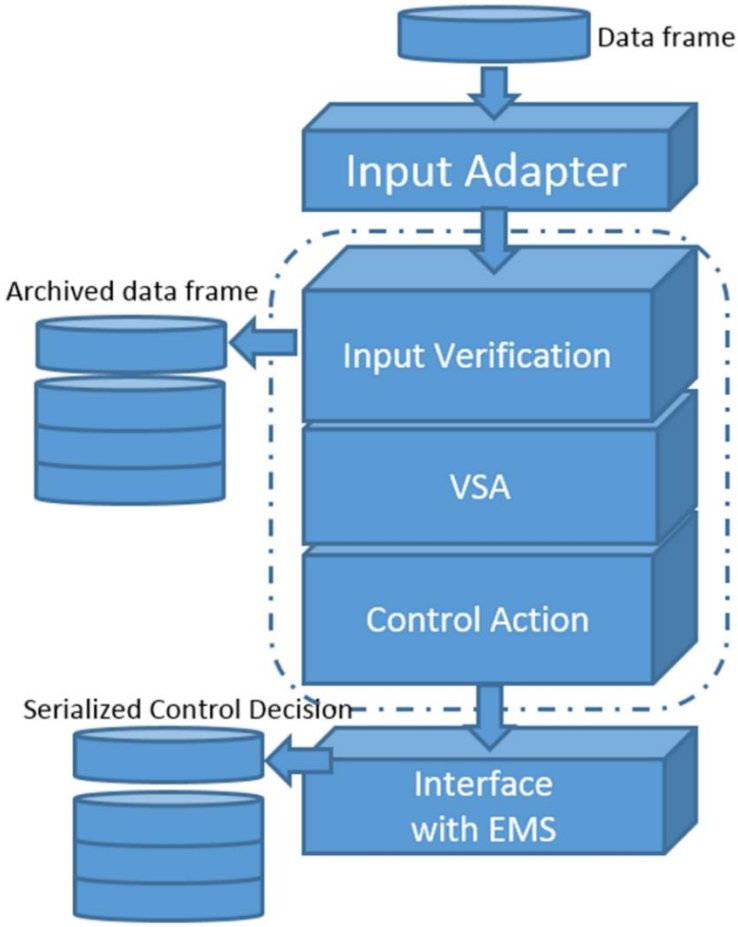


Figure 6-1 Voltage Controller’s software architecture

The voltage control adapter is a user defined application which interface of a class into another interface. The voltage controller adapter provides functions such as input verification, voltage security assessment (VSA), and voltage control action. This adapter also provides a hierarchical modeling for power system network. An example of system modeling is shown in Figure 6-2. There are two layers of the modeling. The first layer is called the Element Layer. This layer includes all elements of a specific power system component. For example, Table 6.1 shows the elements included in a transformer for load tap change.

Table 6.1 Elements of Transformer

Parameters	Description	Status
LocRemV	Local remote	Remote
ScadaSwV	Scada switch	On
HighSideV	Transformer high side breaker	Close
LowSideV	Transformer low side breaker	Close
VoltsV	Transformer low side voltage	Larger than 101.2 or lower than 128.8
Mwv	Real power in transformer	Larger than 0.5 MW
MvrV	Reactive power in transformer	Larger than 0.5 MVAR
TapV	Tap position	Larger than -16 and lower than 16

The second layer is called the Network layer. This layer presents all components of the system network. As can be seen in Figure 6-2, the system network includes several capacitor banks, transformers, substation information, etc.

