Naive Bayes Classifier for Detecting Malformations in Hand Annotated Training/Evaluation Mass Spectrometry Datasets.

Mathew Gutierrez  
University of Montana

Mathewgutierrez00@gmail.com

Abstract

Evaluation of existing mass spectrometry workflows requires a known correct dataset with which to perform quantitative comparison. The fidelity of the evaluation hinges on the fidelity of the hand annotation. To that end, this work presents a Naive Bayes Classifier to validate and detect common malformations of hand annotated features in hand annotated mass spectrometry datasets.

Introduction

Mass spectrometry is an immensely impactful tool in the realm of chemical analysis, and only increasing in importance. It has been the goal of my research to create new software to improve the capacity for researchers to analyze mass spectrometry data. One of the major facets of my work has been to create a workflow to qualitatively evaluate existing solutions for trace detection for mass spectrometry. A trace is a signal in a mass spectrometry data set that represents a specific isotope of a molecule. In order to perform a qualitative evaluation of existing trace finding algorithms there must be a known correct dataset with which to compare to. Obtaining a known correct dataset is non-trivial and required many hours of work on my and others part hand annotating a mass spectrometry dataset1. However, deep analysis of the hand-annotated dataset shows that a small subset (<1%) are annotated incorrectly. With the goal of correcting the errors in the hand annotated data, I trained a naive classifier on a subset of the hand annotated data, and a dataset populated with artificially created incorrectly hand annotated features.

Methods

Incorrect training data employed in machine learning methods has long been an issue that plagues researchers. The fidelity of the algorithm in question, (or in the case of my work, a quantitative evaluation) can only be as good as the dataset being used for learning or evaluation. The concept of using a classifier to validate a known-truth dataset is not a new one2 but will be newly applied to this problem.

A Naive Bayes Classifier was chosen for this problem due to the ability to assume independence for the MSE values in each trace Summary.  While the MSE values for each trace are not actually completely independent, it’s been shown that Naive Bayes Classifiers are robust to some level of dependence3.

The first step in the creation of a model to evaluate and validate hand annotated mass spectrometry datasets is to acquire training data. For this purpose, I created a summary of all traces in the hand annotated dataset by calculating the mean standard error for the m/z RT and intensity of points in each trace (knowing that some subset of the traces are incorrect).  Additionally, modifications are made to the traces to simulate a specific mistake made during hand-annotation (failure to request a new trace ID between grouping traces together), and they are written to the summary file with a 0 to indicate they are a false trace. The original trace is written with a 1 to indicate it is a real trace. A distribution of MSE for real and false traces can be seen in figure 1.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | A screenshot of a cell phone  Description automatically generated | | A screenshot of a social media post  Description automatically generated | |
| Figure 1: Note the severely limited distribution of MSE for real traces, and the highly disparate distribution for fake traces. |

When the summary file is created, it is loaded into the Naive Bayes Model, and train for each set of real (1) and fake (0) traces using a normal distribution as indicated by the distribution of MSEs seen in Figure 1.

Once the model is trained and tested, I ran the entire dataset and recovered the false negatives. The false negatives are important because they represent traces that were labeled as true, but the classifier classified as false. The classification as a false trace casts doubt on the quality of the trace, and its id is recorded for validation. As a last step, the probability of being classified as a real trace is output into a csv for all traces. That .csv was then investigated to determine what more could be done to validate the data.

Results

The Naive Bayes Classifier works exceedingly well in recovering true positives (.99) and true negatives (.98) from the test dataset. With the knowledge that the model will almost certainly recover all true traces, I ran the entire dataset and collected the false negatives. The Model returned a list of ~30 trace IDs for inspection. (Due to the random sample of the train/test split, it is non-deterministic.) Every one of the trace’s output by the model were isolated and analyzed using a data viewer (Figure 2). Observe in Figure 2 the failure of a hand annotator to increment the trace ID between inspecting signals, resulting in two traces quite far apart being labeled as one. Without fail, every single trace returned as a false negative by the model was the result of this failure to distinguish two traces as separate.

|  |
| --- |
| A picture containing tennis  Description automatically generated |
| Figure 2: Two traces incorrectly labeled as one (green), likely due to the failure of a hand annotator to increment the traceID after finishing with one. |

Upon investigation of the csv of probabilities output by the classifier, the majority (~46,000) of the true traces have similar, and high probabilities of belonging to the True class (p = ~.048). Cursory analysis of the traces with that probability show high fidelity annotation (Fig 3)

|  |
| --- |
| A close up of text on a white background  Description automatically generated |
| Figure 3: Example of a trace with P = ~ .048 |

Additionally, investigation into the region of traces that have a relatively high probability of being classified as true (P~= .002), but not as high as .048 showed that the subjective quality of the annotation correlates with the probability of being correctly classified (See Figure 4).

|  |
| --- |
| A close up of a logo  Description automatically generated |
| Figure 4: Note this is a correctly clustered trace, but noise points have been included that affect the quality, and as a result the probability of membership in the real trace class is diminished (P~= .002). |

Discussion and Conclusion

The model described functions to detect the failure of a hand annotator to increment the trace id between annotation quite well by determining the likelihood of belonging to a true class or a false class, as seen in Figure 1. Additionally, the likelihood of belonging to the true class is recorded and can be seen to be surprisingly informative in the subjective quality of a trace as seen in Figure 4.There are certainly edge cases however that are not handled by such a simple model such as this. For instance, it is possible that two traces very close together were grouped as one. In this case, their RT and m/z MSE would be small, and not likely to be caught by the classifier, but would still result in a slightly diminished probability of candidacy in the real traces class.

In conclusion, the Model created has allowed the removal of 30 inappropriately annotated traces from the hand annotated dataset. Further, the csv of probabilities has provided us with a list of suspect traces due to their sub-optimal probability of belonging to the true class. Using this model in conjunction with a skilled hand annotator will result in an improvement in the fidelity and quality of evaluations and function as a training data set for this hand annotated dataset.  To expand on this work, Additional classes could be created to accurately capture the qualities of the less than optimal but still correctly annotated traces.

References

#### Henning J, Tostengard A, Smith R. A Peptide-Level Fully Annotated Data Set for Quantitative Evaluation of Precursor-Aware Mass Spectrometry Data Processing Algorithms Journal of Proteome Research 2019 18 (1), 392-398 DOI: 10.1021/acs.jproteome.8b00659

1. Carla E. Brodley , Mark A. Friedl, Identifying mislabeled training data, Journal of Artificial Intelligence Research, v.11 n.1, p.131-167, July 1999
2. Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3.