Spring Technical Meeting
Eastern States Section of the Combustion Institute
March 6-9, 2022
Orlando, Florida

Analysis of Inlier and Outlier Compounds with respect to Artificial Neural Network Cetane Number Prediction Accuracy

Travis J. Kessler[1,*], **Amina SubLaban**[2], and **J. Hunter Mack**[2]

[1]Department of Electrical & Computer Engineering, University of Massachusetts Lowell, Lowell, MA
[2]Department of Mechanical Engineering, University of Massachusetts Lowell, Lowell, MA
[*]Corresponding author email: Travis_Kessler@student.uml.edu

Abstract

Artificial neural networks (ANNs) are exceptional at forming non-linear correlations between multivariate input and target variables; however, they are often seen as a "black box" approach, since how ANNs form these correlations is somewhat ambiguous. Furthermore, the process underlying how ANNs learn from inlier and outlier samples within the input dataset is not fully understood. Intuitively, it is expected that training ANNs with inlier samples will increase prediction accuracy and training with outlier samples will reduce prediction accuracy; though, in practice, this is not always true. The present work identifies and analyzes inliers and outliers of existing experimental cetane number (CN) data encompassing a variety of compounds and compound groups. It also investigates how ANNs trained to predict CN perform with and without outliers included in the training data, and whether a relationship exists between inliers/outliers and ANN prediction accuracy across the whole dataset and for individual samples. Additionally, individual outlier compounds are analyzed, highlighting how they structurally differ from inlier compounds.

Keywords: Artificial neural network, cetane number, principal component analysis, QSPR/QSAR

1 Introduction

1.1 Fuel Property Prediction

Efforts to reduce carbon emissions and become less dependent on fossil fuels is a broad topic of study, encompassing clean energy production, carbon capture, and many other avenues leading to a cleaner future. One such avenue is the pursuit of cleaner liquid fuels for use in existing engine architectures. Recent efforts explore the use of biomass-derived fuels or drop-in fuel additives, such as those derived from lignocellulosic biomass [1]. Once a compound is synthesized, it must be determined whether it performs optimally in an engine. Various methods exist to standardize and measure the compound's propensity to form soot, its energy density, and its behavior in an engine; however, these methods are more often than not costly and time-consuming, requiring specialized equipment and a significant volume of the compound/fuel in question [2] [3]. To this end, methods for determining characteristics of a compound before it is synthesized or even tested are needed. One such method is to employ predictive models.

A wide variety of predictive models have been utilized in predicting physical/chemical characteristics of fuels, bio-oils, and other hydrocarbons. The cetane number (CN) of a fuel, a measurement of ignition quality of the fuel in a diesel engine, can successfully be predicted using consensus modeling and artificial neural networks (ANNs) trained with cheminformatic descriptors [4] [5]. Further, ANNs trained with quantitative structure-property relationship (QSPR) descriptors, which are numerical representations of a multitude of chemical and physical characteristics of a given compound, have been shown to accurately predict the CN, yield sooting index, energy density, as well as a variety of cold weather characteristics for a wide range of hydrocarbons [6-8].

1.2 Determination of Inliers/Outliers

Using QSPR descriptors for this regression task presents its own challenges. Experimental fuel property databases are relatively small in size (numbering in the hundreds of compounds), and QSPR descriptor genera-

tion software tools often produce descriptors numbering in the thousands. Subsequently, any regression task is considered underdetermined, i.e., there is either no solution or infinitely many solutions. The inherent "flexibility" of underdetermined systems leads to variation in coefficient matrices linking predictors and observations, and it is expected that similarity/dissimilarity between samples greatly affects this variation.

Numerous outlier detection methods aim to quantify the dissimilarity between samples. The most common algorithms leverage a metric of distance (e.g. euclidean, cosine, Chebyshev, etc.) to measure how close two samples are to each other in N-dimensional space. Any sample considered too "far away" from all other samples is considered an outlier. One such method, local outlier factor (LOF), quantifies samples such that samples with an LOF greater than one are considered outliers, and samples with LOF less than or equal to one are considered inliers [9].

