A Comparison of Supervised & Unsupervised Learning for Forest Cover Type Prediction

Pedro Sanchez

Problem & Motivation

- Predicting forest cover type
- Currently the data is labeled for classification into 1-7 cover types
- Can we get rid of labels and expect the same performance?
- Getting labels is costly
    - Require human observation
    - Data is otherwise gathered autonomously
- Can unsupervised learning give a comparable accuracy such that human interaction is not necessary?

Dataset

- # instances = 581,012
  - Training: 15,120
  - Testing: 565,892
- # features = 54 + 1
  - Real: elevation, slope, distance to water, etc.
  - Integer: 0-255 index of Hillshade at {9AM, noon, 3PM} on summer solstice
  - Binary (independent columns): wilderness area, soil type present, etc.
  - Label: 1-7 represents Forest Cover Type designation
- Data gathered “from cartographic variables only (no remotely sensed data)”
  - Labels were provided from RIS data provided by US Forest Service
  - Independent data from US Geological Survey

Existing Techniques & Related Work

- Blackard et. al use an LDA classifier and an artificial neural network
    - Reported LDA accuracy: 58.38% (CV)
    - Reported ANN accuracy: 70.58%
- Crain and Davis use a multi-class SVM, also with PCA
    - SVM accuracy: 78.64 (with 10-fold cross validation)
    - When applying PCA, results only minimally worse
- Crain and Davis also implement k-means clustering
    - Diminishing returns once k > ~30

My Technique

- Using Sci-Kit Learn library for LDA, QDA, K-Means
    - For K-Means, kept lowest distance from members to centroids across 10 runs
- LDA
    - No cross validation
- QDA
    - Perhaps a non-linear classifier would be a better assumption
- K-Means
    - First attempted to set k = 7 and find cluster -> label mapping
        - Poor performance
        - Optimal K is probably not equal to the number of labels
        - Multiple clusters -> one label
    - Then studied the relationship of sum of euclidean distances to as k increases
        - Looking for elbow point