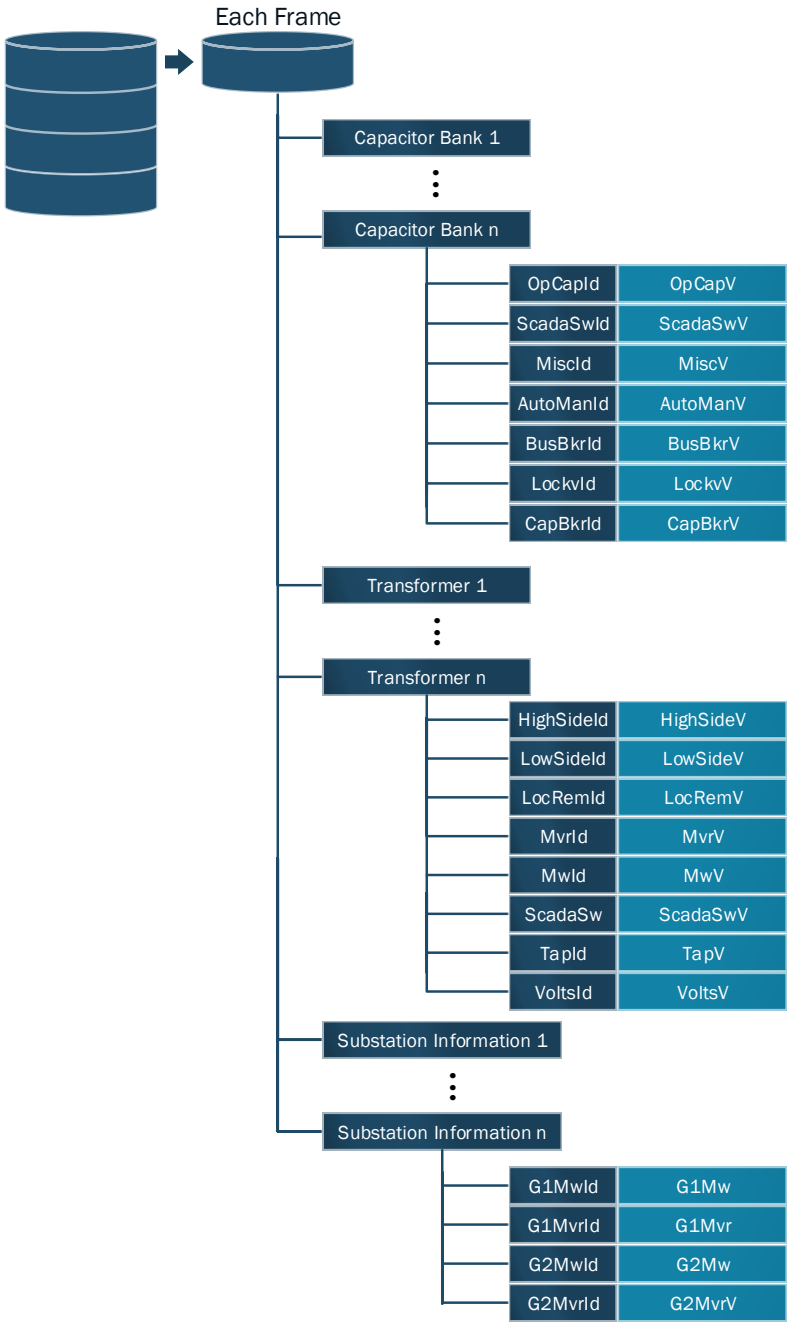


Figure 6-2 Voltage Controller's Data Structure

6.2 Information Flow

The information flow is also presented in Figure 6-1. To begin with, the voltage controller adapter is initialized with an xml based configuration. This configuration represents the system's network components and measurements' default values as well as the keys associated with each measurement.

When the initialization is completed, the synchrophasor data aligned inside the PDC starts streaming through the input adapter. Based on the pre-defined measurement keys, the synchrophasor data will be mapped into the pre-defined data frame with other system information selected from the EMS system. The breaker status selected from the EMS system defines the system topology. In recent PMU development, the breaker status is deigned to be available with each relay connected with a PMU. If the sychrophasor does not provide the breaker status, the breaker status can be also selected from the EMS system, however, the EMS system based selected breaker status is no longer time aligned. To guarantee the effectiveness of the controller, the system network still can be observed by using a technique such as system topology estimation. This technique is well studied in Distribution level, but there will be more research to be conducted in the transmission level.

When the data frame arrived at the voltage controller, the input verification adapter starts the verification. During this process, the controller will verify the bus voltage, breaker status, capbank switch status, LTC tap position, transformer service status, real/reactive power flow through the branches (including transformers). For local voltage controller, the delay counters will be adjusted to guarantee that two capbanks will not be operate within 30 minutes of each other and each capbank will not operate within a minute of a tap control. The tap control is also limited by the delay counter that the control command will not be executed within a minute of a capbank control. For both local and regional voltage controller, the controller will

confirm if the control results satisfy the system's expectation. When the verification procedure is completed, the controller starts the VSA adapter.

For the local voltage controller, the VSA is simply based on the pre-defined threshold value of target bus's voltage, reactive power flow inside control transformers, and reactive power output from the neighboring generators. When the measurement is lower or higher than the limits, the control logics will be triggered and the control signal will be sent back to the EMS system. In regards to the regional voltage control, the VSA is conducted through evaluating PMU voltage magnitude at each bus using one decision tree.

If the VSA adapter indicates "INSECURE" system status, the data frame will flow into the control action adapter. For local voltage controller, the control action adapter is a module with simple pre-defined rules. For regional voltage controller, the control action adapter is designed to select control action using multiple DTs. The control actions associated with DTs that indicating "SECURE" system status will be compared with the current system control status using logical exclusive OR (bitwise XOR). For example, C_1, C_2 and C_3 represented different valid control options; $C_{current}$ denotes the current system capbank status:

$$C_1 = [0,0,0,1,0,0]$$
$$C_2 = [0,0,0,1,1,1]$$
$$C_3 = [1,0,0,0,1,1]$$
$$C_{current} = [0,0,0,0,1,1]$$
$$B_1 = C_1 \; XOR \; C_{current} = [0,0,0,1,1,1]$$
$$B_2 = C_2 \; XOR \; C_{current} = [0,0,0,1,0,0]$$
$$B_3 = C_2 \; XOR \; C_{current} = [1,0,0,1,0,0]$$

It can be seen that B_1 has more $1s$ than B_2 and B_3 which indicates that C_1 requires more capbank to be switched on and off. Therefore, C_2 and C_3 are considered as

better control options. Besides the minimum switching operation, the regional controller also requires the minimum reactive power provided from the control decision. Supposed the third capbank at C_2 has 100 MVAR reactive power while the first capbank at C_3 has only 20 MVAR, C_3 will be selected as the best control option since it can secure the system while keeping the reactive power support minimum.

6.3 A Customized Design of Regional Voltage Controller

A physical view of regional voltage controller is shown in Figure 6-3. The analytic is divided into two parts: online and offline. The offline part includes the EMS system that archives the system snapshots. Each single system snapshot provides system topology, voltage measurements, and power flow information every 5 minutes. An offline adapter which creates/updates decision trees is also included in the offline section.

The online section includes multiple Phasor Measurement Units (PMU) which provide synchronized synchrophasor data with voltage/current magnitude and angle in real time. A Phasor Data Concentrator (PDC) is also presented in the online section. The PDC receives and time-synchronizes phasor data from PMUs to provide output data stream. Since the system might not be covered with PMU at each single bus, the system status still can be observed using state estimator. The Linear State Estimator developed in Virginia Tech back to 2013 is able to provide estimated phasor data and make the system network fully observable. In this study, the PMU estimated phasors are also considered as valid input to the voltage controller.