The present work aims to investigate the role of inlier and outlier samples in training ANNs with QSPR descriptors to predict CN. Moreover, the ANN's behavior upon removing outliers from training data is observed, and the relationship between ANN prediction accuracy and LOF is examined.

2 Experimental Procedure

2.1 Data Preprocessing

Experimental CN data was obtained from a variety of sources including the NREL Compendium of Experimental Cetane Number Data, totalling in 408 unique compounds [10-13]. Simple molecular-input line-entry system (SMILES) strings were gathered/generated for all compounds and validated using compound entries on PubChem [15]. 5305 QSPR descriptors were generated for each compound using the aforementioned SMILES strings and the alvaDesc software package [14].

The data was then normalized and principal component analysis (PCA) was used to reduce the dimensionality of the dataset while retaining nearly all information present in the original dataset and to ensure that each predictor is statistically uncorrelated to all other predictors [16]. In addition to providing ANNs with an appropriate number of input variables with appropriate scales, PCA also allows individual component-response variable relationships to be observed. Given 408 samples, 408 principal components were generated (the maximum possible number of components).

2.2 Artificial Neural Network Training

ANNs were trained using the ECNet Python package, an open-source tool tailored to predict fuel properties [17]. To measure the ability of the ANN to successfully predict all samples in the dataset, the leave one out (LOO) training methodology was utilized; each sample is removed from training to act as a "pseudo-test set". Subsequently, 408 separate train/test splits are created. Additionally, ten ANNs are trained for each train/test split to ensure consistency in results. Results are presented as mean values across all ten ANNs.

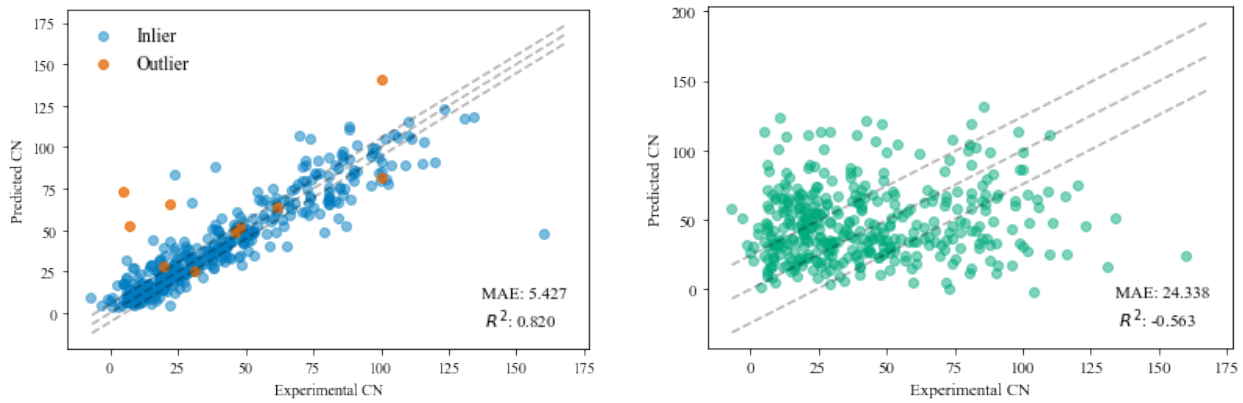
ANN architectures consisted of 408 input neurons (one for each principal component), one hidden layer of 256 neurons, and one output neuron (for the response variable, CN). The rectified linear unit (ReLU) activation function and Adam optimization function were utilized [18], and ANNs were trained for 100 epochs (training iterations).

2.3 Outlier Detection/Removal

Following the generation of principal components, the LOF algorithm was used to determine which samples are outliers. To compare ANN performance with and without outliers, the outliers were removed from the dataset, PCA was performed again, and the aforementioned ANN training procedure was repeated. Median absolute error and the r-squared correlation coefficient were calculated for ANNs with and without outliers.