Experimental Results

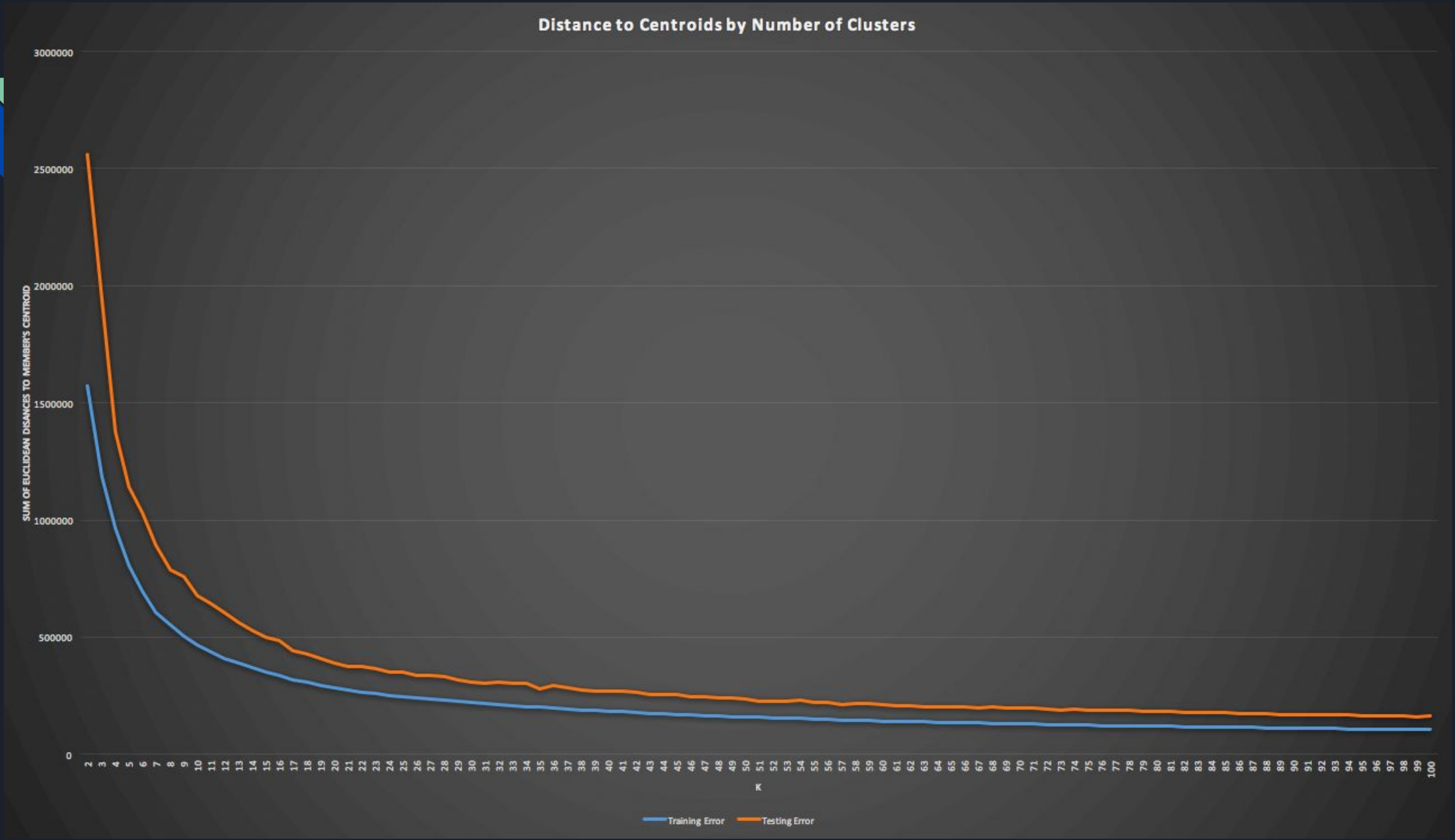
- LDA
  - Training accuracy: 65.0198%
  - Testing accuracy: 58.1252%
    - Decent accuracy, especially when considering that cross validation only added ~.26% more accuracy at the cost of a higher run-time in Crain's work
- QDA
  - Training accuracy: 43.1018%
  - Testing accuracy: 8.0458%
    - Yikes.. Overfitting

Experimental Results

- Clustering
    - Did not calculate error, but sum of euc. distances
        - Used loss function:

    SUM(euc(Xi - mean[c(i)]))

    - From Crain's work, k>60 yielded **BETTER** results than LDA at roughly 60% accuracy.
    - However, LDA model runs much quicker
    - Also hard to establish cluster->label mapping


Distance to Centroids by Number of Clusters
SUM OF EUCLIDEAN DISANCES TO MEMBER'S CENTROID
3000000
2500000
2000000
1500000
1000000
500000
0
K
Training Error
Testing Error

Conclusion

- QDA sucks for this dataset
- LDA can provide quick predictions with labeled data
  - CV can improve this *slightly* at the cost of runtime; still very fast
- K-Means needs less human interaction and can provide acceptable performance at the cost of runtime
  - Runs much faster than ANN
  - Still ~15% worse than ANN accuracy
- Neither perform at the level of NN
- At the very least, human interaction reduced from N to K.

So, at the cost of only some runtime, we can greatly reduce the need for human interaction in this problem and make it completely autonomous!

Future Work

- PCA with LDA
- PCA with K-Means Clustering
  - In Crain's work, PCA was shown to be highly effective at reducing dimensionality while only losing minimal variance.
    - The first 3 principle components retained ~95% of the information
- Determine best method for mapping clusters->labels
- NN

References

1. Blackard, Jock A., Dean, Denis J. “Comparative accuracies of artifical neural networks and discriminant analysis in predicting forest cover types from cartogrpahic variables”. Computers and Electronics in Agriculture (1999).
2. Crain, Kevin, and Graham Davis. “Classifying Forest Type Using Cartographic Features.” Stanford, Stanford, Dec. 2014, cs229.stanford.edu/proj2014/Kevin%20Crain,%20Graham%20Davis,%20Classifying%20Forest%20Cover%20Type%20using%20Cartographic%20Features.pdf.