Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm

Le Xu, *Student Member, IEEE*, Mo-Yuen Chow, *Senior Member, IEEE*, and Leroy S. Taylor, *Member, IEEE*

Abstract—**Power distribution systems have been significantly affected by many outage-causing events. Good fault cause identification can help expedite the restoration procedure and improve the system reliability. However, the data imbalance issue in many real-world data sets often degrades the fault cause identification performance. In this paper, the E-algorithm, which is extended from the fuzzy classification algorithm by Ishibuchi *et al.* to alleviate the effect of imbalanced data constitution, is applied to Duke Energy outage data for distribution fault cause identification. Three major outage causes (tree, animal, and lightning) are used as prototypes. The performance of E-algorithm on real-world imbalanced data is compared with artificial neural network. The results show that the E-algorithm can greatly improve the performance when the data are imbalanced.**

Index Terms—**Data imbalance, data mining, fault cause identification, fuzzy classification, g-mean, neural network, power distribution systems.**

I. Introduction

POWER distribution systems are one vital lifeline of the modern society for maintaining adequate and reliable flows of energy. The geographically dispersed power distribution systems are under various dynamic operating environments; they have been significantly affected by various outage-causing events such as equipment failures, animal contacts, trees, lightning strikes, etc. It is important to diagnose the faults and restore the systems in a timely manner in order to maintain their vitality.

In order to enhance the reliability as well as the availability of power distribution systems, their management systems need to have proper and speedy responses to outages. However, many utilities, for safety reasons, do not restore the faulty sections until they have found the outage causes. The whole restoration process may take from tens of minutes to hours. Linemen may often need to walk along the power distribution lines, which can be miles, in an attempt to find the outage evidences (e.g., burn marks on the pole for possible lightning faults, dead animal bodies for possible animal faults) and to ensure safety (e.g., no down distribution lines) prior to re-energizing the system. Sometimes, linemen need to call the dispatch center for appropriate crews to execute specific advanced tasks. For instance, tree crews may be requested to remove fallen trees in order to restore the system.

Many different methods have been investigated to locate the fault [1], [2]; on the other hand, effective fault cause identification can also provide valuable information to narrow down the areas that have to be searched so as to speed up the restoration and improve the system reliability and availability. For example, the dispatch center can inform the linemen to focus on certain types of outage causes or even dispatch appropriate crews earlier to restore the system. Power distribution systems fault cause identification can be viewed as a classification problem in the sense that operators try to categorize a reported outage into one of the existing fault cause classes that have been carefully arranged by domain experts. Various works make use of the measured currents and voltages to gain fault cause information [3], [4]. With the development of data mining techniques, a large number of researches have studied the applications of data mining approaches to power systems [5], [6]; some researches [7], [8] have utilized the historical power distribution outage data (which usually contain additional information such as environmental factors) to extract fault patterns. However, many real-world outage data are imbalanced [9], i.e., at least one of the classes significantly outnumbers other classes. Since most commonly used classification methods aim to minimize the overall error rate, imbalanced data may cause a biased classification performance [10]: a high accuracy on the majority class but a very low accuracy on the minority class.

The E-algorithm has been extended from the fuzzy classification algorithm proposed by Ishibuchi *et al.* [11] to alleviate the effect of data imbalance [12]. In this paper, the E-algorithm is applied to Duke Energy distribution outage data to illustrate its effectiveness for power distribution system fault cause identification with imbalanced data. Three top customer interruption causes in Duke Energy (and in most utilities [13]) are used as prototypes: tree, animal contact, and lightning strike; the data constitution is imbalanced with respect to each prototyping fault cause. The performance of the E-algorithm is compared with artificial neural network (ANN) that has been studied in our previous works [14], [15] and extensively applied in power systems studies [16], [17].

Section II introduces the data mining-based fuzzy classification E-algorithm. Section III briefly describes power distribution fault cause identification using Duke Energy outage data. Section IV presents the performance of the E-algorithm on fault cause identification and compares it with that of ANN.

Manuscript received May 26, 2006; revised August 5, 2006. This work is supported by the National Science Foundation under Grant ECS-0245383. Paper no. TPWRS-00311-2006.

L. Xu and M.-Y. Chow are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: lexu@ieee.org; chow@ncsu.edu).

L. S. Taylor is with Duke Energy, Distribution Standard, Charlotte, NC 28201 USA (e-mail: lstaylor@duke-energy.com).

Digital Object Identifier 10.1109/TPWRS.2006.888990

II. E-ALGORITHM

Ishibuchi *et al.* have proposed an elegant fuzzy classification algorithm that demonstrates great capabilities to classify well-preprocessed data sets (e.g., little data noises, balanced data, and no outliers); however, it may not achieve a comparably good performance on imbalanced data. Its modification and extension, E-algorithm, is able to effectively alleviate the influence of data imbalance.

A. Fuzzy Sets and Fuzzy Rules

A fuzzy classification system has two key elements: fuzzy sets and fuzzy rules. A fuzzy set can be fully defined by its membership function. Fuzzy rules offer human-like reasoning capabilities and provide transparent inference mechanisms.