Further analysis of outlier samples was performed, such as visualization of principal component-response variable relationships, comparison of compound structures, and whether a trend exists between inlier/outlier samples and ANN prediction error.

3 Results and Discussion



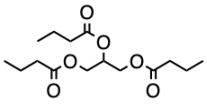
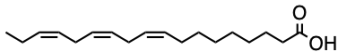
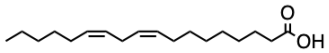
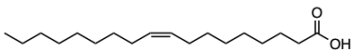
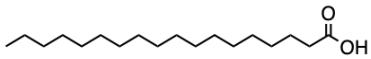
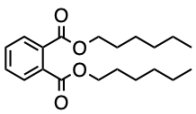
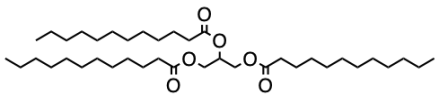
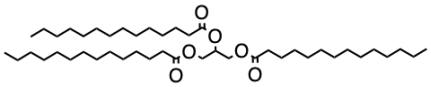
(a) Outliers included during training

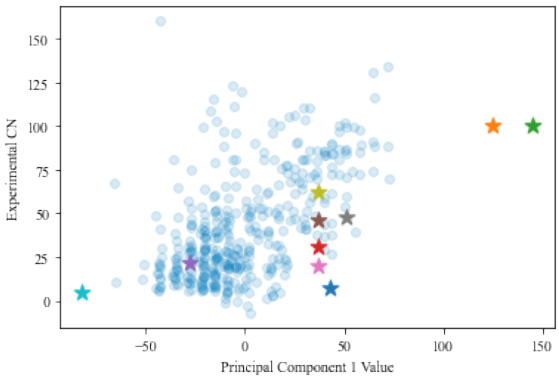
(b) Outliers not included during training

Figure 1: Parity plots, showing predicted CN vs. experimental CN

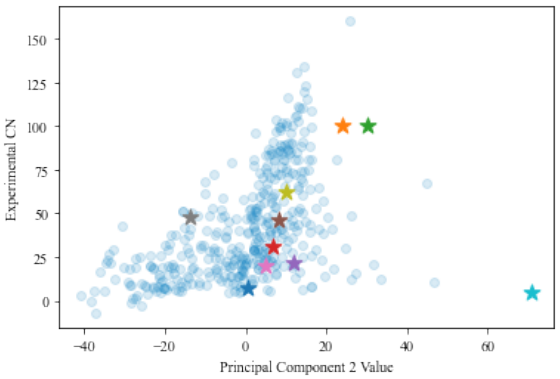
Figure 1.a shows the relationship between predicted CN and experimental CN for all 408 compounds, including outliers (highlighted in orange). In the context of the present work, compounds are considered outliers if their LOF value is greater than or equal to 2.0. The ten compounds deemed to be outliers are listed in Table 1. It is observed that six outlier compounds are predicted adequately (within approximately one median absolute error of parity, denoted by outer dashed lines), while four are predicted poorly. It is worth noting that all but one of these compounds are oxygenated, contrasting with most samples in the original dataset which are hydrocarbons. Additionally, most of these outlier compounds contain long carbon chains.

Table 1: Visualization of outlier compounds; keys correspond to highlighted samples in Figures 2.a-h

