ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

One-class Autoencoder approach to classify Raman spectra outliers

Katharina Hofer-Schmitz, Phuong-Ha Nguyen and Kristian Berwanger *

Fraunhofer Institute for Applied Information Technology FIT
Schloss Birlinghoven - 53754 Sankt Augustin - Germany

Abstract. We present an one-class Anomaly detector based on (deep) Autoencoder for Raman spectra. Omitting preprocessing of the spectra, we use raw data of our main class to learn the reconstruction, with many typical noise sources automatically reduced as the outcome. To separate anomalies from the norm class, we use several, independent statistical metrics for a majority voting. Our evaluation shows a f1-score of up to 99% success.

1 Introduction

Raman spectroscopy is a marker-free analytical method, applied in many research domains [1], especially in biology [2].

Data collection for bio-chemical analysis of small-volume substances is typically tedious. Even slightest, hardly controllable environmental changes can add noise to the recorded signal. One of the advantages of Deep Neural Networks is that, assuming to have enough training data, one might skip preprocessing completely [3].

We are currently observing cell modification processes (electroporation transfection), which are typically only successful in few cases. To detect those, we choose an one-class approach. Although it is not known a-priori how the Raman spectra of the modified cells will differ, we can nonetheless take measure of the original cells. This paper presents our preliminary research on solid and dissolved proteins to demonstrate our method on a less complex spectra dataset.

The paper is structured as follows: In section 2 we present related literature, then we introduce our approach in section 3 and finally we evaluate the described method in section 4.

2 Related Work

Published work for Raman spectra classification include multivariate statistical analysis methods including supervised methods as Partial Least Squares, Linear Discriminant Analysis, Support Vector Machines and unsupervised methods as Principal Component Analysis (PCA) and Cluster Analysis (see [4] for an overview). Classification with supervised Deep learning methods is considered in [5, 6].

*This work was supported by the project 'OptisCell' under the Fraunhofer market-driven prospective research program (MAVO). Moreover, we thank Andreas Pippow for valuable input during our work.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

However, for rarely observed samples of interest, Anomaly Detection approaches are much more suited (see [7, 8] for an overview). In [9] it is shown that Autoencoders are able to detect subtle anomalies, where usual techniques as PCA fail. Since we have a partially labeled dataset, we apply one-class training methods, see e.g. [10, 11]. There, Replicator Neural Networks are trained on one class and the reconstruction score is then used to classify the test dataset.

Our approach differs in the usage of an Autoencoder Neural Network. Instead of reassigning the data into clusters as in Replicator Neural Networks, Autoencoder Neural Networks force the data through layers with less neurons to learn the compressed representation. Moreover, we do not only consider the reconstruction score, but also take into account the distribution of the reconstruction scores of the training data.

3 Method

As mentioned in the introduction, for our biological application, it's very costly to collect spectra of the outlier class. Since the bio-chemical approach to identify and characterize outliers takes months, approaching this as a classical two-class problem is no option. It is however possible to take measure of our normal class (pre-transfection cells) and train an one-class model.

This is done by training an Autoencoder network to learn our normal classes' characteristics by minimizing the reconstruction error (score) with respect to the given loss function, similar to the learnt components of PCA. When using the learnt encodings to reconstruct irregular spectra, we expect a higher reconstruction error.

We use three statistical parameters directly derived from training data to asses the results of the unknown samples: the maximum reconstruction score s_{max}, calculated from the mean absolute error between the original spectra and its reconstruction, a standard deviation threshold $t_{std} = \mu + 2\sigma$, with mean value μ and standard deviation σ and the interquartile threshold t_{80}, which marks the 80% quartile of the reconstruction scores. If at least two of these thresholds are exceeded, a sample's reconstruction is considered as anomaly.

The Deep Learning library `DEEPLEARNING4J` is used for our experiments. Stochastic Gradient Descent is applied, ADAM Updater for the learning rate, L_2-regularization with 0.0001 and the mean absolute error as loss function. As activation function we use leaky ReLU, a modified ReLU activation function, which helps to avoid the problem of vanishing gradients. The weights were initialized with ReLU initialization.

3.1 Datasets

We apply a min-max scaler with the range of $[0, 1]$ to the y-axis and trim the x-axis to the region of interest of the spectrum for all data samples.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

3.1.1 Cysteine on Silicon

Cystein powder was dissolved in hydrochloric acid solution and let to dry on a silicon wafer. Our dataset consists of 5000 Silicon (S) and 5000 Cysteine (Cys) spectra. The S-data is split into a $4500 : 500$ ratio for training and testing. We also use all 5000 Cys-spectra for testing.

3.1.2 Glucose Oxidase in Reaction Buffer

50 μL of 100 U/mL Glucose Oxidase solution in Reaction Buffer were applied to a glass slide and covered to avoid desiccation. This dataset consists of 2000 Reaction Buffer (RB) and 41 Glucose Oxidase (GOx) spectra. We use 1800 RB-spectra for training and the remaining 200 and 41 GOx-spectra for testing.

4 Results

4.1 Cysteine on Silicon

We used 650 input nodes and three inner layers with 64, 16 and 64 nodes, respectively. The learning rate was set to 0.2.

