Shapelet Chapter

1 Introduction

A *shapelet* of a particular time series object within a dataset that is distinctive to the class of which the time series belongs. Distinctive to a class means that it allows us to tell time series of that class apart from time series that are not, by some method. Shapelets were first introduced in Ye and Keogh (2009) and built upon in Mueen et al. (2011). The first paper proposed information theoretic and distance measure based definitions of *distinctness* in a 2-class context, giving an algorithm for Shapelet extraction. The second paper generalised to a multi-class classification context and proposed performance improvements. Both papers suggest a decision tree based classifier using *subsequence distance* (figure **??**) to training cases as features to build training rules.

Shapelets are interesting to the transient classification problem because they should be effective for classification and they are locale-independent. All previous methods explored rely on knowing the start and end points of an event to get good classification accuracy. Some transients (IDVs, XRBs, Flare stars) do not have well defined boundaries, so that is impossible. Additionally, sometimes only a fragment of a signal is available, and even if that fragment is highly characteristic of the class, no previous approach would function well unless it was trained explicitly on that fragment. Shapelets are robust in both of these scenarios and hence are worth exploring.

This chapter examines shapelets in the transient classification problem context, explores the basic algorithm in terms of classification performance, and finally proposes and compares modifications to both shapelet extraction and classification that may improve performance.

2 Experiments

There are some details related to the shapelet algorithm that need to be discussed before making proposals for how to incorporate them into a classifciation approach.

- The Ye 2009 paper determines only one shapelet per class as that having the best information theoretic measure of discrimination. The second paper proposes a find the best single combination of shapelets per class. Neither paper outlines a way to find the best N shapelets, very important for dealing with the high variability of our simulated data and the distortions we apply to it. A potential algorithm involving using clustering of shapelets and a user-defined clustering threshold of *shapelet distinctness* would enable multiple shapelet discovery, with each cluster would be a separate shapelet.

- The distance measures for determining the similarity between time series and shapelet candidate subsequences is a slightly modified Euclidean distance, called the *subsequence distance*. When there is even slight variability (but still a lot of similarity) amongst the phenomena we want the shapelet to represent, the distance measure will give poor results. A good example of such a phenomena is the sharp rise and peak of a Supernova transient. This structure occurs

over many time scales and its peakiness means that unless a flexible distance measure is used, the shapelet will not be as useful as it could be.

- When extracting single shapelets, the multi-class entropy defined in (Mueen et al., 2011) may choose a shapelet giving a good split for a class besides the source class for a shapelet, if such a subsequence happens to exist. This means an entire class will have no highly representative shapelets. A remedy is to use a one-vs-all binary entropy for each class, changing the non-source class labels to, say, B, and having the source class label as A. This forces the algorithm to choose a shapelet that works only for the source class, but may potentially miss useful shapelets for separating the dataset in more general ways. This is however essential in a single shapelet context. If multiple shapelets are used then multi-class entropy is preferable.

- Shapelets extracted from clean, normalised training data will not function well on distorted data. On clean data, even very subtle structures can be highly discriminative so long as they appear, in their subtlety, frequently within a class and not in others. These subtle shapelets become completely useless when noise is introduced (figure **??**. Similarly shapelets chosen from the latter part of light curves are useless if that part of the signal is not observed, and small shapelets with very strong variability are vulnerable to gappy data **??**. Clearly these complications will seriously hinder classifiction performance. A potential solution is to draw shapelets from clean light curves and find their discriminative power *amongst distorted lightcurve datasets*. The algorithm would then ignore shapelets that are sensitive to poor data conditions, and in the case of limited data, would choose shapelets appearing early in the time series.

