Decision tree analysis on the performance of zeolite-based SCR catalysts

Shin Young Bae∗ Jun Seop Shin∗ YeonSoo Kim**
Jong Min Lee†∗

∗ School of Chemical and Biological Engineering, Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Korea (e-mail: jongmin@snu.ac.kr).
** Department of Chemical Engineering, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul, 01897, Republic of Korea (e-mail: kimy3@kw.ac.kr).

Abstract: Urea-based selective catalytic reduction (SCR) is a promising method for removing NOx emissions. In the urea-based SCR method, zeolite-based catalysts are popularly used owing to their applicability over a wide range of temperatures compared to other catalysts. However, they still have several drawbacks such as inferior performance at low temperatures and thermal instability. This has subsequently led to numerous studies and experiments on the development of superior catalysts. While substantial experimental data exist on zeolite-based catalysts, extracting useful information from the data with a simple literature search is difficult owing to not only the amount of data but also complex correlations between feature, including preparation variables such as doped-metal loading and operational variables such as reaction temperature. Recently, extracting insights from a large database has become possible by utilizing machine learning tools. Among them, the decision tree can extract insights on the synthesis of the catalysts as the results derived from the models are intuitive and easy to interpret, unlike those from conventional discriminant machine learning models. In this study, classification models are obtained by training decision tree models with literature data on zeolite-based urea SCR catalysts, particularly the Beta and ZSM-5 types, and several experimental heuristics are extracted from the derived models.

Keywords: Catalysts, Machine learning, Decision tree, Big data, Diesel vehicle, Selective catalytic reduction, Nitrogen oxides.

1. INTRODUCTION

Urea-based selective catalytic reduction (SCR) is a promising method for the removal of NOx emissions. Among urea SCR methods, zeolite-based catalysts are commonly used owing to their wide temperature window. Despite their advantages over other catalysts, zeolite-based catalysts still have flaws such as inferior NOx conversion at low temperatures and the thermal instability regarded as a major drawback in practical applications (Han et al., 2019). To tackle these limitations, various synthesis methods such as doping additional metals or implementing new zeolite supports have been examined under different operating conditions, leading to several articles and related experimental data.

While substantial experimental data exist on zeolite-based catalysts, extracting useful information from the data with a simple literature search is difficult owing to not only large amount of data but also complex correlations between features including preparation variables such as doped-metal loading and preparation methods, and operational variables such as the reaction temperature. Recently, extracting knowledge from a large database by utilizing machine learning tools has become possible. These tools facilitate the understanding of complicated patterns and relations in large databases that cannot easily be analyzed by humans. Hence, machine learning applications have extended to various fields such as biology, materials, and catalysts (Han et al., 2011). In particular, decision trees, which are a popular machine learning tool, have been utilized to extract insights from experimental data in publications regarding various catalysts, such as those involved in water gas shift reaction and reforming of methane (Şener et al., 2018), (Odabaşı et al., 2014). A decision tree has several advantages in acquiring experimental heuristics regarding the development of catalysts compared to other machine learning tools. In particular, unlike conventional discriminant models such as deep neural networks, extracting important experimental features is easy because the classification flowcharts yielded by the decision tree models are explicitly suggested and the interpretation of the results is intuitive. In addition, a decision tree model is non-parametric and does not require any assumptions regarding the distribution of input data. Hence, it is suitable for categorical or encoded discrete numerical input data such as the preparation variables in the catalyst experiments (Friedl and Brodley, 1997). In this study, classification models are obtained by training a model with literature data on zeolite-based urea SCR catalysts, particularly the Beta and ZSM-5 types. The
experimental heuristics are extracted from the derived decision tree models. The remainder of the paper is organized as follows: In Section 2, the database of the zeolite-based catalysts for training and the computational details of the decision tree models and data preprocessing are indicated. In Section 3, the settings for decision tree models and the analysis of the results derived from the models are suggested. Finally, concluding remarks are provided in Section 4.

