ARCHITECTURAL STUDY OF EFFECTS OF KERNEL SIZES ON CHEMOMETRIC DATA

Nikhitha Gudur

OSTBAYERISCHE TECHNISCHE HOCHSCHULE REGENSBURG

Sensorik-ApplikationsZentrum

WOMEN IN DATA SCIENCE REGENSBURG

INTRODUCTION

Chemometrics can be explained as finding a way of extracting relevant information from chemical data, representing and displaying that information, and also finding out how to get such information into data [1].

It has been observed that Convolutional Neural Networks (CNN) have great scope in the area of chemometrics. However, efficiently training the CNN models by identifying the right set of parameters and hyperparameters has become a major challenge and one of the problems which need to be tackled in order to encourage its application more extensively in chemometrics

DATA

Two datasets namely were used in this work, Mango dataset [2] and Melamine dataset [3].

Mango Dataset

- Samples of Keitt mangoes across four harvest seasons.
- Spectra for minimum and maximum DM values of mango dataset is shown in figure 1.

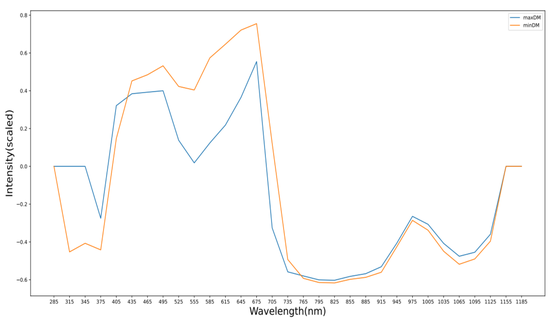

Fig 1: Spectra for minimum and maximum DM

Melamine Dataset

- Dataset consists of Fourier-transform near infrared (FT-NIR) absorbance spectra and turbidity point readings of Melamine Formaldehyde obtained from Metadynea GmbH.

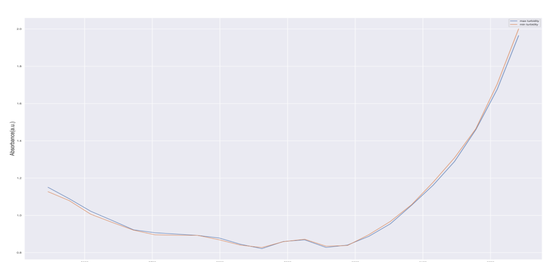

Fig 2: Spectra for minimum and maximum turbidity point for X1 input

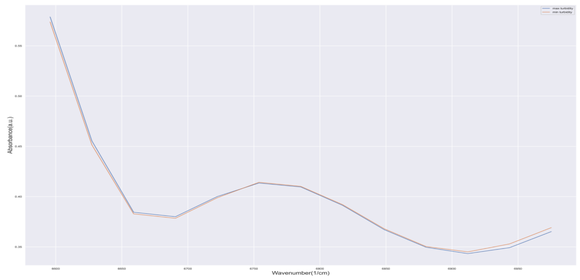

Fig 3: Spectra for minimum and maximum turbidity point for X2 input

METHODS

Several methods were used in this work and they are shown in figure 4.

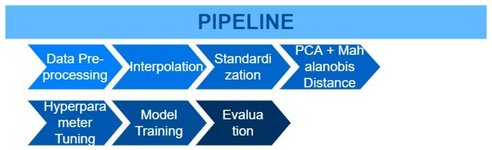

Fig 4: Pipeline of methods followed in this work for mango and melamine datasets

Outliers detection was done using PCA and Mahalanobis distance and the visualization for mango dataset is shown in figure 5.

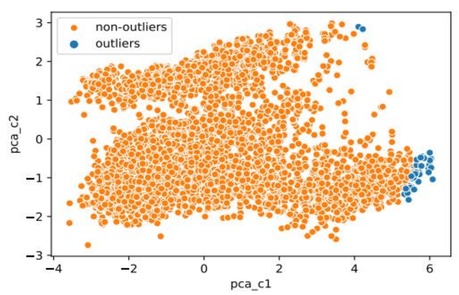

Fig 5: Outliers detected in Train partition of Mango dataset through PCA and Mahalanobis distance

MODELS

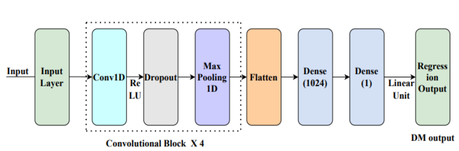

Fig 6: Architecture of 1D CNN model trained on Mango dataset

Model used for training on mango dataset is shown in figure 6.