- Speed concerns, use limited dataset

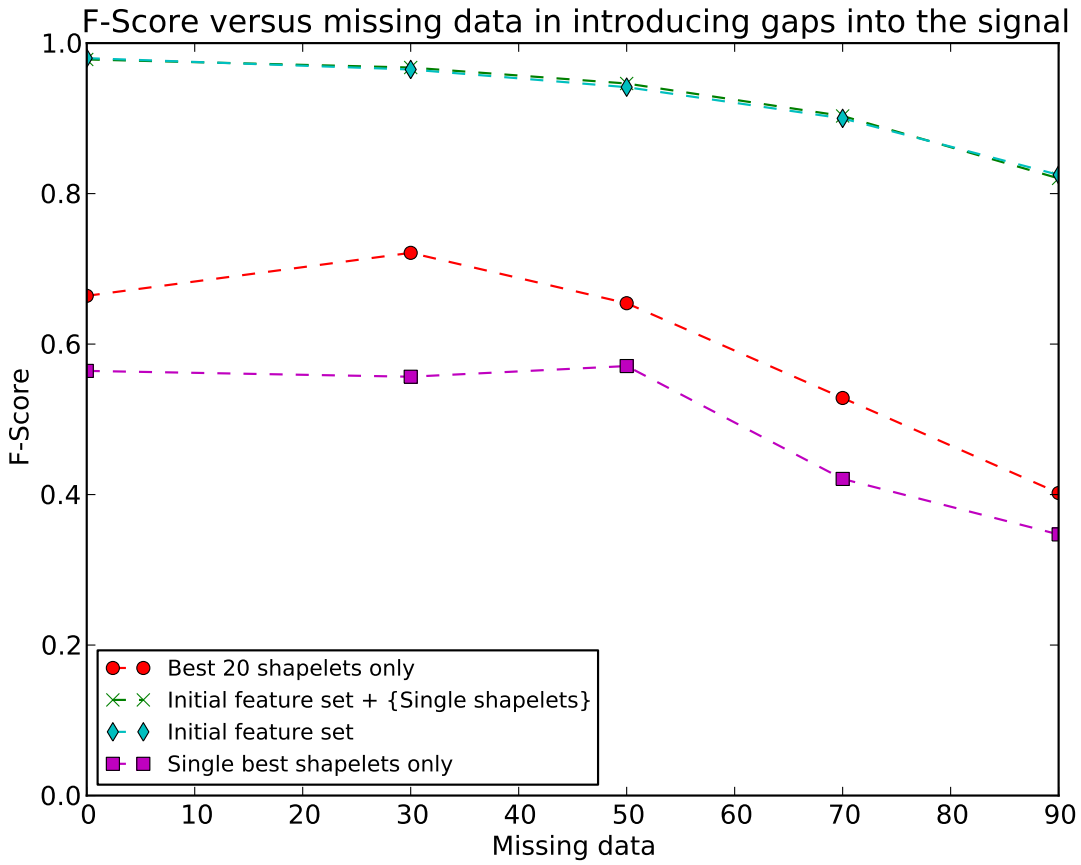
The following sections outline the experiments leading out of the above discussion.