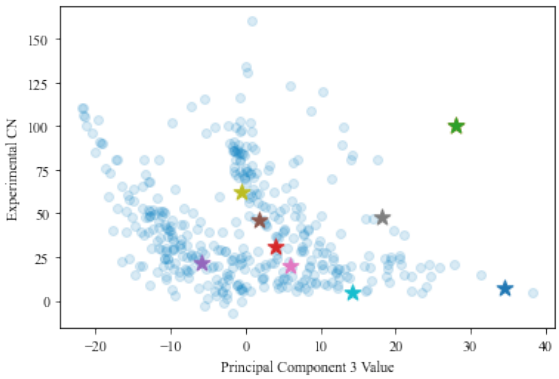
Name	Formula	CAS ID	Structure	Exp. CN	Key
Methanol	CH_4O	67-56-1		5.0	
Hept-1-yne	C_7H_{12}	628-71-7		22.0	
Tributyrin	$C_{15}H_{26}O_6$	60-01-5		7.0	
Linolenic acid	$C_{18}H_{30}O_2$	463-40-1		20.0	
Alpha-linoleic acid	$C_{18}H_{32}O_2$	60-33-3		31.0	
(Z)-octadec-9-enoic acid	$C_{18}H_{34}O_2$	112-80-1		46.0	
Stearic acid	$C_{18}H_{36}O_2$	57-11-4		62.0	
Dihexyl phthalate	$C_{20}H_{30}O_4$	84-75-3		48.0	
Trilaurin	$C_{39}H_{74}O_6$	538-24-9		100.0	
Trimyristin	$C_{45}H_{86}O_6$	555-45-3		100.0	



