Prediction prokaryotic incubation times from genomic features

Maeva Fincker

Problem

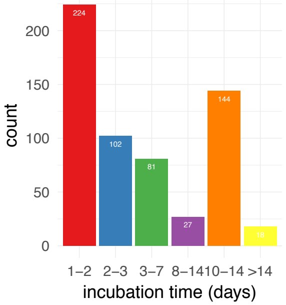
- only 2% of know microorganisms can be grown under laboratory conditions
- Low cost of sequencing technology has made the genomes of these uncultivable microbes available.
- Goal of the project: **predict incubation times of microorganisms from genomic features**

Data extraction

Labels:

- Incubation times scraped from BacDive database
- 6 classes: 1-2, 2-3, 3-7, 8-14, 10-14 and >14 days

Total: 596 examples



Features:

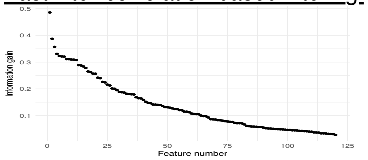
- Counting occurrences of proteins belonging Pfam families (as a proxy for function) in 596 genomes + genome length + number of 16S RNA operon
- Removing features that do not appear in at least 3 genomes
- Extremely sparse and redundant dataset

Total: 7535 features

Feature selection

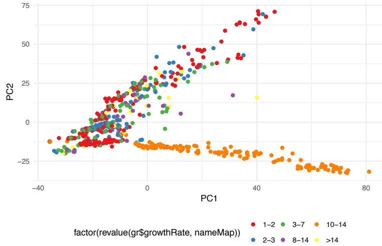
- Too many redundant features
- Tried different feature selection: AUC, information gain, fast-filter correlation

Fast filter-correlation based filtering:



- remove redundant features that are more correlated with each other than with the level using symmetric uncertainty
- 120 features selected

PCA data projection:



- 95% of the variance is explained by 392 features
- extremely clear separation between microorganisms with a 10-14 incubation time and the other

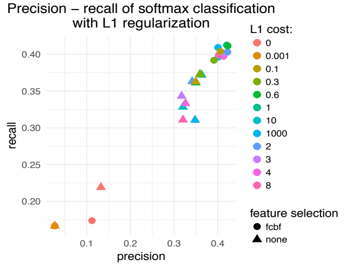
Classification

Model selection:

- Split the dataset 59 / 537 examples between test and training set
- Generalized precision/recall (sum over all classes)
- Parameter validation via 10-fold cross validation on the training set

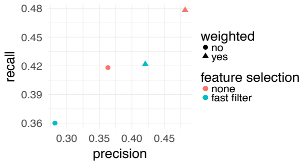
Softmax classification with L1 regularization

L1 cost optimization on full data



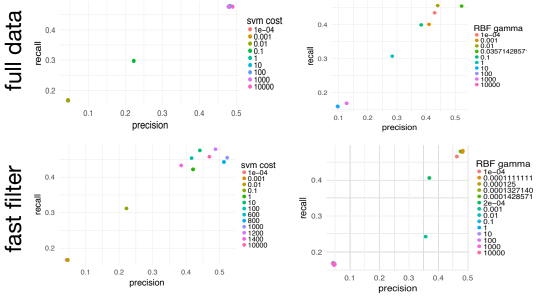
SVM one vs. one with RBF kernel

Weighted classes on full data

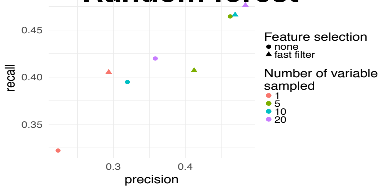


Cost selection

RBF gamma selection



Random forest



Comparison

Full data | Fast filter

Softmax

```
         true                         true
pred  0  1  2  3  4  5       pred  0  1  2  3  4  5
   0 12  4  2  0  0  1          0 14  5  2  0  0  0
   1  3  5  3  0  0  0          1  2  3  3  1  0  0
   2  1  3  2  2  0  0          2  0  5  0  3  0  2
   3  0  1  0  1  0  2          3  0  1  2  0  0  1
   4  0  1  0  1 14  0          4  0  0  0  1 14  0
   5  0  0  0  1  0  0          5  0  0  0  0  0  0
```

SVM

```
         true                         true
pred  0  1  2  3  4  5       pred  0  1  2  3  4  5
   0 14  6  2  1  0  0          0 13  6  2  0  0  0
   1  2  5  2  0  0  1          1  3  6  2  1  0  0
   2  0  3  3  2  0  1          2  0  1  2  3  0  2
   3  0  0  0  1  0  1          3  0  1  1  1  0  1
   4  0  0  0  1 14  0          4  0  0  0  0 14  0
   5  0  0  0  0  0  0          5  0  0  0  0  0  0
```

Random forest

```
         true                         true
pred  0  1  2  3  4  5       pred  0  1  2  3  4  5
   0 14  5  2  1  0  0          0 15  5  2  1  0  0
   1  2  4  2  0  0  1          1  1  4  2  0  0  0
   2  0  5  3  2  0  1          2  0  5  3  2  0  2
   3  0  0  0  1  0  1          3  0  0  0  1  0  1
   4  0  0  0  1 14  0          4  0  0  0  1 14  0
   5  0  0  0  0  0  0          5  0  0  0  0  0  0
```

Conclusion

- Random forest and SVM produce similar results (RF does slightly better on class 1 and 2).
- Results from the full data and the filtered data are similar too.

- No algorithm was able to properly tease apart medium-fast growing organism

- This might be due to:
  - the imbalance of class examples
  - mislabeling of the data (due to unknown nutrient requirements)

- Slow growing organisms (10-14 days) have a marked signal that differentiate them from faster organisms.