Fuzzy rules are usually expressed as an *if-then* form. Assuming a fuzzy classification system with K rules, m inputs, and n outputs (in this paper, we only consider the case with single output attribute, so $n = 1$), the kth rule is expressed as

$$R_k : \text{IF } x_1 \text{ is } A_{1,k} \text{ AND } \ldots \text{AND } x_m \text{ is } A_{m,k} \quad \text{THEN } y \text{is } B_k \tag{1}$$

where $k = 1, \ldots, K, A_{i,k}, i = 1, \ldots, m$ is the fuzzy set for input attribute x_i in rule R_k, and B_k is the fuzzy set for output attribute y in rule R_k. Fuzzy rules can also be expressed in vector forms

$$\mathbf{A}_k \Rightarrow B_k, \quad \text{where } \mathbf{A}_k = (A_{1,k}, \cdots, A_{m,k}). \tag{2}$$

The determination of both fuzzy membership functions and fuzzy rules usually requires sufficient domain knowledge from experts. It can be a challenging task to develop a good fuzzy classification system without adequate domain knowledge. However, the E-algorithm extracts both fuzzy set membership functions and fuzzy rules from data by utilizing the statistical information revealed by the normalized fuzzy versions of data mining measures *support* and *confidence* so that an effective fuzzy classification system can be developed even in short of domain knowledge [12].

B. Support and Confidence

The association analysis in data mining discovers meaningful relationships among attributes in the form of association rules $\mathbf{X} \Rightarrow Y$ [18] (only one consequent attribute is considered in this paper). Association rules have the same format as fuzzy rules; they indicate that the data satisfying the antecedent part $\mathbf{X}$ are also likely to satisfy the consequent part Y.

Support measures how often the antecedent attributes $\mathbf{X}_k$ ($k = 1, \ldots, K, K$ is the number of association rules) and the consequent attribute Y_k occur together. *Confidence* measures how likely it is that the consequent attribute Y_k occurs when the antecedent attributes $\mathbf{X}_k$ have occurred

$$\text{support}(\mathbf{X}_k \Rightarrow Y_k) = P(\mathbf{X}_k \cap Y_k) \tag{3}$$

$$\text{confidence}(\mathbf{X}_k \Rightarrow Y_k) = P(Y_k|\mathbf{X}_k) = \frac{P(\mathbf{X}_k \cap Y_k)}{P(\mathbf{X}_k)} \tag{4}$$

where $P(\cdot)$ is the probability operator.

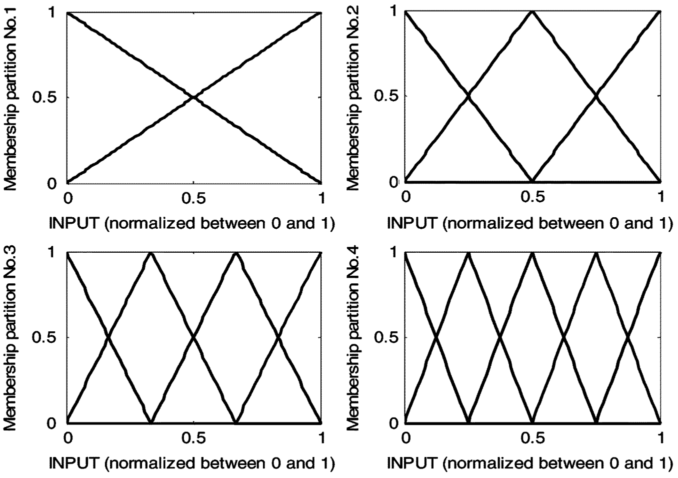

Fig. 1. Four fuzzy partitions for each attribute membership function.

These two measures are extended into fuzzy versions using the compatibility grade [19] of a data sample with a rule. The *compatibility grade* of $\mathbf{X}_l$ with the kth rule is denoted as $\mu_{\mathbf{A}_k}(\mathbf{x}_l)$

$$\mu_{\mathbf{A}_k}(\mathbf{x}_l) = \mu_{A_{1,k}}(x_{1,l}) \times \cdots \times \mu_{A_{m,k}}(x_{m,l}) \tag{5}$$

where $\mathbf{X}_l = (x_{1,l}, \ldots, x_{m,l})$ is the lth data sample, $l = 1, \ldots, N$, where N is the total number of data samples, m is the number of attributes in each data sample, and $\mu_{A_{i,k}}(x_{i,l}), i = 1, \ldots, m$ represents the membership of the attribute $x_{i,l}$ in relation to the antecedent fuzzy set $A_{i,k}$ of the rule R_k.

The normalized fuzzy version of *support*, $s(\mathbf{A}_k \Rightarrow B_k)$, is defined as the ratio of the normalized sum of the compatibility grades of class B_k data with the kth rule to the number of data samples

$$s(\mathbf{A}_k \Rightarrow B_k) = \frac{\frac{\sum_{l \in B_k} \mu_{\mathbf{A}_k}(\mathbf{x}_l)}{N_{B_k}/N}}{N}. \tag{6}$$

The normalized fuzzy version of *confidence*, $c(\mathbf{A}_k \Rightarrow B_k)$, is defined as the percentage of the normalized sum of compatibility grades of class B_k data with the kth rule to the sum of compatibility grades of all data samples with the kth rule

$$c(\mathbf{A}_k \Rightarrow B_k) = \frac{\frac{\sum_{l \in B_k} \mu_{\mathbf{A}_k}(\mathbf{x}_l)}{N_{B_k}/N}}{\sum_{l=1}^{N} \mu_{\mathbf{A}_k}(\mathbf{x}_l)}. \tag{7}$$

C. Membership Functions

Since the appropriate fuzzy set partitions for each attribute is unknown *a priori*, the E-algorithm simultaneously uses four fuzzy set partitions for each attribute, as shown in Fig. 1. As a result, each antecedent attribute is initially associated with 14 fuzzy sets generated by these four partitions as well as a special "do not care" set (i.e., 15 in total).