Instead of re-writing a new input adapter, an open source synchrophasor data platform called openECA [50] is employed to process the input measurements. This

platform provides an open source API to support analytics interface between synchrophasr data and user defined applications.

The regional voltage controller is developed based on the openECA platform. It is able to provide system status inspection, VSA, and regional voltage control. The decision trees provided VSA and control decisions are mapped from the offline adapter.

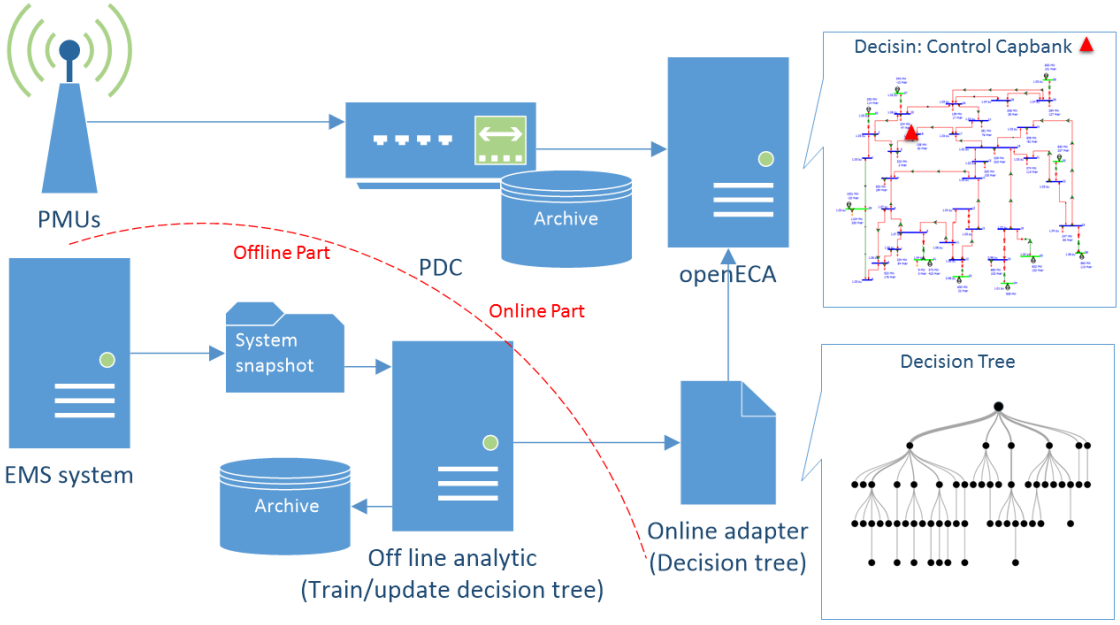


Figure 6-3 Concept of Software Framework

6.4 Software Testing

Before implementing the controller into the field, the application should be evaluated in simulation environment. In this study, dynamic simulation is not required, the voltage measurements are yielded from the power flow simulation. The developed controller is initially tested with PSS/E based model. The PSS/E model is able to continuously simulate the power system and to provide measurements as input signals for the voltage controller written in C#.

When the logic is triggered inside the voltage controller, the control decision will be sent back and executed in PSS/E. This cross-platform setup provides an alpha version simulation environment. Though this environment is not able to provide real time simulation, it still provide satisfactory test environment to evaluate different functions of analytics.

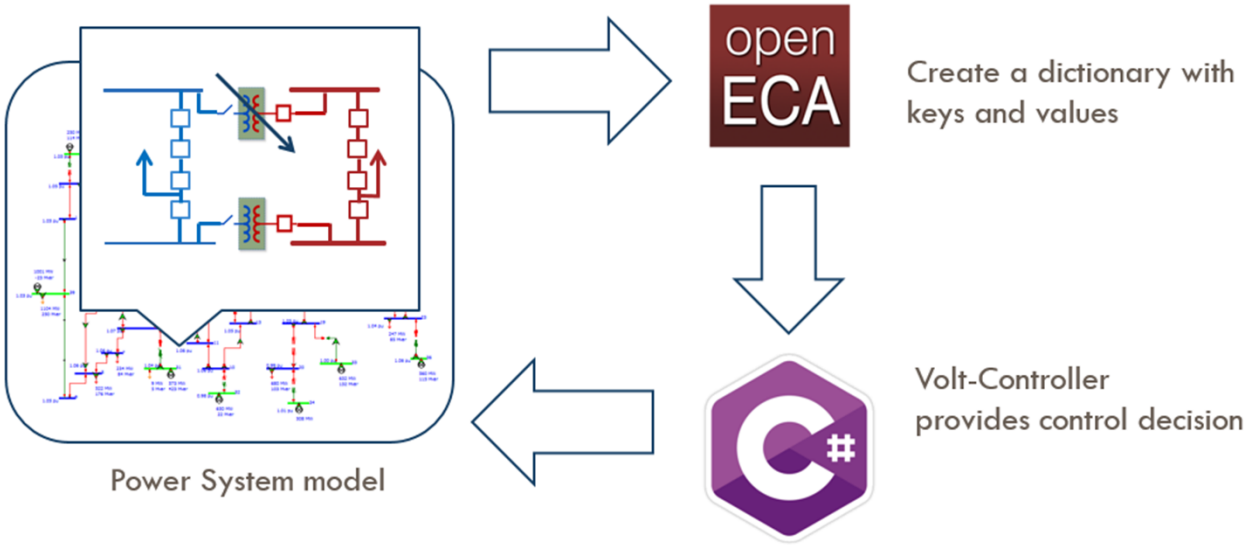


Figure 6-4 Cross-platform testing

Figure 6-5 shows a typical power system model for local voltage control. The substation local voltage controller has both Capacitor Bank and LTC control schemes. Besides the voltage at bus 314691 and voltages at Pamp 115 kV substation and Crew 115 kV substation, the reactive power flow inside transformers in Farm substation and reactive power output from Clov generator are also considered as input signal.

