Advances in Machine Learning for Credit Card Fraud Detection

May 14, 2014

Alejandro Correa Bahnsen
Introduction

Europe fraud evolution
Internet transactions (millions of euros)

- 2007: €500
- 2008: €600
- 2009: €700
- 2010: €800
- 2011E: €
- 2012E: €

Source: SNT Security and Trust

Université du Luxembourg
Introduction

US fraud evolution
Online revenue lost due to fraud (Billions of dollars)
Introduction

- Increasing fraud levels around the world
- Different technologies and legal requirements makes it harder to control
- Lack of collaboration between academia and practitioners, leading to solutions that fail to incorporate practical issues of credit card fraud detection:
  - Financial comparison measures
  - Huge class imbalance
  - Low-latency response time
 Agenda

- Introduction
- Database
- Evaluation
- Algorithms
  - Cost-sensitive logistic regression
  - Bayes Minimum Risk
  - Example-dependent cost-sensitive decision tree
- Conclusions & Future Work
Simplify transaction flow
Data

- Larger European card processing company
- Jan2012 – Jun2013 card present transactions
- 1,638,772 Transactions
- 3,444 Frauds
- 0.21% Fraud rate
- 205,542 EUR lost due to fraud on test dataset
### Data

**Raw attributes**

<table>
<thead>
<tr>
<th>TRXID</th>
<th>Client ID</th>
<th>Date</th>
<th>Amount</th>
<th>Location</th>
<th>Type</th>
<th>Merchant Group</th>
<th>Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2/1/12 6:00</td>
<td>580</td>
<td>Ger</td>
<td>Internet</td>
<td>Airlines</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2/1/12 6:15</td>
<td>120</td>
<td>Eng</td>
<td>Present</td>
<td>Car Rent</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2/1/12 8:20</td>
<td>12</td>
<td>Bel</td>
<td>Present</td>
<td>Hotel</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3/1/12 4:15</td>
<td>60</td>
<td>Esp</td>
<td>ATM</td>
<td>ATM</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3/1/12 9:18</td>
<td>8</td>
<td>Fra</td>
<td>Present</td>
<td>Retail</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3/1/12 9:55</td>
<td>1210</td>
<td>Ita</td>
<td>Internet</td>
<td>Airlines</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Derived attributes

<table>
<thead>
<tr>
<th>Trx ID</th>
<th>Client ID</th>
<th>Date</th>
<th>Amount</th>
<th>Location</th>
<th>Type</th>
<th>Merchant Group</th>
<th>Fraud</th>
<th>No. of Trx – same client – last 6 hour</th>
<th>Sum – same client – last 7 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2/1/12 6:00</td>
<td>580</td>
<td>Ger</td>
<td>Internet</td>
<td>Airlines</td>
<td>No</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2/1/12 6:15</td>
<td>120</td>
<td>Eng</td>
<td>Present</td>
<td>Car Renting</td>
<td>No</td>
<td>1</td>
<td>580</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2/1/12 8:20</td>
<td>12</td>
<td>Bel</td>
<td>Present</td>
<td>Hotel</td>
<td>Yes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3/1/12 4:15</td>
<td>60</td>
<td>Esp</td>
<td>ATM</td>
<td>ATM</td>
<td>No</td>
<td>0</td>
<td>700</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3/1/12 9:18</td>
<td>8</td>
<td>Fra</td>
<td>Present</td>
<td>Retail</td>
<td>No</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3/1/12 9:55</td>
<td>1210</td>
<td>Ita</td>
<td>Internet</td>
<td>Airlines</td>
<td>Yes</td>
<td>1</td>
<td>760</td>
</tr>
</tbody>
</table>

### By Group Last Function

<table>
<thead>
<tr>
<th>By</th>
<th>Group</th>
<th>Last</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client</td>
<td>None</td>
<td>hour</td>
<td>Count</td>
</tr>
<tr>
<td>Credit Card</td>
<td>Transaction Type</td>
<td>day</td>
<td>Sum(Amount)</td>
</tr>
<tr>
<td>Merchant</td>
<td>week</td>
<td></td>
<td>Avg(Amount)</td>
</tr>
<tr>
<td>Merchant Category</td>
<td>month</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merchant Country</td>
<td>3 months</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data