2.1 Single shapelet per class

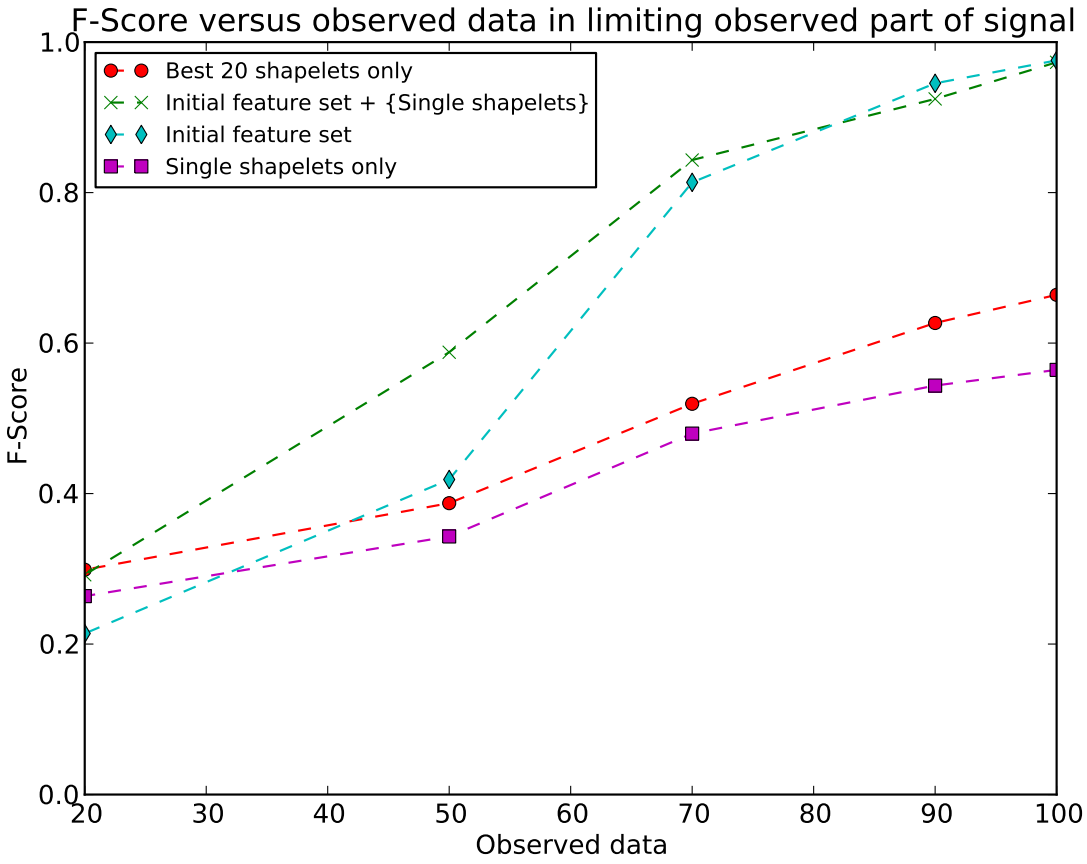
2.2 Multiple shapelets

3 Results

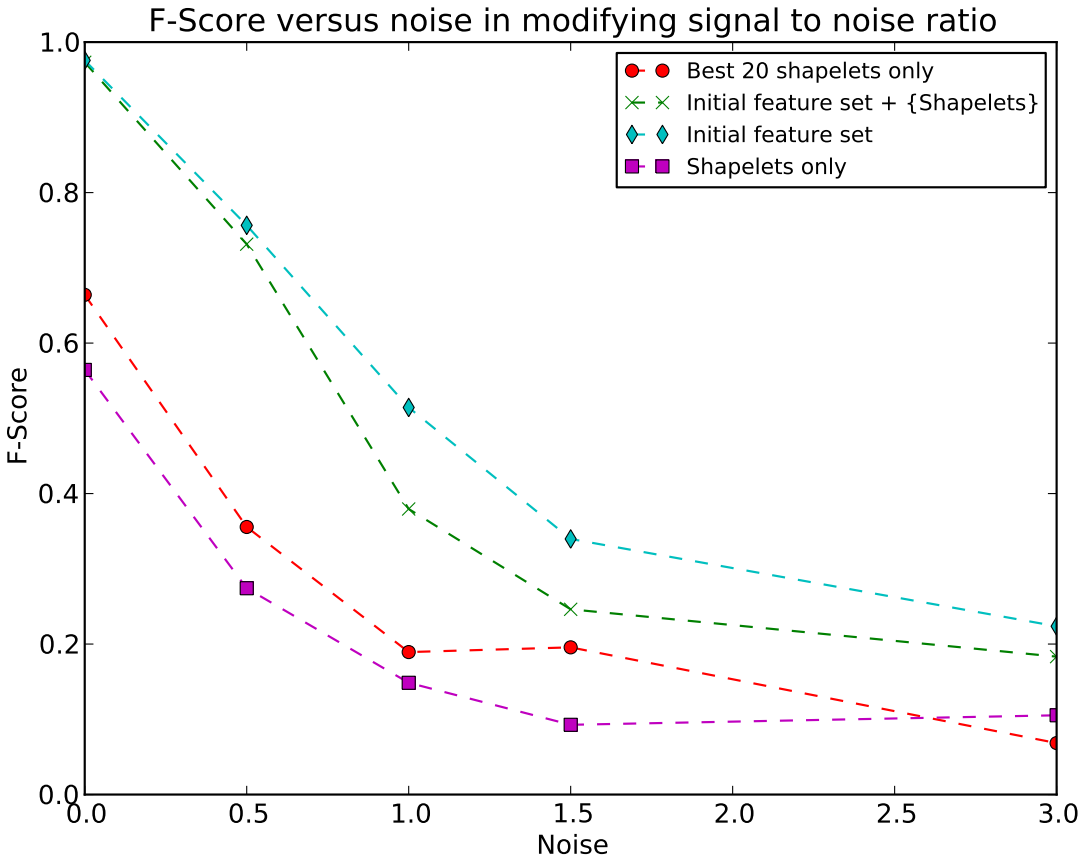
3.1 Introducing gaps into the light curve



