Classification Trees for Complex Synchrophasor Data

Anamitra Pal1,*, J. S. Thorp1, Taufiquar Khan2, and S. Stanley Young3

1Virginia Tech, Blacksburg, USA.
2Clemson University, Clemson, South Carolina, USA.
3National Institute of Statistical Sciences (NISS), Research Triangle Park, North Carolina, USA.

* Corresponding Author. Address: Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA-24061 USA. E-mail address: anam86@vt.edu (Anamitra Pal).

Abstract—Classification & Regression Trees (CART) have been used for various applications in power systems. In most of these applications, phasor data obtained from PMUs are being used for building the decision tree. However, the splits in CART are based on a single attribute or a combination of variables chosen by CART itself rather than the user. But as PMU data are complex numbers, both the attributes, real and imaginary, should be considered simultaneously for making critical decisions. For example, changing the reference bus in situations where the split is only on real or imaginary part of a complex voltage (or current) measurement can cause the performance of the tree to degrade significantly. An algorithm is proposed in this paper to allow splits on complex synchrophasor data. The methodology is implemented on two systems – a detailed model of the California Power System where it is used for developing an adaptive protection scheme, and the IEEE-118 Bus system where it is used to classify dynamic events based on trajectories of voltage measurements obtained from PMUs. The MATLAB® implementation of CART (classregtree.m) has been used for performing both the analysis.

Keywords—Classification & Regression Trees (CART), Decision Trees, Fisher’s Linear Discriminant (FLD), Linear Discriminant Analysis (LDA), Synchrophasors, Wide Area Measurement System (WAMS).

1. INTRODUCTION

Wide Area Measurement System (WAMS) using synchronized Phasor Measurement Units (PMUs) have been extensively used for monitoring power system networks all over the world. PMU-based
measurements provide new methods for achieving real-time control, stability enhancement, and transfer capacity improvement of the power network. In the US and China especially, these measurements have been widely used for state estimation, protection, and control based on situational awareness for operational decision making. PMUs, when placed at a bus provide time synchronized measurements of the voltage phasor and branch current phasors of all the branches emerging from that bus. Various data mining techniques have been employed to make use of the collected information [1-5], with Decision Trees being the most popular approach [6, 7]. Decision trees extract information from large sets of data and intuitively represent the gained knowledge through a series of if-else statements. Classification & Regression Trees (CART), a non-parametric decision tree learning technique, is especially suited to power systems due to the complex non-linear behavior of the system [8, 9]. A commercial implementation of this technique (CART®) has also been developed by Salford Systems [10].

CART data is in the form of an array with rows being the events/outcomes and columns being the measurements. In its simplest form, CART picks one measurement at a time for performing the splits. While this is very effective in handling data having univariate attributes, difficulties are observed where the predictors are multivariate. Multi-modal classification problems involve pattern recognition from disjoint regions in feature space. Synchronophasor data are an example of a multi-modal classification problem in the domain of power systems. PMU measurements are generally complex. When CART picks one column, it uses either the real or the imaginary part of the measurement, but not both. Thus, it is not able to address the complete phasor in a single split. This also creates problems when there is a change of reference [11]. Moreover, although CART allows splitting on Linear Combinations (LCs) involving as many as 6 attributes, these are chosen as “p chooses d” [12]. Hence, for performing the split, CART is not particularly likely to select a linear combination that includes both the real and the imaginary part of a single complex number.

Techniques addressing multi-modal classifications include decision trees [13-14], Artificial Neural Networks (ANNs) [15, 16], Support Vector Machines (SVMs) [17], etc. with decision trees being the most commonly used method. The algorithm described in this paper addresses multi-modal classifications
by adding a pre-processing step in the input to the decision tree so as to represent high dimensional data by a single variable. Two examples are provided to illustrate the proposed concept. Each complex measurement is treated as a single attribute in the first example whereas a trajectory of complex measurements is treated as a single attribute in the second example. The resulting single attribute is then used for performing the splits. Since it is only the inputs to CART that are modified and not the logic based on which CART does the splitting, the technique developed here is expected to be used in solving a wide variety of engineering problems.