**Date of transaction**

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/03/2012</td>
<td>03:14</td>
</tr>
<tr>
<td>07/03/2012</td>
<td>00:47</td>
</tr>
<tr>
<td>07/03/2012</td>
<td>02:57</td>
</tr>
<tr>
<td>08/03/2012</td>
<td>02:08</td>
</tr>
<tr>
<td>14/03/2012</td>
<td>22:15</td>
</tr>
<tr>
<td>25/03/2012</td>
<td>05:03</td>
</tr>
<tr>
<td>26/03/2012</td>
<td>21:51</td>
</tr>
<tr>
<td>28/03/2012</td>
<td>03:41</td>
</tr>
</tbody>
</table>

Arithmetic Mean = \(\frac{1}{n} \sum t\)

Periodic Mean = \(\tan^{-1}\left(\frac{\sum \sin(t)}{\sum \cos(t)}\right)\)

Periodic Std = \(\sqrt{\ln\left(\frac{1}{\left(\frac{1}{n} \sum \sin(t)\right)^2 + \left(\frac{1}{n} \sum \cos(t)\right)^2}\right)}\)

\(t \sim \text{vonmises}(k \approx \frac{1}{\text{std}})\)

\(P(-zt < t < zt) = 0.95\)
## Data

<table>
<thead>
<tr>
<th>Date of transaction</th>
<th>Time Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/03/2012</td>
<td>03:14</td>
</tr>
<tr>
<td>07/03/2012</td>
<td>00:47</td>
</tr>
<tr>
<td>07/03/2012</td>
<td>02:57</td>
</tr>
<tr>
<td>08/03/2012</td>
<td>02:08</td>
</tr>
<tr>
<td>14/03/2012</td>
<td>22:15</td>
</tr>
<tr>
<td>25/03/2012</td>
<td>05:03</td>
</tr>
<tr>
<td>26/03/2012</td>
<td>22:15</td>
</tr>
<tr>
<td>28/03/2012</td>
<td>03:41</td>
</tr>
<tr>
<td>02/04/2012</td>
<td>02:02</td>
</tr>
<tr>
<td>03/04/2012</td>
<td>12:10</td>
</tr>
</tbody>
</table>

**New Features**

- Inside CI(0.95) last 30 days
- Inside CI(0.95) last 7 days
- Inside CI(0.5) last 30 days
- Inside CI(0.5) last 7 days
Confusion matrix

<table>
<thead>
<tr>
<th>True Class ((y_i))</th>
<th>Fraud ((y_i=1))</th>
<th>Legitimate ((y_i=0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted class ((p_i))</td>
<td>Fraud ((c_i=1))</td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>Legitimate ((c_i=0))</td>
<td>FN</td>
</tr>
</tbody>
</table>

- **Misclassification** = \(1 - \frac{TP+TN}{TP+TN+FP+FN}\)
- **Recall** = \(\frac{TP}{TP+FN}\)
- **Precision** = \(\frac{TP}{TP+FP}\)
- **F-Score** = \(2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\)
## Evaluation - Financial measure

### Motivation:

- Equal misclassification results
- Frauds carry different cost

<table>
<thead>
<tr>
<th>TRX ID</th>
<th>Amount</th>
<th>Fraud</th>
<th>Algorithm 1 Prediction (Fraud?)</th>
<th>Algorithm 2 Prediction (Fraud?)</th>
<th>Algorithm 3 Prediction (Fraud?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>580</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>1210</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Results:

