When is Early Classification of Time Series Meaningful? (Extended Abstract)

Renjie Wu, Audrey Der, Eamonn J. Keogh
Computer Science & Engineering Department
University of California, Riverside
rwu034@ucr.edu, ader003@ucr.edu, eamonn@cs.ucr.edu

I. Introduction

The problem of *early classification of time series* (ETSC) generalizes classic time series classification to ask if we can classify a time series subsequence with sufficient accuracy and confidence after seeing only some prefix of a target pattern. The idea is that the earlier classification would allow us to take immediate actions, such as sounding an alarm or applying the brakes in an automobile. In this work, we make a surprising claim. In spite of the fact that there are dozens of papers on ETSC, it is not clear that any of them could ever work in a real-world setting. The issue is not with the algorithms per se, but with the vague and underspecified problem definition.

II. ETSC is Much Harder Than it Appears

Most ETSC papers consider only data in the UCR format, as shown in Fig. 1, assuming that all exemplars are of the same length and at least approximately aligned in time [1].

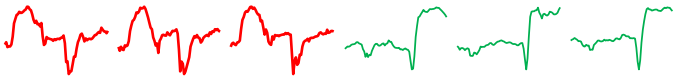

Fig. 1. Samples of data in the UCR format. The exemplars are of the same length and carefully aligned. The exemplars are utterances of the words cat and dog, spoken by a female in English, represented in MFCC Coefficient 2.

It is important to note that while our examples used natural language for simplicity, we have observed the following three issues in datasets containing gestures, writing, electrical power demand, etc., and in almost everywhere we looked.

A. The Prefix Issue

The prefix problem is the assumption that the pattern to be early classified is not a prefix of a longer innocuous pattern.

Consider what would happen when we test a ETSC model of {cat, dog} on the utterance "*It was said that Cathy's dogmatic catechism dogmatized catholic doggery*", as shown in Fig. 2.

This sentence will produce six false positives: three in each class. Note that we cannot simply recant the classifications *after* we see the rest of the longer word. The whole point of ETSC is to take *immediate* actions, otherwise in no sense are we doing *early* classification – we are just doing classification.

We believe that the prefix problem may be essentially insurmountable in many domains.

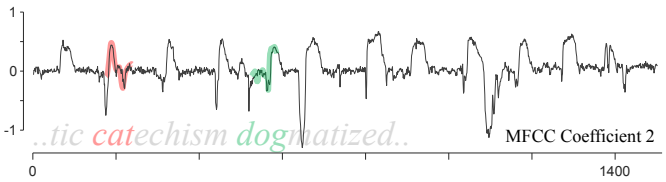

Fig. 2. A snippet of the phrase "*It was said that Cathy's dogmatic catechism dogmatized catholic doggery*". On this sentence, any ETSC method will make confident and early predictions, all of which will later have to be recanted.

B. The Inclusion Issue

The inclusion problem is the assumption that the pattern to be early classified is not comprised of smaller atomic units that are frequently observed on their own.

Suppose we learn a model for early classification of the vocalization of {lightweight, paperweight}. We can do very well after seeing the first 10% to 20% of these utterances.

However, suppose the universe contains sentences such as "*In the morning light, I could see that I got a papercut from the paper that the light was wrapped in.*" This sentence would give us two false positives for each class. Clearly, sub-patterns could be vastly more common than the full modeled pattern.

C. The Homophone Issue

The homophone problem is the assumption that two semantically different events will have different shapes in the time series representation.

Suppose that we train a model for early classification of the vocalization of {flower, wither}. Assume that *any* word containing the target word is also a true positive. This means we are completely free of the prefix and inclusion problems.

However, what about the following sentence from Leviticus 2:1 "*Whither anyone presents a grain offering as an offering to the Lord, his offering shall be of fine flour, and...*"? This sentence does not contain either of the target words, but it contains two near-perfect homophones, *flower* vs. *flour* and *wither* vs. *whither*, which would give us false positives.

III. Peeking into the Future

Because UCR datasets are z-normalized, almost all papers on ETSC suffer from a logical flaw that causes their accuracy to plunge when used on the streaming data. In a streaming

environment, you cannot do z-normalization until *after* you have seen all the data, otherwise it is not *early* classification.

Let us visit the ETSC community's favorite dataset, GunPoint [1]. As shown in Fig. 3, we produced a "denormalized" version of the testing data by adding to each instance a random number in the range [-1, 1]. It is important to understand how small of a change this is: approximately equivalent to tilting the camera randomly up or down by about 1.9 degrees.

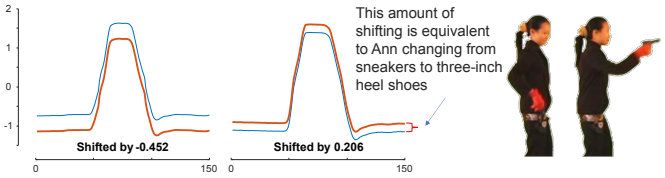

Fig. 3. Original examples from the GunPoint dataset together with denormalized versions, which have been slightly shifted in the Y-axis.

