Automatic classifiers for enhancing atmospheric data analysis

Jussi Tiira
Institute for Atmospheric and Earth System Research / Physics
University of Helsinki, Finland
jussi.tiira@helsinki.fi

ABSTRACT

In this project, examples of both supervised and unsupervised classification of atmospheric data are presented. The adventages and disadventages of automating classification tasks in atmospheric research are discussed.

KEYWORDS

classification, clustering, atmosphere

ACM Reference Format:
Jussi Tiira. 2019. Automatic classifiers for enhancing atmospheric data analysis. In *Data Science for Natural Sciences Seminar course reports*. University of Helsinki, Finland, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Analysis on atmospheric data often involves searching or classifying certain events or features in multidimensional data from large data sets. Practical examples include separating new-particle formation (NPF) events from non-events [Joutsensaari et al. 2018; Zaidan et al. 2018], classifying particle shapes in particle imager data [Lindqvist et al. 2012] and looking for signals of precipitation processes in weather radar data [Tiira and Moisseev 2018]. Due to noise and artifacts in measurement data, manual visualization method provides the best classification accuracy for each of the aforementioned tasks. However, manual classification of thousands of items would be intensely time consuming, and in case the class boundaries would need to be changed, one would have to start from the beginning. Additionally, human subjectivity reduces the repeatability of the classification. In these aspects, an automated classification method has potential of bringing major advantages over manual processing.

If classes are predefined, the term "supervised classification" can be used to dissociate these methods from clustering, which is often referred to as unsupervised classification. Unsupervised classification is be especially useful for discovering patterns in a large dataset or when theoretical knowledge on the classification subjects and the sought regularities is incomplete.

A typical classification pipeline consists of data cleaning and feature engineering and scaling followed by the actual classification algorithm. Selecting and configuring the individual parts of the pipeline appropriately is a critical but not trivial part of the analysis. In this report, examples of supervised and unsupervised classification are shown in the context of atmospheric research.

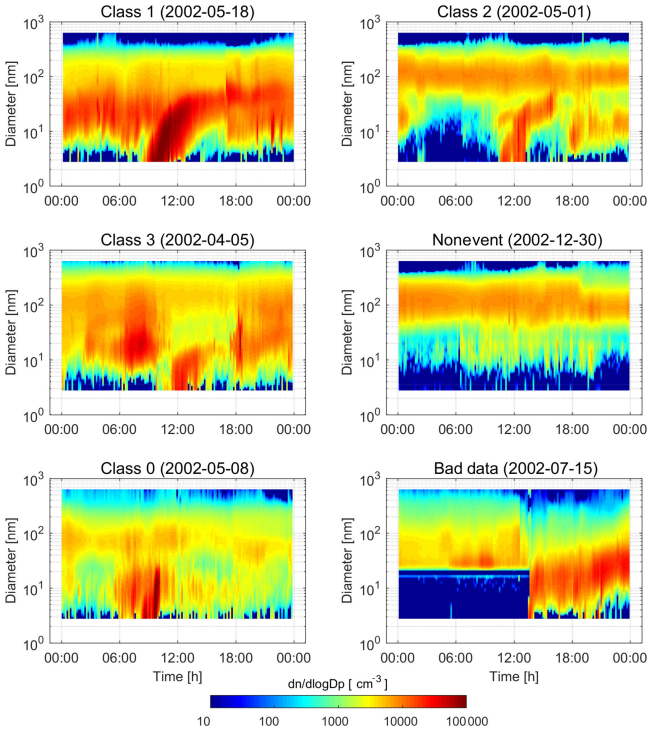

Figure 1: Examples of NPF classes. Adopted from Joutsensaari et al. [2018].

2 SUPERVISED CLASSIFICATION: IDENTIFYING NPF EVENTS

Supervised classification is used when class boundaries are predefined. Notable examples of supervised classification methods include support vector machines, nearest neighbors, naive bayes and decision trees. In this work, It is reviewed how a deep convolutional neural network (CNN) can be used for classification of NFP events, as shown by Joutsensaari et al. [2018]. They used a pretrained CNN called AlexNet [Krizhevsky et al. 2017] by transfer learning it to recognize NFP events. The pretraining had been conducted using millions of images of common items such as cars, fruits and animals, from the ImageNet database [Deng et al. 2009]. Given the pretraining, the model can be modified for NFP event classification using only hundreds of images for transfer learning.

Data Science for Natural Sciences, Spring 2019, Helsinki
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

An NPF event involves generation of nanometer scale aerosol particles through nucleation, which then grow to larger sizes. Traditionally, NPF event classification is done with 6 classes, visualized in Fig. 1:

- Class 1: A clear NPF event
- Class 2: Like class 1, but with weaker signal
- Class 3: There are signs of nucleation, but it is hard to say if the particles continue to grow or the growth seems to stop
- Nonevent: No NPF event occurs during the whole day
- Class 0: Days that don't meet the criteria to be classified to other classes
- Bad data: There was a problem in the measurement

Conventionally, the manual classification is done visually using similar plots as in Fig. 1. Thus, Joutsensaari et al. [2018] chose to use a CNN method as it mimics the visual method of human classifiers.

The overall accuracy of the classification for all classes was 63%. The classification model had difficulties in distinguishing between classes 1 and 2. The total classification accuracy is increased to 75% if these two classes are combined. Thus, Joutsensaari et al. [2018] recommend combining these classes when applying the classification method.