- **Miss-Class**: 2 / 6
- **Cost**: 1222 / 1212 / 14
Cost matrix

<table>
<thead>
<tr>
<th>Predicted Positive</th>
<th>Actual Positive $y_i = 1$</th>
<th>Actual Negative $y_i = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_i = 1$</td>
<td>$C_{TP_i}$</td>
<td>$C_{FP_i}$</td>
</tr>
<tr>
<td>$c_i = 0$</td>
<td>$C_{FN_i}$</td>
<td>$C_{TN_i}$</td>
</tr>
</tbody>
</table>

where the cost associated with two types of correct classification, true positives and true negatives, and the two types of misclassification errors, false positives and false negatives, are presented.
Evaluation

- As discussed in [Elkan 2001], the cost of correct classification should always be lower than the one of misclassification. These are referred to as “reasonableness” conditions.

\[ C_{FPi} > C_{TNi} \quad \text{and} \quad C_{FNi} > C_{TPi} \]

- Using the “reasonableness” conditions, the cost matrix can be scaled and shifted to a simpler one with only one degree of freedom

\[
\begin{array}{c|c}
\text{Negative} & C^*_{FNi} = \frac{(C_{FNi} - C_{TNi})}{(C_{FPi} - C_{TNi})} \\
\text{Positive} & C^*_{TPi} = \frac{(C_{TPi} - C_{TNi})}{(C_{FPi} - C_{TNi})}
\end{array}
\]
Cost-sensitive problem definition

- Classification problem cost characteristic:

\[ b_i = C^*_{FN_i} - C^*_{TP_i} - 1 \]

with mean \( \mu_b \) and std \( \sigma_b \)

- A classification problem is defined as:

<table>
<thead>
<tr>
<th>cost-insensitive</th>
<th>( \mu_b = 0 ) and ( \sigma_b = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>class-dependent</td>
<td>( \mu_b \neq 0 ) and ( \sigma_b = 0 )</td>
</tr>
<tr>
<td>cost-sensitive</td>
<td>( \sigma_b &gt; 0 )</td>
</tr>
<tr>
<td>example-dependent</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

Cost matrix: Fraud detection

<table>
<thead>
<tr>
<th></th>
<th>Actual Positive ( y_i = 1 )</th>
<th>Actual Negative ( y_i = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Positive ( c_i = 1 )</td>
<td>( C_a )</td>
<td>( C_a )</td>
</tr>
<tr>
<td>Predicted Negative ( c_i = 0 )</td>
<td>( Amt_i )</td>
<td>0</td>
</tr>
</tbody>
</table>

\( C_a \) refers to the administrative cost and \( Amt_i \) to the amount of transaction \( i \).
Evaluation

Cost-sensitive problem evaluation

- Cost of applying a classifier to a given set

\[ C(S) = \sum_{i=1}^{N} \left( y_i (c_i C_{TP_i} + (1 - c_i) C_{FN_i}) + (1 - y_i)(c_i C_{FP_i} + (1 - c_i) C_{TN_i}) \right) \]

- Savings are:

\[ C^*(S) = \frac{C_s(S) - C(S)}{C_s(S)} \]

where

\[ C_s(S) = \min \left\{ C_0(S), C_1(S) \right\} \]

and \( C_0 \), \( C_1 \) refers to special cases where for all the examples, \( c_i \) equals to 0 and 1 respectively.
Agenda

- Introduction
- Database
- Evaluation
- Algorithms
  - Cost-sensitive logistic regression
  - Bayes Minimum Risk
  - Example-dependent cost-sensitive decision tree
- Conclusions & Future Work
Logistic Regression

- Model

\[
\log \left( \frac{p}{1-p} \right) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n
\]

- Cost Function

\[
J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[ -y_i \log(p_\theta(x_i)) - (1 - y_i) \log(1 - p_\theta(x_i)) \right]
\]
Cost Sensitive Logistic Regression

- **Cost Matrix**

<table>
<thead>
<tr>
<th>Predicted Positive $c_i = 1$</th>
<th>Actual Positive $y_i = 1$</th>
<th>Actual Negative $y_i = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Negative $c_i = 0$</td>
<td>$Amt_i$</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Cost Function**

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[ y_i \left( p_{\theta}^*(x_i) Ca + (1 - p_{\theta}^*(x_i)) Amt_i \right) + (1 - y_i) p_{\theta}^*(x_i) Ca \right]$$

