Catalysis Today xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Catalysis Today

journal homepage: www.elsevier.com/locate/cattod

Statistical review of dry reforming of methane literature using decision tree and artificial neural network analysis

Ayşe Neslihan Şener[a], M. Erdem Günay[b], Aybüke Leba[a], Ramazan Yıldırım[a,*]

[a] Department of Chemical Engineering, Boğaziçi University, 34342, Bebek, Istanbul, Turkey
[b] Department of Energy Systems Engineering, Istanbul Bilgi University, 34060, Eyup, Istanbul, Turkey

ARTICLE INFO

Keywords:
Dry reforming of methane
Data mining
Knowledge extraction
Decision trees
Artificial neural network

ABSTRACT

The aim of this work was to extract knowledge for dry reforming of methane (DRM) reaction from experimental data using data mining tools such as decision trees and artificial neural networks. An extensive database containing 5521 data points depending on 63 catalyst preparation and operational variables was constructed from 101 papers published between 2005 and 2014; the output variables were CH_4 conversion, CO_2 conversion and H_2/CO ratio of the product stream. Then, the database, as a whole or as subsets for different base metals were analyzed using decision trees (DT) to develop heuristics for high performance and artificial neural networks (ANN) to determine relative importance of input variables and predict the performance under unstudied conditions; mostly CH_4 conversion, which is the most frequently reported output variable, were used in analysis. The testing accuracy of the decision tree was about 80% leading to four heuristics (i.e. four possible courses of action) for high CH_4 conversion over Ni based catalyst. The first decision point to separate these heuristics is the reaction temperature as can be expected. This is followed by the other variables such as support type, W/F and reduction temperature. ANN analysis revealed that operational variables have higher relative importance (55%) compared to catalyst preparation variables (45%). The most important operational variable was found to be the reaction temperature while the active metal and the support are the most important catalyst preparation variables. ANN model was also tested to predict the data, which was not seen by the model before, and the data in 65 papers out of 101 were predicted within 15% error while 76 papers had the error rate of less than 20%.

1. Introduction

Dry reforming of methane (DRM) has been investigated significantly over the last few decades to convert two greenhouse gases (CO_2 and CH_4) into synthesis gas with a H_2/CO ratio desirable for the synthesis of long chain hydrocarbons and oxygenated chemicals through Fischer-Tropsch like process [1]. However, there are some serious problems like coke formation and metal sintering that decrease the efficiency and long term stability of the catalysts, and therefore prevent the commercialization of this process. To overcome these problems, significant amount of work has been devoted in recent years by testing various combinations of active metal, promoter, support and catalyst preparation method as well as various configuration of reaction systems and operational conditions [2].

The most common catalyst material that has been studied for DRM is supported nickel. However, Ni based catalysts have a tendency to form coke on catalyst surface; this can result in deactivation of the catalyst and/or plugging of tubes inside reactor. Although noble metals such as ruthenium and rhodium have been also studied and they have been found to be more active and resistant to carbon deposition, their high cost and low availability are preventing them to become a practical alternative to Ni [3]. Significant amount of work has been also reported aiming to improve the performance of Ni catalyst using some promoters. The situation is similar in the use of other catalyst materials and preparation and reaction conditions; although some alternatives (like γ-Al_2O_3 for support, incipient wetness impregnation as the preparation method and fixed bed as the reactor configuration) have received more attention, many other alternatives were also tested. As a result, there are more than 1000 article in Web of Science database on dry reforming of methane. In these work, the effects of very large number of catalyst preparation variables (like active metal type and loading, promoter type and loading, support, preparation method, calcination conditions and reduction conditions) and reaction variables (like temperature, feed compositions, flow rate and catalyst weight) were investigated creating large amount of data that cannot be analyzed using simple literature search. Data mining tools can be used in order to do this more effectively and extract valuable knowledge from the database constructed from these publications, and this knowl-

* Corresponding author.
E-mail address: yildirra@boun.edu.tr (R. Yıldırım).

http://dx.doi.org/10.1016/j.cattod.2017.05.012
Received 29 October 2016; Received in revised form 8 March 2017; Accepted 4 May 2017

edge can be used to improve the future experimental work.

Data mining is the process of extracting useful information from the large databases using some tools like k-means clustering, decision trees, artificial neural networks, support vector machines and so on [4]. It aims to identify patterns, trends and correlations that may be hidden within the network of data points, and that may not be observed with naked eyes. There are several applications of these tools in the literature for modeling the experimental data in various fields of science or engineering including dry reforming reactions. For example, Omata et al. [5] studied the effect of preparation parameters of Co-MgO catalysts for the dry reforming of methane using an experimental design procedure coupled with artificial neural networks; they also implemented a grid search algorithm to find the conditions for maximum catalytic activity. Similarly, Ayodele and Cheng [6] used artificial neural networks and Box Behnken design together to model and optimize the input variables (reaction temperature, reactant feed ratio and CH_4 partial pressure) for desired CH_4 and CO_2 conversions and syngas ratio. In another study, Hossain et al. [7] applied artificial neural networks for the same reaction over Ni/$CaFe_2O_4$ catalysts to model H_2 and CO yields as well as CH_4 and CO_2 conversions.