This approach can help define fuzzy membership functions for the problems lack of domain knowledge, with the trade-off of increasing the computational demand though.

D. Fuzzy Rules

The E-algorithm first enumerates all the possible combinations of antecedent fuzzy sets and then assigns each combination a consequent part to generate a rule; all these rules form an initial rule population. Since each antecedent attribute corresponds to 15 possible fuzzy sets, the total number of possible combinations of antecedent fuzzy sets for m attributes is 15^{m}, which increases exponentially with m. In order to reduce the computational demand, only the rules with less than or equal to three antecedent attributes are generated in this paper; it is also a common practice in Ishibuchi *et al.*'s work [11]. As shown in (5), the compatibility grade is the product of several membership values. Since membership values are always no larger than one, the longer the rule is, the more membership values are included to calculate the compatibility grade of a data sample with the rule, and then the smaller the compatibility grade usually is. Thus, only short rules are logically included to reduce the computing requirement while keeping the reasonable performance.

Each antecedent fuzzy set's combination corresponds to one fuzzy rule once its consequence is specified. The consequence is determined by (8): the class that gets the maximum confidence value given the antecedent fuzzy sets combination $\mathbf{A}_k$ is assigned as the rule consequence

$$B_k = \arg \max_{p \in \{1,2,\ldots,M\}} c(\mathbf{A}_k \Rightarrow \text{class } p) \tag{8}$$

where M is the total number of classes.

The E-algorithm further assigns each rule a certainty grade CF_k as the rule weight. The certainty grade is the difference between the maximum confidence value and the second largest confidence value c_{sec} given $\mathbf{A}_k$

$$\mathrm{CF}_k = c(\mathbf{A}_k \Rightarrow B_k) - c_{\mathrm{sec}} \tag{9}$$

where $c_{\mathrm{sec}} = \max_{q \in \{1,2,\cdots,M\}; q \neq p} c(A_k \Rightarrow \text{class } q)$.

Using the product of $s(\mathbf{A}_k \Rightarrow B_k)$ and $c(\mathbf{A}_k \Rightarrow B_k)$ as the measure, a user-defined number N_s rules for each class are chosen by trial and error from the initial rule population (in this paper, $N_s = 30$). These winning rules form the fuzzy classification rule base S extracted from the data and are responsible to make decisions in classification tasks.

E. Fuzzy Classification

When implementing fuzzy classification tasks on test data, the single winner rule method [19] is employed. For any test data $\mathbf{x}_r$, the single winner rule method chooses from the fuzzy classification rule base S a rule that yields the maximum product of the compatibility grade with the test data $\mu_{A_k}(\mathbf{x}_r)$ and the certainty grade CF_k. This winner rule determines the class to which $\mathbf{x}_r$ belongs.

III. Power Distribution System Fault Cause Identification

A. Data Selection

In this paper, Duke Energy distribution outage data are used to illustrate the fault cause identification. Every time an outage in Duke Energy distribution systems is detected as a result of the activation of protective devices (e.g., a circuit breaker, a fuse), the relative information is recorded into the data collection system. Each outage record consists of 33 information fields, of which six are considered as the most essential and influential factors based on the suggestions from Duke Energy senior distribution engineers and statistical significance test [20]. These six fields are: *circuit ID, weather, season, time of day, phases affected,* and *protective devices activated*. The attribute *cause* entered by the crew during the restoration process records the actual root cause of the outage; it is used as the class label. Three major customer interruption causes (tree, animal contact, and lightning strike) are used in this paper for illustration purposes.

TABLE I
Overview of Elements of Each Influential Factor

Circuit ID	all the circuit identification numbers under consideration, e.g., 19031208
Weather	fair, cold, rain, wind, wind & lightning, lightning, hail, snow, ice, hot
Season	spring, summer, fall, winter
Time of day	midnight, morning, afternoon, evening
Phases Affected	X, Y, Z, XY, XZ, YZ, XYZ, no info
Protective Device Activated	Transmission Device, Station Circuit Breaker, Line Recloser, Primary Fuse, Transformer Fuse, Transformer CSP, Panel Base, SEC/SVC Self Clearing, Manual Device, Primary Self Clearing

Based on domain experts' suggestions and considerations of different geographical features and system status, seven of Duke Energy's 32 service regions in North Carolina and South Carolina are selected as reasonable service area representations: *Chapel Hill (CH), Clemson (CS), Durham (DH), Greenville (GV), Hickory (HC), Lancaster (LC),* and *Winston-Salem (WS)*. These seven regions cover metropolitan areas, cities, towns, rural areas, and wooded areas; these regions also embody both old systems and new systems.

B. Fault Cause Identification Scheme

All the six selected factors are categorical variables as shown in Table I. The categorical variables are transformed into numerical variables using the likelihood measure [21] so that they can be used in the E-algorithm, which requires numerical inputs in order to determine attributes' fuzzy memberships in relation to different antecedent fuzzy sets.

