Determining NDMA Formation during Disinfection using Treatment Parameters

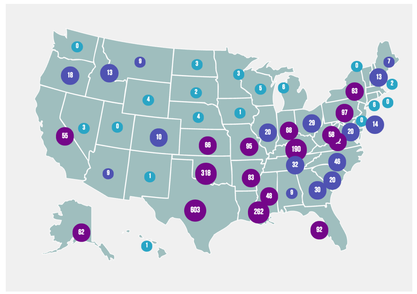
Aleksandra Szczuka aszczuka@stanford.edu

CS229 Final Project Fall 2016

Motivation

In order to get rid of harmful viruses and bacteria, drinking water must be disinfected. However, when water is disinfected, harmful carcinogenic disinfection byproducts (such as *Nitrosodimethylamine*, NDMA) may form. Once formed, reducing NDMA concentrations is costly for drinking water utilities, and preventing NDMA formation is highly desirable. The goal of this project is to predict whether or not NDMA will form during water disinfection given available treatment parameters from water treatment plants in the United States.

Data



Total DBP violations in drinking water plants in 2016 (USEPA).

The complete data set was obtained by merging datasets for water treatment utilities available through the EPA. In total, 10 data sets were compiled for 108,604 utilities located throughout the US. Features such as facility size, type of water disinfected, point of sampling, disinfectant used in the facility Features were chosen based on literature knowledge of what parameters may affect NDMA formation. Data was preprocessed for the ease of use in Matlab, for instance, categorical variables such as plant size were assigned numerical categorical values instead of string values of 'XL' and 'S'. The resulting data set was highly skewed of the 108,604 available data points the breakdown was as follows:

1906 samples where NDMA > detection limit
106698 samples below detection limit (0)

Methodology

Feature	Feature	…	Feature	Label
a_{11}	a_{21}	…	a_{N1}	0
a_{12}	a_{22}	…	a_{N2}	1
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

B →

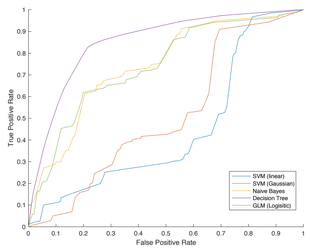
Feature	Feature	…	Feature	Label
0.5 (a_{11} + a_{12})	0.5 (a_{21} + a_{22})	…	0.5 (a_{N1} + a_{N2})	0.5
0.5 (a_{11} + a_{12})	0.5 (a_{21} + a_{22})	…	0.5 (a_{N1} + a_{N2})	0.5
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

A ↓

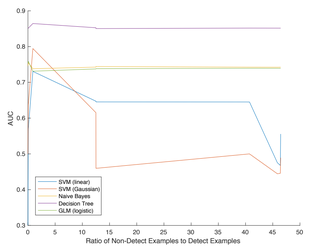
Feature	Feature	…	Feature	Label
a_{12}	a_{22}	…	a_{N2}	1
…	…	…	…	0
a_{1N}	a_{2N}	…	a_{NN}	0

Two different data processing methods were applied in order to reduce data skew and several algorithms were compared for each:
(A) Random Deletion of Non-Detect Samples: Logistic Regression, Naïve Bayes, SVM, Decision Trees (Classification)
(B) Randomized Data Augmentation: Linear Regression, Decision Trees (Regression Algorithms)
70% of the data was used for training and 30% was the test data

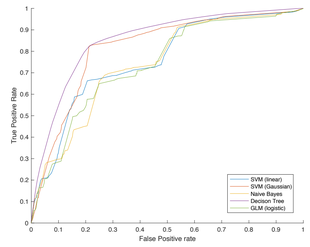
Results and Discussion



ROC curves for unaltered data set



AUC for random data deletion



ROC curves for most optimal Non-detect: Detect data deletion ratio

For the original data set, the highest AUC value for the test set occurred when decision trees were implemented. SVM (Linear and Gaussian kernel) performed the worst.
When non-detect samples are randomly deleted, the AUC's for the training data peak at a 2:1 Non-detect and Detect samples. SVM(Linear and Gaussian kernel) are most affected.

Fitting method	All data	With augmentation
Generalized Linear Model	0.664	0.638
Decision Tree	0.533	0.564

AUC values for the training set using randomized data augmentation that yielded the highest AUC values.

For Randomized Data augmentation, best results were achieved when data was averaged 10 times (2, 10, and 100 tested). However, the AUC's were worse than classification with random deletion of non-detect samples. Therefore, only the classification algorithms were implemented on the test data. For the test set, the SVM (Gaussian kernel) predicted the most of the detect samples, however, it also showed a more false positives than other fitting methods. Decision Trees may be the more appropriate classification to use.

Fitting Method	Train All Data	Train With Random Deletion	Test Correctly Classified	Test Incorrectly Positive
Generalized Linear Model (logistic)	0.739	0.718	67	11608
Naive Bayes	0.742	0.735	52	7608
SVM (linear)	0.556	0.749	74	9744
SVM (gaussian)	0.511	0.802	357	14285
Decision Tree	0.851	0.855	274	830
Random Guess			172	16274

AUC values for training data sets for the classification algorithms used. For the test data, the number of correctly classified detect samples is shown (out of 358), and the incorrectly classified non-detect samples (out of 32582).

How Much will form?

In order to predict what the amount of NDMA formed in a treatment plant would be, two fitting methods were used: decision tree regression and a general linear model. The error associated with those fitting methods is shown below. Decision regression trees outperformed the general linear model, however, analytical method results usually have lower errors.

Fitting method	Error
Generalized Linear Model	34.4%
Decision Regression Tree	11.7%

Mean error for regression models.

Future work

The results achieved in this project are a good starting point for future work. However, the data set that I was able to compile this quarter was very limited. Even though the EPA has been monitoring NDMA formation throughout the united states, they do not provide data for water quality parameters associated with the treatment plants, rather, they provide parameters such as disinfection technique used. While this is important information, it does not paint the full picture for NDMA formation. The main goal would now be to map the water quality parameters available online with the correct water treatment plants by either names used by the EPA or GPS coordinates. Further processing of obtained data to reduce skew would also be appropriate, using techniques such as stratified sampling. Cross validation would also likely improve the results.

References

"Third Unregulated Contaminant Monitoring Rule." Environmental Protection Agency. Web. www.epa.org

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology* 143.1(1982):29-36.