It is also important to note what effect this would have on normal nearest neighbor classification: *none*.

In Table I, we compute the accuracy of six ETSC algorithms on both normalized and denormalized GunPoint. We tested many settings and reported only the *best* results.

TABLE I
THE ACCURACY OF SIX EARLY CLASSIFICATION ALGORITHMS

Algorithm	Normalized	DeNormalized
(*min. support* = 0) ECTS [2]	86.7%	68.7%
(*min. support* = 0) RelaxedECTS [2]	86.7%	68.7%
EDSC-CHE [3]	94.7%	62.7%
EDSC-KDE [3]	95.3%	58.7%
(τ = 0.1) Rel. Class. [4]	90.0%	70.0%
(τ = 0.1) LDG Rel. Class. [4]	91.3%	71.3%

These results show that the algorithms can do apparently very well on GunPoint. However, when applying to streaming data, the accuracy will plunge. Distance measures are *brittle* to changes in the mean (and standard deviation) of the exemplars.

It is critical not to misunderstand this result. It is not that these algorithms forgot a step, and we can just add it back in. When the algorithms see a value, they are assuming that it is z-normalized based on other values that do not yet exist!

IV. Does Early Classification *Ever* Make Sense?

In our long search for a dataset that might work under ETSC assumptions, our best match was a dataset that consists of more than 12.5 billion datapoints of chicken behavior, measured using an accelerometer, as shown in Fig. 4 (*right*).

Consider the time series shown in Fig. 4 (*left*). It is an excellent template to detect chicken's behavior of dustbathing.

The time series shown in Fig. 4 (*center*) is a prefix of the first template. Classifying with this shorter template can achieve an accuracy that is not statistically significantly different from the accuracy achieved with the longer template.

However, this dataset cannot justify ETSC. We did not need any special algorithms to understand that the shorter template

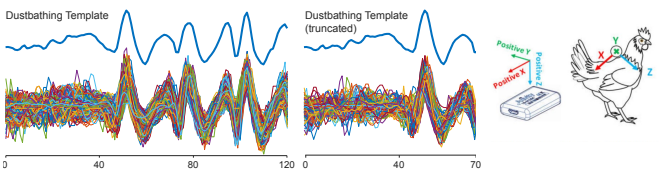

Fig. 4. (*left*) A template for dustbathing and its 500 nearest neighbors. (*center*) A truncated version of the template and its 500 nearest neighbors. (*right*) The data was obtained from a backpack sensor.

is as effective as the longer template. This took common sense and a few minutes of low-code exploration of the data.

Let us revisit the GunPoint dataset. As shown in Fig. 5, due to how GunPoint was created, the last one to two seconds are non-class discriminating sections. The difference between two classes in GunPoint mostly happens at the *beginning*.

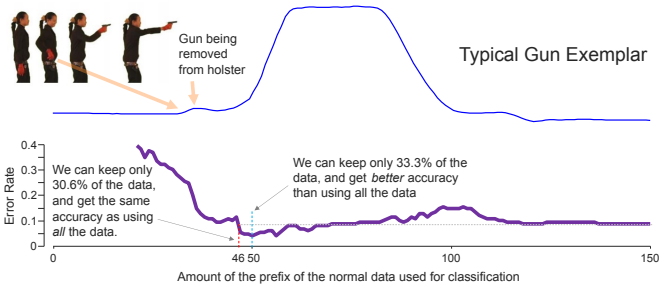

Fig. 5. (*top*) A typical example from GunPoint annotated to show where the discriminating region is. (*bottom*) The holdout classification error-rate of every prefix of the GunPoint data from lengths 20 to 150 (the full length).

A large number of UCR datasets have similar formatting conventions. Thus, it seems possible that some (possibly a very large) fraction of the apparent success of ETSC may be due to nothing more than a formatting convention: *padding*.

V. Conclusions

The commonly understood ETSC task may not be a meaningful problem to solve. All current research efforts that address this problem will be condemned to being overwhelmed by false positives if actually deployed in a real-world setting. Virtually all the algorithms are making the assumption that the data they are seeing *now* is normalized relative to data that only exists in the *future*. We believe that the issue is not with the proposed algorithms per se, but the intrinsically underspecified and vague definition of the problem itself.

References

[1] H. A. Dau *et al.* (2018) The UCR time series classification archive. [Online]. Available: https://cs.ucr.edu/~eamonn/time_series_data_2018/

[2] Z. Xing, J. Pei, and P. S. Yu, "Early classification on time series," *Knowledge and Information Systems*, vol. 31, no. 1, pp. 105–127, 2012.

[3] Z. Xing, J. Pei, P. S. Yu, and K. Wang, "Extracting interpretable features for early classification on time series," in *Proc. 2011 SIAM Intl. Conf. Data Mining*, 2011, pp. 247–258.

[4] N. Parrish, H. S. Anderson, M. R. Gupta, and D. Y. Hsiao, "Classifying with confidence from incomplete information," *J. Machine Learning Research*, vol. 14, no. 76, pp. 3561–3589, 2013.