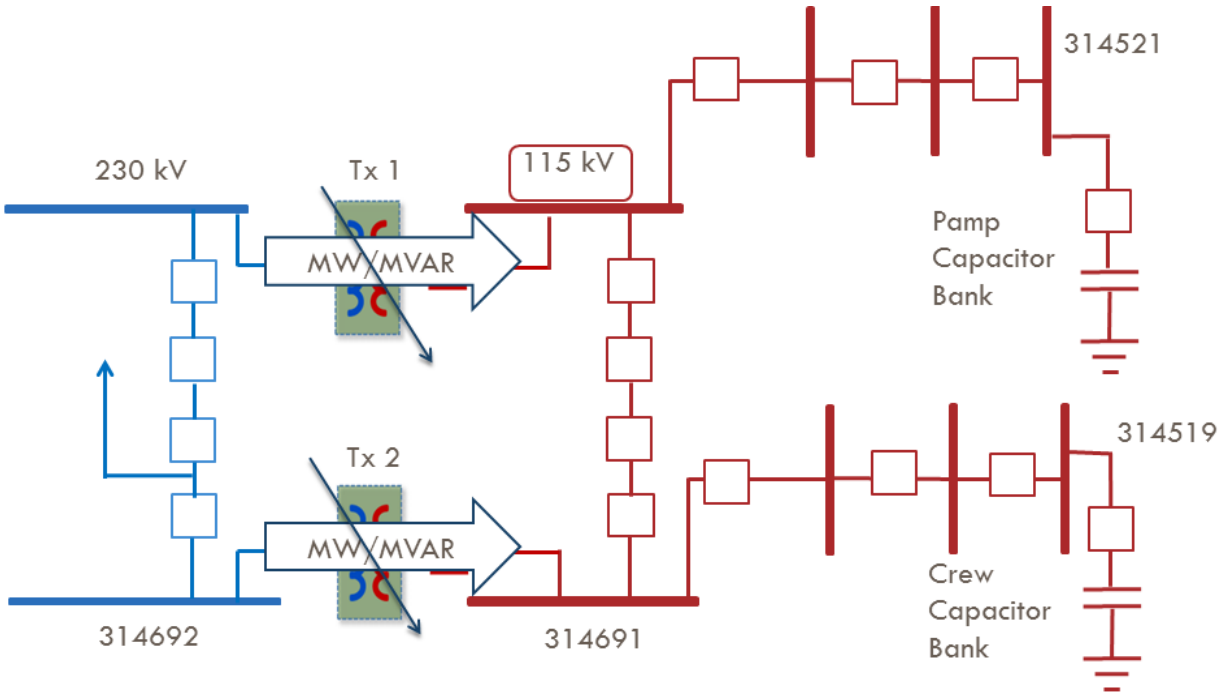


Figure 6-5 Local voltage controller

As the load increase, the voltages at the load buses drop. Figure 6-6has shown 3 times of touches of the lower limit 114 kV at time instances 11, 20,and 26, each of which has triggered tap changing in both transformers due to the sufficient spare amount to the highest tap position.

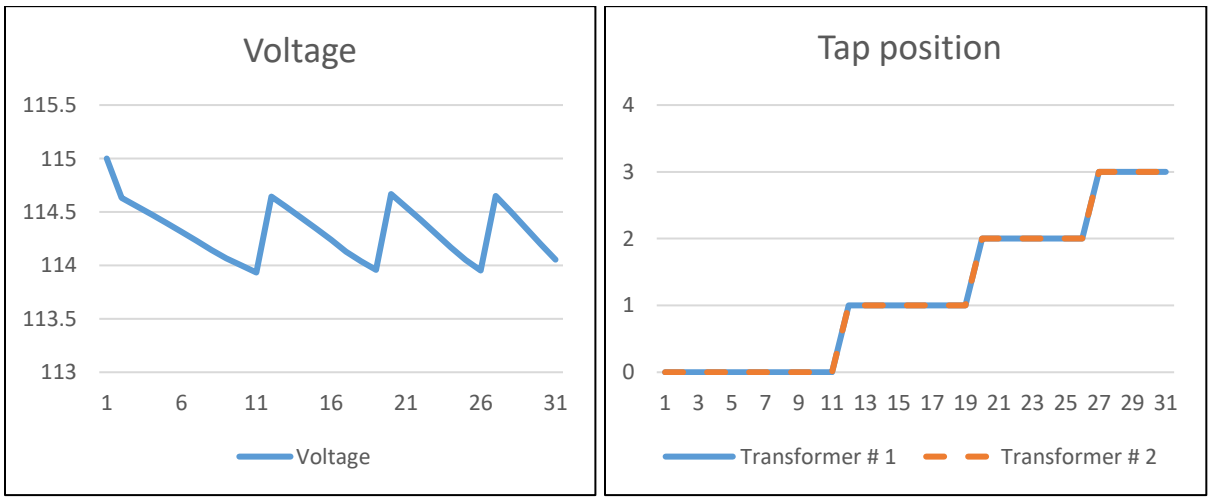


Figure 6-6 (a) Both transformers' voltages reach lower limits (b) Both transformers' voltages reach lower limits

Figure 6-7 shows the simulation conducted to demonstrate the control mechanism for two capacitor banks located at buses 314519 and 31452. At time instance 2, due to the high-load setting, the voltage at the controlled bus of the capacitor bank 1 has significantly dropped to 111.11kV, then the voltage controller decided to close one of the capacitor bank breaker and raised up the voltage at the time instance 3. Such process occurred again at the time instance 29, the voltage controller closed the capacitor bank 2's breaker, after the voltage at the controlled bus of capacitor bank 2 dropped to 113.49kV (< 113.5kV). In addition, at time instance 6, because the tap-changing operation occurred after a certain amount of time delay, the voltages are dropped at both controlled buses.

