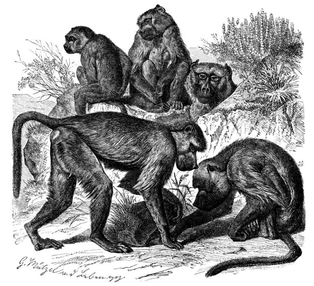
Applying Machine Learning to Predict and Explain Primate Consortship

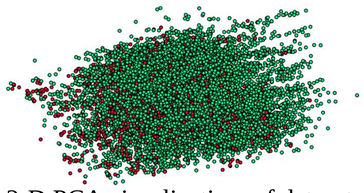
Josh King, Vayu Kishore, Filippo Ranalli {jking9,vayu,franalli}@stanford.edu



Predicting

We investigate the reasons for success and failure of mating between wild yellow baboon pairs. Our analysis applies classification methods to examine whether successful consortships can be predicted, and whether certain behavioral or genetic features are especially relevant in determining consortship.

Data



2-D PCA visualization of dataset

Our input dataset, from Tung et al (2012) is a set of genetic and behavioral features between potentially mating pairs, and a label indicating whether consortship occurred.

- 12,000 observations, 115 females, 121 males
- 1648 consorts, 10493 non-consorts

Additional preprocessing applied to standardize features and remove points that had conflicting labels. We also tried PCA whitening and data augmentation but found that they did not improve results.

Features

- **Genetic/Biological**: two measure of genetic diversity of each of the pair, estimated genetic distance of the pair, age, conceptiveness of female.
- **Social/Behavioral**: male rank, males/females present during consortship.
- **Transformations**: non-linear transformations of genetic distance rank, and age, as well as indices derived from the combination of genetic diversity of the pair.
- **Graph-Based**: PageRank, HITS.

Models

- Gaussian SVM

$$f(w,b) = [\frac{1}{n}\sum_i max(0, 1 - y_i(w \cdot \phi(x_i) + b))] + \lambda ||w||^2$$

- AdaBoosting with Decision Stumps

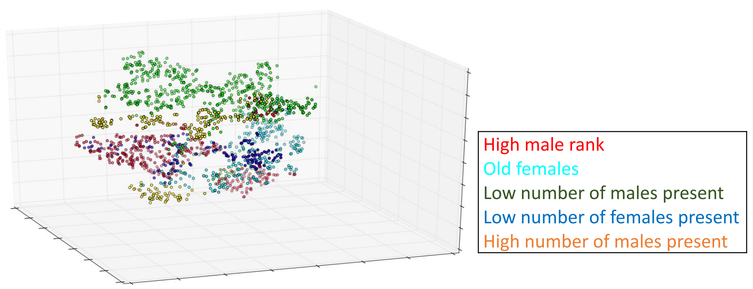
$$E_t = \sum E[F_{t-1}(x_i) + \alpha_t h(x_i)]$$

$$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{1+\epsilon_t}) \qquad \epsilon_m = \frac{\sum_{y_i \neq k_m(x_i)} w_i^m}{\sum_i w_i^m}$$

- Random Forest
- Edge Prediction

For all models, Classes were weighted to address imbalanced dataset.

Discussion



- Could predict consortship to some extent, but high false positive rate, even after trying to account for data imbalance.
- Both social and genetic factors contribute to models.
- AdaBoosting: female hybrid score, male genetic diversity, female age, male rank, males present.
- Gaussian SVM: male/female hybrid scores, male/female genetic diversity, female age, male/female rank, males present, females present.
- Difficult to cluster non-consorting pairs.
- No gains from using researcher-added transformed features from initial dataset.
- Graphical features helped marginally.
- Relationships in social mammals are messy, even machine learning can only do so much.

Results

AdaBoosting Feature Selection

Features	Error	Precision	Recall	Fischer	AUC
1	0.410	0.177	0.565	0.269	0.578
4	0.394	0.183	0.563	0.277	0.587
6	0.382	0.210	0.660	0.318	0.637
7	0.375	0.211	0.652	0.318	0.636
9	0.368	0.214	0.656	0.323	0.641

Gaussian SVM Feature Selection

Features	Error	Precision	Recall	Fischer	AUC
2	0.458	0.161	0.561	0.249	0.552
3	0.483	0.169	0.641	0.267	0.575
4	0.406	0.190	0.627	0.291	0.605
5	0.396	0.202	0.660	0.309	0.627
6	0.362	0.228	0.689	0.343	0.663
7	0.352	0.231	0.680	0.344	0.664
8	0.339	0.237	0.661	0.349	0.665
9	0.324	0.242	0.664	0.355	0.671

Cross-Validation Training and Test Metrics

	Train Error	Test Error	Train F1	Test F1
Gaussian SVM	0.242	0.362	0.491	0.357
AdaBoosting	0.313	0.355	0.381	0.3312
Random Forest	0.019	0.145	0.925	0.229
Edge Prediction	0.131	0.167	0.31	0.236

Normalized Confusion Matrix for AdaBoosting

	Predicted Consort	Predicted Non-Consort
True Consort	0.647	0.3525
True Non-Consort	0.363	0.637

Normalized Confusion Matrix for SVM

	Predicted Consort	Predicted Non-Consort
True Consort	0.686	0.314
True Non-Consort	0.31	0.691

Future Work

- Model groups separately.
- Continue to investigate graphical methods and other ways to augment features for existing dataset.
- Find or generate a larger dataset with additional features.

References

Jenny Tung, Marie J. E. Charpentier, Sayan Mukherjee, Jeanne Altmann, and Susan C. Alberts (2012) Genetic Effects on Mating Success and Partner Choice in a Social Mammal. The American Naturalist, 2012 180:1, 113--129.