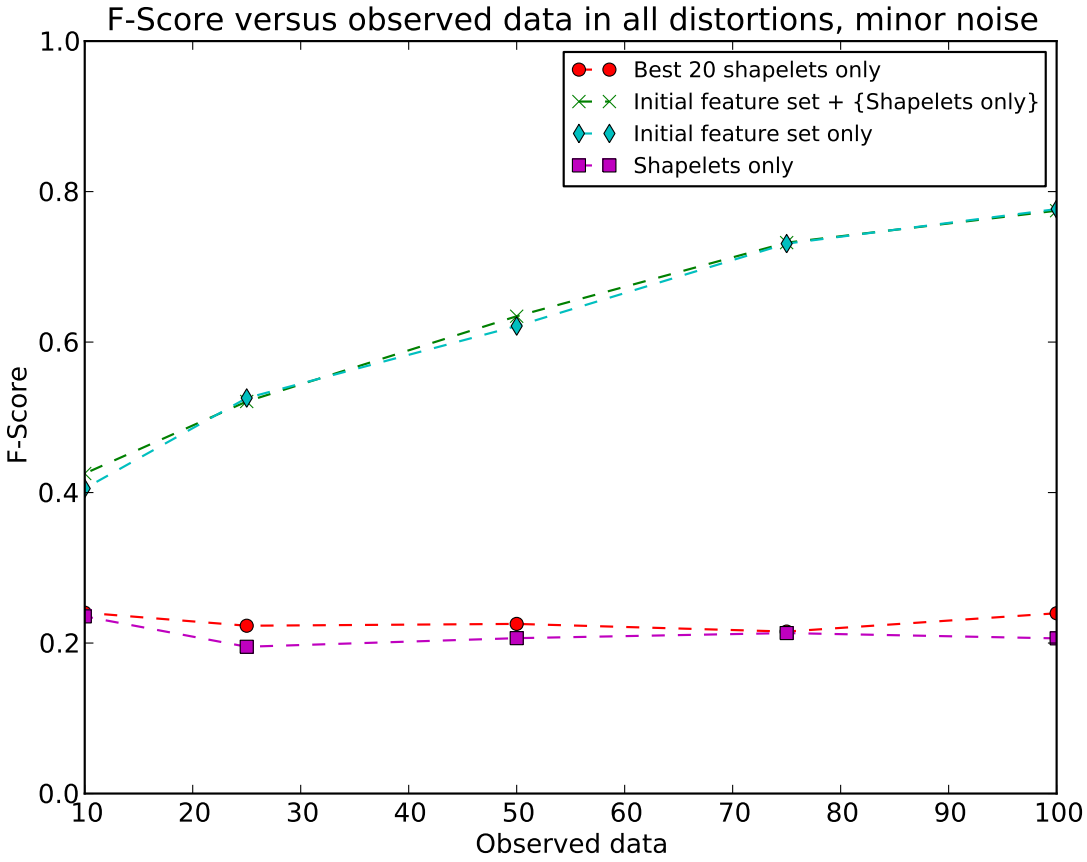
3.2 Limiting the amount of the light curve observed



