Feature engineering for chemical compound prediction

Mikolaj Kacki

July 2019

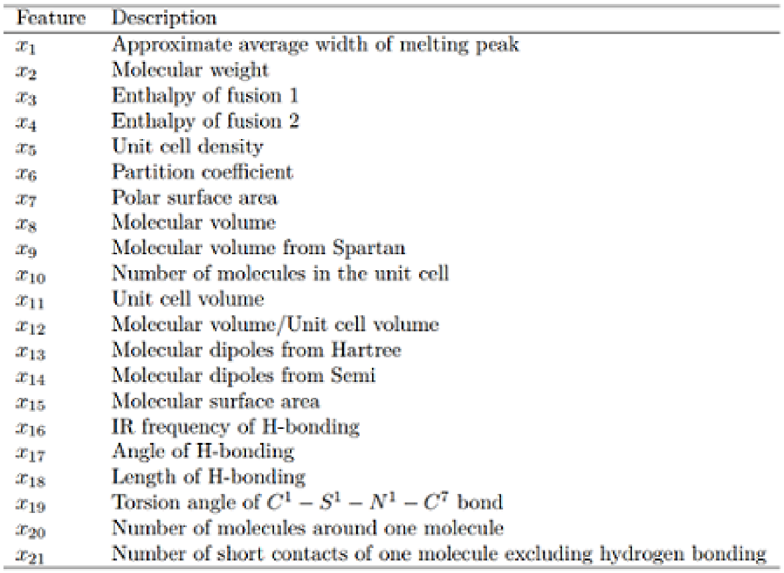
1 Abstract

The purpose of this analysis is to identify which variables have a significant effect on melting point of a chemical compound given 21 numerical variables.

1.1 Melting temperature

The melting point of a chemical compound is the temperature at which the compound melts. It is an important property to chemists who aim to develop new compounds that have particular thermophysical behaviour.

In the experiment 60 observations are made with melting point ranging from **68.17** to **220.22**, its median equals **142.07** while mean **142.05**.



Feature	Description
x_1	Approximate average width of melting peak
x_2	Molecular weight
x_3	Enthalpy of fusion 1
x_4	Enthalpy of fusion 2
x_5	Unit cell density
x_6	Partition coefficient
x_7	Polar surface area
x_8	Molecular volume
x_9	Molecular volume from Spartan
x_{10}	Number of molecules in the unit cell
x_{11}	Unit cell volume
x_{12}	Molecular volume/Unit cell volume
x_{13}	Molecular dipoles from Hartree
x_{14}	Molecular dipoles from Semi
x_{15}	Molecular surface area
x_{16}	IR frequency of H-bonding
x_{17}	Angle of H-bonding
x_{18}	Length of H-bonding
x_{19}	Torsion angle of $C^1 - S^1 - N^1 - C^7$ bond
x_{20}	Number of molecules around one molecule
x_{21}	Number of short contacts of one molecule excluding hydrogen bonding

Figure 1: Variables description

2 Analysis

2.1 Correlation plot

The very simplest way of observing dependencies between variables is a correlation matrix. Plotting the matrix clearly shows correlation coefficients between a set of variables. The values of coefficients range from -1 to 1, correlations of 0.7 and bigger indicate highly correlated variables.

Correlation plot shows that melting point is highly correlated (0.82) with enthalpy of fusion 2 (x_4) and significantly correlated with enthalpy of fusion 1 and polar surface area - respectively 0.66 and 0.55.

From the plot we can also observe correlations between other factors what can be useful for further feature engineering if we aim to predict melting point.