In the work cited above, the models were constructed using the experimental data generated by the researchers themselves, hence their applicability is limited. However, more general models and heuristics, which can be valid in wider range, could be developed if these methods are applied to a larger database constructed from the entire literature, and they can help the researchers to benefit from the large experience accumulated in the literature over the years. Indeed, we utilized such approach in various reaction systems starting from selective CO oxidation over Cu-Ce catalysts [8] and noble metal catalysts [9] as well as water gas shift reactions [10]. Then, we moved to more complex systems like direct alcohol fuel cells [11] and biodiesel production [12]. Recently we successfully implemented the decision tree analysis for steam reforming of methane [13], which is quite closer to methane dry reforming process. Similar analysis were also conducted by other researchers on the data acquired from other sources for steam reforming [14], photocatalysis over titania [15] and oxidative coupling of methane [16,17].

In this communication, we reported the construction of an extensive database from past publications in the literature for dry reforming of methane and analysis of this database using decision tree and artificial neural network methods to develop models and heuristics to improve the future experimental work.

2. Material and methods

2.1. Database construction

The database was constructed by extracting the experimental data on dry reforming of methane from published articles in various databases (such as ACS, ScienceDirect and Wiley) between 2005 and 2014. The annual numbers of papers related to DRM is presented in Fig. 1 (more than 1000 in total); this figure was drawn using Web of Science search for papers containing “dry reforming of methane (or CH_4)”, “carbon dioxide (or CO_2) reforming of methane (or CH_4)” or “methane (or CH_4) dry reforming” in their title directly; this number is probably much larger since some articles may not use these words in their titles (value for 2016 covers only the publications up to mid-October). 403 of these research publications were reviewed but 101 articles containing 5521 data points were used to construct the data set because remaining papers were not suitable to extract data or contained variables that are not repeated in sufficient number of papers for a reliable model. The output (or performance) variables were chosen as CH_4 conversion (5521 data points; all publications reported this variable), CO_2 conversion (3930 data points; some publications did not report this variable) and H_2/CO ratio (2323 data points; some publications did not report this variable as well). These outputs were found to be influenced by 63 independent variables related to catalyst preparation and reaction tests as given in Table 1 with their applicable ranges.

2.2. Computational details

MATLAB R2013b was used for computational work. Decision tree analysis was used to develop heuristics for higher catalytic performance while artificial neural networks were used to determine the relative importance of input variables and to predict the outcome of unstudied conditions. CH_4 conversion for DRM reaction was used as the output variable with 5521 data points. CO_2 was not modeled because the number of instances is small (3930), and CH_4 conversion is already sufficient to determine performance; however, decision tree analysis was also developed for H_2/CO even though the number of data points was also small (2323) because there are no other output variables that may substitute this.

In decision tree analysis, “classregtree” function of MATLAB in classification mode was used. The CH_4 conversion values are classified as *low* (0–50%), *medium* (51–75%) and *high* (76–100%); this division is not only meaningful physically but also suitable to avoid class imbalance problem because each class had approximately equal number of data points [4]. For H_2/CO ratio, the classes were determined as 0–0.89, 0.90–1.09 and > 1.1 (or 1.1–3.3). Gini's diversity index was used to select the optimal split. The optimum decision tree was found by minimizing the error rate of individual leaf nodes (the fraction of wrong classifications in that node) and the total error rate of the entire tree (the weighted average error of all the leaves). Large trees may have small training errors; however, they may overlearn, and may give hints that are too complex for practical purposes. Hence, the optimal tree size was decided by minimizing the error obtained by using the tree to classify a dataset not seen before (testing). For this purpose, the total data set was randomly divided into two sets, one containing 2/3 of the data points while the other has 1/3. The larger set was used as training and constructing the decision tree and the smaller one was used for testing and evaluating its generalization ability. Then, the tree size with an acceptable error rate for training and the smallest error rate for testing was chosen as optimum [9,18].

In the artificial neural network modeling, various network topologies were investigated first to determine the optimal topology using root mean square error (RMSE) of testing. Levenberg-Marquardt (trainlm) was used for training while Bayesian Regularization (trainbr) algorithm was used for testing together with tangent sigmoid (tansig) as transfer function. Average root mean square error (RMSE) for testing was used as the performance indicator for the neural network by implementing k-fold cross validation method. The entire database was randomly divided into k sets (k was chosen as 4 in this study), k-1 sets were used to train the network while remaining one set was used for testing the ability of network for unseen data. This procedure was repeated 4 times and average RMSE was used to select the optimum topology. Each network topology was trained 10 times in order to eliminate the effects of random initialization of the neural network weights. Early stopping technique was used in order to prevent the model overfitting [19].

The change of root mean square error technique was used to find the relative importance of input variables. The procedure was as follows: one of the input variables was removed from the data set, and then network was trained with the remaining variables. The RMSE value of the model calculated in the absence of this input variable was compared with the original RMSE calculated in the presence of all input variables. The difference was used as the significance of this input group [19].

One input neuron was assigned for each individual continuous variable (such as temperature, W/F ratio, feed compositions), one input neuron was assigned for each option for each categorical variable (for example one for incipient to wetness impregnation, one for sol-gel and so on.) with a value of 1 (if the method was used for that data) and 0 (if