HUMBOLDT-UNIVERSITÄT ZU BERLIN

MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT

INSTITUT FÜR INFORMATIK

Methods on Open Set Recognition for time series data

Study Project Exposé

Melanie Wüstner - melanie.wuestner@student.hu-berlin.de

15. September 2022

Supervisors: Dr. rer. nat. Patrick Schäfer

Contents

1 Introduction 2

2 Background 2

2.1 Open set recognition 2

2.2 Time series classification 4

3 Objectives 4

4 Related Work 5

5 Approach 5

6 References 7

1 Introduction

The most common approaches for supervised machine learning operate under the closed world assumption. This means it is assumed, that every possible or existing class or label has been seen in the training data. If an unseen class appears in the test data or during the application of the machine learning system, it will predict one of the known classes with a high confidence. In contrast, open world or open set recognition methods aim to eliminate this issue of misidentified unknown classes, by identifying classes that have not been seen in training data and reacting accordingly. Regarding this matter much research has been done on open set recognition in the fields of image and object recognition. In their paper "Generalized Out-of-Distribution Detection: A Survey" Yang et al. [1] introduce a set of methods that among others implement approaches for open set recognition. However, these methods and most of the research address image recognition tasks. There barely are methods and extensions that can explicitly be used on time series data. Due to the different structure of time series data in comparison to image data the extension of the already existing methods to application on time series data is not trivial.

In this student research paper, the 34 methods on open set recognition in Yang et al.'s publication shall be filtered on to at most five methods that possibly are applicable on or easily extendable to time series data. For this purpose, two filtering stages, a prefiltering phase and a detailed filter, will be gone through. During the detailed filtering the implementations of the prefiltered methods will be executed on a benchmark of time series data. The final selection of methods will be explained in detail, the applied filter stages will be presented and the applicability of the methods or possible extensions per methods will be discussed.

2 Background

2.1 Open set recognition

The common supervised learning approaches learn from a given fixed labelled train data set without considering other classes that may appear during test time and therefore use the closed world assumption.

Closed World Classification: Classification under the closed world assumption assumes that all possible classes have been seen during training. Possibly occurring samples during testing phase would erroneously be classified into one of the known classes with high confidence [2].

However, this assumption is unrealistic in most real-world applications. Using a song recognizing application like Shazam as an example this assumption will not be correct with a significant probability in real world applications. In addition to the possibility of missing songs because of unpopularity new publications might appear daily. In addition, there is no guarantee of users only using the system on songs. The system could be exposed to humming of a human, animal noises or several other noise. These examples of unknown samples can be classified into two categories.

1) Semantic Shift: The semantic shift describes unknown samples that are part of the context in which the system operates [1]. Using the above example this would mean new or unpopular songs.

2) Covariate Shift: The covariate shift includes all unknown samples being out of context for the system, that means data that is from a different domain [1]. In the above example this would be equal to noises or recordings that are not songs.

As systems using the closed world assumption would classify these unknown samples wrong, this behaviour is not desirable. In contrast the open set recognition problem tries to solve this issue.

Open Set Recognition (OSR): OSR aims on the one hand to classify known samples that have been in the training data correctly and on the other hand identifying samples that have not been seen in the training data and therefore do not match into any known class [1].

This behaviour can be achieved using different approaches. Existing methods for open set recognition can be split up into three main categories, explained in the following.

1) Classification-based methods: The initial approach on identifying unknown instances simply uses the softmax confidence score as an indicator to determine which instances are known or unknown and therefore are classification-based.

2) Distance-based methods: Distance-based methods perform category-based clustering and prototyping on the data set in advance. Afterwards distances within clusters are calculated and used to expose samples that do not match into the clusters. Yang et. al [1] does not categorize distance-based methods further, however many of the listed methods use gaussian approaches for embedding or encoding.

3) Reconstruction-based methods: Reconstruction-based methods are based on the behaviour of embeddings or encoder-decoder architectures to yield measurable differences for samples it has been trained for and samples it has never seen. Reconstructing the original sample from these results then causes different behaviour which makes unseen samples detectable. For detecting the unseen samples two common approaches are used. Sparse representation methods encode samples and use their dense representation as an indicator for differences while reconstruction-error methods use the error between the reconstructed sample and its original, assuming that known samples produce higher quality reconstructions [1].

2.2 Time series classification

Time series: We define a time series as an ordered sequence of data points over time. This order must not necessarily be by time, thus time series are sometimes also called data series. If each data point consists of one single value, the series is called univariate. In contrast if each data point is a vector and therefore consists of multiple values the series is multivariate.

To classify time series data many techniques and algorithms have been introduced by previous research. Differentiating algorithms for classifying time series data by the feature types they aim to find in the data, as purposed by Bagnall et al. [3], yields into five basic method categories which will be explained in the following.

1) Shapelets: One approach for classifying time series data is to search for certain subsequences defining a class in the data, so called shapelets. The shapelets may appear at any point or phase of the data. A classification is then determined by the existence or non-existence of one or more shapelets.

2) Dictionary-based: Rather than classifying data based on the absence or presence of a pattern, dictionary-based approaches count reoccurring patterns and classify the data by the counted occurrences of patterns.

3) Distance-based: Instead of searching for exact patterns in a time series, distance-based algorithms compare two data series using a distance measure like the Euclidean distance to determine decision boundaries.