The likelihood measure as shown in (10) represents the conditional probability of the occurring of an outage caused by a specific fault given a certain condition (e.g., the likelihood of an observed outage caused by tree given icy weather condition)

$$L_{i,j} = \frac{N_{i,j}}{N_j} \tag{10}$$

where i refers to fault cause (e.g., tree, lightning), j refers to fault-related event or condition (e.g., windy weather, fuse activated), $N_{i,j}$ is the number of outages caused by fault i under condition j, N_j is the number of outages under condition j, and $L_{i,j}$ is the likelihood measure of fault i given condition j.

The likelihood measure can provide useful information for fault cause identification; it is logically used as the inputs to the E-algorithm. However, the likelihood measure depends on both fault type i and influential condition j; the same data are

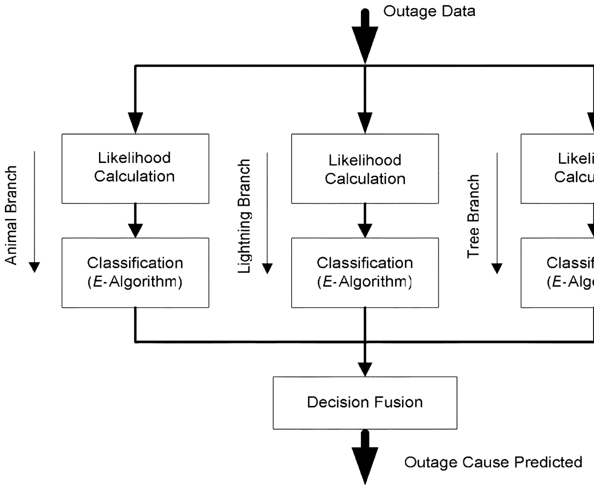

Fig. 2. Power distribution fault cause identification schematic diagram.

TABLE II
PROPORTIONS OF TREE-CAUSED, ANIMAL-CAUSED, AND LIGHTNING-CAUSED FAULTS IN DIFFERENCE REGIONS

	Training Data			Test Data		
	Lightning (%)	Animal (%)	Tree (%)	Lightning (%)	Animal (%)	Tree (%)
CH	7.31	15.35	29.58	5.71	14.58	31.18
CS	9.85	22.66	30.82	4.46	22.56	34.58
DH	6.34	11.69	22.14	1.89	9.67	23.58
GV	11.10	18.53	27.87	2.93	16.87	30.65
HC	15.50	16.86	19.26	2.87	18.14	24.41
LC	10.13	8.70	27.32	2.92	7.46	35.89
WS	9.53	12.94	21.66	9.72	14.53	21.90

mapped to different sets of likelihood measures for different fault causes, even under the same influential condition. It means that the likelihood measures change along with fault causes. As a result, the power distribution fault cause identification scheme shown in Fig. 2 consists of three identical branches in parallel: the tree branch, the animal branch, and the lightning branch. Each branch identifies its designated fault cause. It can be expanded to identify more fault causes.

Within each branch, the outage data are first transformed by the likelihood calculation module. The generated likelihood measures are then passed to the classification module, where the E-algorithm is applied to determine the class of the inputted outages. Since each branch is only responsible for its designated fault cause, it faces a binary classification task. For instance, the tree branch categorizes the cause of an inputted outage as either tree or nontree, which can be animal contact, lightning strike, or others. The number of the outages by one particular fault cause may only account for a small percentage of the total outages due to the diversity of fault causes (the proportions of tree-caused faults, animal-caused faults, and lightning-caused faults are shown in Table II). Thus, the classifiers always face imbalanced data classification tasks.

The decision fusion module combines the results from different branches into a final classification decision. When different branches reach consistent outage cause estimations, the decision fusion model can make a straightforward decision. When conflicting results happen occasionally, this module compares the compatibility grades of the test data with the winner rule from each branch chosen by the single winner rule method to determine the outage cause.

TABLE III
CONFUSION MATRIX

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

IV. RESULTS AND DISCUSSIONS

The Duke Energy outage data from 1994 to 2002 are used in this paper. In each representative region, the data are divided into training data and test data by year: the outage data from 1994 to 1999 are used as the training set and the remaining data (2000 to 2002) form the test set.

Table II shows proportions of tree-caused, animal-caused, and lightning-caused faults in each region. Lightning-caused faults account for an average proportion of 9.97% in training data and 4.36% in test data. The region of DH only has 1.89% lightning-caused faults in its test dataset. Animal-caused faults have an average proportion of 15.25% in training data and 14.83% in test data. Tree-caused fault is one of the largest outage categories: tree-caused faults have an average proportion of 25.52% in training data and 28.88% in test data. Comparing with the lightning fault and animal fault, it is relatively balanced between tree-caused outage and nontree-caused outage.

The performance of the E-algorithm for outage cause identification is compared with a three-layer feed-forward ANN using back-propagation algorithm as investigated in our previous works. The ANN-based fault cause identification employs the same scheme as the one in Fig. 2. Due to its randomness property, ANN is run 30 times for each case in order to generate statistically representative results. In this paper, only the results from individual branches are presented in order to demonstrate the performance of two methods on imbalanced data.

A. Performance Measure

When the data are imbalanced, the conventional performance measure of the overall classification accuracy can be misleading. Take a two-class imbalanced data set Q as an example. Assume that 95% of the data are from the majority class R, while only 5% of the data are from the minority class T. If a classifier blindly categorizes every case into class R, it can still achieve an overall accuracy as high as 95% without even processing the data, which is certainly undesirable. Kubat *et al.* have proposed the g-mean [22] to evaluate the classification performance on imbalanced data sets. The g-mean is developed from confusion matrix as shown in Table III (assuming the tree/animal/lightning faults are positive classes and nontree/nonanimal/nonlightning faults as negative classes).