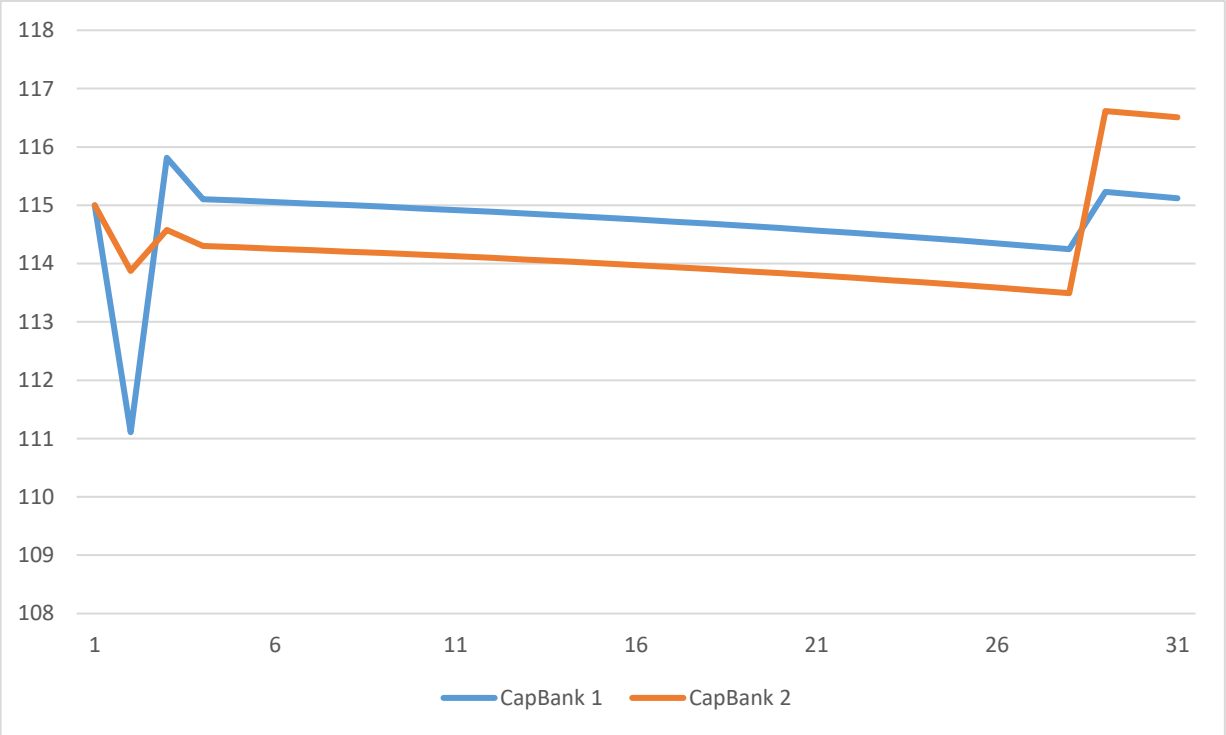


Figure 6-7 Voltage measurement when capbanks are switched on

The regional voltage controller is tested with IEEE 118 bus system. The test results are shown in section 3.4.

6.5 Summary

This chapter presents a generic software framework for both local and regional voltage controller. This framework contains 3 major components: input adapter, voltage controller, and EMS interface. The voltage controller adapter can be customized as a local voltage control or regional voltage controller. A detail explanation of regional voltage controller's customized design is given. Finally, this chapter presents the software testing methodology. A local voltage controller's test results are shown.

Chapter 7
Conclusion & Future Research

Synchrophasors provide a new paradigm in the way that transmission operators and owners monitor the electric network. The widely deployment of PMU in the recent decade starts bringing benefits to utilities and their electric customers. The utilities' strong commitment to improving the network security and electric power quality has enabled them to step out of the realm of research and development of real time system monitoring and into the questions on how to extract more benefit from their investment in synchrophasor technology: "Besides the system monitoring technique, is there any other application that can be developed based on synchrophasor?" or "What kind of information can we get through mining the EMS data and synchrophasor measurements?" With the hope of continued innovation in these directions and answer these questions, the work presented in this dissertation has proposed a data-driven framework for Energy Management System to provide voltage security assessment/control and system network partition.

7.1 Main Contributions:

Overall Contributions

The work presented in this dissertation proposes, develops and implements a data analytic based framework for regional voltage control with real time synchrophasor data. This was accomplished by developing a voltage security assessment method combined with parallel decision trees to implement a regional voltage controller. An advanced online boosting technique is used to increase model training efficiency of the proposed methodology and a data analytic based system partition approach is developed to allow for a system wide voltage controller.

A Data-driven Regional Voltage Assessment and Control

This dissertation presents a DT based voltage security assessment approach. Based on this assessment approach, a regional voltage controller using parallel DTs was developed. The developed voltage controller was evaluated using IEEE 118 bus system. The simulation results show that the proposed control scheme is able to secure the system while minimizing the number of control candidates as well as the reactive power injection.