4) Interval-based: Algorithms based on intervals extract subsequences of random but fixed offsets and compute summary statistics such as mean or variance. Finally, a tree-based learner is used. These methods include a feature selection to determine expressive intervals.

5) Convolution-based: Similar to Shapelets convolutional-based algorithms search for convolutions in the data. The main difference between these two approaches is the space the convolutions or shapelets are searched in. While Shapelets are found in instances of the training data, convolutions are determined by searching the hole space of possible data.

3 Objectives

Considering the large set of methods for open set recognition on the image and object recognition problem referenced in the base paper by Yang et al. this student research paper aims to filter those methods that potentially are applicable to time series data and run as well as evaluate the implementations of the filtered methods. Prior to filtering and testing methods this paper shall explain the motivation of open set recognition and give a light introduction into classifying time series data in order to provide a base for the following main objectives:

1. From the large set of the 34 available methods for open set recognition filter down to up to five methods that are applicable to time series data and for which Python code is available.
2. Give a detailed explanation on the backbone and the functionality of the filtered methods.
3. Try to execute the existing implementations of the filtered methods on time series data.
4. Elaborate possible necessary changes on the methods or the implementation to be applicable to time series data.

From the aimed goals described above in addition the following side objectives derive, that shall be fulfilled in the process to the main objectives:

1. Introduce open set recognition and classify it into the machine learning context.
2. Point out the difference of time series classification to especially image recognition and provide a rough overview of to time series applicable machine learning concepts.
3. Determine and present a measure of the term "applicability of machine learning concepts to time series data"

4 Related Work

In their paper "Generalized Out-of-Distribution Detection: A Survey" Yang et al. [1] provide among others a list of 34 methods for open set recognition. This paper is used as a basis in order to get a set of methods that can be filtered and evaluated regarding their applicability to time series data. To the best of my knowledge, there is no paper yet giving an overview of extendable or adjustable open set recognition methods for time series data. However, some research has been done on providing a specific method for open set recognition on time series data [4]. Also the paper on an intra class-splitting method mentioned in the basis paper already provides an improved version of the backbone of the method, that is extendable on time series data [5]. In A. Bagnall et al. [3] a set of methods for general time series classification is introduced, which can be used to compare backbones and determine if the methods listed in the basis paper are possibly applicable to time series data.

5 Approach

In their paper Yang et al. [1] reference 34 methods on open set recognition and multi-class novelty detection. During the student research project these methods will be filtered down to at most five methods in two stages, that potentially are applicable on time series data. In the first stage the methods will be observed theoretically and tested on certain criteria, while in the second practical stage a test and training dataset for a) covariate shift and b) semantic shift will be created, and implementations will be executed. The two stages and their goals will be explained in detail in the following.

During the first stage, the methods will be prefiltered by the following criteria:

- Code is available in the preferred programming language Python and can be run on the data intended for the method.
- There is a reference to a possible extension to time series in the paper or the backbone of the method is generally applicable on time series data.
- The backbone of the method is provided by a common Python library like Scikit-Learn or SciPy.
- As many as possible different method categories are covered after consideration of the above criteria.

The goal of this stage is to identify at most ten methods that would potentially be suitable for the research project.

In the second stage, only the in the first stage prefiltered methods are considered and tested regarding the following process:

- Two benchmarks of time series data will be created using the UCR time series classification repository (http://www.timeseriesclassification.com/) containing a covariate and semantic shift.
- Found implementations of the prefiltered methods will be executed on the created benchmarks.

To create an expressive benchmark covariate as well as semantic shift will be included in the data. To achieve this first a main topic will be chosen for the benchmark, afterwards on the one hand samples will be included in the test data benchmark that have not been in the training data but belong to the topic chosen before. On the other hand, data from other domains than the chosen main topic will be included in the test data to achieve a semantic shift. In addition to the latter, it could be tried to even include non-time series data like serialized images in the test data.

The goal of the second stage is to determine methods for which the code provided is executable on the benchmark or for which the code can be made executable by small adjustments. After the second stage at most five methods should be left as the final set of methods applicable to time series data.

6 References

[1] J. Yang, K. Zhou, Y. Li, und Z. Liu, „Generalized Out-of-Distribution Detection: A Survey“. arXiv, 21. Oktober 2021. Zugegriffen: 5. Juli 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/2110.11334

[2] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, und T. E. Boult, „Toward Open Set Recognition“, *IEEE Trans. Pattern Anal. Mach. Intell.*, Bd. 35, Nr. 7, S. 1757–1772, Juli 2013, doi: 10.1109/TPAMI.2012.256.

[3] A. Bagnall, J. Lines, A. Bostrom, J. Large, und E. Keogh, „The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances“, *Data Min. Knowl. Discov.*, Bd. 31, Nr. 3, S. 606–660, Mai 2017, doi: 10.1007/s10618-016-0483-9.

[4] T. Akar, T. Werner, V. K. Yalavarthi, und L. Schmidt-Thieme, „Open Set Recognition for Time Series Classification“, in *Advances in Knowledge Discovery and Data Mining*, Cham, 2022, S. 354–366. doi: 10.1007/978-3-031-05936-0_28.

[5] P. Schlachter, Y. Liao, und B. Yang, „Open-Set Recognition Using Intra-Class Splitting“, in *2019 27th European Signal Processing Conference (EUSIPCO)*, Sep. 2019, S. 1–5. doi: 10.23919/EUSIPCO.2019.8902738.