Project Research Paper - Tapping into the Data

Aidan Sawyer

October 28, 2016

A Pint-by-Pint Analysis of Craft Beer Styles by Characteristics and Region

Precedents

The craft beer world is no stranger to the important prospects that digital technologies offer to the discovery and discussion of the sacred drink.

A plethora of tools exist extolling its praises and aiding its description and creation, echoing the simultaneous and vigorous rise of web and mobile technologies and craft/home brewing.

From apps like RateBeer and untappd, to professional reports on the health and growth of the craft and homebrewing beer industry by groups such as the **beer association** to the more niche studies on the specific components of beer such as the *hop*
studies of the BarthhaasGroup, and the *barley* studies of the **American Malting Barley Association**.

From the online databases of **BreweryDB**, **CraftCans**, **OpenBeerDB**, and the open-sourced applications for homebrewing created by friends (**OpenBrews**), the podcasts created by strangers (**The Beer Genome Project**), and even technical white papers [Predictive analysis of beer quality],

Challenges

One of the best things about the explosion of craft beer in the America's has been the intense diversity in what is being offered. From IPL (Imperial Pale Lagers) to Black IPAs and other mashups, it's difficult to bracket beers to a singular style, and new beers are being created and expanded every month.

Such inventive and creative recipes tend to blur the lines of what delineates one style from another, and thus complicates my task of predicting and classifying a beer based on a limited set of characteristics.

As these differentiations become more distinct and the quantity of craft beers increases, it helps to have a growing list of various measures to inter-compare and create features from, but unfortunately, single databases containing bulk information about the same beers can be difficult to fine.

This requires that someone attempting to create a project out of a limted dataset must either combine data from multiple sites, or select a list of only 2-3 attributes to a large number of beers, which is what I've chosen.

An additional challenge is to decide how one wishes to delineate between styles given the hierarchy. How to conclude the "correct" set of standards with which to compare one's values and the degree of specificity with which one will consider a belgian whitte, for instance, a Pale Ale, can have significant impact on the final results and evaluation between algorithms.

Ethical Considerations

The possible ethical considerations of this project involve the prospect that the work of the project could assist in identifying the characterstics critical to classifying beer styles.

If this were to be entirely successful, the potential benefit to all current (and potential) beer lovers is high enough to warrant the project as moral necessity to the author.

Business Cases

This project has the capacity to assist with and verify the standards of the the style guides and official standards of the homebrew and craft beer industries, validating or contradicting the metrics established by the judges.

Since there are no ratings or sales data associated with the clusters in the data I've selected, the direct business considerations are abstract, but if the conclusions of this project are compared against other projects, there is a potential for the project to be monetarily useful.

If it could be identified, for instance, that a particular style of beer happened to be more popular and/or cheaper to mass produce, the economic effects in a booming economy may be severe.

Moreover, if the published style guides were to be proved empirically falsifiable according to the accepted classifications of the breweries themselves, popular organizations may lose the credibility and funding they need to survive.

Data and Data Cleaning

The data, coming from the afore-mentioned craftcans database, comes exclusively from those beers which have been canned, and have a slight bias in the annoyingly hipster direction.

Apart from this unfortunate defect, there are additional considerations that must be taken into account. Namely: - *Location* is currently set to 'City, State' and needs to be converted to State, since there aren't enough from each city to gather meaningful data.

- _Empty Fields_: there are multiple fields in which there is no listed ABV or IBU. these fields must be removed to ensure a clean and consistent comparator.

- _Create Subsets_: the most difficult data cleaning issue lies in how best to determine what constitutes an honest-to-charlie-mopps 'beer style' and where to draw the line of difference between styles and sub-styles. As this is difficult to determine programmatically, a sufficient amount of doman research is required.

- _Remove ID_: a classic and persistent problem in data mining is to remove any ID fields in order to prevent over-fitting. This dataset is no different.

- _feature creation: strength_: it may very well be the case that ABV or IBU alone are insufficient measures for the style determination. One possible combined feature consists of a relation between them, which compares them as a sort of product.

Algorithms

As there are no direct correlates to this project, I will have to construct my own general methodology, though I can find some guiding info from the white papers.

I intend to

Distance Metrics

Since I want to maintain consistency between algorithmic results, the same metric of a '*centroid*' will be used to campare the data records.

This ensures that all relevant points used in calculation will be actual values from the database.

Strategy

As stated in the proposal, the stated intention of the project is to compare the clustering accuracy of

Validation

Given the size of the database I've found (~2700 entries), I should be able to set aside a small percentage of the data to use as validation data.

Beyond this, I intend to use N-Fold Validation to compare the accuracy of my clustering algorithms, and ROC Curves to compare the classification algorthims.

Algorithms

I generally intend to differentiate between impartial/random and seed-based strategies when breaking up the data. This will allow me to acheieve an objective measure of their efficency against other methods (via domain knowledge and research), as well as permit me to compare the methods against each other.

Clustering

- K-means
  - random, impartial
  - seed points of a few items
- shared property: center-based values
  - ABV = x +- y
  - IBU = x +- y
- otsu's method: see which clusters break down and are diferentiated first

Classification

- decision trees
- zero attribute
- one attribute
- naive bayes
- k-nearest neighbors

References

Dong, Jian-Jun, Qing-Liang Li, Hua Yin, Cheng Zhong, Jun-Guang Hao, Pan-Fei Yang, Yu-Hong Tian, and Shi-Ru Jia. "Predictive Analysis of Beer Quality by Correlating Sensory Evaluation with Higher Alcohol and Ester Production Using Multivariate Statistics Methods." Food Chemistry 161 (2014): 376-82. Web.

. . . (the references referred to by their URL alone)