2. DATABASE AND METHODS

2.1 Database construction

The database was constructed by collecting the experimental data on NOx conversion from numerous publications with a total of 1918 data points. These databases can be categorized according to the zeolite type, Beta, which has a large pore size (He et al., 2009), (Gao et al., 2016), (Wang et al., 2019a), (Pereda-Ayo et al., 2014), (Xu et al., 2014), (Xu et al., 2018), (Xu et al., 2016), (Krivoruchenko et al., 2014), (Zhu et al., 2019), (Xia et al., 2016), (Kwak et al., 2012), (Wang et al., 2019b), (Rahkunnaa-Tolonen et al., 2005), (Krivoruchenko et al., 2015), (Zhu et al., 2016), (Nedyalkova et al., 2013) and ZSM-5, which has a medium pore size (He et al., 2009), (Qi et al., 2008), (Brandenberger et al., 2011), (Xiaoyan et al., 2012), (Park et al., 2006), (Ye et al., 2012), (Zhang et al., 2014), (Salker and Weisweiler, 2000), (Jouini et al., 2018), (Sultana et al., 2013), (Panahi et al., 2015), (Li et al., 2016), (Brandenberger et al., 2010), (Lou et al., 2014), (Kwak et al., 2012), (Wang et al., 2019b). In these publications, the efficiency of NOx conversion is selected as the target variable to be enhanced, where measurement data on the efficiencies were provided depending on the preparation or operational variables. In organizing the initial dataset for the decision tree models, NOx conversion was chosen as the output variable and other variables were selected as input features. Table 1 shows that the input features comprise both continuous and categorical data, where their applicable ranges or identities according to the data types are suggested.

2.2 Computational details on decision tree models

Decision trees are popular classification models that sequentially divide datasets into subsets according to the splitting criteria. Among the several training algorithms of decision trees, the Classification and Regression Tree (CART) algorithm is used herein. CART is a binary tree wherein each internal node yields two succeeding nodes and the post-pruning method called cost complexity pruning is used to simplify the model structure (Friedl and Brodley, 1997).

To obtain decision tree models with high accuracy, several parameters need to be considered. First, the number of classes in training data should be balanced. Otherwise the data belonging to a minority group are more often misclassified. In case of NOx conversion dataset, the samples where the NOx conversion corresponds to the values of 0~75% and 75~100% were assigned to class “low” and “high,” respectively. These divisions are not only meaningful physically but also suitable for addressing the class imbalance problem. Next, the best subset of input features to ensure model performance were selected. The features that are redundant and irrelevant to the target variable might degrade the accuracy of decision tree models and could make understanding the results difficult. To prevent these limitations, recursive feature elimination (RFE), a popular feature selection method, was utilized, which is explained in the next section. Finally, setting the hyperparameters of the decision tree modes is critical. Decision tree models include several hyperparameters such as maximum depth of trees and maximum number of leaf nodes. We herein performed a grid search for these hyperparameters to discover the hyperparameters yielding the best performance.

Table 1. Catalyst features and their ranges

| Category         | Features                          | Ranges for continuous variables or identities for discrete variables |
|------------------|-----------------------------------|---------------------------------------------------------------------|
| Zeolite type     | type                              | ZSM-5, Beta                                                         |
|                  | Si/Al                             | 4.6 ~ 40.0                                                          |
| Metal species    | species and composition (wt%)     | Fe (0.0 ~ 10.0), Cu (0.0 ~ 14.4), Mn (0.0 ~ 30.0)                   |
| Preparation      | methods                           | Incipient to wetness impregnation (IWI), Ion-Exchange (IE) Solid-state-ion-exchange (SSIE), IE + IW Homogeneous deposition precipitation (HDP) |
|                  | calcination time (hr)             | 2.0 ~ 20.0                                                          |
|                  | calcination temperature (°C)      | 200 ~ 700                                                           |
| Hydrothermal aging condition | O₂ (%)                     | 0.0 ~ 20.0                                                          |
|                  | H₂O (%)                           | 0.0 ~ 10.0                                                          |
|                  | CO₂ (%)                           | 0.0 ~ 5.0                                                           |
|                  | N₂, He, Ar (%)                    | 0.0 ~ 100.0                                                         |
|                  | temperature (°C)                  | 0.0 ~ 900.0                                                         |
|                  | time (hr)                         | 0.0 ~ 48.0                                                          |
| Reaction condition | NO (ppm)                        | 350.0 ~ 1200.0                                                     |
|                  | NH₃ (ppm)                         | 350.0 ~ 1200.0                                                     |
|                  | O₂ (%)                            | 2.0 ~ 14.0                                                         |
|                  | H₂O (%)                           | 0.0 ~ 10.0                                                          |
|                  | CO₂ (%)                           | 0.0 ~ 5.0                                                           |
|                  | N₂, He, Ar (%)                    | 76.0 ~ 98.0                                                         |
|                  | GHSV (hr⁻¹)                       | 0.12 ~ 6.7                                                          |
|                  | temperature (°C)                  | 50.0 ~ 772.0                                                        |