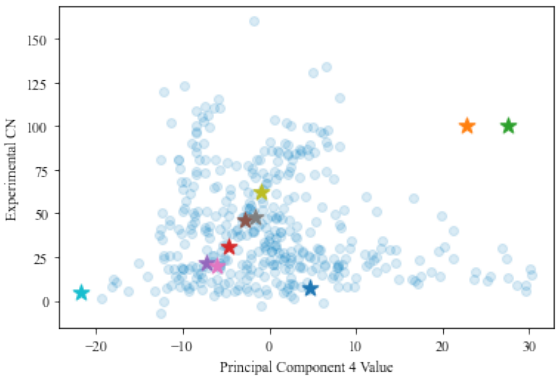
(a) Principal component 1



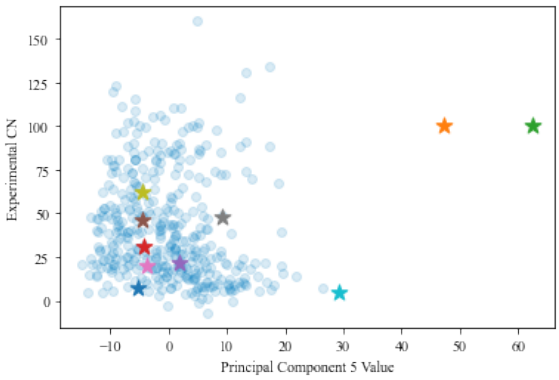
(b) Principal component 2



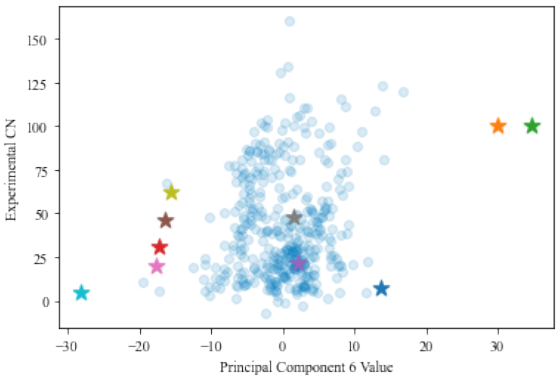
(c) Principal component 3



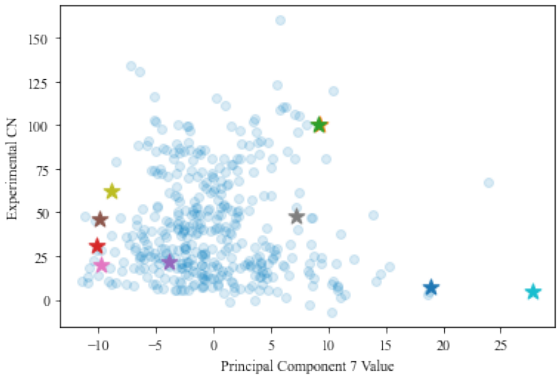
(d) Principal component 4



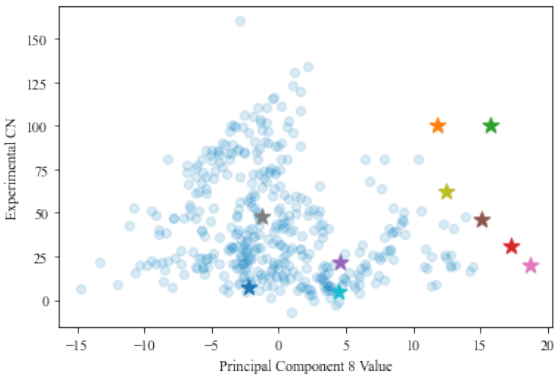
(e) Principal component 5



(f) Principal component 6



(g) Principal component 7



(h) Principal component 8

Figure 2: Relationships between CN and first eight principal components

After the outliers were removed from the dataset (398 remaining compounds), PCA was re-performed, and the ANN training procedure was repeated. The resulting relationship between predicted CN and experimental CN is observed in Figure 1.b. It is evident that ANNs trained after removing outliers perform inadequately compared to ANNs trained with outliers included in the training data.

These results indicate a dependence on outlier samples to successfully regress on a given dataset; Blatná states that outliers with respect to predictors, or "leverage points", can be beneficial to regression if the predictor values are relatively removed from the majority of samples but lie close to the line/curve of regression defined by the majority of samples [19]. In short, outliers such as these represent "extremes", but generally adhere to trends defined by inlier samples.

This phenomenon can be visualized by examining the relationship between principal components and the response variable, CN, for outlier compounds. Figures 2.a-h show the relationship between the first eight principle components and CN. Outliers are indicated using stars of various colors. It is observed that every outlier, for one or more principal components, exists both inside and outside general inlier trends. These results, in addition to sub-par ANN prediction accuracy without using outliers in training data, suggests that these outlier compounds are beneficial to ANN training.

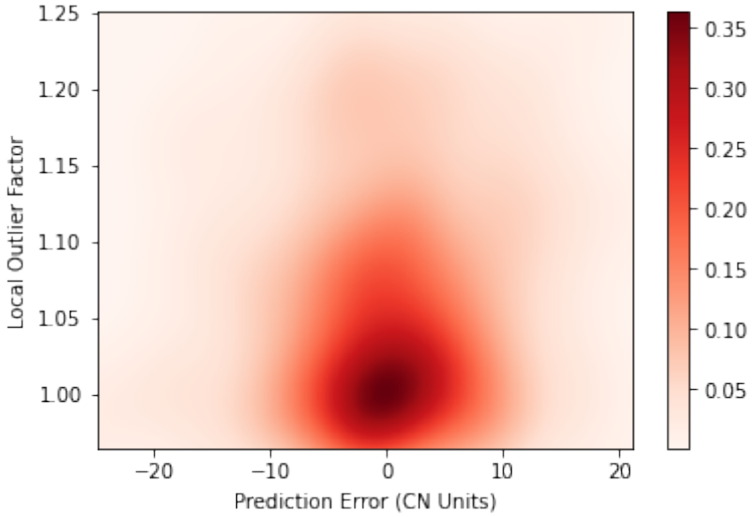


Figure 3: Density map showing the relationship between local outlier factor and ANN prediction error

Figure 3 shows a density map which illustrates the relationship between LOF and ANN prediction error. It appears that samples that are deemed outliers (higher LOF) tend to have lower prediction error; as previously discussed, if outliers are beneficial to ANN training, then it can be expected that outliers would be key samples leveraged heavily by the ANN during training to form correlation(s).

4 Conclusions

The present work analyzes how inliers/outliers affect ANN training, the relationship between outlier metrics and ANN prediction error, and investigates individual outliers with respect to compound structure and principal component-response variable relationships. In summary:

- Outliers are beneficial in ANN training, ultimately providing ANNs with information that is necessary to form predictor-target correlations during regression
- A weak relationship exists between LOF and ANN prediction error; i.e., outliers are not consistently predicted poorly
- Outlier compounds (in general) are oxygenated hydrocarbons, with long carbon chains

5 Acknowledgements

This work was funded by the U.S. Department of Energy under award DE-EE0008479 as part of the Co-Optimization of Fuels Engines (Co-Optima) project sponsored by the U. S. Department of Energy (DOE)

Office of Energy Efficiency and Renewable Energy (EERE) Bioenergy Technologies Office (BETO).