The remaining part of the paper is structured as follows. Section 2 gives a brief summary of the uses of CART in power systems thereby highlighting the potential applications of the technique developed here. The proposed algorithm is explained in Section 3. Section 4 illustrates how the algorithm performs on two test scenarios. Section 5 describes the simulation results when it is implemented on two systems – a detailed model of the California Power System where it is used for developing an adaptive protection scheme, and the IEEE-118 Bus system where it is used to classify dynamic events based on trajectories of voltage measurements obtained from PMUs. Section 6 concludes the paper.

2. APPLICATIONS OF CART IN POWER SYSTEMS

A CART tree is a binary decision tree that is constructed by splitting the parent node and subsequent nodes into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The logic is based on choosing the best split among all possible splits at each parent node so that the child nodes are purest. The CART algorithm initially grows a decision tree as large as possible and then selectively prunes it upwards. Cost complexity criterion is used in the pruning process. The objective is to attain a minimum sized tree with minimized cost complexity. Cost complexity criterion and number of branches vary depending on the application. More details about the computational aspects of the CART methodology – splitting criteria, structural complexity, etc. can be found in [18].

Decision Trees have been used extensively in power systems for performing different types of analysis. A security-dependability adaptive protection scheme separately using voltage angles and current...
magnitudes is developed in [7]. In [19], a real-time transient stability prediction scheme using voltage angles and decision trees is investigated. Voltage angles are again used in decision tree processing for response-based discrete event control [20]. A fast online voltage security monitoring scheme using PMU measurements with decision trees built using voltage angles is developed in [21]. Phasor magnitude and angle have (separately) been used for real-time transient instability detection [22]. In [23], power system security assessment is done using decision trees built from voltage angles obtained from PMUs. A wide-area response-based control using phasor measurements and decision trees based on voltage angles is developed in [24]. Splitting of decision trees on a single attribute is done in [25-27]. Voltage angles have again been used for power system transient stability forecasting in [28].

CART is best suited for such applications because of its apparent robustness and efficiency of use. It requires little data for its preparation, works well even if some of the assumptions made during data generation are violated and performs very well with large data in a comparatively short time [29]. Moreover, the resulting model is easy to understand and implement. However, the drawback of CART is that only univariate splits are allowed in it, i.e. each split in CART depends on the value of a single predictor variable. Hence, it is unable to perform an optimum split with a single node when applied on data having many attributes. In most of the applications mentioned above, decisions were being made based on one measurement obtained from a PMU. As such, CART was making decisions based on a single attribute. However, a PMU placed on a bus records the complete voltage phasor (magnitude and angle) and the complete branch current phasor (real and imaginary components) of all the branches emerging from that bus. Although using one attribute (for instance, voltage angle) works at times, it is not always a good strategy. As will be illustrated in this paper, on many occasions, it is more appropriate to use all the data that is made available through the placement of the PMU for making decisions, rather than just one attribute. Moreover, since the complex number is a single entity, it should be treated as such. Splitting only on the real or only on the imaginary component of a PMU measurement does not address the complete phasor. Similarly, a linear combination involving the real part of some variables and the imaginary part of others is not physically meaningful and is inefficient for placing PMUs. For example, if
complex currents are considered separately then CART might choose to do a split based on the real current flowing through one line and the imaginary current flowing through a different line, in which case at least two PMUs would be needed for measuring the currents. However, if the complex currents are treated as a single entity, then CART will decide based on the real and imaginary components of the current simultaneously and so even a single PMU might suffice.

