Aiding the Detection of Explosive Materials with Machine Learning: Extracting Relevant Features from Multivariate Sensor Data

Yannick Meuser

Hochschule Darmstadt – Fachbereiche Mathematik und Naturwissenschaften & Informatik

Motivation

A reliable detection of explosives is of utmost importance in places of public life such as airports and train stations, or critical infrastructure like nuclear power plants. Often times with terrorist motives, an improvised explosive device is disguised as a harmless object. The explosive substance itself can be found as powdered, liquid or as a solid material, difficult to distinguish from a benign substance. In order to decide whether a suspicious object poses a greater risk to the general public, fractions of the unknown substance can be sampled for further analysis. For this purpose, sensor data from a special device for the detection of explosive and harmless substances is processed within the scope of this thesis. As a general goal, an already existing approach for distinguishing between explosives and benign substances [3] will be extended by two new use cases:

I) Detection of high-energetic, explosive materials

II) Identification of specific subgroups, based on the chemical structure of a substance

Both use cases provide an additional, higher level of detail about an unknown, potentially dangerous substance in the decision-making process. In the thesis, emphasis is placed on the extraction of features from sensor data, in order to be able to implement new Machine Learning models for each use case, based on a unified feature extraction process.

Methodology

In order to work with the provided sensor data, certain pre-processing steps are necessary. This includes the correction of outliers due to signal interference, a downsampling, and an averaging [4] procedure for the recorded measurements. For the feature extraction part, the time series are being transformed by a method called „Bag of Symbolic Fourier Approximation Symbols“ [5]. The resulting feature set represents each time series as a Bag-of-Words model. The following figures visualize examples of the features (*symbols*) extracted for an identical sensor but different substances. It is apparent that signals (*right*) with similar patterns show a resemblance in the distribution of symbols (*left*).

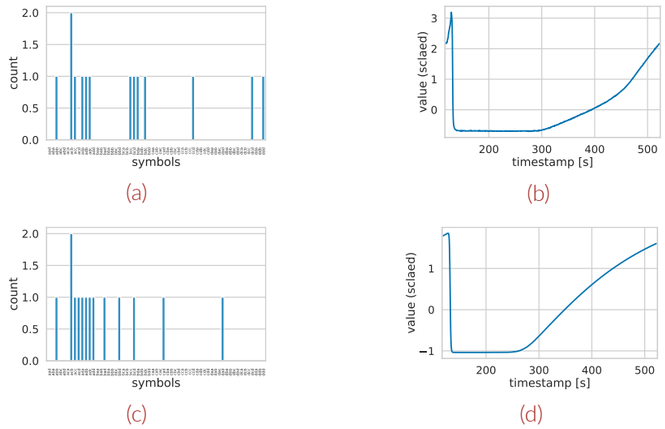

Figure 1. Histograms of the BOSS transformation (*a* & *c*) for similar sensor responses of two different substances (*b* & *d*).

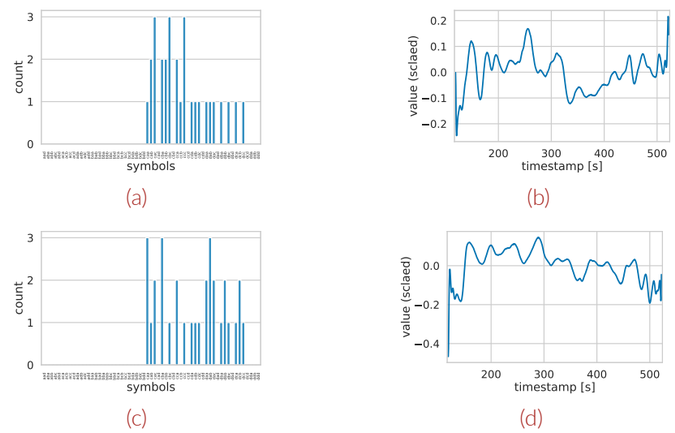

Figure 2. Histograms of the BOSS transformation (*a* & *c*) for similar sensor responses of two different substances (*b* & *d*).

This procedure is applied on various types of sensors (9 in total), which leads to a multivariate property for each sample of a substance. Using this BoW feature set allows for the application of various kinds of Machine Learning methods, like e. g. Latent Dirichlet Allocation [2]. In the work of the thesis, this method has been chosen as a dimension reduction technique for the BOSS feature set. In addition, the topic modeling approach is able to extract latent features from the data for a new feature representation. Finally, in a following step, K-Means++ [1] has been applied to cluster the obtained topics for each substance. This last clustering step is used as a predictor for the various use cases by utilizing available class labels with metrics to determine the optimal number of clusters in a Grid Search.

Results & Conclusion

The results of the implementation show different performance capabilities for each of both use cases. In the first one, good results could be achieved, while being able to isolate high-energetic substances from benign ones. The second use case was more difficult to implement, since the provided class labels are often times not as deterministic as for e. g. the binary classification of explosives and benign materials. Therefore, some of the cluster members overlap with other classes. Nevertheless, it is possible to separate some groups of substances from the rest. It can be shown that the framework of selected methods is able to perform well on the dataset. Regarding performance improvements, one of the most contributing factors would be to increase the size of samples for the available dataset.

References

[1] D. Arthur and S. Vassilvitskii.
K-Means++: The Advantages of Careful Seeding.
In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan.
Latent Dirichlet Allocation.
Journal of Machine Learning Research, 3:993–1022, 2003.

[3] K. Konstantynovski, C. Hammer, G. Njio, N. Wenzel, G. Holl, and T. M. Klapotke.
Library Free Bulk Detection of Explosives – Combining Simple Sensors for Resolving a Complicated Issue.
Unpublished (received on Oct. 2020), n. d.

[4] F. Petitjean, A. Ketterlin, and P. Gançarski.
A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering.
Pattern Recognition, 44(3):678–693, 2011.

[5] P. Schäfer.
The BOSS Is Concerned with Time Series Classification in the Presence of Noise.
Data Mining and Knowledge Discovery, 29(6):1505–1530, 2015.