The true positive rate $(Acc^{+} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}))$ indicates the classification accuracy of the positive class, while the true negative rate $(Acc^{-} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}))$ indicates the classification accuracy of the negative class. The g-mean examines classification accuracies on both positive and negative classes

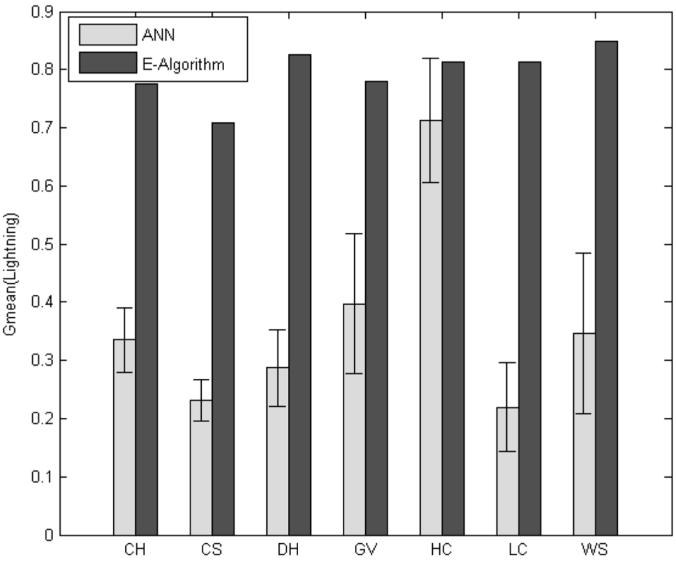

Fig. 3. G-means for lightning fault identification (test data).

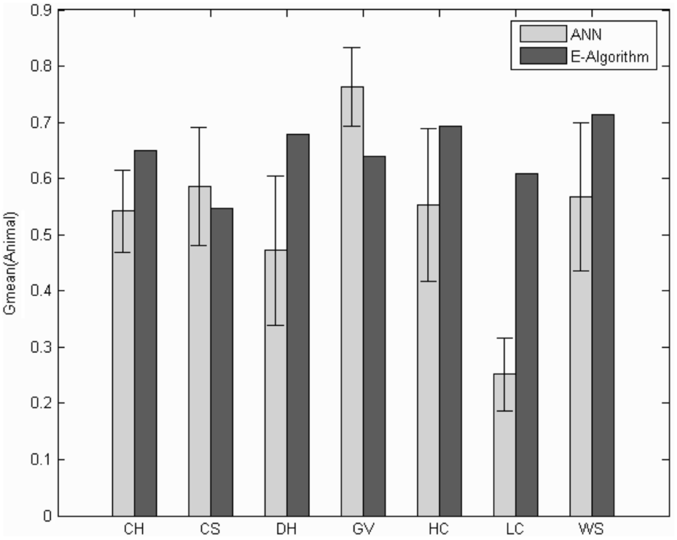

Fig. 4. G-means for animal fault identification (test data).

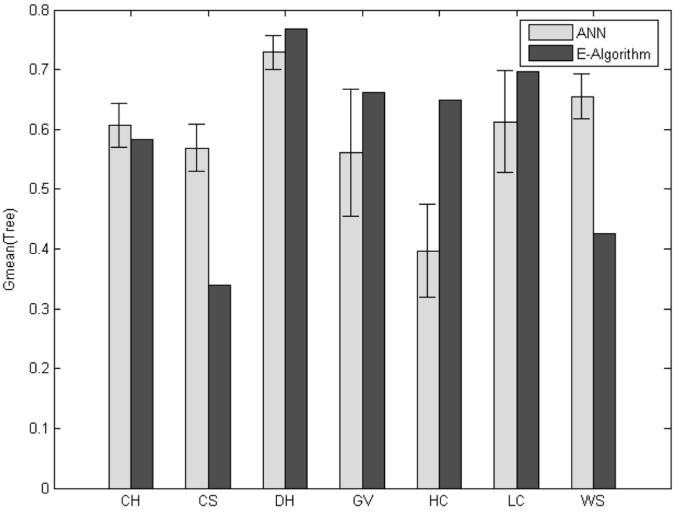

Fig. 5. G-means for tree fault identification (test data).

TABLE IV
P-VALUES OF ONE-SAMPLE TESTS OF HYPOTHESIS ON G-MEANS

	Lightning	Animal	Tree
CH	<0.0001	0.0028	0.9158
CS	<0.0001	0.7711	1.0000
DH	<0.0001	0.0019	0.0053
GV	<0.0001	0.9995	0.0310
HC	0.0013	0.0227	<0.0001
LC	<0.0001	<0.0001	0.0271
WS	<0.0001	0.0164	1.0000

and punishes large disparities between them; it is mathematically described as

$$\mathrm{g-mean} = \sqrt{Acc^{+} \times Acc^{-}}. \tag{11}$$

The basic idea behind the g-mean is to maximize the accuracies on both classes: the g-mean is large when both Acc^{+} and Acc^{-} are large and the difference between Acc^{+} and Acc^{-} is small, i.e., the classification accuracies on both positive and negative classes are high and there is no large disparity between them, which represents a balanced performance.