Reference [11] tries to address this problem and perform a split by providing a reference angle to the measurements. For instance, if the load flow is used to generate the data, then the swing bus is used as the reference. However, the actual application requires a physical reference (as the measured angles will have the reference angle subtracted from them), which becomes a problem when the reference is changed. Typically this happens when the PMU on the reference bus fails or when the utility decides to install the needed PMU elsewhere. It was observed in [11] that performing the split on the real or imaginary part of a complex measurement caused the performance of the tree to degrade significantly. However, there was no guarantee that an optimum split would be obtained by changing the reference for all kinds of situations/distributions. The technique illustrated in this paper shows how the original CART algorithm can be used to make decisions while considering measurements having multiple attributes in a single split without needing a reference. Therefore, the method proposed in this paper is expected to find use in the areas of –

- Adaptive Protection Schemes
- Event classification
- Real-time power system transient stability/instability predictions
- Online voltage security monitoring
- Online dynamic security assessment
- Response based control using phasor measurements
- Optimal PMU placement, etc.
3. PROPOSED ALGORITHM

A. Strategy for handling high dimensional data having two classes

Consider a set of observations \( \bar{p} \) for each sample of an event with known classes \( q_b \) and \( q_r \); the subscripts \( b \) and \( r \) denoting the “blue” and “red” data sets which are to be separated. The objective of the classification problem is to then find a good predictor for the class \( q \) of any sample given only an observation \( \bar{p} \). By performing simulations and analyzing the results, it was realized that for similar distributions (skewed or otherwise), a simple and efficient way of doing the split was by making the two distributions spherical (based on LDA which is a simplification of FLD). This is done by multiplying the individual distributions with the Cholesky-Decomposition of the inverse of its covariance matrix. This results in the new distributions becoming spherical with an identity covariance matrix. The advantage is that the perpendicular bisector of the line joining the two centroids of the new (spherical) distributions then becomes the optimum split between the two data sets.

Let the two classes of observations have centroids \( \bar{\mu}_{q_b} \) and \( \bar{\mu}_{q_r} \) and covariance \( C_q \). Then the optimum split for this new distribution is obtained by solving for \( \bar{w}_{\text{new}} \) such that,

\[
S(\bar{\mu}_{q_b} - \bar{\mu}_{q_r})^T \left( \bar{w}_{\text{new}} - \frac{\bar{\mu}_{q_b\text{new}} + \bar{\mu}_{q_r\text{new}}}{2} \right) = 0
\]

where,

\[
\begin{align*}
\bar{\mu}_{q_b\text{new}} &= S\bar{\mu}_{q_b} \\
\bar{\mu}_{q_r\text{new}} &= S\bar{\mu}_{q_r} \\
S^T S &= C_q^{-1}
\end{align*}
\]

It is to be noted here that \( \bar{w}_{\text{new}} \) defines the hyper-plane which perpendicularly bisects the line joining the two new centroids. Therefore, (1) holds true in the new co-ordinate system where the two distributions are spherical with identity covariance matrices. The data points to be segregated are now projected onto \( \bar{w}_{\text{new}} \). For example, let \( \bar{w}_{\text{new}} \) be parameterized by a linear equation in 3-d space as shown in (3),

This is an Author's Accepted Manuscript of an article published in Electric Power Components and Systems, 41:14, 1381-1396, Date of Online Publication: 20 Sep 2013. Copyright: Taylor & Francis, available online at: http://www.tandfonline.com/doi/10.1080/15325008.2013.824048
\[ \bar{w}_{\text{new}}: ax + by + cz + d = 0 \]  

(3)

Then, if \( \bar{x}_i = (x_i, y_i, z_i) \) be the \( i^{th} \) data point belonging to the set of observations \( \bar{p} \), then the shortest distance from \( \bar{x}_i \) to the hyperplane defined by \( ax + by + cz + d = 0 \) is given by,

\[ D(i) = \frac{ax_i + by_i + cz_i + d}{\sqrt{a^2 + b^2 + c^2}} \]  

(4)

The split can now be performed on the following basis –

\[
D(i) \leq 0: \text{Red} \\
D(i) > 0: \text{Blue}
\]

(5)

From (3)-(5) it becomes clear that, since the distance vector will be a one-dimensional quantity irrespective of the number of dimensions the hyperplane has, this logic can be extended to address high dimensional data with ease.

