REBAGG: REsampled BAGGing for Imbalanced Regression
LIDTA 2018 - Dublin, Ireland

Paula Branco, Luis Torgo and Rita P. Ribeiro

LIAAD - INESC TEC
DCC - FCUP
Dalhousie University

10 September 2018
Learning from Imbalanced Domains in Regression Tasks

Problem context

- Learning from imbalance domains is a key issue
- Characteristics:
  - non-uniform preferences of the user;
  - poor representation of the most important cases, in the available data
- This problem has received much attention which has been mostly focused on classification problems
- Fewer solutions exist for dealing with imbalanced regression problems
- Existing solutions for regression are essentially pre-processing methods that change the data distribution by over-sampling and/or under-sampling
Problem Definition

Regression Tasks

Given: a training sample, $\mathcal{D} = \{(x, y)\}_{i=1}^{N}$
Obtain: a model that approximates the regression function $y = f(x)$
Problem Definition

Regression Tasks

Given: a training sample, $\mathcal{D} = \{\langle x, y \rangle\}_{i=1}^{N}$
Obtain: a model that approximates the regression function $y = f(x)$

Imbalanced Regression Tasks

The goal is the predictive accuracy on a particular subset of the domain of the target variable $Y$ - the rare and important values.
An Illustrative Example of an Imbalanced Regression Task

Relevance Function of data set a1

\[ \phi(Y) \]

Y

0% 59%
An Illustrative Example of an Imbalanced Regression Task

Relevance Function of data set a1

user defined threshold

$\phi(Y)$

0% 59%

Y

P.Branco, L.Torgo and R.P.Ribeiro

REBAGG

10 September 2018
An Illustrative Example of an Imbalanced Regression Task

Relevance Function of data set a1

\[ \phi(Y) \]

User defined threshold

0% 59%

Y
An Illustrative Example of an Imbalanced Regression Task

Relevance Function of data set a1

\[ \phi(Y) \]

user defined threshold

\[ Y \]

0% 59%

\[ D_N \quad D_R \]

P.Branco, L.Torgo and R.P.Ribeiro

REBAGG

10 September 2018 4 / 21
A Solution: Pre-processing Strategies

Pre-processing strategies

- **Change the data distribution on the training sample** to make the learners focus on the important cases.
- The change carried out usually **aims at balancing the distribution of the rare (but more important) cases with the normal (frequent and less important) observations.**
- **Advantages:**
  - **flexibility:** use of any learning algorithm
  - **simplicity:** only involve manipulating the data distribution
- Several solutions for both classification and regression
Ensembles and Pre-processing Strategies

- Ensemble methods involve building several different models that are combined using a certain aggregation strategy.
- It is important to have diversity in the models that compose the ensemble.
- The diversity among the ensemble members is achieved using different data samples obtained by a given pre-processing strategy.
- The combination of ensemble methods with pre-processing strategies had promising results for the class imbalance problem.
- As far as we known, no similar attempt was made for imbalanced regression.
Ensembles and Pre-processing Strategies

- Have shown advantages in imbalanced binary classification
- Extended to other problems such as multiclass
- In particular, bagging-based strategies have shown interesting advantages
- Motivated the extension of bagging-based methods to regression problems
Main Contributions of our work

- Propose REBAGG (REsampled BAGGing), the first ensemble method proposed for tackling the problem of imbalanced regression
- REBAGG integrates data pre-processing strategies with bagging in imbalanced regression tasks
- REBAGG generates diversity in the models by applying different resampling methods to the training set
- Demonstrate the clear advantage of REBAGG in a diversity of domains and for multiple learning algorithms
Bagging (bootstrap aggregating)\(^1\) consists of building models using bootstrap samples of the original training data.