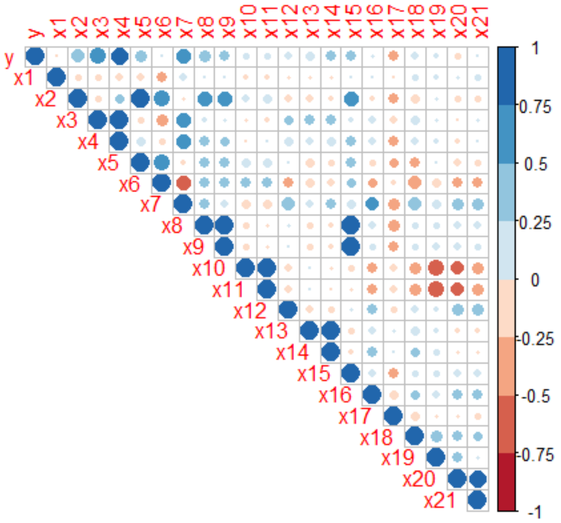


Figure 2: Correlation matrix

2.2 Regression analysis

Statistical approach used here will be to select the best regression model from all models that could be fitted. The chosen model would balance complexity and goodness of fit excluding irrelevant variables.

Stepwise algorithm is performed (using forward selection). The algorithm starts with the simplest (intercept) model and then compares the simplest model with others +1 variable models, chooses the one with the smallest Aikake Information Criterion (AIC) to balance complexity and goodness of fit:

$$AIC_i = nlog(\frac{RSS_i}{n}) + 2k \tag{1}$$

Hence the fitted model is:

$$y = -341.65 + 0.65x_1 + 0.26x_2 + 2.53x_4 + 1.04x_{10} + 4.03x_{14} + 97.611x_{18} \tag{2}$$

Note that the algorithm included x_4 but excluded x_3 that has significant correlation with response variable. It may be because of bigger variance of x_3 ($\text{var}(x_3) = 1023.528$, while $\text{var}(x_4) = 98.57$) or high correlation between x_3 and x_4 ($\text{cor}(x_3, x_4) = 0.915$), the algorithm didn't want to overcomplicate the model including two highly correlated variables. However, the plot clearly shows relationship between both x_3, x_4 and response variable.

Note also that according to p-values x_2 is the second most important variable in the model (its p-value equals $4.72 * 10^{-6}$), the most important is x_4 (its p-value equal to $4.37 * 10^{-15}$).

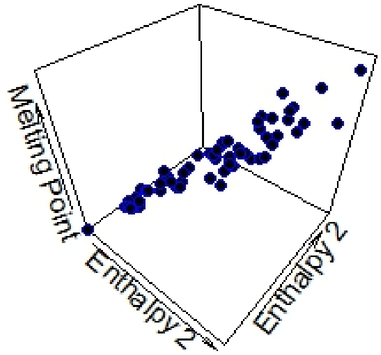


Figure 3: Melting point/enthalpies plot

It's also worth adding why x_7 was excluded despite significant correlation. Despite one outlier it shows linear relation with dependence variable (look Figure 4), so it's significant when prediciting melting point. However, it could have been excluded due to the fact that there are only 8 unique values so it could be represented as a factor variable not numerical.

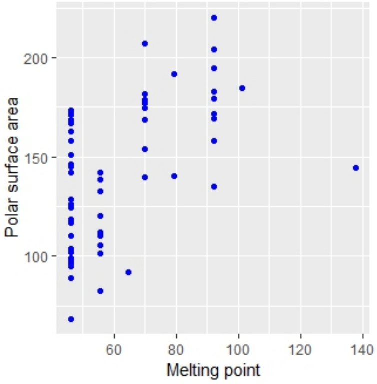


Figure 4: Melting point/polar surface area plot

We have a few tools to compare this model to different ones. Checking the summary of the model we see that $R^2 = \mathbf{0.82}$ and $adj.R^2 = \mathbf{0.80}$. We can moreover check the root mean square error:

$$RMSE(\theta) = \sqrt{E(\hat{\theta} - \theta)^2} \tag{3}$$

In this case - RSME = **15.15**. This values describe how well model fits, it could be compared for example with RMSE of different model to choose which fits better (for example RMSE of intercept only model is **35.76**, and full model's **11.84** so RMSE of fitted model is not that big compared with full model bearing in mind that it includes 6 variables not 21). However, prediction and checking goodness of fit may not be adequate since the dataset is small and outliers are harder to detect (and remove). If the dataset was bigger we could also use **cross validation** and so obtain better measure of predictive power of the model.