When the two distributions are very different, an optimum split can be performed by adding the covariances of the two data sets and using it on both the data (based on traditional FLD). Let the two classes of observations have centroids \( \bar{\mu}_{q_b} \) and \( \bar{\mu}_{q_r} \) and experimental covariances \( C_{q_b} \) and \( C_{q_r} \). Then, FLD defines a performance index \( \rho \) which maximizes the projected class differences relative to the sum of the projected within-class variability [15, 30]. Mathematically, this is stated as –

Maximize \( \rho: \rho = \frac{(m^T(\mu_{q_b} - \mu_{q_r}))^2}{m^T(C_{q_b} + C_{q_r})m} \)  

(6)

where the vector \( m \) is the normal to the discriminant hyper-plane. On solving (6), it is realized that the maximum separation occurs when,

\[ m = (C_{q_b} + C_{q_r})^{-1}(\bar{\mu}_{q_b} - \bar{\mu}_{q_r}) \]  

(7)

Now, any vector perpendicular to \( m \) is given by,

\[ \frac{1}{2}(\bar{\mu}_{q_b} - \bar{\mu}_{q_r}) + \alpha \bar{w} \]  

(8)

where \( \alpha \) is the optimizing variable. In order to find the optimum hyper-plane, we have to minimize:

\[ \left\| \bar{p} - \frac{1}{2}(\bar{\mu}_{q_b} - \bar{\mu}_{q_r}) - \alpha \bar{w} \right\|^2 \]  

(9)

This is an Author's Accepted Manuscript of an article published in Electric Power Components and Systems, 41:14, 1381-1396, Date of Online Publication: 20 Sep 2013, Copyright: Taylor & Francis, available online at: http://www.tandfonline.com/doi/10.1080/15325008.2013.824048
On solving (9), using Weighted Least Squares (WLS) algorithm, we get the expected value of $\alpha$ as,

$$\hat{\alpha} = (\bar{w}^T\bar{w})^{-1}(\bar{p} - \frac{1}{2}(\bar{\mu}_q_b - \bar{\mu}_q_r))$$

(10)

Using this value of $\hat{\alpha}$, we obtain the splitting variable $\bar{D}$ as:

$$\bar{D} = [I - \bar{w}(\bar{w}^T\bar{w})^{-1}(\bar{w}^T)](\bar{p} - \frac{1}{2}(\bar{\mu}_q_b - \bar{\mu}_q_r))$$

(11)

From (11), it can be inferred that the original multivariate data can be replaced in CART by the single variable $\bar{D}$. It is also easy to show that the hyperplane perpendicularly bisecting the line joining the two centroids in the new co-ordinate system is equivalent to a rotation of the hyperplane perpendicularly bisecting the line joining the two centroids in the original co-ordinate system, thereby making it no longer perpendicular. This is the geometric interpretation of Fisher’s algorithm. Therefore, for any unknown distribution a direct application of FLD is sufficient for performing an optimum split. As such, a suitable name for this method is Fisher’s Linear Discriminant applied to Synchrophasor Data (FLDSD).

B. Strategy for handling high dimensional data having multiple classes

In the previous sub-section, it was proved that the proposed approach is able to perform binary splits on high-dimensional data. However, this method could also be used to separate multiple classes, by taking two classes at a time. In order to do so, distances of the data points from all the hyperplanes must be initially computed. Taking two classes at a time, (11) is used to compute for the distances. For $n$-class distribution, we have,

$$\text{Number of Hyperplanes} = \frac{n \times (n - 1)}{2}$$

(12)

The distances to the hyperplanes must then be fed into CART for selecting the optimum distance variable for performing the split. Since all the input variables have single attributes, CART can directly select the distance variable that will result in the best possible split. By selecting two classes at a time, and proceeding until all the class combinations have been covered, even a tree of depth $n - 1$ can successfully separate $n$ data classes. The flowchart of the proposed algorithm is shown in Figure 1.
Find means and compute covariances of all the distributions

Taking two distributions at a time, multiply them by inverse of the sum of the covariances

Compute the equation of the line joining the centroids of the resulting distributions

Solve for the optimal hyperplane perpendicular to the line joining the centroids

Replace data of original distributions by distances of individual points from this optimally selected hyperplane

All distribution pairs covered

Yes

Set the resulting distance variables as inputs to CART for training the tree

Given an actual event, compute distances to the relevant hyperplanes and follow the tree to the corresponding terminal node