On analyzing the misclassifications of the automated method, Joutsensaari et al. [2018] found multiple examples where the misclassification was actually made by the human classifier and not by the CNN. They concluded that when combining classes 1 and 2, the CNN may be even more reliable than a human classifier.

3 UNSUPERVISED CLASSIFICATION: IDENTIFYING SNOW PROCESSES FROM RADAR PROFILES

Sometimes theoretical knowledge on the processes behind a dataset is not complete enough for defining the classes by hand. In such cases unsupervised learning can be used for discovering features and patterns in data. In this report we look at how a simple K-means clustering algorithm is used for classifying polarimetric weather radar profiles as reported by Tiira and Moisseev [2018]. The study aims to use simple machine learning methods for studying the structure of the profiles and to identify precipitation processes. Automated identification of patterns in radar profiles have potential for developing tools for weather services, such as filling gaps in radar measurements caused by objects such as buildings blocking the radar beams.

The vertical profiles of three polarimetric weather radar variables, equivalent reflectivity factor (Z_e), differential reflectivity (Z_{dr}) and specific differential phase (K_{dp}), are used in the analysis. The combination of these variables holds information on snowfall intensity and average snow particle sizes and shapes inside the measurement volume. In the analysis the variables are cleaned, normalized and combined. The resulting 577-dimensional vectors are dimensionally reduced using principal component analysis (PCA) such that each profile is described with 30 principal components. These components are used as classification features in K-means clustering to create a classification model.

Ambient temperature, along with humidity, is a major factor determining which precipitation processes occur in the atmosphere. Profiles are cut at the top of melting layer in order to standardize the profile base temperature to melting temperature. In case there is no melting layer, i.e. it is snowing on the surface, surface temperature is included as an extra classification parameter along with the 30 principal components described above. As a result, there are two separate classification models and the choice between them is made based on if it is raining or snowing on the surface.

The Tiira and Moisseev [2018] study is the first attempt to automate classification of vertical profiles of radar measurements. As such, it aims to use simple, repeatable and interpretable methods. Domain knowledge is used sparingly to avoid over-fitting and to allow wide applicability of the method e.g. in other locations.

The usage of unsupervised classification allowed Tiira and Moisseev [2018] to discover and interpret previously undocumented types of snow process signatures in the vertical profile measurements, such as dual dendritic growth layers in presence of an inversion layer. The method also produced profile classes which represent previously documented fingerprints of snow processes. Therefore it is suitable for automated analysis of snow processes in large time series of vertical profile data.

4 CONCLUSIONS

Classification problems are very common in atmospheric sciences, and automating these processes may save time and bring advantages in repeatability and even reliability [Joutsensaari et al. 2018].

Whether automating classification saves time compared to manual methods depends on multiple factors, such as size and complexity of the data set. The classification pipeline involves multiple steps that need to be implemented, optimized and tested. Feature extraction is one of the crucial steps. In the presented studies, this step was largely automated. AlexNet, used by Joutsensaari et al. [2018], had been pretrained to extract meaningful features in image data. Tiira and Moisseev [2018], in turn, used PCA components as features in their classification. Manually defining features to extract might be a time consuming process, but gives finer control. As a downside, with increased control often comes increased risk of overfitting and involving personal biases in the process.

If nothing else, automating classification brings repeatability. This is not only important for others to be able to repeat the results, but it may prove to be useful even in the original research. For example, the classification criteria might need to be modified for some reason. With manual classification, the whole process might need to be repeated, in contrast to just making the modifications to the automated classification pipeline and rerunning the computations. Repeatability does not always imply transparency and interpretability, as seen with black box classifiers.

When implemented in a meaningful way, machine learning classifiers mostly make misclassifications in cases that are difficult to interpret even for humans. Joutsensaari et al. [2018] and many others have found that human errors may result in flagrant misclassifications of even textbook example quality objects. Therefore, even a simple automated classifier with poor overall accuracy may be useful in a hybrid approach for spotting the most obvious mistakes of a human classifier.

REFERENCES

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern*

recognition. Ieee, 248–255.

J. Joutsensaari, M. Ozon, T. Nieminen, S. Mikkonen, T. Lähivaara, S. Decesari, M. C. Facchini, A. Laaksonen, and K. E. J. Lehtinen. 2018. Identification of new particle formation events with deep learning. *Atmospheric Chemistry and Physics* 18, 13 (July 2018), 9597–9615. https://doi.org/10.5194/acp-18-9597-2018

A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. https://doi.org/10.1145/3065386

H. Lindqvist, K. Muinonen, T. Nousiainen, J. Um, G. M. McFarquhar, P. Haapanala, R. Makkonen, and H. Hakkarainen. 2012. Ice-cloud particle habit classification using principal components. *Journal of Geophysical Research* 117, D16 (Aug. 2012). https://doi.org/10.1029/2012JD017573

J. Tiira and D. N. Moisseev. 2018. Fingerprints of precipitation processes revealed by unsupervised classification of profiles of polarimetric radar variables. In *10th European Conference on Radar in Meteorology and Hydrology (ERAD 2018)*. Wageningen, The Netherlands, 52.

M. A. Zaidan, V. Haapasilta, R. Relan, H. Junninen, P. P. Aalto, M. Kulmala, L. Laurson, and A. S. Foster. 2018. Predicting atmospheric particle formation days by Bayesian classification of the time series features. *Tellus B: Chemical and Physical Meteorology* 70, 1 (Jan. 2018), 1–10. https://doi.org/10.1080/16000889.2018.1530031