Feature Reduction and Visualization

John Cooper, Brandon Palomino
MATH 250, Section 01, Dr. Chen
05/23/2022

TABLE OF CONTENTS

I DATA BACKGROUND AND PROCESSING

Trees come in a variety of different types and sizes. They provide many benefits on improving health and the environment. Over time, questions arise on how to differentiate between different tree types. What type of trees grow in an area based on surrounding characteristics? Which tree types can grow in more diverse environments? These questions lead to solving problems in regards to separating different tree types.

The data set we used comes from UVI Machine Learning Repository and is called Forest Cover Type Data Set. The data set consists of 581,021 observations and has 55 distinct features. The data set contains tree observations from four areas of the Roosevelt National Forest in Colorado. All observations are cartographic variables, meaning no remote sensing. Each tree is portrayed by 30 meter by 30 meter sections. The data set includes information on tree type, shadow coverage, distance to nearby landmarks, soil type, topography, and more.

For the purpose of this analysis on the tree coverage data set, MATLAB will be used to process the data set and visualize them under several dimensionality reduction techniques. The toolbox Stats is used throughout each model.

Before visualizing the data, prepossessing and parameter values must be set in order for the algorithms to be used. Given the vast amount of features, features were grouped as variable triplet that were visualized simultaneously, grouped by Cover Type. In addition, features were also grouped as variable couples that were again visualized simultaneously under a Cover Type. In the end, there were seven cover types that can be classified among all the trees.

Looking at Figure 1 (1), each tree was plotted on a graph in relation to Elevation, Slope, and Aspect. Across all perspective of the data, there is a form of separation among all cover types with all the data points being distributed in a similar curve. However, in Figure 2 (2) and in Figure 3 (3), Hill shades and Distances shows a scattered distribution among all tree data, meaning that it is difficult to classify a tree with certain features in mind.

The lack of separation among features is likely due most trees having the same measurements in regards to those categorical variables. This can be shown using Figure 4 (4) where a parallel plot is used to visualize the mean distribution among all features. While Elevation shows a good amount of separation, a majority of the other features are piled together to denote a lack of separation. In addition, high dimensional data such as the tree cover data set can be visualized as An Andrews plot where value x is a high-dimensional data point that under non-integer values.