No

Figure 1. Flowchart of Proposed Algorithm

4. ILLUSTRATION

Since PMU data are usually complex numbers, the goal is to express such multivariate data by a single entity. This single entity/variable would be used for performing the split in CART. But real-time PMU measurements offer some challenges of their own. It was observed that data sets often had very skewed distributions (as seen in Figure 2). The reason for this was that voltage magnitudes and steady state
currents expressed in “per unit” were near unity, but the angles expressed in degrees and transient current values were not so. Thus, when such diverse measurements were combined together to make decisions, the resulting cloud became ellipsoidal in shape. This issue was addressed by transforming the ellipsoid into a sphere using the proposed algorithm as shown in Figure 3. It becomes obvious from Figure 3 that this technique has resulted in the two distributions becoming circular (spherical, for higher dimensions) from their earlier elliptical (ellipsoidal, for higher dimensions) shapes. This was further verified by computing the covariances of the new distributions, both of which were found to be identity.

![Figure 2](image1.png)

Figure 2. Dotted line shows best primary (single column) split for data sets having similar distributions

![Figure 3](image2.png)

Figure 3. Dotted Line shows the optimum split on data sets having similar distributions using proposed algorithm

This is an Author's Accepted Manuscript of an article published in Electric Power Components and Systems, 41:14, 1381-1396, Date of Online Publication: 20 Sep 2013, Copyright: Taylor & Francis, available online at: http://www.tandfonline.com/doi/10.1080/15325008.2013.824048
Another problem was the apparent randomness in some of the distributions when, for example, a fault occurred in the system. Similar to what is seen in Figure 4; it was observed that such distributions followed no set pattern. The application of the proposed approach to the example in Figure 4 is shown in Figure 5. It becomes obvious from Figure 5 that using this method a perfect split between the data points has been obtained. The next section illustrates the application of the proposed logic in making decisions based on synchrophasor data consisting of complex numbers as well as trajectories of complex numbers, by representing them as a single variable.

![Figure 4](image1.png)

**Figure 4.** Dotted line shows best primary (single column) split for data sets having random distributions

![Figure 5](image2.png)

**Figure 5.** Dotted Line shows the optimum split on data sets having random distributions using proposed algorithm

This is an Author's Accepted Manuscript of an article published in Electric Power Components and Systems, 41:14, 1381-1396, Date of Online Publication: 20 Sep 2013. Copyright: Taylor & Francis, available online at: http://www.tandfonline.com/doi/10.1080/15325008.2013.824048
5. **Simulation and Results**

A. *An adaptive protection scheme for the California Power System*

In [7], an algorithm was developed which used CART to categorize a system as “safe” or “stressed” based on data collected from PMUs. The main goal of that algorithm was to partition the power system state space intelligently in order to develop decision rules to adjust security/dependability balance of relay protection schemes. Using [7] as a case study, classifications based on measurements having two attributes is discussed here. The algorithm proposed in this paper is applied to the detailed model of the California Power System for two scenarios, Heavy Winter and Heavy Summer. The data sample used for the simulations is obtained from [7]. The total number of training cases in the Heavy Winter case was 4150, whereas in the Heavy Summer case it was 11367. The total number of “out-of-sample” cases was 660 for the Heavy Winter case and 1155 for the Heavy Summer case.

The data presented to the CART implementation program in MATLAB® (classregtree.m) [31] were all the 500kV voltages and currents present in the system – a total of 42 voltage angle and 90 complex current measurements. Every 500kV current was measured as a complex number and each complex number was treated as a single entity. Ten-fold cross-validation was done to improve the accuracy of the prediction. In selecting the splitting nodes CART picked the substations where PMUs would be placed. Thus, it was the CART algorithm which chose the nodes that would result in an optimum split as seen in Figures 6 and 7. Figure 6 depicts the decision tree obtained for the Heavy Winter case. An overall accuracy of 99.46% was obtained for this case with $x_{65}$, $x_{83}$, and $x_{18}$ denoting currents and/or voltage angles of specific lines and buses. Figure 7 depicts the decision tree for the Heavy Summer case. It had an overall accuracy of 99.38%. The results indicate that PMUs placed on lines and buses selected by the tree can classify the system as “safe” (blue) or “stressed” (red) with very high accuracy.