References

[1] D. C. Elliott, T.R. Hart, G. G. Neuenshwander, L. J. Rotness, A. H. Zacker. Catalytic hydroprocessing of biomass fast pyrolysis bio-oil to produce hydrocarbon products. Env. Prog. & Sustainable Energy, vol. 28(3), pp. 441-449, 2009

[2] ASTM D6890-16e1, Standard Test Method for Deter- mination of Ignition Delay and Derived Cetane Number (DCN) of Diesel Fuel Oils by Combustion in a Constant Volume Chamber. ASTM International, West Conshohocken, PA, 2016, www.astm.org

[3] ASTM 613-16a, Standard Test Method for Cetane Number of Diesel Fuel Oil. ASTM International, West Conshohocken, PA, 2016, www.astm.org

[4] E.A. Smolenskii, V.M. Bavykin, A.N. Ryzhov, O.L. Slovokhotova, I.V. Chuvaeva, A.L. Lapidus. Cetane numbers of hydrocarbons: calculations using optimal topological indices. Russ. Chem. Bull. 57 (3) (2008) 461–467.

[5] H. Yang, C. Fairbridge, Z. Ring. Neural network prediction of cetane numbers for isoparaffins and diesel fuel. Petrol. Sci. Technol. 19 (5–6) (2001) 573–586.

[6] T. Kessler, E.R. Sacia, A.T. Bell, J.H. Mack. Artificial neural network based predictions of cetane number for furanic biofuel additives. Fuel 206 (2017) 171–179.

[7] T. Kessler, P.C. St. John, J. Zhu, C.S. McEnally, L.D. Pfefferle, J.H. Mack. A comparison of computational models for predicting yield sooting index. Proceedings of the Combustion Institute 38 (1) (2020) 1385-1393.

[8] T. Kessler, T. Schwartz, H.W. Wong, J.H. Mack. Predicting the Cetane Number, Yield Sooting Index, Kinematic Viscosity, and Cloud Po9int for Catalytically Upgraded Pyrolysis Oil Using Artificial Neural Networks. Internal Combustion Engine Division Fall Technical Conference (2020).

[9] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD international conference on Management of data (2020) 93-104.

[10] J. Yanowitz, M.A. Ratcliff, R.L. McCormick, J.D. Taylor, M.J. Murphy. Compendium of Experimental Cetane Numbers. NREL/TP-5400-61693, 2014.

[11] D.A. Saldana, L. Starch, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, B. Creton. Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. Energy and Fuels, Vol. 25, No. 9, 3900-3908, 2011.

[12] J. Taylor, R. McCormick, W. Clark. Report on the relationship between molecular structure and compression ignition fuels. NREL Technical Report, 2014.

[13] M. Dahmen, W. Marquardt. A Novel Group Contribution Method for the Prediction of the Derived Cetane Number of Oxygenated Hydrocarbons. Energy and Fuels, vol. 29(9), 5781-5801, 2015.

[14] A. Mauri. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In: Roy K. (eds) Ecotoxicological QSARs. Methods in Pharmacology and Toxicology (2020) 801-820.

[15] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton. Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019 Jan 8; 47(D1):D1102-1109.

[16] S. Wold, K. Esbensen, P. Geladi. Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2(1-3) (1987) 37-52.

[17] T. Kessler, J.H. Mack. ECNet: Large scale machine learning projects for fuel property prediction. Journal of Open Source Software, 2(17) (2017) 401.

[18] D.P. Kingma, J.Ba. Adam: A Method for Stochastic Optimization. International Conference for Learning Representations, 2015.

[19] D. Blatná. Outliers in Regression. Trutnov 30 (2006) 1-6.