*GHSV (hr⁻¹) : Gas hourly space velocity (hr⁻¹)
2.3 Computational details on data preprocessing

Encoding categorical data into numerical data is critical in machine learning. There are two typical encoding methods: (i) label encoding that assigns an arbitrary integer to each category and (ii) one-hot encoding that represents the feature as a one-hot vector where the corresponding category is given a value of one and others are given a value of zero. In this study, the former method was used because the latter method can induce a sparse tree structure where the trees tend to grow in one direction. Among various feature selection methods, we adopted RFE that iteratively removes one feature with the smallest ranking from the original feature set at a time. This method has the advantage in that it prevents the models from becoming sub-optimal when several features are removed simultaneously. Here, the ranking scores of features are given as Gini importance that are calculated in training decision tree models.

3. RESULTS AND DISCUSSIONS

Before training the decision tree models, the total dataset was split into two subdivisions: Beta-based catalysts and ZSM-5-based catalysts. When the total dataset was used for training, judging whether the extracted heuristics were general or specific inference regarding zeolite type was difficult. These unclear results from the decision tree model are caused by the complex correlations between zeolite type and the other features. Therefore, the decision tree models were trained based on each dataset divided by zeolite type, leading to more clearly interpretable decision tree models. This approach is also reasonable in that numerous zeolite-based catalyst studies have been conducted around a specific zeolite type.

In the case of a Beta-based catalyst, the number of data samples corresponding to “high” and “low” were 561 and 432, respectively. The numbers of ZSM-5-based data samples corresponding to each class were 452 and 473, respectively. The numbers of these class is quite comparable for both catalysts, hence, the class imbalance problem can be prevented.

RFE was executed for the original feature subset except zeolite type as given Table 1. Nominal decision tree models without pruning are used to obtain information gain, which is utilized as ranking criterion. The result of RFE about Beta is illustrated in Figs. 1, but that about ZSM-5 is not presented here. Accuracy denotes the average classification accuracy of the model when the model is trained by a selected feature subset with five-fold cross validation. The solid line and shaded regions indicates the mean of classification accuracy and one standard deviation above and below the mean, respectively. The optimal number of selected features for Beta and ZSM-5 was 8 and 13, respectively and that of Beta is represented by the dotted line in the Fig. 1.

To prevent overfitting and to obtain an interpretable size of tree, setting hyperparameters to prune the decision tree model is necessary. In addition, it is important in ensuring the model performance. To acquire the best hyperparameters, we performed a grid search on several critical hyperparameters whose values are within a reasonable range. The best hyperparameters derived by grid search are listed in Table 2. For Beta-based catalysts, the classification accuracy of the models with the best and the worst hyperparameters was 0.89 and 0.83, respectively. In the case of ZSM-5 based catalysts, the accuracy was 0.85 and 0.74, respectively. These results demonstrate that more precise models can be generated by appropriately selecting hyperparameters.

3.1 Decision tree analysis on Beta-based catalysts

The optimal decision tree for the Beta-based catalyst is illustrated in Fig. 2. The reaction temperature was selected as the attribute of the root node, top of the decision trees, as it is the most deterministic factor in splitting the NOx conversion of the catalyst. In general, NOx conversion is heavily dependent on the reaction temperature. According to the criterion of the root node, the total dataset was partitioned into the two subdivisions specified by the feature of high and low reaction temperatures. Based on the tree structure, we extracted a few significant heuristics for each temperature region, as given in Table 3, where the recommended heuristics are denoted as R, while unadvisable heuristics are denoted as U.