Two main steps are required:

- generation of \(k\) different models using bootstrap samples of the training set
- aggregation of the models prediction

\(^1\)Leo Breiman. “Bagging predictors.” Machine Learning, 24(2):123–140, 199
### Key Idea
- build a number of models using **pre-processed samples** of the training set
- use the trained models to obtain predictions on unseen data by applying an averaging strategy

### Pre-processed Samples of the Train Set
We developed four main types of resampling methods to apply:
- balance
- balance.SMT
- variation
- variation.SMT
REBAGG variants: Balancing the train sets

Original training set

Build sets of normal and rare cases

Build new balanced training sets using bootstrap or SMOTER

Obtain m models

Average the m models predictions

M*
REBAGG variants: Varying the normal to rare cases ratio

- Build sets of normal and rare cases
- Build new training sets with random ratio of rare to normal cases
- Obtain $m$ models
- Average the $m$ models predictions
## Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>N</th>
<th>tpred</th>
<th>p.nom</th>
<th>p.num</th>
<th>nRare</th>
<th>% Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>servo</td>
<td>167</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>34</td>
<td>20.4</td>
</tr>
<tr>
<td>a6</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>33</td>
<td>16.7</td>
</tr>
<tr>
<td>Abalone</td>
<td>4177</td>
<td>8</td>
<td>1</td>
<td>7</td>
<td>679</td>
<td>16.3</td>
</tr>
<tr>
<td>machCpu</td>
<td>209</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>34</td>
<td>16.3</td>
</tr>
<tr>
<td>a3</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>32</td>
<td>16.2</td>
</tr>
<tr>
<td>a4</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>31</td>
<td>15.7</td>
</tr>
<tr>
<td>a1</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>28</td>
<td>14.1</td>
</tr>
<tr>
<td>a7</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>27</td>
<td>13.6</td>
</tr>
<tr>
<td>boston</td>
<td>506</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>65</td>
<td>12.8</td>
</tr>
<tr>
<td>a2</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>22</td>
<td>11.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set</th>
<th>N</th>
<th>tpred</th>
<th>p.nom</th>
<th>p.num</th>
<th>nRare</th>
<th>% Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>a5</td>
<td>198</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>21</td>
<td>10.6</td>
</tr>
<tr>
<td>fuelCons</td>
<td>1764</td>
<td>38</td>
<td>12</td>
<td>26</td>
<td>164</td>
<td>9.3</td>
</tr>
<tr>
<td>availPwr</td>
<td>1802</td>
<td>16</td>
<td>7</td>
<td>9</td>
<td>157</td>
<td>8.7</td>
</tr>
<tr>
<td>cpuSm</td>
<td>8192</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>713</td>
<td>8.7</td>
</tr>
<tr>
<td>maxTorq</td>
<td>1802</td>
<td>33</td>
<td>13</td>
<td>20</td>
<td>129</td>
<td>7.2</td>
</tr>
<tr>
<td>bank8FM</td>
<td>4499</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>288</td>
<td>6.4</td>
</tr>
<tr>
<td>dAiler</td>
<td>7129</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>450</td>
<td>6.3</td>
</tr>
<tr>
<td>ConcrStr</td>
<td>1030</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>55</td>
<td>5.3</td>
</tr>
<tr>
<td>Accel</td>
<td>1732</td>
<td>15</td>
<td>3</td>
<td>12</td>
<td>89</td>
<td>5.1</td>
</tr>
<tr>
<td>airfoild</td>
<td>1503</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>62</td>
<td>4.1</td>
</tr>
</tbody>
</table>

**Table:** Used data sets and characteristics ($N$: nr of cases; $tpred$: nr predictors; $p.nom$: nr nominal predictors; $p.num$: nr numeric predictors; $nRare$: nr. cases with $\phi(y) > 0.8$; %Rare: $nRare/N \times 100$).
**Experimental Evaluation**

Regression Algorithms

<table>
<thead>
<tr>
<th>Learner</th>
<th>Parameter Variants</th>
<th>R package</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPART</td>
<td>( \text{minsplits} = {20, 50, 100, 200} ), ( \text{cp} = {0.01, 0.05} )</td>
<td>rpart</td>
</tr>
<tr>
<td>MARS</td>
<td>( \text{nk} = {10, 17} ), ( \text{degree} = {1, 2} ), ( \text{thresh} = {0.01, 0.001} )</td>
<td>earth</td>
</tr>
<tr>
<td>SVM</td>
<td>( \text{cost} = {10, 150, 300} ), ( \text{gamma} = {0.01, 0.001} )</td>
<td>e1071</td>
</tr>
<tr>
<td>RF</td>
<td>( \text{mtry} = {5, 7} ), ( \text{ntree} = {500, 750, 1500} )</td>
<td>randomForest</td>
</tr>
<tr>
<td>GBM</td>
<td>( \text{distribution} = \text{gaussian} ), ( \text{n.trees} = {300, 450, 600} ), ( \text{shrinkage} = {0.01, 0.1} ), ( \text{interaction.depth} = {1, 2} )</td>
<td>gbm</td>
</tr>
</tbody>
</table>
Experimental Evaluation