B. Results

The g-means achieved by the E-algorithm and ANN on test data for lightning/animal/tree fault are presented in Figs. 3–5, respectively. The performance of ANN is based on its 30-run results: the height of the vertical bars in the figures represents the mean value, and the "whisker" represents its 95% confidence intervals. The E-algorithm has deterministic results once N_s, the number of rules to be included in the fuzzy classification rule base S, has been decided, so only the actual value is presented as the height of the corresponding vertical bar. The region names have been defined in Section III-A.

One-sample tests of hypothesis are also performed on the experimental data to determine whether the E-algorithm provides higher g-means statistically than ANN does. That is

$$\mathrm{g-mean}_{\mathrm{ANN}} = \mathrm{g-mean}_{E} \text{ (the null hypothesis)} \tag{12}$$

is tested against

$$\mathrm{g-mean}_{\mathrm{ANN}} < \mathrm{g-mean}_{E} \text{ (the alternative hypothesis).} \tag{13}$$

The decision is made based on P-values of the tests that show the probability of obtaining the existing experimental data given the null hypothesis [23]; so a low P-value leads to the rejection of the null hypothesis. The commonly used level of significance 0.95 is chosen in this paper; so a P-value under 0.05 will reject the null hypothesis in favor of the alternative hypothesis. Table IV presents the P-values of the one-sample tests of hypothesis on g-means achieved by the E-algorithm and ANN.

Fig. 3 clearly indicates that the E-algorithm has a significant dominance of g-means when identifying lightning-caused faults. In the region of LC, the average value of g-mean by the E-algorithm even exceeds ANN by as much as 271%. In Table IV, the P-values of the one-sample tests of hypothesis for lightning-caused faults are smaller than 0.05 in all seven regions. Thus, the null hypothesis $(\mathrm{g}-\mathrm{mean}_{\mathrm{ANN}}=\mathrm{g}-\mathrm{mean}_{E})$ is rejected in favor of the alternative hypothesis $(\mathrm{g}-\mathrm{mean}_{\mathrm{ANN}}<\mathrm{g}-\mathrm{mean}_{E})$; in another words, it can be concluded that the g-mean by the E-algorithm is larger than the average g-mean by ANN.

Although the E-algorithm consistently achieves higher performance than ANN for lightning-caused faults in all the seven selected regions, the similar clear-cut conclusions cannot be drawn for the other two fault causes.

As shown in Fig. 4, the g-means by the E-algorithm for animal-caused fault are larger than the average g-means by ANN in five regions but are smaller in two regions: CS and GV. The one-sample tests of hypothesis also indicate that the E-algorithm outperforms ANN in five regions and ANN actually has a larger g-mean value than the E-algorithm in the region of GV; the conclusion cannot be drawn for the remaining region CS at the significance level of 0.95.

Similarly, the E-algorithm for tree-caused faults has larger g-means in four regions but smaller g-means in CH, CS, and WS comparing with ANN, as shown in Fig. 5. The results of one-sample tests of hypothesis in Table IV show that the E-algorithm outperforms ANN in four regions and gets smaller g-means in CS and WS. The conclusion cannot be made for the region of CH at the significance level of 0.95.

C. Discussions

When implementing classification tasks, a standard ANN with back-propagation algorithm aims to minimize the overall error rate. For an imbalanced data set, the majority class has dominant influence on the overall error since ANN tends to prioritize the different classes in favor of the class with more training data examples in order to achieve a high overall accuracy. This biased favor may sacrifice the performance on classifying minority class and achieves a high accuracy on the majority class but a very low, sometimes unacceptable, accuracy on the minority class.

The E-algorithm extracted qualified classification rules from the data based on the statistical information revealed by two normalized fuzzy versions of data mining concepts: *support* $s(\mathbf{A}_k \Rightarrow B_k)$ and *confidence* $c(\mathbf{A}_k \Rightarrow B_k)$; the normalization of these measures, as shown in (6) and (7), alleviates the effect of data imbalanced constitution [12].

Based on the comparisons, the E-algorithm achieves larger g-mean values than ANN in most of the cases. It is also noticed that with the increase of outage proportion, the dominance of the E-algorithm over ANN is weakened: the average proportion of lightning faults, animal faults, and tree faults is in an ascending order as shown in Table II, while E-algorithm outperforms ANN in all the seven regions for lightning faults, five regions for animal faults, and only four regions for tree faults.

We further examine the individual cases in Table II. It is observed that animal faults do not necessarily account for more percentage than lightning faults; neither do tree faults with animal faults. For example, the region of LC only has 7.46% animal faults, while the region of WS has 9.72% lightning faults; the region of WS only has 21.90% tree faults, while the region of CS has 22.56% of animal faults. Therefore, we disregard the categories of fault causes to further investigate the relationship between the increase of g-mean values and the percentage of positive classes (lightning/animal/tree faults), as shown in Fig. 6.

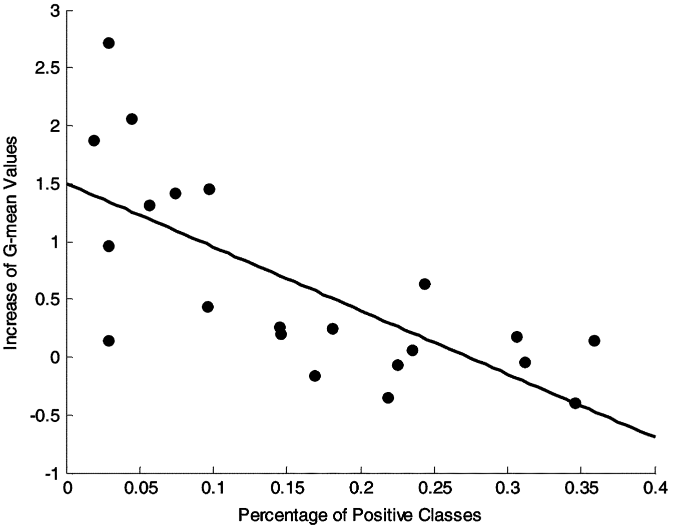

Fig. 6. Relationship between the increase of g-means and the percentage of positive classes.