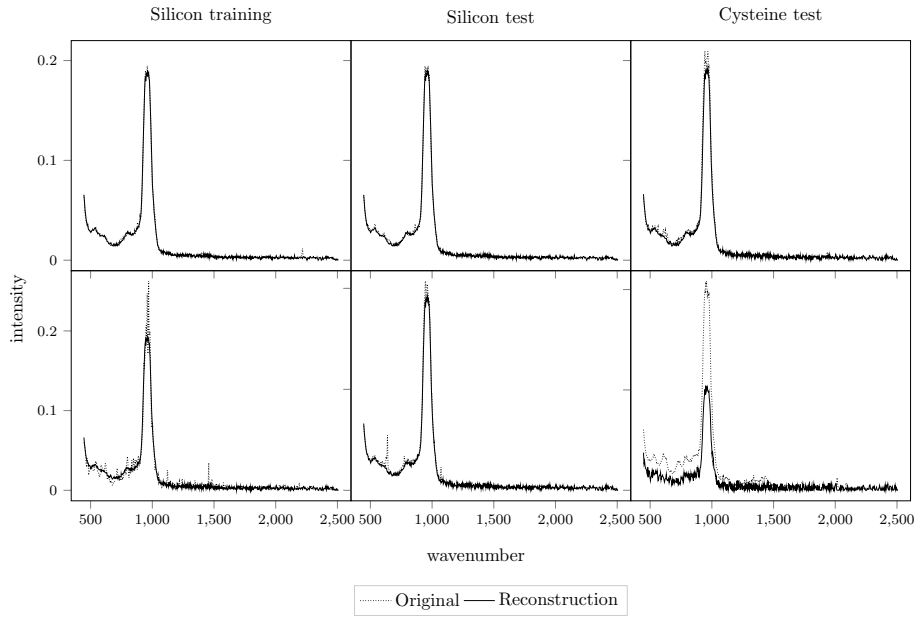

Fig. 1: Good (top row) and bad (bottom row) reconstruction examples

Some spectra and their reconstruction are given in figure 1. Some noise in the spectra and furthermore a variance between each class can be observed. For the reconstructions we observe, that especially spikes are removed and white noise is reduced.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

The distribution of these scores is plotted in figure 2. As expected it shows different distributions for the S- and Cys-data. While the scores of Cys and S from the test data hardly overlap, we can see a bigger overlap between S training data scores and Cys scores. Nevertheless, most spectra can be separated clearly.

The boxplot in figure 2 shows $s_{max} = 0.0836$ for S. The other two thresholds are given by $t_{std} = 0.0815$ and $t_{80} = 0.0813$ (for a comparison see table 1), while all Cys-scores are located much higher.

Applying the different thresholds and considering our equally weighted method we get the overall confusion matrix in table 2 showing that our method can detect all Cys spectra, while only 9 Silicon spectra were wrongly classified as Cys, i.e., a 1.83% false positive rate.

Threshold	*f1-score*
s_{max}	98.69 %
t_{std}	99.91 %
t_{80}	99.03 %
overall	99.91 %

Table 1: Comparison of different thresholds

	Cys	*S*
Cys	5000	0
S	9	491

Table 2: Overall confusion matrix

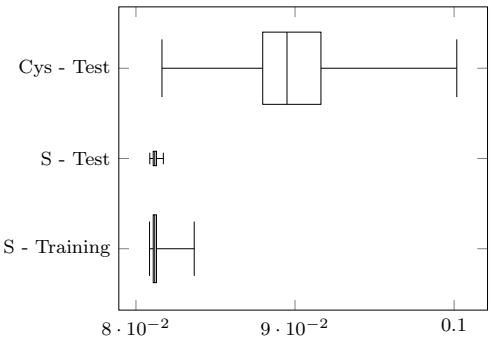

Fig. 2: Boxplots of reconstruction scores for Silicon/Cysteine

4.2 Glucose Oxidase in Reaction Buffer

Here, we also use 650 input nodes, and 64, 4 and 64 nodes, respectively, in the hidden layers. The learning rate was set to 0.3.

Some spectra and their reconstructions are plotted in figure 3. Compared to the S/Cys-dataset, there is much more noise (peaks and white noise), which is mainly filtered out by the reconstruction. Figure 5 shows different distributions for the classes' reconstruction scores. However, the overlap between the RB scores and the GOx scores are much higher than for the S/Cys-dataset.

Moreover, considering the training data has $s_{max} = 0.1374$, we cannot separate any GOx spectra by that value, since the highest GOx score is 0.1276. The other two thresholds are given by $t_{80} = 0.1139$ and $t_{std} = 0.1182$. The overall evaluation is given in table 4 and table 3.