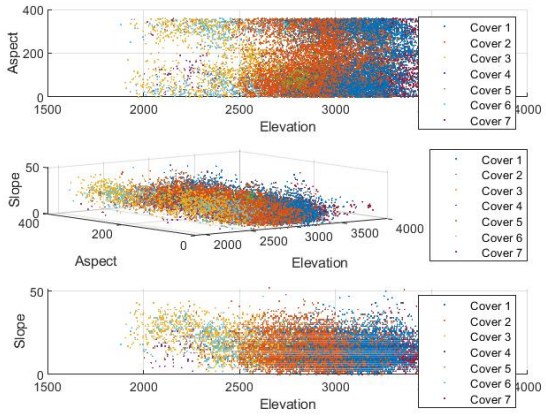

Figure 1: Elevation, Slope, and Aspect

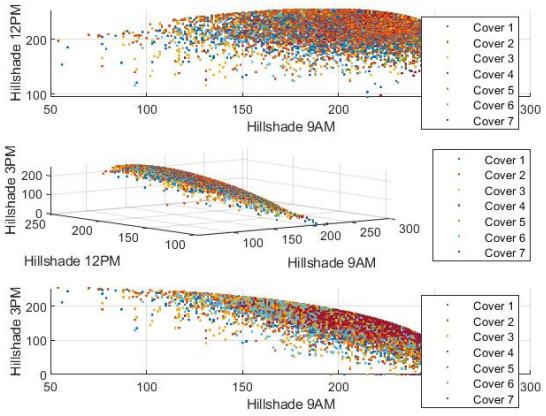

Figure 2: Hillshades

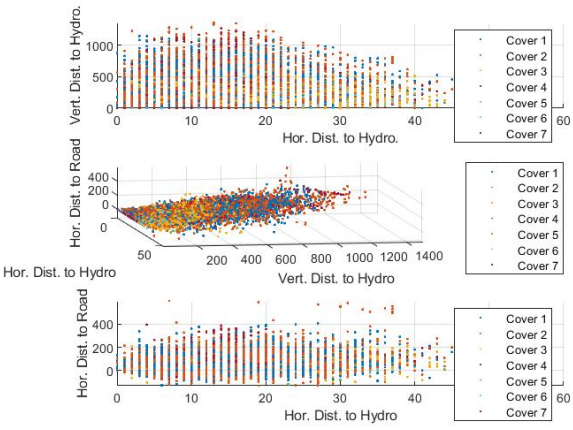

Figure 3: Distances to Hydrology

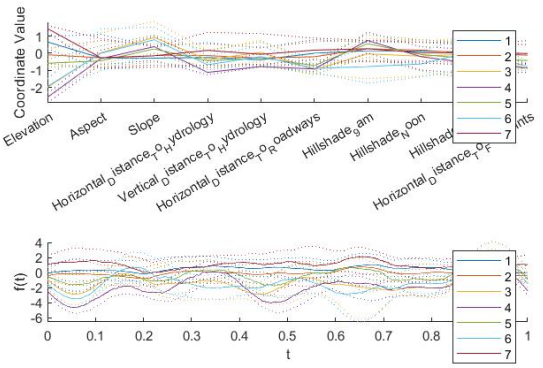

Figure 4: Andrew's Plot/Parallel Plots

II PCA

The point of PCA is to linearly project, with as much variance as possible, our 10D data onto a 3D and 2D coordinate system in which the axes represent the maximum-variance directions of the data. Looking at Figure 5 (5) two separate groupings are created along opposing directions in the 2D projection space, identified by the "wings" of the projected data. This indicates that by using a general maximum-variance projection, we are able to discriminate not between classes, but between two groups which differ in the size of their numerical characteristics. This indicates that there is another variable unaccounted for here which could possibly discriminate between these classes. These plots also indicate a possible non-linearity in the data, with observations potentially arising from a more parabolic structure. To quantify the poorness of fit here, the 2D maximum variance directions only captured 46% of the total variance in the data, while the 3D maximum variance directions captured an additional 17% for a total of 63% of the total variance.

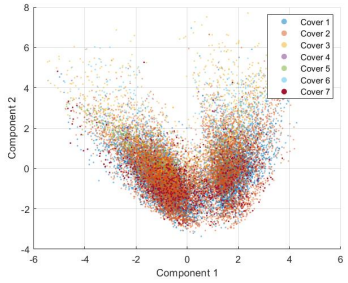

Figure 5: PCA 3D Projection

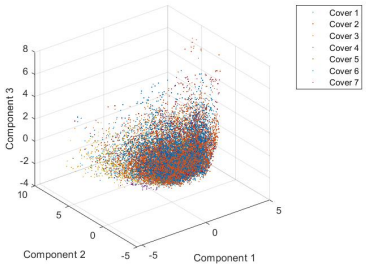

Figure 6: PCA 2D Projection

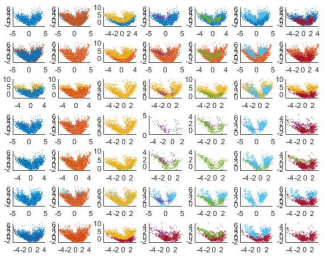

Figure 7: PCA Pairwise Plots

III LDA

By discriminating between in-class and out-of-class sources of variation in the data, LDA performed much better than PCA if we use class separation as the sole criterion. The pairwise plots in Figure 10 (10) are most representative of the fact that classes can become linearly separable when additional sources of variation are accounted for. In fact most classes, when compared in a pairwise, can be identified against one another. This isn't the most efficient way to classify cover types, but it introduces a clear path for understanding what it would take to classify this data on numerical features alone. Moreover, we see relatively clear discriminating lines in the 2D and 3D projections. This is interesting, since this idea seems to compete with the apparent non-linearity in PCA, but LDA, in fact, seems to be the best best for separating classes. Also, due to the within-scatter matrix being non-singular, PCA was not run prior to LDA.