Table I compares the results obtained here with those obtained in [7]. From the table, it is realized that by using the proposed approach, a smaller tree (indicating requirement of lesser number of PMUs) is able to provide higher classification accuracy. Table II compares the performance of the decision trees.
obtained using this technique with that obtained by rotating along a reference as was done in [11]. From Tables I and II, it becomes clear that the proposed methodology is more accurate in making decisions involving complex synchrophasor data. There are applications where only the voltage angle, for example, is relevant for making decisions. However, relaying in particular deals with complex quantities and the technique developed here is necessary for making the optimal split. Simulation cases developed in [7] were used here because it provided a system big enough to test the robustness of this technique. In the next example, application of this technique to high-dimensional, multi-class data is illustrated.

![Decision Tree](image)

**Figure 6. Decision Tree for Heavy Winter case**

![Decision Tree](image)

**Figure 7. Decision Tree for Heavy Summer case**
TABLE I

COMPARING SIZE OF DECISION TREES (FOR TRAINING PURPOSES)

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Using algorithm developed in [7]</th>
<th>Using Proposed Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Nodes</td>
<td>Misclassification Rate (%)</td>
</tr>
<tr>
<td>Heavy Winter</td>
<td>6</td>
<td>1.00</td>
</tr>
<tr>
<td>Heavy Summer</td>
<td>6</td>
<td>1.00</td>
</tr>
</tbody>
</table>

TABLE II

COMPARING OVERALL (TRAINING AND TESTING) PERFORMANCE OF DECISION TREES

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Errors</td>
<td>Accuracy (%)</td>
<td>Number of Errors</td>
</tr>
<tr>
<td>Heavy Winter</td>
<td>4810</td>
<td>48</td>
<td>99.00</td>
</tr>
<tr>
<td>Heavy Summer</td>
<td>12522</td>
<td>172</td>
<td>98.63</td>
</tr>
</tbody>
</table>

B. Classification of dynamic events based on voltage measurements obtained from PMUs

A three-phase PMU-only state estimator has been created for Dominion Virginia Power’s 500kV network as part of a DOE Demonstration Project. The estimator will update every 1/30th of a second with time-tagged measurements of high voltage buses as its outputs. Being a linear estimator, no iterations will be involved. It will be a truly dynamic estimator fast enough to observe events that have never been captured before. The IEEE-118 Bus System which has eleven 345 kV buses is used as the test system. A typical output of this estimator for a dynamic event [32] is shown in Figure 8. The figure shows the trajectories of the eleven complex voltages (positive sequence), for one second, for a three-phase fault on line 26-30 followed by an unsuccessful high speed reclose. The display has been split into four windows.
for clarity. From Figure 8 it can be realized that unless a label is provided to the plots that the estimator generates, it will be very difficult to identify, just by looking at the plots, the event that has occurred. The algorithm proposed in this paper provides a solution to this problem by treating the trajectory of complex numbers as a single attribute which is then used for classifying different dynamic events.

Figure 8. Eleven 345 kV bus voltage trajectories of IEEE-118 bus system at 30 times a second plotted in complex plane for a Three-phase fault on Line 26-30 with Unsuccessful High Speed Reclose

The following procedure was adopted for generating the data to train the trees. The dynamic events used for classification were Single Line-to-Ground (SLG) fault, Three Phase-to-Ground (TPG) fault and Zone II operations. The possible classes under line-to-ground faults and three phase-to-ground faults were No-Reclose (NR), Successful High Speed Reclose (SHSR) and Unsuccessful High Speed Reclose (USHSR). Zone II operation with an over-reach of up to 150% was assumed to occur because of a stuck breaker following a Single Line-to-Ground fault or a Three Phase-to-Ground fault. As such, the possible classes under Zone II operation were Single Line-to-Ground (SLGZ2) and Three Phase-to-Ground (TPGZ2). Since there were ten 345kV lines present, ten parallel decision trees were created (one tree for each line), with the time-tagged breaker statuses used to identify the relevant tree in the case of an event.
270 cases were created for all the ten lines, for line-to-ground faults and three-phase-to-ground faults, respectively. The number of cases for Zone II operation varied from line to line (depending on how the line-under-test was connected to the rest of the system). A total of 6674 cases were created for the IEEE 118-Bus system. Ten-fold cross-validation was done to improve the accuracy of the prediction.