TABLE V
CAUSE-AVERAGED OVERALL ACCURACY BY ANN AND E-ALGORITHM

	ANN	E-algorithm
Lightning	0.9516	0.9069
Animal	0.8767	0.8748
Tree	0.7527	0.7633

A simple linear regression model in Fig. 6 demonstrates the relationship between the percentage of positive classes (which can be animal fault, tree fault, or lightning fault) and the increase of g-mean values of the E-algorithm over ANN. This straight-line function in the figure is expressed as

$$\begin{aligned} &Y(\text{increase of G} - \text{mean values}) \\ &\quad = 1.50 - 5.49 \times X(\text{Percentage of Positive Classes}). \end{aligned} \tag{14}$$

As the percentage of positive classes gets larger, the increase of the g-mean values gets smaller. It means that when the data balance issue gets less severe (i.e., the data are more balanced), the improvement of the E-algorithm over ANN gets smaller; when the data constitution is relatively balanced, the E-algorithm does not show significant superiority over ANN.

G-mean emphasizes the balance in classifying different classes rather than the overall classification accuracy. Table V further presents the average overall accuracy rates (over seven regions) of the prototyping outage causes identification. The E-algorithm has an accuracy rate as high as 0.9069 for lightning-caused fault while maintaining much more balanced performance on different classes, though it is lower than the value of 0.9516 achieved by ANN. The E-algorithm has an overall accuracy of 0.8748 for animal-caused faults, which is close to the rate of 0.8767 achieved by ANN. The E-algorithm even achieves a higher average accuracy of 0.7633 for tree-caused faults than the rate of 0.7527 by ANN.

Comparing with lightning strike and animal contact, tree fault diagnosis is more complicated: the tree fault category in Duke Energy outage database includes a few subcategories such as trees, danger trees, and vines, which diversifies the fault patterns; tree faults are closely related with many influential factors (such as weather condition, geographic information, tree trimming cycle, and human activities), many of which are not captured by the measurements or not presented in the outage database. As we can see in Table V, the accuracy of tree-caused outage identification is lower than the other two outage causes. Considering that the results are based on real-world data, the E-algorithm does achieve a satisfactory performance. Dispatchers can use the algorithm with more confidence on lightning-caused faults, followed by animal-caused faults, and then tree-caused faults.

Other than its advantage in classifying imbalanced data, the E-algorithm can extract fuzzy rules from the data to help explain the inference mechanism of outage cause identification; this is the generally acknowledged advantage of a fuzzy classification technique. A few rule examples for lightning-caused faults are as follows.

1) IF the likelihood measure of weather is medium (according to membership partition No. 4 in Fig. 1) AND the likelihood measure of phases affected is high (according to membership partition No. 1 in Fig. 1), THEN the outage is likely to be caused by lightning.
2) IF the likelihood measure of weather is low (according to membership partition No. 1 in Fig. 1), THEN the outage is not likely to be caused by lightning.

As presented in Section II-C, the E-algorithm simultaneously applies four membership partitions, as shown in Fig. 1, to each attribute and then chooses the antecedent fuzzy set combinations that have more statistical support to implement the classification task. The attributes within a chosen combination do not have to follow the same partition. For example, the first antecedent attribute in rule example 1 uses membership partition No. 4 (as labeled in Fig. 1) while the second antecedent attribute uses membership partition No. 1.

V. Conclusion

Effective power outage cause identification can help expedite the restoration procedure and improve the distribution system reliability and availability. However, the data imbalance issue encountered in many real-world data sets often affects the performance of fault cause identification, especially for minority-class causes, since most commonly used methods aim to minimize the overall error rate. In this paper, the E-algorithm that is extended from the elegant fuzzy classification algorithm by Ishibuchi *et al.* for imbalanced data is applied to Duke Energy distribution outage data for cause identification. Its performance in terms of g-mean, the performance measure for imbalanced classification, is compared with ANN investigated in our previous works. The results show that the E-algorithm can achieve better performance when the data are imbalanced; the superiority is proportional to the severeness of the data imbalance.

References

[1] C.-F. Chien, S.-L. Chen, and Y.-S. Lin, "Using Bayesian network for fault location on distribution feeder," *IEEE Trans. Power Del.*, vol. 17, no. 3, pp. 785–793, Jul. 2002.

[2] D. Thukaram, H. P. Khincha, and H. P. Vijaynarasimha, "Artificial neural network and support vector machine approach for locating faults in radial distribution systems," *IEEE Trans. Power Del.*, vol. 20, no. 2, pp. 710–721, Apr. 2005.

[3] O. Dag and C. Ucak, "Fault classification for power distribution systems via a combined wavelet-neural approach," in *Proc. Int. Conf. Power System Technology*, 2004, pp. 1309–1314.