According to the heuristic R1, the catalysts that comprise fairly low amounts of Si/Al ≤ 10 show high NOx conversion at low temperatures. These behaviors were demonstrated by the experiments conducted by (Xu et al., 2014), wherein a unique zeolite-synthesizing approach was introduced that produces Beta containing low amounts of Si/Al. Owing to the distinct properties of catalysts derived by this approach, the amount of isolated Cu$^{+}$ increases and the proximity of the Cu ions is closer, leading to the enhancement of active sites. In addition to Cu, high NOx

![Fig. 1. Recursive feature elimination on Beta-based catalysts](image-url)
conversion can be achieved for the Fe-doped Beta zeolite synthesized by the suggested method (Zhu et al., 2016). In particular, the Beta-based catalysts that have at least 1.0 wt% of Cu ions have high NOx conversion at a relatively low temperature of 137.0 °C.

Except employing the suggested zeolite synthesizing approach, obtaining heuristics to ensure high NOx performance at severely low temperature is difficult because the NH$_3$-NOx reactions are activated as the reaction temperature increases. However, several heuristics such as R2 can be observed for temperature ranging from 194.0 °C to 248.0 °C, which are still relatively low temperatures. R2 suggests that high NOx conversion can be obtained when Cu(wt%) is higher than 5.2 in the case where Si/Al is 10~22. On the contrary, achieving NOx conversion for the catalyst is difficult where Cu(wt%) is lower than 5.2. These trends can be demonstrated by the experiments in (Pereda-Ayo et al., 2014), where a tentative explanation is suggested that the presence of CuO enhanced via high copper content promotes NOx reduction at lower temperatures.

According to the decision tree models, the data samples satisfying criterion such as Si/Al $\leq$ 22.0, are interpreted to have high NOx conversion. However, this conclusion might be misleading because the dataset corresponding to Si/Al $>$ 22.0 is limited to the specific catalyst containing high metal content, particularly iron (Xia et al., 2016), it was unreasonable to draw a conclusion that the condition of Si/Al $>$ 22.0 is the only splitting criterion which guarantees high NOx conversion. Therefore, we added another critical features such as the Fe(wt%) content into the recipes of heuristic R2.

Additional analysis, as described above, was conducted for extracting the heuristic U1. As shown in Fig. 3, most samples satisfying the condition of H$_2$O(%) $>$ 5.5 are classified into "low" class label. However, most catalysts corresponding to this dataset were had a low Cu content that is regarded as the important feature in determining the performance of NOx conversion. Therefore, we replaced the previous criterion with the criteria Cu(wt%) $\leq$ 1.7. This rearranged heuristic U1 implies that the synthesis of a catalyst containing a small amount of metals, such that Fe(wt%) $\leq$ 0.05 and Cu(wt%) $\leq$ 1.7, should be avoided because high NOx conversion cannot be achieved using these catalyst, even at a high reaction temperature.

Heuristic U2 shows that performance of catalysts containing high amount of Cu can degrade when the reaction temperature is considerably high. Perdeda et al. explained that the presence of isolated Cu$^{2+}$ ions enhancing NOx conversion at high temperature can be lessened as the amount

Table 3. Heuristics for high NOx conversion for Beta-based catalysts

| Reaction Temperature | Heuristics | Preparation variables | Operational variables | Accuracy of classification |
|----------------------|------------|------------------------|------------------------|---------------------------|
| Low                  | R1         | Si/Al $\leq$ 10.0      | 175.0 $\leq$ $T_r$ $\leq$ 248.0 | 24/26                     |
|                      |            | Si/Al $\leq$ 10.0, 1.0 $\leq$ Cu(wt%) $\leq$ max | 137.0 $\leq$ $T_r$ $\leq$ 248.0 | 7/9                       |
|                      | R2         | 10 $< Si/Al \leq 22$, 5.2 $\leq Cu(wt\%) \leq$ max | 194.0 $\leq$ $T_r$ $\leq$ 248.0 | 4/5                       |
|                      |            | 22.0 $< Si/Al \leq max$ | $T_r$ $>$ 248.0 | 5/6                       |
| High                 | U1         | Fe(wt%) $\leq$ 0.05, Cu(wt%) $\leq$ 1.7 | $T_r$ $>$ 248.0 | 0/37                      |
|                      | U2         | Fe(wt%) $\leq$ 0.05, 4.23 $< Cu(wt\%) \leq$ max, 8.6 $< Si/Al \leq max$ | $T_r$ $>$ 435.0 | 0/7                       |

