Predicting respiratory diseases from lung sounds using machine learning

Richard Anniko, Computer Science Bachelor, year III, University of Tartu, Institute of Computer Science
Supervisor: Dmytro Fishman

The purpose of this project was to create a freely accessible codebase that would make it easier for machine learning researchers to create innovation in the medical field. Combining machine learning with the medical field would decrease false diagnoses and save lives.

This project focused on predicting lung diseases from a breathing sounds dataset. 8 experiments were conducted on 5 different machine learning models. In addition, a novel data augmentation algorithm was performed and tested. This in total took 223 hours.

All of the results are packaged in a freely accessible codebase.

Implementation

The following experiments were conducted on 4 classical machine learning methods:
- using all features to train the models,
- using less complex models to decrease overfitting,
- using class weights to counter dataset unbalancedness,
- using fewer features to decrease noise in the data.

Experiments on the deep learning model:
- using all features to train the model,
- using class weights to counter dataset unbalancedness,
- using a novel data augmentation algorithm.

The following models were used: decision tree classifier, random forest, support vector machine, XGBoost and CNN.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Decision tree</th>
<th>Random forest</th>
<th>XGBoost</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>0.4956 +/- 0.12</td>
<td>0.4195 +/- 0.13</td>
<td>0.4461 +/- 0.18</td>
<td>0.1140 +/- 0.003</td>
</tr>
<tr>
<td>Less complex</td>
<td>0.4875 +/- 0.11</td>
<td>0.4175 +/- 0.13</td>
<td>N/A</td>
<td>0.1129 +/- 0.002</td>
</tr>
<tr>
<td>Class weights</td>
<td>0.5716 +/- 0.10</td>
<td>0.4338 +/- 0.18</td>
<td>N/A</td>
<td>0.1423 +/- 0.08</td>
</tr>
<tr>
<td>Fewer features</td>
<td>0.5749 +/- 0.12</td>
<td>0.4499 +/- 0.15</td>
<td>0.4693 +/- 0.21</td>
<td>0.2629 +/- 0.06</td>
</tr>
</tbody>
</table>

The results show that on the particular dataset, which was highly unbalanced (6 times more COPD patients than pneumonia patients), the simplest classical machine learning model – the decision tree – came out ahead. In addition, decreasing the number of features and implementing class weights helped.

The data augmentation algorithm did not work as intended, perhaps because it was used too excessively.

All the scores are the macro-average of f1-scores, meaning it includes each diagnosis class equally when calculating the f1-score.