2.3 Boruta algorithm

The algorithm is designed as a wrapper around a random forest classification algorithm. It iteratively removes the features which are proved by a statistical test to be less relevant than random probes (i.e. having smaller Z-scores). [1]

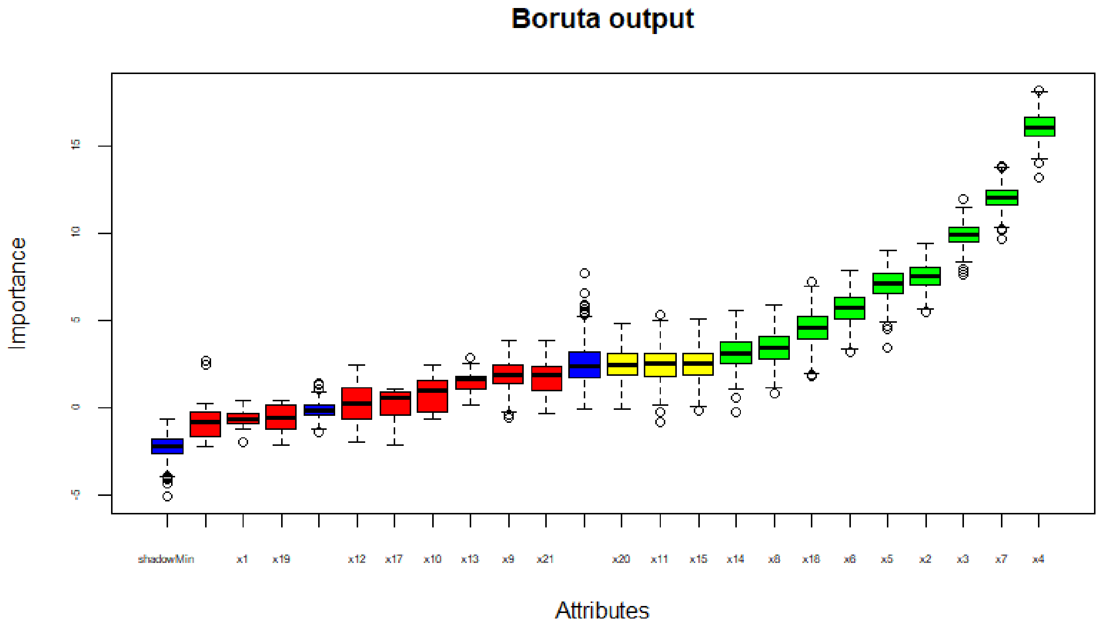


Figure 4: Boruta used on crystal data

The algorithm confirmed what was concluded in the correlation matrix plot. Namely, the plot states that the most important variables are x_4, x_7 and x_3, we have also note x_2, x_5 are quite important.

3 Conclusion

All in all, I observed that the biggest effect on melting point have enthalpies (especially enthalpy 2) and polar surface area. Except those, also molecular weight and unit cell density have some effect on melting point.

If the aim was to predict the response variable it could be useful to use principal component analysis. PCA reduces number of dimensions by 10 and so 11 components explain 96.3% of variability.

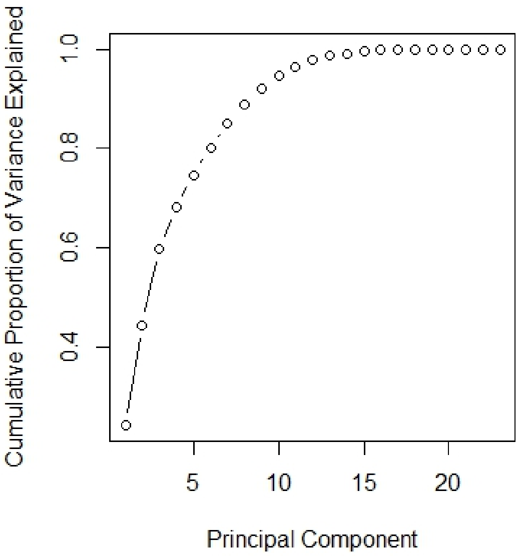


Figure 5: Principal component analysis plot

References

[1] Miron B Kursa, Aleksander Jankowski, and Witold R Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.