*Accuracy of classification = high/(high + low), $T_r$: Reaction temperature
of Cu increase (Pereda-Ayó et al., 2014). To synthesize a catalyst covering a wide range of high temperature, the recipes identified by heuristic U2 should be avoided.

3.2 Decision tree analysis on ZSM-5-based catalysts

The illustration of the decision tree on ZSM-5-based catalyst is not presented here, but the heuristics extracted from the models are summarized in Table 4. In the decision tree of ZSM-5-based catalyst, the reaction temperature was selected as an attribute of the root node whose value is slightly higher than that of Beta-based catalyst. R3 suggests that high NOx conversion at low temperature can be obtained when the catalysts are doped by bi-metals such as Fe and Cu or with a high content of Mn. According to (Zhang et al., 2014), ZSM-5-based catalysts containing both Fe and Cu exhibited excellent performance owing to their improved redox ability and larger acid sites over the catalyst surface. The catalyst containing a high content of Mn demonstrated high NOx conversion (Lou et al., 2014). These results are supported by the superiority of Mn at low temperature owing to its property regarding variable valences and detect sites of various oxidation states (Suib, 1998).

Heuristic R4 shows that high NOx conversion at low temperature can be achieved for catalysts containing Cu( wt%) > 2.7 and prepared under calcination temperature(°C) ≤ 525.0. These catalysts also have strong thermal stability. Thus, high NOx conversion are still guaranteed after the process of aging under temperature(°C) ≤ 750. In addition, a notable inference is the criterion of the calcination temperature. The performance of catalysts that calcined at temperature (°C) ≥ 525 are degraded so that the corresponding samples are assigned to “low” class. Lou et al. observed that active surface of Mn-doped catalysts was reduced when the catalysts are prepared under calcination temperature(°C) ≥ 600 (Lou et al., 2014). Similar to the case of Mn, it is thought that the calcination at severely high temperature has a negative effect on the performance of catalysts doped by other metals such as Fe or Cu.

As can be seen in the heuristic U3, catalysts containing small amount of metals also exhibit low NOx conversion even under the high reaction temperature in similar to trends mentioned above. Hence, it is recommended to avoid synthesis of the catalysts based on these recipes.

4. CONCLUSION

In this study, the classification model with decision tree was obtained by training the literature data on zeolite-based urea SCR catalysts. Also, experimental heuristics leading to superior catalysts can be extracted from the models. However, few interpretable rules are yielded from decision tree owing to the sparsity of training data. This fundamental limitation can be tackled by increasing the number of data. Otherwise, further statistical analysis on correlations among features or data visualizing are required to clarify the extracted rules.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (MSIP) (NRF-2016R1A5A1009592).

REFERENCES

The reference list includes various sources related to the study of nitrogen oxide reduction using catalysts.

Table 4. Heuristics for high NOx conversion for ZSM-5-based catalysts

| Reaction Temperature | Heuristics | Preparation variables | Operational variables | Accuracy of classification |
|----------------------|------------|-----------------------|-----------------------|---------------------------|
| Low                  | R3         | 2.0 ≤ Fe(wt%) ≤ 8.0, Cu(wt%) = 4.0 | 177.0 < T_r ≤ 300.0 | 33/39                     |
|                      | R4         | 2.7 ≤ Cu(wt%) ≤ max, T_c ≤ 750.0 °C, calcination temperature ≤ 525.0 | 172.0 ≤ T_r ≤ 300.0 | 21/27                     |
| High                 | U3         | Fe(wt%) ≤ 0.64, Cu(wt%) ≤ 1.3 | 300.0 ≤ T_r ≤ 575.0 | 6/18                      |

*Accuracy of classification = high/(high + low), T_r: Reaction temperature, T_c: Aging temperature.