Hyperparameter tuning was performed with 2, 3, 5, and 10 kernel sizes to identify best hyperparameters that can be used for all kernel sizes. During tuning, Kernel size 10 obtained least RMSE value.

Model used for training on melamine dataset is shown in figure 7.

Hyperparameter tuning and training processes followed for melamine dataset were similar. During tuning, Kernel size 2 obtained least RMSE value.

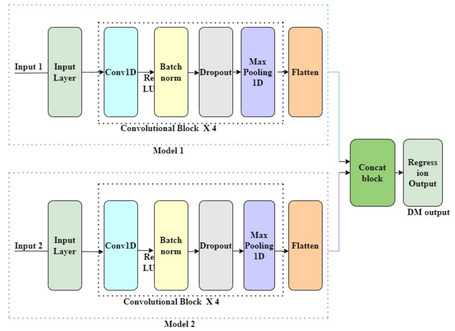

Fig 7: Architecture of 1D CNN Concatenated model trained on Melamine dataset

In addition to this, stride and batch normalization were added to the same concatenated model configuration and trained for same kernel sizes on melamine data.

RESULTS

Results of the 1D CNN model and 1D CNN Concatenated model (with stride and batch normalization) are shown in figures 8 and 9 for mango and melamine datasets respectively.

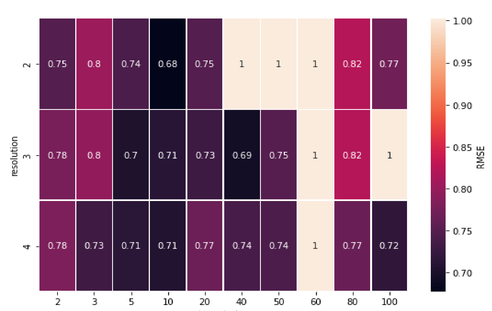

Fig 8: Heatmap of Mango dataset Test set RMSE values

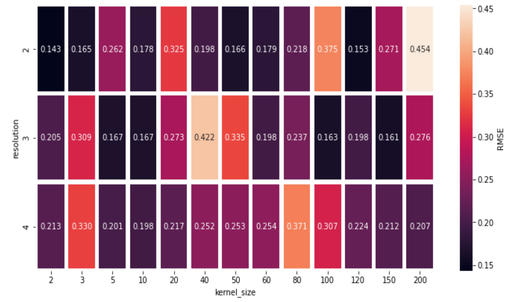

Fig 9: Heatmap of Melamine dataset Test set RMSE values

It can be observed from figure 8 that the model showed good performance for lower kernel sizes, i.e., kernel sizes in the range of 2 – 20 and attained error values lesser than the higher kernel sizes i.e., kernel sizes in the range of 30 – 100.

For the melamine dataset, no clear pattern is visible in figure 9 regarding the influence of smaller or larger kernel sizes. Nevertheless, adding stride and batch normalization layers to the convolutional blocks of the 1D CNN concatenated model for all the kernel sizes reduced the error values significantly.

CONCLUSION

The main agenda of this work was to investigate the effects of kernel sizes for 1D CNN models on chemometric datasets.

In summary, with this study, it was concluded that kernel sizes have importance in influencing the performance of CNN models on chemometric data. Moreover, other hyperparameters like stride are also important and can help fine-tune the models to achieve better results.

REFERENCES

1. Wold, Svante. "Chemometrics; what do we mean with it, and what do we want from it?." *Chemometrics and intelligent laboratory systems* 30, no. 1 (1995): 109-115.
2. Anderson, N. T., K. B. Walsh, P. P. Subedi, and C. H. Hayes. "Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content." *Postharvest Biology and Technology* 168 (2020): 111202.
3. Nikzad-Langerodi, Ramin, Werner Zellinger, Susanne Saminger-Platz, and Bernhard A. Moser. "Domain adaptation for regression under Beer–Lambert's law." *Knowledge-Based Systems* 210 (2020): 106447.