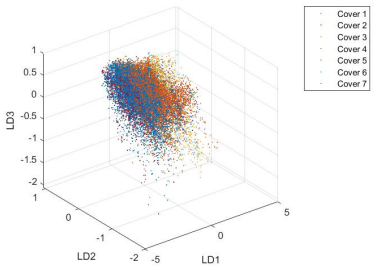

Figure 8: LDA 3D Projection

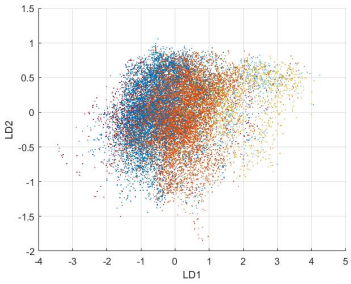

Figure 9: LDA 2D Projection

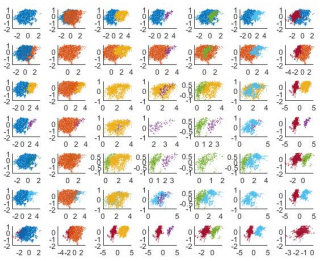

Figure 10: LDA Pairwise Projection

IV MDS

The results of applying MDS with a cosine dissimilarity to the data are poor, as indicated by a Kruskal Stress of about 0.21 for both the two-dimensional and three-dimensional reductions. That is, approximating and projecting angles from a high-dimensional space into a two-dimensional and three-dimensional space is not a good fit for the data - this suggests, perhaps, that the structure of the data is simpler, and perhaps linear, like

in the LDA figures. Despite this, the MDS projections are interesting in that they seem to agree with the PCA projections. Classes are not able to be distinguished, but groups of classes possessing similar numerical features are. This seems to suggest that when general distances are preserved from the original space, we obtain poor separation between classes, but good separation between groups possessing similar numerical qualities. When we begin to account for in-class and out-of-class differences, we seem to obtain good separation between classes with no serious distinctions between general groups of data.