- **Objective**

Find $\theta$ that minimized the cost function
Cost Sensitive Logistic Regression

- **Cost Function**
  
  \[
  J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[ y_i \left( p_\theta^*(x_i) Ca + (1 - p_\theta^*(x_i)) Am t_i \right) + (1 - y_i) p_\theta^*(x_i) Ca \right]
  \]

- **Gradient**
  
  \[
  \frac{\partial J(\theta)}{\partial \theta(j)} = \frac{1}{m} \sum_{i=1}^{m} \left[ -y_i Am t_i + Ca - y_i Ca - y_i \right] \left( \frac{-e^{-\sum_{j=1}^{n} \theta(j) x(i,j)}}{1 + e^{-\sum_{j=1}^{n} \theta(j) x(i,j)}} \right)^2 \left( -x(i,j) \right)
  \]

- **Hessian**
  
  \[
  \frac{\partial^2 J(\theta)}{\partial \theta(j1) \partial \theta(j2)} = \frac{1}{m} \sum_{i=1}^{m} \left[ -y_i Am t_i + (1 - y_i) Ca \right] \left( (-x(i,j1))(x(i,j2))^2 \right) \left( 1 - p_\theta^*(x_i) \right)^3 \left( p_\theta^*(x_i) \right)^3
  \]
Experiments – Logistic Regression

Sub-sampling procedure:

Select all the frauds and a random sample of the legitimate transactions.

* OLD Dataset
Experiments – Logistic Regression

Results

* OLD Dataset
Experiments – CS Logistic Regression

Results

* OLD Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Cost</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Model</td>
<td>€ 148,562</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>€ 31,174</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>€ 37,785</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>€ 66,245</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>€ 67,264</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>€ 73,772</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>€ 85,724</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments – CS Logistic Regression

- **Savings**
  - Logistic Regression
  - Cost-Sensitive Logistic Regression
  - Comparison between Training and Under-sampling

- **F1-Score**
  - Logistic Regression
  - Cost-Sensitive Logistic Regression
  - Comparison between Training and Under-sampling
Introduction
Database
Evaluation
Algorithms
  • Cost-sensitive logistic regression
  • Bayes Minimum Risk
  • Example-dependent cost-sensitive decision tree
Conclusions & Future Work
Bayes Minimum Risk

- Decision model based on quantifying tradeoffs between various decisions using probabilities and the costs that accompany such decisions

- Risk of classification

\[
R(c_i = 0|x_i) = C_{TN_i} (1 - \hat{p}_i) + C_{FN_i} \cdot \hat{p}_i.
\]

\[
R(c_i = 1|x_i) = C_{TP_i} \cdot \hat{p}_i + C_{FP_i} (1 - \hat{p}_i)
\]
Bayes Minimum Risk

• Using the different risks the prediction is made based on the following condition:

\[ c_i = \begin{cases} 
0 & R(c_i = 0|X_i) \leq R(c_i = 1|X_i) \\
1 & \text{otherwise}
\end{cases} \]

• Example-dependent threshold

\[ t_{BMR_i} = \frac{C_{FP_i} - C_{TN_i}}{C_{FN_i} - C_{TN_i} - C_{TP_i} + C_{FP_i}} \]

Is always defined taking into account the “reasonableness” conditions
Probability Calibration

- When using the output of a binary classifier as a basis for decision making, there is a need for a probability that not only separates well between positive and negative examples, but that also assesses the real probability of the event [Cohen and Goldszmidt 2004]
Probability Calibration

- Reliability Diagram

\( \pi_1 \) is the positive rate and \( \hat{p}_i \) is the predicted probability
Probability Calibration

- ROC Convex Hull calibration [Hernandez-Orallo et al. 2012]

<table>
<thead>
<tr>
<th>Class (y)</th>
<th>Prob (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>0</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Probability Calibration