Taking 345kV line between buses 38 and 65 as an example, the complex voltages (real and imaginary) of the two buses were obtained for different classes. Since both buses 38 and 65 were connected to other 345kV buses, Zone II operation could be detected on either end. A total of 722 cases were identified for this line which fell under 10 classes – three classes for line-to-ground fault, three classes for three phase-to-ground fault, two classes for Zone II operation on bus 38 end, and two classes for Zone II operation on bus 65 end. A second’s worth of data starting from the time of the fault was used for classification purposes. This resulted in a 10-class classification of 300-dimensional data for each of the 722 cases. The dimension of the data is based upon 30 complex (real and imaginary) voltages for five 345kV buses – 38, 65, 30 (which is connected to 38), 64 and 68 (which are connected to 65). The proposed algorithm was applied to this data to calculate the distances to the \( 45 \left( = \frac{10 \times (10 - 1)}{2} \right) \) possible hyperplanes. These distances were fed into the CART implementation program in MATLAB\textsuperscript{®} (classregtree.m) [31] for selecting the optimal distance variables for performing the splits. The resulting decision tree is shown in Figure 9, where “\( d_i - j \)” denotes the distance of the individual points from the hyperplane separating classes \( i \) and \( j \). From the figure, it is observed that for line 38-65, CART chose the distances \( d1 - 2, d1 - 3, d1 - 9, d2 - 8 \) and \( d9 - 10 \) as the optimal variables for performing the splits. Given an actual event, the distances from each data point to the five hyper-planes need to be computed and the tree followed to the respective terminal node to identify the event that has been captured. Similar decision trees were obtained for the other nine lines. It was found that the decision trees developed in this way was able to successfully distinguish between the different dynamic events with 100% accuracy for all the lines. The results indicate that the proposed methodology of using the distance variable for separating high-dimensional, multi-class data is a logical approach that yields good solutions.
6. **Conclusion**

PMU data have been used for making critical decisions for quite some time now, but the true potential of these devices has not yet been realized. When placed at a bus, they provide real-time measurements of voltages and currents connecting that bus with the rest of the system. But, being complex quantities, these measurements have not been addressed completely until now. This paper proposes an algorithm for making decisions based on complex synchrophasor data using CART. When only the real or the imaginary attribute of the phasor is used for making the split, the result is not optimal as the complete phasor is not considered. Others who have tried to solve this problem have done so by rotating the axes along a reference, but that approach does not provide a guarantee for a successful result in all situations.

In this paper, the complex number/trajectory of complex numbers is represented using a single variable which is then used to perform the splits. Although the idea of representing complex variables by a single attribute is not new, the novelty of this paper lies in the extension of this idea to higher dimensions involving multiple classes. Moreover, since it is only the inputs to CART which have been modified, and not the algorithm on which CART operates, this technique can be readily applied to any engineering
problem which involves decision making based on multivariate data. The areas in power systems where this technique can be readily used are also summarized. The technique is found to be simple, robust, and easily applicable to large, real-life systems.

7. **Acknowledgements**

This work was partly supported by the California Institute for Energy and the Environment (CIEE) Contract TRP-08-06, funded by the California Energy Commission’s Public Interest Energy Research Program and partly by the Lockheed Martin Corporation (LMC). This paper does not necessarily represent the views of the Energy Commission, its employees or the State of California. It has not been approved or disapproved by the Energy Commission nor has the Energy Commission passed upon the accuracy or adequacy of the information.

Many faculties and students of the Power Lab at Virginia Tech have contributed greatly towards this research. The authors would like to especially thank Dr. Arun G. Phadke and Dr. Virgilio Centeno for their valuable inputs.

8. **References**