[4] K. L. Butler and J. A. Momoh, "A neural net based approach for fault diagnosis in distribution networks," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, 2000, pp. 1275–1278.

[5] S. Santoso and J. D. Lamoree, "Power quality data analysis: From raw data to knowledge using knowledge discovery approach," in *Proc. IEEE Power Eng. Soc. Summer Meeting*, 2000, vol. 1, pp. 172–177.

[6] H. M. Dola and B. H. Chowdhury, "Data mining for distribution system fault classification," in *Proc. Annu. North Amer. Power Symp.*, 2005, pp. 457–462.

[7] J. T. Peng, C. F. Chien, and T. L. B. Tseng, "Rough set theory for data mining for fault diagnosis on distribution feeder," *Proc. Inst. Elect. Eng., Gen., Transm., Distrib.*, vol. 151, no. 6, pp. 689–697, Nov. 2004.

[8] L. Xu, M.-Y. Chow, and X. Z. Gao, "Comparisons of logistic regression and artificial neural network on power distribution systems fault cause identification," in *Proc. IEEE Mid-Summer Workshop Soft Computing Industrial Applications*, 2005, pp. 128–131.

[9] L. Xu and M.-Y. Chow, "Power distribution systems fault cause identification using logistic regression and artificial neural network," in *Proc. Int. Conf. Intelligent Systems Application Power Systems*, 2005, pp. 163–168.

[10] C. Chen, A. Liaw, and L. Breiman, Using Random Forest to Learn Imbalanced Data. [Online]. Available: www.stat.berkeley.edu/users/chenchao/666.pdf.

[11] H. Ishibuchi and T. Yamamoto, "Comparison of heuristic rule weight specification methods," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2002, pp. 908–913.

[12] L. Xu, M.-Y. Chow, and L. S. Taylor, "Data mining based fuzzy classification algorithm for imbalanced data," in *Proc. IEEE World Congr. Computational Intelligence*, 2006, pp. 4216–4221.

[13] R. E. Brown, *Electric Power Distribution Reliability*. Boca Raton, FL: CRC, 2002.

[14] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 53–60, Feb. 2006.

[15] M.-Y. Chow, S. O. Yee, and L. S. Taylor, "Recognizing animal-caused faults in power distribution systems using artificial neural networks," *IEEE Trans. Power Del.*, vol. 8, no. 3, pp. 1268–1274, Jul. 1993.

[16] D. Niebur and A. J. Germond, "Power flow classification for static security assessment," in *Proc. Int. Forum Applications Neural Networks Power Systems*, 1991, pp. 83–88.

[17] Y. Hayashi, S. Iwamoto, S. Furuya, and C.-C. Liu, "Efficient determination of optimal radial power system structure using Hopfield neural network with constrained noise," *IEEE Trans. Power Del.*, vol. 11, no. 3, pp. 1529–1535, Jul. 1996.

[18] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann, 2001.

[19] H. Ishibuchi, T. Nakashima, and T. Murata, "Voting in fuzzy rule-based systems for pattern classification problems," *Fuzzy Sets Syst.*, vol. 103, no. 2, pp. 223–238, 1999.

[20] L. Xu, M.-Y. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," in *Proc. North Amer. Power Symp.*, Oct. 20–21, 2003.

[21] M.-Y. Chow and L. S. Taylor, "Analysis and prevention of animal-caused faults in power distribution systems," *IEEE Trans. Power Del.*, vol. 10, no. 2, pp. 995–1001, Apr. 1995.

[22] M. Kubat, R. Holte, and S. Matwin, "Machine learning for the detection of oil spills in radar images," *Mach. Learn.*, vol. 30, pp. 195–215, 1998.

[23] R. E. Walpole, R. H. Myers, S. L. Myers, K. Ye, and K. Yee, *Probability and Statistics for Engineers and Scientists*, 7th ed. Englewood Cliffs, NJ: Prentice-Hall, 2002.

Le Xu (S'01) received the B.Eng. degree in automation from Tsinghua University, Beijing, China, in 2001 and the M.S. degree in electrical engineering from North Carolina State University, Raleigh, in 2003. He is currently pursuing the Ph.D. degree with the Advanced Diagnosis Automation and Control Laboratory at North Carolina State University, Raleigh.

His research interests include intelligent health monitoring of power distribution systems.

Mo-Yuen Chow (S'81–M'82–SM'93) received the B.S. degree in electrical and computer engineering from the University of Wisconsin, Madison, in 1982 and the M.Eng. and Ph.D. degrees from Cornell University, Ithaca, NY, in 1983 and 1987, respectively.

Upon completion of the Ph.D. degree, he joined the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, and has held the rank of Professor since 1999. His core technology is diagnosis and control, artificial neural network, and fuzzy logic with applications to areas, including motors, process control, power systems, and communication systems. He has established the Advanced Diagnosis Automation and Control (ADAC) Laboratory at North Carolina State University.

Leroy S. Taylor (M'93) was born in 1949 in Greenville, NC. He received the B.A. degree in physics from the University of North Carolina in 1971

He is a Senior Distribution Engineer for Duke Energy, Charlotte, NC. Joining Duke Power in 1977, he acquired extensive experience in distribution system engineering, operation, and construction. Since 1987, he has conducted intensive investigation on the cause of power quality disturbances that originate in the distribution system. He has also redesigned several Duke Power mainframe reporting systems used to evaluate and improve distribution system reliability and power quality.

Dr. Taylor is a registered professional engineer in North Carolina.