Applying Unsupervised Learning to Constitution Based Data For Geographical and Temporal Grouping

Tyler Hartwig [1]

Abstract

Machine Learning is applied to a database of various variables applied to various constitutions from the various countries. In particular, unsupervised learning used to naturally group these constitutions by geographical region, and by the time frame in which they are written. Due to a large number of variables, several tactics were employed to reduce dimensionality and make the learning problem feasible.

1. Introduction

Dr. Curt Nichols of Baylor University posses a database of various (mostly binary) variables on constitutions from nations around the world. Dr. Nichols proposed the idea of grouping the countries based on a subset of these variables, with an end goal of naturally grouping the data according to geographical region and the time period in which the constitution was written. Realistically when constitutions are written, they are based on other constitutions. With this in mind, it is likely a particular country will base their constitution on the constitutions from the countries surrounding them, or recently created constitutions. This is the main motivation for looking for geographical and temporal groupings in this paper.

The database itself has over 1000 variables for each country (including geographical and temporal information) making this problem practically infeasible if one desires to consider all present variables. Many tactics are used to combat this, such as: reducing the number of variables used, principle component analysis, and uniquely defined difference functions which define a smaller set of dimensions. Additionally, no geographical or temporal variables are used, as this would remove the statistical implications of the countries naturally being grouped by region or time period.

This problem very naturally lends itself to hierarchical clustering, and is the subset of unsupervised learning which will be used in the remainder of the paper. This kind of grouping is natural, as it can easily depict the closeness of one country to another, both in time or space.

2. Related Research

Greg Ver Steeg and Aram Galstyan present a method for finding the most informative hierarchical clustering for high dimensional data (**?**). The paper illustrates how the variables can be estimated using multivariate probability theory in order to obtain a similarity measurement. This leads way to a neural network which can then maximize this similarity. Additionally a computationally feasibly method is presented for using this similarity definition for hierarchical clustering.

3. Experiment Setup

One of the main focuses of this experiment is to reduce the dimensionality of the problem into a useful size. This is done in a few ways to start. According to Dr. Nichols, our domain expert, the variables of particular interest are the Executive, Legislative, and Judiciary branch variables, along with those pertaining to Criminal Procedures and Rights. Many variables are removed as well, particularly those without numeric values. This initially cuts the data to be about 700 variables. This is clearly still an incredibly large number of dimensions.

The data is initially clustered off of this data in euclidean distance, which unsurprisingly does not yield acceptable results. This is not surprising, as the data is generally binary, however these variables often take on large values to indicate the absence of the variable (rather than true or false). It is obvious that a better distance metric is necessary to this problem. One of the first metrics used is a Hamming distance between each data point. This is a much more natural evaluation of binary variables. How this works specifically is each variable for each country is compared against each other country, when the variables are the same, their "similarity" is increased. This allows the data to again be clustered, yielding more favorable results

[*] Equal contribution [1] Baylor University, Waco, Texas. **AUTHORERR: Missing \icmlcorrespondingauthor.**

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP. Copyright 2017 by the author(s).

geographically.

One other method is used (currently) which is essentially redefine the variables used in this experiment. Dr. Nichols also has interest in how similar each constitution is to that of several “base” constitutions. Particularly, The United States, British and Austrian constitutions serve as a standard to compare against. A new coordinate system can then be defined, in which each axis is a “base” constitution, and each value is the similarity of a given constitution to each axis. This measurement does not give outstanding results either, and the domain expert will not be contacted to further this project.