An Online Boosting Approach for Voltage Security Assessment and Control

An adaptive decision tree methodology using online boosting was developed to provide a more efficient update for tree training. This approach was evaluated based on IEEE 118 bus system model with a topology change scenario. Simulation results show that the proposed method is able to reduce the computation burden and have lower misclassification errors compared with the default decision tree training method. Based on the online boosting method, a Hoeffding tree based algorithm is proposed to provide a more efficient update for weak learners. A simulation was conducted using the same data set generated for online boosting. The results how

that the Hoeffding tree based induction scheme provides better performance than the default decision tree training method.

A Power System Network Partition Methodology

A system network partition approach is introduced in this dissertation to decompose the system into sub-areas and thus reduce the computational burden for operating conditions generation and control candidates selection. This partition approach is achieved by the following procedures: 1. Provide a modal based algorithm to identify the VCAs. 2. Conduct a QV analysis based clustering to determine groups of buses that sharing the similar QV minimums and neighboring generators' reactive power reserve basin. 3. Find out the best set if VCAs by comparing different set of identified VCAs and QV analysis based clusters. 4. Decompose the system based on the final VCAs. The simulation results based on IEEE 118 bus system show that the decomposed system are sharing the similar voltage security pattern as the original system.

7.2 Future Research

A Bayesian Approach for Online Boosting

Though the new tree training method introduced in Chapter 4 shows promising result, the trained model has a risk of overfitting since it does not consider the prior information. An alternative approach called Bayesian Addictive Regression Tree (BART) [36] will be studied in a new framework.

Similar as the boosting method, BART is developed as an ensemble model and continually evolve in time based on new events. Different from the online boosting method, each tree stump is constrained by a regularization prior to be a weak learner, and fitting and inference are accomplished through an iterative Bayesian back-fitting MCMC algorithm [52] that generates samples from a posterior. The proposed new

framework based on BART will be developed for voltage security assessment/control. The first Alpha version will be tested and evaluated using IEEE benchmark model. Eventually, the ultimate commercialization of this framework will be implemented with field measurements.

A Real Time Power System Simulation Platform for Synchrophasor based Applications

In this dissertation, the performance of synchrophasor based applications is evaluated through simulation. The accuracy of the simulation result is highly depending on the pre-developed system model. For planning problem or problem doesn't require real time simulation, the evaluation process can be conducted offline using PSSE, PSCAD, EMTP, PSLF, etc [53] [54] [55]. For simulation requires real time testing, RTDS is an alternative solution. It provides a transient simulator with time step less than 50 us. RTDS is also able to incorporate field measurements by directly connect with PDC, this feature enables RTDS to provide a better simulation environment that imitates the true pattern of the real system in real time. However, RTDS is not designed to run with complicated user-defined analytics. To incorporate RTDS into simulation environment for synchrophasor based applications, a new software platform is required. This software platform is designed as an interface that and enables the user to develop their own algorithm. Figure 7-1 reveals the concept about the real time simulation platform. Through this platform, user's algorithm is able to analyze RTDS data and provide control signal back to RTDS in real time. Different communication protocol can be studied using this platform[56]. This new software platform not only shortens the time required for application evaluation, but also provides a standardized environment for different types of applications.

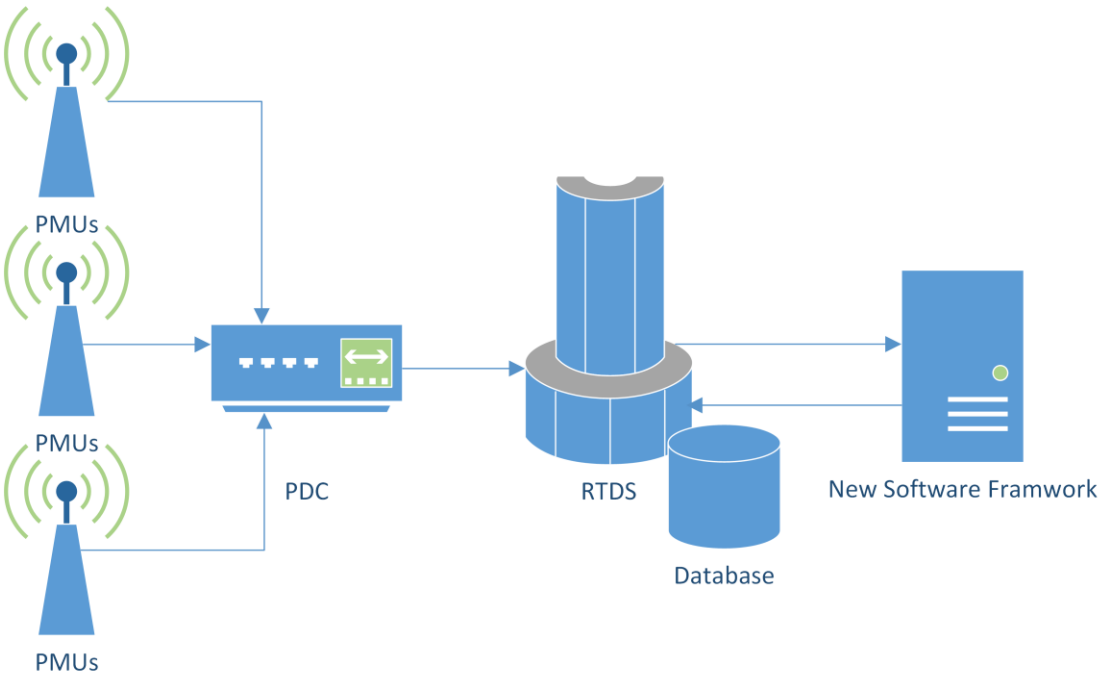


Figure 7-1 Concept of New Real Time Power System Simulation Platform

Bibliography

[1] T. V. Cussem and C. Vournas, *Voltage Stability of Electric Power Systems*. Kluwer Academic Publishers, 1998.