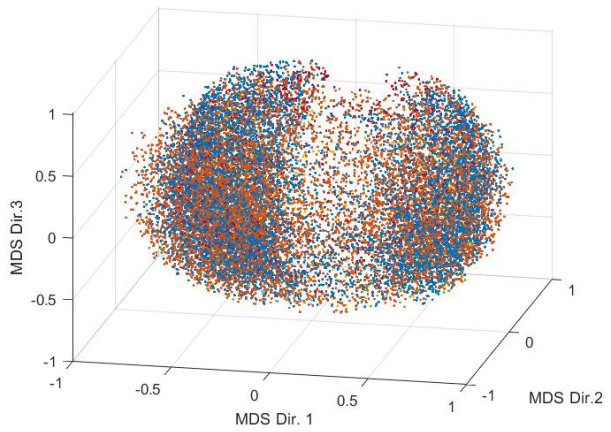

Figure 11: MDS in 3D

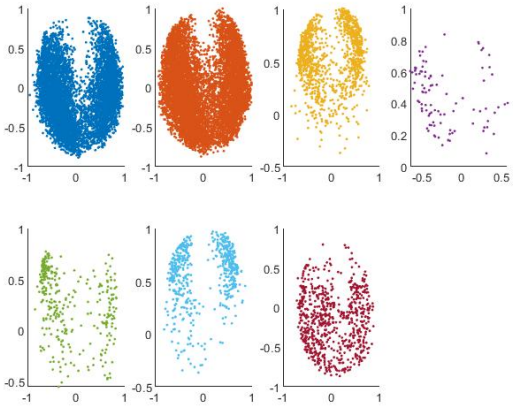

Figure 12: MDS in 2D by class

V ISOMap

The ISOMap projections are similarly confusing. Assuming this data is sampled from a manifold, the 3D ISOMap projections in Figure 13 below tell us that it's not the classes that are necessarily grouped together, but that two groups, each containing all classes, are closely grouped together and are seemingly distinguished by something other than their numerical characteristics. The apparent difference between ISOMap and MDS is the use of weighted neighborhood graphs which are able to measure in-class and out-of-class dissimilarity similar to LDA. This suggests one of two things: (1) either our assumption that this data arises from some complicated structure is false, or (2) the apparent pairwise linear separability

arising from LDA is false. The more obvious of these two is the first - if the data is pairwise linearly separable by LDA, it would seem that the assumption of a complicated data structure that segregates the data by general groupings instead of classes is wrong.

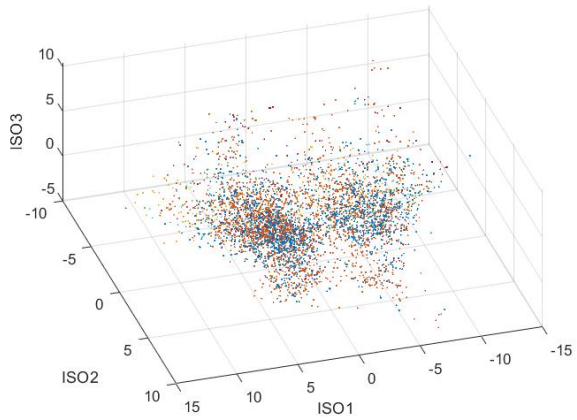

Figure 13: ISOMap 3D

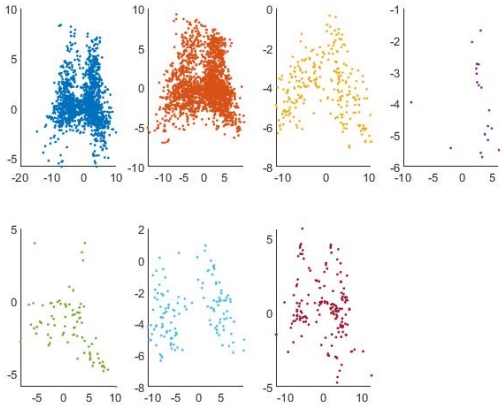

Figure 14: ISOMap 2D

VI Conclusions

In summary, all dimension reduction techniques showed a form of separation between cover types to some extent. All cover types overlapped with one another majority with one another majority of the time throughout each algorithm. By evaluating the results, we can see that each cover type separated in a similar shape or curve when compared to one another. In conclusion, all tree types can be considered similar in features. However, they can all be classified differently due to other factors as well.

If we were to continue this project under a longer time frame, improvements could be made to the project. For example, if other categorical variables were to be incorporated into the algorithms, separation results would drastically improve as it will be easier to discern different cover types on the plot graphs. Regardless, the features used in the project provided insight on how certain variables contribute to the overall separation of data. A deeper investigation into ISOMap and LDA projections are warranted, since they seem to suggest different things. However, if one

were attempting to crudely classify cover types, LDA is the only method which seems to provide a reasonable way to do this. It should also be noted that LDA is the only supervised method among these. It's possible that an unsupervised method of separation is too omnibus for our specific data, which seems nuanced in its potential to distinguish cover types through its numerical features.

VII REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/covertype

[2] https://www.kaggle.com/uciml/forest-cover-type-dataset

[3] https://www.mathworks.com/help/stats/

[4] utkarshtrivedi(2022).Isomap(D,n_fcn,n_size,options)(https://www.mathworks.com/matlabcentral/fileexchange/62449-isomap-d-n_fcn-n_size-options),MATLABCentralFileExchange.RetrievedMay23, 2022.

[5] G. Chen, Introduction to Matrix-Based Data Science: Mathematics, Computing and Data, vol. 1. Springer Nature, 2022.