- ROC Convex Hull calibration

<table>
<thead>
<tr>
<th>Class (y)</th>
<th>Prob (p)</th>
<th>Cal Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>0.333</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.333</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0.333</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0.666</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>0.666</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>0.666</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The calibrated probabilities are extracted by first grouping the probabilities according to the points in the ROCCH curve, and then the calibrated probabilities are equal to the slope for each group.
Probability Calibration

- Reliability Diagram

![Reliability Diagram](image)
Experiments – Bayes Minimum Risk

• Estimation of the fraud probabilities using one of the following algorithms:
  1. Random Forest
  2. Decision Trees
  3. Logistic Regression

• For each algorithm comparison of:
  • Raw prediction
  • Bayes Minimum Risk
  • Probability Calibration and Bayes Minimum Risk

• Trained using the different sets
  • Training
  • Under-sampling
Experiments – Bayes Minimum Risk

**Savings**

<table>
<thead>
<tr>
<th></th>
<th>-</th>
<th>BMR</th>
<th>CAL BMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0%</td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td>Under-sampling</td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
</tr>
</tbody>
</table>

**F1-Score**

<table>
<thead>
<tr>
<th></th>
<th>-</th>
<th>BMR</th>
<th>CAL BMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Under-sampling</td>
<td>0.05</td>
<td>0.15</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Experiments – Bayes Minimum Risk

**Savings**

- Decision Trees: - BMR CAL BMR
- Logistic Regression: - BMR CAL BMR

**F1-Score**

- Decision Trees: - BMR CAL BMR
- Logistic Regression: - BMR CAL BMR

- Training
- Under-sampling
Agenda

- Introduction
- Database
- Evaluation
- Algorithms
  - Cost-sensitive logistic regression
  - Bayes Minimum Risk
  - Example-dependent cost-sensitive decision tree
- Conclusions & Future Work
EDCS – Decision trees

Decision trees

Classification model that iteratively creates binary decision rules \((x^j, v^j_m)\) that maximize certain criteria

Where \((x^j, v^j_m)\) refers to making a rule using feature \(j\) on value \(m\)
EDCS – Decision trees

Decision trees - Construction

- Then the impurity of each leaf is calculated using:

  \[
  \begin{align*}
  \text{Misclassification} & : I_m(\pi_1) = 1 - \max\{\pi_1, (1 - \pi_1)\} \\
  \text{Entropy} & : I_e(\pi_1) = -\pi_1 \log \pi_1 - (1 - \pi_1) \log (1 - \pi_1) \\
  \text{Gini} & : I_g(\pi_1) = 2\pi_1(1 - \pi_1)
  \end{align*}
  \]

- Afterwards the gain of applying a given rule to the set $S$ is:

  \[
  \text{Gain}((x^j, l^j_m)) = I(\pi_1) - \frac{|S^l|}{|S|} I(\pi_1^l) - \frac{|S^r|}{|S|} I(\pi_1^r)
  \]
EDCS – Decision trees

Decision trees - Construction

• The rule that maximizes the gain is selected

$$(best_x, best_l) = \arg\max_{(i,m)} Gain((x^j, l^j_m))$$

• The process is repeated until a stopping criteria is met:
EDCS – Decision trees

Decision trees - Pruning

• Calculation of the Tree error and pruned Tree error

\[ \epsilon(Tree, S) = \frac{\epsilon(EB(Tree, branch), S) - \epsilon(Tree, S)}{|Tree| - |EB(Tree, branch)|} \]

• After calculating the pruning criteria for all possible trees. The maximum improvement is selected and the Tree is pruned.