[2] E. E. Bernabeu, J. S. Thorp, and V. Centeno, "Methodology for a Security/Dependability Adaptive Protection Scheme Based on Data Mining," *IEEE Trans. POWER Deliv.*, vol. 27, no. 1, pp. 104–111, 2012.

[3] R. Diao Kai Sun, Vijay Vittal, Robert J.O'Keefe, Michael R.Richardson, Navin Bhatt, Dwayne Stradford, and Sanjoy K.Sarawgi., "Decision tree-based online voltage security assessment using PMU measurements," *IEEE Trans. Power Syst.*, pp. 832–839, 2009.

[4] I. Kamwa S. R.Samantaray, and Geza Joós, "On the accuracy versus transparency trade-off of data-mining models for fast-response PMU-based catastrophe predictors," *IEEE Trans. Smart Grid*, vol. 3, no. 1, pp. 152–161, 2012.

[5] S. Rovnyak Stein Kretsinger, James Thorp, and Donald Brown, "Decision trees for real-time transient stability prediction," *IEEE Trans. Power Syst.*, vol. 4, no. 1417–1426, p. 9, 1994.

[6] S. Rovnyak Chih-Wen Liu, Jin Lu, Weimin Ma, and James Thorp, "Predicting future behavior of transient events rapidly enough to evaluate remedial control options in real-time," *IEEE Trans. Power Syst.*, vol. 3, no. 1195–1203, p. 10, 1995.

[7] C. et al Liu, "A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees," *Power Syst. IEEE Trans.*, vol. 29, no. 2, pp. 717–730, 2014.

[8] P. Kessel and H. Glavitsch, “Estimating the Voltage Stability of a Power System,” *IEEE Trans. Power Deliv.*, vol. PWRD-1, no. 3, pp. 346–354, 1986.

[9] M. Hasan, S. Lefebvre, M. Saad, and D. Asber, “A decentralized control of partitioned power networks for voltage regulation and prevention against disturbance propagation,” *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1461–1469, 2013.

[10] D. Yang, Z. Nie, K. D. Jones, and V. A. Centeno, “Adaptive Decision-Trees-Based Regional Voltage Control,” in *IEEE*, Morgantown, WV, USA, 2017.

[11] Z. Nie, D. Yang, and V. Centeno, “Hoeffding-Tree-Based Learning from Data Streams and Its Application in Voltage Security Assessment,” in *IEEE*, San Antonio, Texas, USA, 2017.

[12] D. Yang, “A Power System Network Reduction Framework for Data-driven Regional Voltage Control,” in *IEEE*, Morgantown, WV, USA, 2017.

[13] A. G. Phadke and J. S. Thorp, “HISTORY AND APPLICATIONS OF PHASOR MEASUREMENTS,” in *2006 IEEE PES Power Systems Conference and Exposition*, 2006, pp. 331–335.

[14] G. Andersson *et al.*, “Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance,” *IEEE Trans. Power Syst.*, vol. 20, no. 4, pp. 1922–1928, 2005.

[15] D. Yang, H. Latchman, D. Tingling, and A. A. Amarsingh, “Design and Return on Investment Analysis of Residential Solar Photovoltaic Systems,” *IEEE Potentials*, vol. 34, no. 4, pp. 11–17, Jul. 2015.

[16] D. Yang, R. Sun, J. D. L. Ree, and M. Mcvey, “Capacitor bank model validation with particle swarm optimization algorithm,” in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[17] K. D. Jones, J. S. Thorp, and R. M. Gardner, “Three-phase linear state estimation using Phasor Measurements,” in *2013 IEEE Power Energy Society General Meeting*, 2013, pp. 1–5.

[18] A. Stankovic, M. Ilic, and D. Maratukulam, “Recent results in secondary voltage control of power systems,” *IEEE Trans. Power Syst.*, vol. 6, no. 1, pp. 94–101, Feb. 1991.

[19] T. E. Dy-Liacco, “Enhancing power system security control,” *IEEE Comput. Appl. Power*, vol. 10, no. 3, pp. 38–41, Jul. 1997.

[20] L. Wehenkel, T. V. Cutsem, and M. Ribbens-Pavella, “An Artificial Intelligence Framework for On-Line Transient Stability Assessment of Power Systems,” *IEEE Power Eng. Rev.*, vol. 9, no. 5, pp. 77–78, May 1989.

[21] K. Sun, S. Likhate, V. Vittal, V. S. Kolluri, and S. Mandal, “An Online Dynamic Security Assessment Scheme Using Phasor Measurements and Decision Trees,” *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1935–1943, Nov. 2007.

[22] R. Sun and V. A. Centeno, “Wide Area System Islanding Contingency Detecting and Warning Scheme,” *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2581–2589, Nov. 2014.

[23] C. A. Jensen, M. A. El-Sharkawi, and R. J. Marks, “Power system security assessment using neural networks: feature selection using Fisher discrimination,” *IEEE Trans. Power Syst.*, vol. 16, no. 4, pp. 757–763, Nov. 2001.

[24] L. S. Moulin, A. P. A. da Silva, M. A. El-Sharkawi, and R. J. Marks, “Support vector machines for transient stability analysis of large-scale power systems,” *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 818–825, May 2004.