The results show that while 98.4% of RB spectra are correctly classified, only 63.41% of the GOx spectra are detected as anomalies. The problem here is that the training data scores are much wider spread (see figure 5) - likely due to the

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

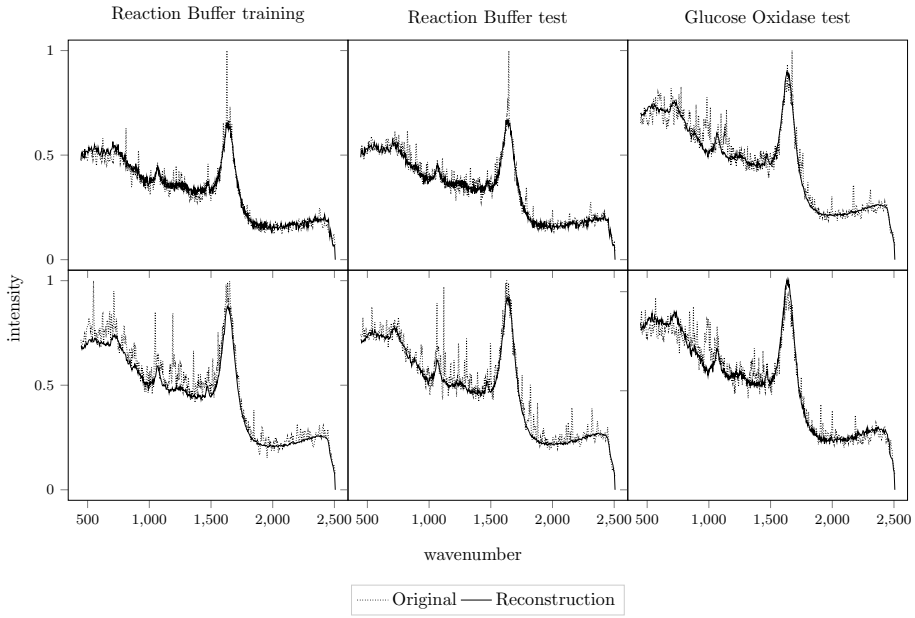

Fig. 3: Good (top row) and bad (bottom row) reconstruction examples

Threshold	*f1-score*
t_{std}	74.29 %
t_{80}	74.51%
overall	74.29 %

Table 3: Comparison of different thresholds

	GOx	*RB*
GOx	26	15
RB	3	197

Table 4: Overall confusion matrix

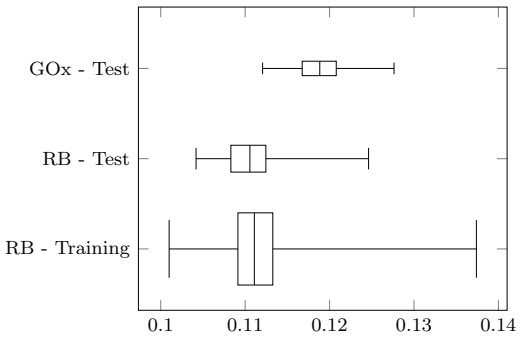

Table 5: Boxplots of reconstruction scores

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

dominant contribution of noise to the spectra. However, the reconstruction of RB spectra looks good and furthermore removes most of the noise.

5 Conclusion

Our results lead to good reconstructions of Raman spectra with promising classification results. The reconstructed spectra show that the model was able to learn the shape of major peaks of the spectrum where characteristic chemical information is contained. We consider to train our models with more data to improve anomaly detection on our RB/GOx dataset, as the overall available training data consisted of less than 2000 samples. We might also investigate the usage of (deep) Denoising Autoencoder for the future and define new, independent and possibly weighted thresholds especially to evaluate our approach on complex spectra of the electroporation transfection process.

References

[1] Kazuo Nakamoto. *Infrared and Raman Spectra of Inorganic and Coordination Compounds, Theory and Applications in Inorganic Chemistry*. Wiley Online Library, 2008.

[2] Joke De Gelder, Kris De Gussem, Peter Vandenabeele, and Luc Moens. Reference database of raman spectra of biological molecules. *Journal of Raman Spectroscopy*, 38(9):1133–1147, 2007.

[3] Jinchao Liu, Margarita Osadchy, Lorna Ashton, Michael Foster, Christopher J. Solomon, and Stuart J. Gibson. Deep convolutional neural networks for raman spectrum recognition: A unified solution. *Analyst*, 142(21):4067–4074, 2017.

[4] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2(1):8, Jun 2015.

[5] Sigurdur Sigurdsson, Peter Alshede Philipsen, Lars Kai Hansen, Jan Larsen, Monika Gniadecka, and Hans-Christian Wulf. Detection of skin cancer by classification of raman spectra. *IEEE transactions on biomedical engineering*, 51(10):1784–1793, 2004.

[6] Claire Lifan Chen, Ata Mahjoubfar, Li-Chia Tai, Ian K Blaby, Allen Huang, Kayvan Reza Niazi, and Bahram Jalali. Deep learning in label-free cell classification. *Scientific reports*, 6:21471, 2016.

[7] Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, Oct 2004.

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[9] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2Nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, pages 4:4–4:11, New York, NY, USA, 2014. ACM.

[10] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins, and Lifang Gu. A comparative study of rnn for outlier detection in data mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 709–712. IEEE, 2002.

[11] Hoang Anh Dau, Vic Ciesielski, and Andy Song. *Anomaly Detection Using Replicator Neural Networks Trained on Examples of One Class*, pages 311–322. Springer International Publishing, Cham, 2014.