Caleb Pudvar

Jockers/Witten Questions

1. According to the authors, the two most important factors from a machine learning perspective involve the choice of features and the selection of an appropriate and effective classification technique.
2. The most consistently reliable choices in terms of authorship attribution are the analysis of high frequency words and n-grams.
3. Studies focused on authorship attribution have determined that the Bayesian networks are generally the most effective classifiers.
4. Benchmarking: An established point of reference against which computers or programs can be measured in tests comparing their performance or reliability.
5. Tuning Parameters: The different features present in the vector, which are used for analysis.
6. Closed Set: A set that includes all the data relevant to that set. All data for a given situation is within the set.
7. The three objectives of the study are:
   1. To expose the authorship attribution community to classification methods that have not been previously applied to authorship problems
   2. To compare the relative performance of these methods
   3. To apply these methods to a classic authorship attribution problem

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Delta | K-Nearest Neighbors | Support Vector Machine | Nearest Shrunken Centroids | Regularized Discriminant Analysis |
| Tuning Parameters Used | Number of features used in classification | Number of nearest neighbors in classification | Determine the cost of violating vector constraints | Number of features used | Control number of features used |

1. The raw set contains every feature that appears at least once in texts by each of the authors. The difference with the pre-processed features set is that it only includes words or bi-grams that have a relative frequency of .05%. This narrows down the analysis matrix from 85x2907 (raw set) down to a more manageable 85x298 matrix for the pre-processed set.
2. The matrix dimensions correspond to the frequency of words or bi-grams for each paper in the set of Federalist papers. For instance, 85x298 means that for the 85 papers, there are 298 features to check.
3. Cross-validation: a sample of the population data is checked for validity by applying the method to a different sample from the same population.
4. The training data consists of the samples authored by Hamilton, Madison, and Jay. The test data consists of the coauthored and disputed texts.
5. Among the fifteen papers in the test set:
   1. There are 3 papers in the coauthored group. They are papers #18, 19, and 20.
   2. There are twelve papers in the disputed group. They are papers #49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 62 and 63.
6. According to the results, most of the papers are likely written by Madison, and only a couple may be from Hamilton.
7. While NSC is confident that Madison authored the papers, NSC is sensitive to the parameters used to fit the model (feature set and training sample choice). Instead of using this data as verification, it should be looked at in terms of probability that Madison was the author of the disputed texts.