[25] M. He, Z. Junshan, and V. Vijay, “Robust online dynamic security assessment using adaptive ensemble decision-tree learning,” *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4089–4091, 2013.

[26] M. V. Suganyadevia and C. K. Babulalb, “Estimating of loadability margin of a power system by comparing Voltage Stability Indices,” in *Communication and Energy Conservation 2009 International Conference on Control, Automation*, 2009, pp. 1–4.

[27] S. Shukla and L. Mili, “A hierarchical decentralized coordinated voltage instability detection scheme for SVC,” presented at the North American Power Symposium, 2015.

[28] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*, 2nd ed. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2014.

[29] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Taylor & Francis, 1984.

[30] H. Grabner and Horst Bischof, “On-line boosting and vision,” presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[31] Y. Freund and S. Robert E, “Experiments with a new boosting algorithm,” *Icml*, vol. 96, pp. 148–156, 1996.

[32] Y. Freund and R. E. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” in *Computational Learning Theory*, 1995, pp. 23–37.

[33] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors),” *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, Apr. 2000.

[34] P. Domingos and G. Hulten, “Mining High-speed Data Streams,” in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2000, pp. 71–80.

[35] “A Very Fast Decision Tree Algorithm for Real-Time Data Mining of Imperfect Data Streams in a Distributed Wireless Sensor Network - Semantic Scholar.” [Online]. Available: /paper/A-Very-Fast-Decision-Tree-Algorithm-for-Real-Time-Hang-Fong/ec71bc20442cb839f1a01417fe11446a8a4f61a8. [Accessed: 27-Apr-2017].

[36] A. Bifet and R. Kirkby, *DATA STREAM MINING A Practical Approach*. .

[37] P. Lagonotte, J. C. Sabonnadiere, J.-Y. Leost, and J.-P. Paul, “Structural analysis of the electrical system: Application to secondary voltage control in France,” *IEEE Trans. Power Syst.*, vol. 4, no. 2, pp. 479–486, 1989.

[38] V. Alimisis and P. C. Taylor, “Zoning Evaluation for Improved Coordinated Automatic Voltage Control,” *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2736–2746, Sep. 2015.

[39] H. Sun, Q. Guo, B. Zhang, W. Wu, and and B. Wang., “An adaptive zone-division-based automatic voltage control system with applications in China,” *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1816–1828, 2013.

[40] R. A. Schlueter, I.-P. Hu, M.-W. Chang, J. C. Lo, and A. Costi, “Methods for determining proximity to voltage collapse,” *IEEE Trans. Power Syst.*, vol. 6, no. 1, pp. 285–292, 1991.

[41] R. A. Schlueter, “A voltage stability security assessment method,” *IEEE Trans. Power Syst.*, vol. 13, no. 4, pp. 1423–1438, Nov. 1998.

[42] M. K. Verma and S. C. Srivastava, “Approach to determine voltage control areas considering impact of contingencies,” *Transm. Distrib. IEE Proc. - Gener.*, vol. 152, no. 3, pp. 342–350, May 2005.

[43] Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” in *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1997, pp. 21–34.

[44] A. Ahmad and L. Dey, “A K-mean Clustering Algorithm for Mixed Numeric and Categorical Data,” *Data Knowl Eng*, vol. 63, no. 2, pp. 503–527, Nov. 2007.

[45] R. E. Kass and L. Wasserman, “A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion,” *J. Am. Stat. Assoc.*, vol. 90, no. 431, pp. 928–934, Sep. 1995.

[46] D. Pelleg and A. W. Moore, “X-means: Extending K-means with Efficient Estimation of the Number of Clusters,” in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, 2000, pp. 727–734.

[47] B. Hancock, *Notes on Bayesian Information Criterion Calculation for X-Means Clustering*. 2017.

[48] “Dynamic REI equivalents for short circuit and transient stability analyses (PDF Download Available),” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/222697832_Dynamic_REI_equivalents_for_short_circuit_and_transient_stability_analyses. [Accessed: 26-Apr-2017].

[49] F. Li and R. P. Broadwater, “Software framework concepts for power distribution system analysis,” *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 948–956, May 2004.

[50] *openECA: Open Source Extensible Control & Analytics*. Grid Protection Alliance (GPA), 2017.

[51] H. A. Chipman, E. I. George, and R. E. McCulloch, “BART: Bayesian additive regression trees,” *Ann. Appl. Stat.*, vol. 4, no. 1, pp. 266–298, Mar. 2010.

[52] T. Hastie and R. Tibshirani, “Bayesian backfitting (with comments and a rejoinder by the authors,” *Stat. Sci.*, vol. 15, no. 3, pp. 196–223, Aug. 2000.

[53] H. Liu *et al.*, “ARMAX-Based Transfer Function Model Identification Using Wide-Area Measurement for Adaptive and Coordinated Damping Control,” *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1105–1115, May 2017.

[54] X. Zheng, N. Tai, Z. Wu, and J. Thorp, “Harmonic current protection scheme for voltage source converter-based high-voltage direct current transmission system,” *Transm. Distrib. IET Gener.*, vol. 8, no. 9, pp. 1509–1515, Sep. 2014.

[55] C. Mishra, K. D. Jones, A. Pal, and V. A. Centeno, “Binary particle swarm optimisation-based optimal substation coverage algorithm for phasor measurement unit installations in practical systems,” *Transm. Distrib. IET Gener.*, vol. 10, no. 2, pp. 555–562, 2016.

[56] A. A. Amarsingh, H. A. Latchman, and D. Yang, “Narrowband Power Line Communications: Enabling the Smart Grid,” *IEEE Potentials*, vol. 33, no. 1, pp. 16–21, Jan. 2014.