Exploratory Analysis of Massive Data for Distribution Fault Diagnosis in Smart Grids

Yixin Cai, Mo-Yuen Chow
Electrical and Computer Engineering,
North Carolina State University
July 2009
Outline

- Introduction
- Integrate data
- Evaluate a single feature
- Build a fault cause classifier
- Summary
Introduction

- Intelligent fault management
  - detection, recording, location, diagnosis, restoration, …

- Problem of interest
  - Diagnosis: predict what the root cause is based on the available information before the engineers go on-site
  - Help (not replace) the engineers to identify the root cause faster

- Challenges
  - Stochastic nature of faults
  - Noisy data with errors
  - More and more incoming data in Smart Grids
Data Integration

- Data sources
  - Utility OMS database
Data Integration

- Data sources
  - Utility OMS database
  - Public database on weather, environment, geographic features, etc.

http://landcover.usgs.gov/usgslandcover.php
Data Integration

- Data sources
  - Utility OMS database
  - Public database on weather, environment, geographic features, etc.
  - Private vendors

Data Integration

- Data integration under GIS framework
  - Spatial relation
Data Integration

- Data integration under GIS framework
  - Spatial relation
  - Spatial-temporal relation
Evaluate a Single Feature

- Data preprocess
  - Define the root cause of interest
  - Clean errors and noises
  - Extract features

- Categorical features
  - Likelihood measure: \( L_{i,j} = P(o_i \mid X = x_j) = \frac{N_{i,j}}{N_j} \)
  - (Mosaic) plot

Weather vs. Tree Faults

Weather Conditions:
1: Clear Weather, 2: Extreme Temperature, 4: Raining, 6: Thunderstorm, 8: Windy

Season vs. Tree Faults
Season:
1: Spring, 2: Summer, 3: Fall, 4: Winter

Land Use vs. Tree Faults
Land Use:
Continuous features

- Likelihood measure
  \[ L_{i,j} = P(o_i \mid X \geq x_j) = \frac{N_{i,j}}{N_j} \]
- plot

Distance to Trees vs. Tree Faults

Distance to Roads vs. Tree Faults

Wind Speed vs. Tree Faults
Build a Fault Cause Classifier

- Linear discriminant analysis (LDA)

\[ D = w^T f = \sum_{i=1}^{N} w_i f_i \quad D = [\Sigma^{-1}(\mu_1 - \mu_0)]^T f \]

- Logistic regression (LR)

\[ \logit(c = 1) = \ln \frac{P(c = 1)}{P(c = 0)} = \alpha + \beta^T f \quad P(c = 1) = \frac{1}{1 + e^{-\alpha - \beta^T f}} \]

- Comparison

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>linear classifier</td>
<td>non-linear classifier</td>
</tr>
<tr>
<td>Data assumption</td>
<td>normal distributed with equal variance</td>
<td>none</td>
</tr>
<tr>
<td>Computation</td>
<td>matrix manipulation</td>
<td>maximum likelihood</td>
</tr>
</tbody>
</table>
Case Study

- **Data sources**
  - Progress Energy Carolinas outage database
  - NC Climate Office
  - NC State Univ. GIS data service

- **Fault causes of interest**
  - Tree-caused
  - Animal-caused
  - Other

- **Features**
  - 7 categorical
  - 5 continuous

- **Classifiers**
  - LDA
  - LR

### Classification Performance Using LDA on Sample Dataset

<table>
<thead>
<tr>
<th></th>
<th>6 Features</th>
<th>12 Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>training</td>
<td>testing</td>
</tr>
<tr>
<td>Tree fault</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.75(0.01)</td>
<td>0.76(0.01)</td>
</tr>
<tr>
<td>POD</td>
<td>0.32(0.03)</td>
<td>0.34(0.03)</td>
</tr>
<tr>
<td>FAR</td>
<td>0.34(0.03)</td>
<td>0.32(0.03)</td>
</tr>
<tr>
<td>Animal fault</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.84(0.02)</td>
<td>0.83(0.01)</td>
</tr>
<tr>
<td>POD</td>
<td>0.31(0.04)</td>
<td>0.29(0.04)</td>
</tr>
<tr>
<td>FAR</td>
<td>0.42(0.05)</td>
<td>0.43(0.05)</td>
</tr>
</tbody>
</table>

### Classification Performance Using LR on Sample Dataset

<table>
<thead>
<tr>
<th></th>
<th>6 Features</th>
<th>12 Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>training</td>
<td>testing</td>
</tr>
<tr>
<td>Tree fault</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.76(0.02)</td>
<td>0.76(0.02)</td>
</tr>
<tr>
<td>POD</td>
<td>0.32(0.03)</td>
<td>0.32(0.03)</td>
</tr>
<tr>
<td>FAR</td>
<td>0.30(0.04)</td>
<td>0.30(0.04)</td>
</tr>
<tr>
<td>Animal fault</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.83(0.02)</td>
<td>0.83(0.02)</td>
</tr>
<tr>
<td>POD</td>
<td>0.30(0.03)</td>
<td>0.31(0.03)</td>
</tr>
<tr>
<td>FAR</td>
<td>0.42(0.04)</td>
<td>0.41(0.06)</td>
</tr>
</tbody>
</table>
Summary

- Methods for exploratory data analysis
  - Integrate data from multiple sources under GIS framework
  - Use likelihood measure to evaluate both categorical and continuous features
  - Apply LDA and LR as fault cause classifiers

- Findings
  - LDA and LR performs similar
  - Adding new features helps fault diagnosis

- Future work
  - Systematic feature selection methods
  - Advanced fault diagnosis algorithms
  - Novel sampling strategy