rethnicity
An R Package for Predicting Ethnicity From Names

Fangzhou Xie
Department of Economics, Rutgers University
August 26, 2022
Introduction

- Race information is often missing but we need a way to predict/impute for analysis.
- Surnames have been used to predict race (Fiscella and Fremont, 2006), but first names are also correlated (Fryer and Levitt, 2004).
- Other ways to improve accuracy, e.g. geo-coding, but name-only approach will be more versatile.
- I use both first names and last names for prediction.
Huge literature on race/ethnicity prediction (Fiscella and Fremont, 2006; Elliott et al., 2009; Ye et al., 2017; Lee et al., 2017; Sood and Laohaprapanon, 2018) and many available packages/services (Sood (2021), NamePrism, nationalize.io etc).

My package is neural network based (v.s. SSA name list matching), name-only (v.s. geocoding), and R package (v.s. online services)

The scope of rethnicity is close to (Sood and Laohaprapanon, 2018, ethnicolr).
Methodology

- Data and Undersampling
- Character-level Dictionary
- Bidirectional LSTM
- Knowledge Distillation
- Export to C++ and Create R Package
Methodology

- Data and Undersampling
- Character-level Dictionary
- Bidirectional LSTM
- Knowledge Distillation
- Export to C++ and Create R Package
Methodology

- Data and Undersampling
- Character-level Dictionary
- Bidirectional LSTM
- Knowledge Distillation
- Export to C++ and Create R Package
Methodology

- Data and Undersampling
- Character-level Dictionary
- Bidirectional LSTM
- Knowledge Distillation
- Export to C++ and Create R Package
Methodology

- Data and Undersampling
- Character-level Dictionary
- Bidirectional LSTM
- Knowledge Distillation
- Export to C++ and Create R Package
Data and Undersampling

- Most classification algorithms assume a relatively balanced dataset and equal misclassification cost. Classifiers trained on imbalanced dataset will introduce bias that disproportionately favors the majority class.
- One way to deal with this is to use oversampling, e.g. SMOTE algorithm (Japkowicz, 2000; Chawla et al., 2002; Fernandez et al., 2018). But this is done by KNN on feature space, which is infeasible in high-dimensional NLP problems.
- Florida Voter Registration dataset, extracted from Florida Voter Registration System with officially registered Florida voters as of 2017. Undersampling to adjust imbalance and to reduce training time.
Character-level Dictionary

- NLP algorithms consider “tokens” to be the unit and the first step is to create a “dictionary”. However, dictionary is dependent on the training data and is often large. Not very practical in analyzing names.
- Character-level dictionary (Zhang et al., 2015; Sutskever et al., 2011) can reduce dictionary size and prevent out-of-vocabulary problem.
Bidirectional LSTM

- Long short-term memory (Hochreiter and Schmidhuber, 1997, LSTM) is well-known in sequence modeling. Bidirectional LSTM (Graves and Schmidhuber, 2005, Bi-LSTM) captures context even better.
- Using Bi-LSTM over LSTM will improve accuracy in predicting races.
Knowledge Distillation

- Character level dictionary requires a larger model and more parameters. Knowledge distillation (Hinton et al., 2015) is used to compress the models.
- Train a large model on the dataset, then train a small model with similar architecture to simulate the results of the larger model. Use the small model for production.
Export to C++

- Training and distilling in Keras
- Export to C++ by *frugally-deep* project and reduce dependencies
- Wrapper in Rcpp and multi-threading by RcppThread
- Result in R package: *rethnicity*
Comparison with Other Packages/Services

- Availability (ethnicolr, NamePrism, nationalize.io)
- Accuracy (ethnicolr)
- Performance (ethnicolr)
### Availability

<table>
<thead>
<tr>
<th></th>
<th>Ethnicity</th>
<th>Ethnicolr</th>
<th>NamePrism</th>
<th>nationalize.io</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>free</td>
<td>free</td>
<td>free*</td>
<td>paid**</td>
</tr>
<tr>
<td>Rate Limit</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Dependency</td>
<td>Low</td>
<td>High</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Language</td>
<td>R</td>
<td>Python</td>
<td>API</td>
<td>API</td>
</tr>
</tbody>
</table>

**Table:** Comparison across some publicly available services/packages for predicting ethnicity from names. `rethnicity` provides a free and light-weight package for the R community without rate-limiting.

*: NamePrism requires filling application form.

** nationalize.io is free up to 1000 requests.
### Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Fullname</th>
<th></th>
<th>Lastname</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rethnicity</td>
<td>ethnicolr</td>
<td>rethnicity</td>
<td>ethnicolr</td>
</tr>
<tr>
<td>asian</td>
<td>0.79</td>
<td>0.60</td>
<td>0.73</td>
<td>0.54</td>
</tr>
<tr>
<td>black</td>
<td>0.73</td>
<td>0.55</td>
<td>0.67</td>
<td>0.32</td>
</tr>
<tr>
<td>hispanic</td>
<td>0.85</td>
<td>0.75</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>white</td>
<td>0.68</td>
<td><strong>0.90</strong></td>
<td>0.55</td>
<td><strong>0.88</strong></td>
</tr>
<tr>
<td>total</td>
<td>0.76</td>
<td><strong>0.83</strong></td>
<td>0.69</td>
<td><strong>0.78</strong></td>
</tr>
</tbody>
</table>

**Table:** Comparison of accuracy between rethnicity and ethnicolr.

- rethnicity significantly improves accuracy on minority groups.
“Your classifier managed to correctly identify 92% of individuals in my dataset and 79% of Europeans in my dataset. Worth taking into account that the sample I tried only consists of 1000 individuals of which only 200 were European.

Well done. Out of the 7 different ”ethnicity from name classifiers” I’ve tried, yours is the only one to score anything above 45% in Northern EU. 79% is amazing.” – MrMatsson on Github

Although the package is trained in U.S. context, it seems to generalize well on European names. But be cautious when used in other countries’ names.
Performance

Comparison of Elapsed Time

Sample Size vs. Time Elapsed (Seconds)

- ethnicolr
- rethnicity_0
- rethnicity_2
- rethnicity_4
- rethnicity_8
Citations and Usages


References II