- 40 learning approaches (8 RPART + 8 MARS + 6 SVM + 6 RF + 12 GBM)
- 20 regression data sets
- Original learning approaches
- 2 variants of Bagging (BAGG) (using 10 and 40 models)
- 8 variants of REBAGG (using 10 and 40 models combined with the 4 base variants of REBAGG)
How to evaluate regression models for these tasks?
Utility-based Regression

- A framework allowing the calculation of **precision and recall for regression tasks** considering the issue of numeric accuracy (Torgo and Ribeiro, 2007; Ribeiro, 2011)
How to evaluate regression models for these tasks?

Utility-based Regression

- A framework allowing the calculation of **precision and recall for regression tasks** considering the issue of numeric accuracy (Torgo and Ribeiro, 2007; Ribeiro, 2011)

- Framework based on the concept of **utility-based regression**
How to evaluate regression models for these tasks?

Utility-based Regression

A framework allowing the calculation of **precision and recall for regression tasks** considering the issue of numeric accuracy (Torgo and Ribeiro, 2007; Ribeiro, 2011)

Framework based on the concept of **utility-based regression**

Usefulness of a prediction is a function of both the **numeric error** of the prediction and the **relevance of both the predicted** $\hat{y}$ and **true** $y$ **values**:

$$U(y, \hat{y}) = f(L(y, \hat{y}), \phi(y), \phi(\hat{y}))$$
A framework allowing the calculation of **precision and recall for regression tasks** considering the issue of numeric accuracy (Torgo and Ribeiro, 2007; Ribeiro, 2011)

Framework based on the concept of **utility-based regression**

Usefulness of a prediction is a function of both the **numeric error** of the prediction and the **relevance of both the predicted** $\hat{y}$ and **true y values**: $U(y, \hat{y}) = f(L(y, \hat{y}), \phi(y), \phi(\hat{y}))$

**Automatic methods for estimating the values of relevance** on extreme values tasks (Ribeiro, 2011).
How to evaluate regression models for these tasks?

Utility-based Regression

- A framework allowing the calculation of precision and recall for regression tasks considering the issue of numeric accuracy (Torgo and Ribeiro, 2007; Ribeiro, 2011)
- Framework based on the concept of utility-based regression
- Usefulness of a prediction is a function of both the numeric error of the prediction and the relevance of both the predicted $\hat{y}$ and true $y$ values:

$$U(y, \hat{y}) = f(L(y, \hat{y}), \phi(y), \phi(\hat{y}))$$

- Automatic methods for estimating the values of relevance on extreme values tasks (Ribeiro, 2011).
- We use this evaluation framework in the experimental analysis of our proposal.
Experimental Evaluation

- Evaluation according to the $F_1^\phi$ calculated using the set-up described
- 2 repetitions of a 10-fold stratified cross validation process
- Statistical significance of the observed paired differences was measured using the non-parametric Friedman Test and the post-hoc Nemenyi Test
Experimental Evaluation

Results

CD diagrams overall results for ensembles built with 10 models (left) and 40 models (right)
Globally, the advantage of using REBAGG algorithm is clear for all learners.

For RPART learner, our proposal presents a less striking advantage.

REBAGG algorithm displays a better performance with statistical significance in the majority of the situations.

REBAGG results are better when using 40 models, but are also good when using only 10 model.

Globally, no variant of REBAGG stands out.
Main Conclusions

- We propose **REBAGG** a new method for tackling imbalanced regression tasks that combines ensembles with pre-processing strategies;
- **REBAGG** is the *first ensemble method proposed for imbalanced regression*
- We show the **clear advantage of REBAGG strategy** when compared to:
  - using the imbalanced data set;
  - bagging strategy;
REBAGG: REsampled BAGGing for Imbalanced Regression
LIDTA 2018 - Dublin, Ireland

Paula Branco, Luis Torgo and Rita P. Ribeiro

LIAAD - INESC TEC
DCC - FCUP
Dalhousie University

10 September 2018

Full results and code for reproducing the experiments
https://github.com/paobranco/REBAGG