• Later the process is repeated until there is no further improvement.
EDCS – Decision trees

- Maximize the accuracy is different than maximizing the cost.
- To solve this, some studies had been proposed method that aim to introduce the cost-sensitivity into the algorithms [Lomax and Vadera 2013].
- However, research have been focused on class-dependent methods [Draper et al. 1994; Ting 2002; Ling et al. 2004; Li et al. 2005; Kretowski and Grzes 2006; Vadera 2010]
- We propose:
  - Example-dependent cost based impurity measure
  - Example-dependent cost based pruning criteria
EDCS – Decision trees

Cost based impurity measure

- The impurity of each leaf is calculated using:

\[ I_c(S) = C_s(S) = \min \left\{ C_0(S), C_1(S) \right\} \]

\[ f(S) = \begin{cases} 0 & \text{if } C_0(S) \leq C_1(S) \\ 1 & \text{otherwise} \end{cases} \]

- Afterwards the gain of applying a given rule to the set \( S \) is:

\[ Gain_c((x^j, l^j_m)) = I_c(S) - (I_c(S^l) + I_c(S^r)) \]
Weighted vs. not weighted gain

\[
Gain((x^j, l^j_m)) = I(\pi_1) - \frac{|S^l|}{|S|} I(\pi^l_1) - \frac{|S^r|}{|S|} I(\pi^r_1)
\]

\[
Gain_c((x^j, l^j_m), S) = I_c(S) - (I_c(S^l) + I_c(S^r))
\]

- Using the not weighted gain, when booths left and right leafs have the same prediction, the gain is equal 0

if

\[
f(S^l) = f(S^r)
\]

then

\[
I_c(S) = (I_c(S^l) + I_c(S^r))
\]
EDCS – Decision trees

Cost sensitive pruning

\[ PC_c = \frac{C(S,f(S,Tree))-C(S,f(S,EB(Tree,branch)))}{|Tree|-|EB(Tree,branch)|} \]

- New pruning criteria that evaluates the improvement in cost of eliminating a particular branch
Experiments - EDCS – Decision trees

• Comparison of the following algorithms:
  • Decision Tree – not pruned
  • Decision Tree – error based pruning
  • Decision Tree – cost based pruning
  • EDCS-Decision Tree – not pruned
  • EDCS-Decision Tree – error based pruning
  • EDCS-Decision Tree – cost based pruning

• Trained using the different sets:
  • Training
  • Under-sampling
  • Cost-proportionate Rejecting-sampling
  • Cost-proportionate Over-sampling
Experiments - EDCS – Decision trees

% Savings

- DT not pruned
- DT error pruning
- DT cost pruning
- EDCSDT not pruned
- EDCSDT error pruning
- EDCSDT cost pruning
Experiments - EDCS – Decision trees

F1-Score

- DT not pruned
- DT error pruning
- DT cost pruning
- EDCSDT not pruned
- EDCSDT error pruning
- EDCSDT cost pruning
Experiments - EDCS – Decision trees

% Savings

- Training
- Under sampling
- Rejection sampling
- Over sampling

- DT error pruning
- DT cost pruning
- EDCSDT cost pruning
Experiments - EDCS – Decision trees

**Tree size**

- Training
- Under sampling
- Rejection sampling
- Over sampling

- DT error pruning
- DT cost pruning
- EDCSDT cost pruning

**Training time (m)**

- Training
- Under sampling
- Rejection sampling
- Over sampling

- DT error pruning
- DT cost pruning
- EDCSDT cost pruning
Experiments – Comparison

% Savings

F1-Score

Cost-Sensitive Logistic Regression
RF - CAL-BMR
EDCSDT cost p

Cost-Sensitive Logistic Regression
RF - CAL-BMR
EDCSDT cost p
Conclusions

• New framework for defining cost-sensitive problems

• Including the cost into Logistic Regression increases the savings

• Bayes minimum risk model arise to better results measure by savings and results are independent of the base algorithm used

• Calibration of probabilities help to achieve further savings

• Example-dependent cost-sensitive decision trees improves the savings and have a much lower training time than traditional decision trees
Future work

• Boosted Example Dependent Cost Sensitive Decision Trees

• Example-Dependent Cost-Sensitive Calibration Method

• Reinforced Learning (Asynchronous feedback)
Contact information

Alejandro Correa Bahnsen
University of Luxembourg
Luxembourg

al.bahnsen@gmail.com
http://www.linkedin.com/in/albahnsen
http://www.slideshare.net/albahnsen
http://www.slideshare.net/albahnsen
References


References