Constrained Classification of Large Imbalanced Data

Martin Hlosta, R. Stríž, J. Zendulka, T. Hruška

Brno University of Technology, Faculty of Information Technology
Božetěchova 2, 612 66 Brno
ihlosta@fit.vutbr.cz

22.10.2013, Ostrava
Annual Conference of IT4Innovations
Problem

- Data from Security research, data from malware detection
- Binary classification of highly imbalanced data (about 1:33)
- Large data – >5 mil. records (vs medical datasets)
- Accuracy constraints for minority class, 99%
- Fast classification of unknown data
• **Imbalance ratio of our dataset**
Imbalanced Classification

- Typical classification algorithms don’t work well for imbalanced data
- Larger imbalance ratio, worse performance, e.g. 1:99 – all data to the majority class
- New methods to deal with them
  - **Sampling** – random undersampling, oversampling, informed methods – SMOTE (artificial generation of new data).
  - **Cost-sensitive learning** – penalizing misclassification on minority classes more.
  - **Ensemble methods**
Imbalanced Classification

• New evaluation metrics (vs. accuracy)
  • Geometric mean = G-mean
    • $G - \text{mean} = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$
  • Harmonic mean = F-measure
  • Area under ROC curve = AUC

• Sensitivity = TPR = accuracy on minority class
• Specificity = TNR = accuracy on majority class
• **Maximize specificity with constraint on sensitivity**
  • Objective function: specificity
Our approach

- Cost-sensitive Logistic Regression (CS-LR) combined with optimization algorithms
  1. **CS-LR** – generation of candidate solutions, different costs (fast, vs. SVM, NN)
  2. **Optimization** – optimizing candidates from CS-LR using stochastic methods – GA, PSO

- Resulting model: weight vector (fast classification of new data)
Our approach

**Constrained Classification of Large Imbalanced Data (22.10.2013)**

**Input data**

**Initial costs**

**CS-LR block**

- LR<sub>c1</sub> cost-sensitive LR threshold moving
- LR<sub>c2</sub>
- ... LR<sub>cn</sub>

Order models by G-mean descending

**Optimization block**

- Select top models as initial candidates
- Find optimum satisfying the minority accuracy constraint

- GA
- PSO

**Final model**

<table>
<thead>
<tr>
<th>threshold</th>
<th>TPR</th>
<th>TNR</th>
</tr>
</thead>
</table>
• Logistic regression – machine learning method, weight model, logistic function $[0;1]$ thresholding to determine class

• Imbalanced data
  a) threshold moving
  b) case-sensitive learning, determine the cost for misclassifying minor class data

• combination of both to find initial solutions for optimization block

• optimization criterion – G-mean
  a) Unconstrained solutions
  b) Constrained – satisfying the accuracy on minority class
Stochastic optimization algorithms

• Optimization of weights from CS-LR block
• Stochastic optimization algorithms
  • Genetic algorithm – GA
  • Particle Swarm Optimization - PSO
Genetic Algorithm

- Optimization algorithm, natural selection and biological evolution
  - Selection, cross-over, mutation
- Initial chromosomes – random or output of CS-LR
- Fitness function definition – includes the accuracy constraint on minor class – 99%

$$\text{fitness} = (\text{Sens} \times C1 + \text{Spec} \times C2) \times \text{IsConstr} + (\text{Sens} \times C3 + \text{Spec} \times C4)$$

- $C_i$ – importance ratio between class accuracies
Particle Swarm Optimization

- Swarm intelligence - social behavior of birds/insect
- PSO – birds searching for food, updating velocity
  - Towards personal best
  - Towards the best of the neighborhood
- Imbalanced data with constraints
  1. Penalty function – penalizing solutions that do not satisfy the constraint
  2. Strategy with modified updating – the personal best is updated only in some cases
    - Higher spec. and some particle satisfy constraint
    - Higher sensitivity and no particle satisfy constraint
- Initial solution – random or from CS-LR
Experiments

- 5 000 000 records, 120 binary attributes
- Class ratio: 1:33
- Minor class accuracy constraint: 99%
- Initial solutions for Optimization
  - Constrained/Unconstrained/Random

1. Cross-validation
2. Training and testing on the same dataset - where the algorithms converge after reasonable number of iterations
3. Behavior of algorithms
## CS-LR best solutions

<table>
<thead>
<tr>
<th>Cons./Uncons.</th>
<th>Specificity (majority)</th>
<th>Sensitivity (minority)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconstrained</td>
<td>0.783</td>
<td>0.869</td>
</tr>
<tr>
<td>Constrained</td>
<td>0.480</td>
<td>0.990</td>
</tr>
</tbody>
</table>
Cross-validation

- 5 times - 5-fold stratified CV
- PSO1 = PSO with penalty, PSO2 = modified update
- PSO with penalty and initial candidates from CS-LR outperforms others
- Initial solutions from CS-LR lead to better solutions than strating from random
- Sensitivity constraint - sensitivity didn’t drop under 0.9899
Train-test on the same data

- 10 x 1000 iterations
- Best mean solution also for PSO with penalty function and candidates from CS-LR
- Best overall for PSO with modified strategy
GA - behavior

![Graph showing the behavior of GA over iterations. The graph compares RAND MEAN, UNC MEAN, and CONS MEAN.](image)
PSO Penalty - behavior
PSO modified strategy - behavior

Constrained Classification of Large Imbalanced Data (22.10.2013)
Conclusion

- CS-LR + Stochastic Optimization method to solve Constrained Classification of Large Imbalanced Data
- Maximize specificity with high sensitivity constraint – 99%
  - CS-LR - generate initial candidate models for Optimization
  - GA/PSO – optimization of weights
- PSO with Penalty function with candidates from CS-LR outperformed others
- ~10% higher specificity than the best from CS-LR

- FUTURE
  - Utilize unlabeled data – semi-supervised learning

• Hlosta, M., Stríž, R., Zendulka, J., Hruška, T.: PSO-based Constrained Imbalanced Data Classification, 2013, s. 6 (accepted for the conference Informatics 2013)
Thank you for your attention.