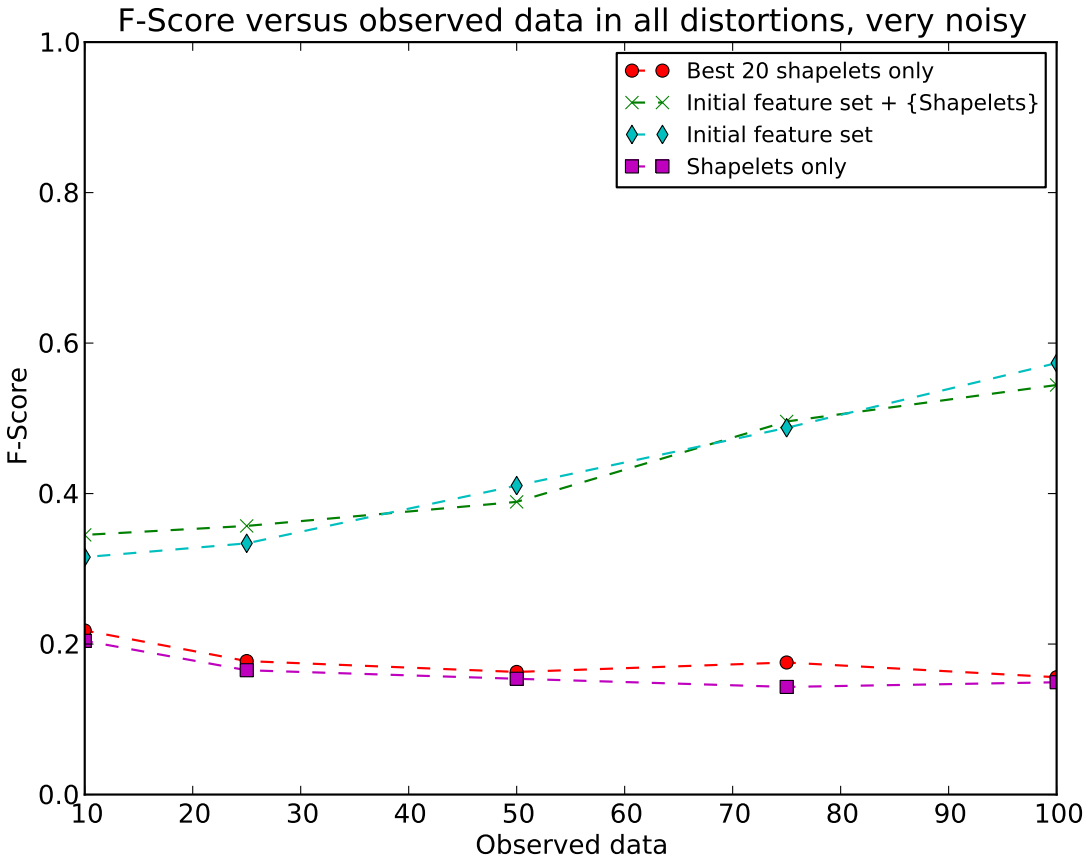
3.3 Introducing noise into the light curve



3.4 Light curves with 1 in 2 datapoints missing, power law applied, clear signal within noise



3.5 Light curves with 1 in 2 datapoints, power law applied, noisy signal



4 Discussion

4.1 Undistorted data

The performance of both sets of shapelets on the undistorted set of lightcurves is shown in figures **??**, **??** and **??** with an F-Score of approximately 0.58. This is a lower than expected result because there is so little deviation within the time series classes of the undisorted light curves. There are two possible explanations for this outcome:

1. There are some classes that have no subsequences (subject to our length constraints [15, 40, 65, 90, 105]) that are separated by the distance measure from all the other classes.

2. The shapelet extraction algorithm, trained on a subset of the training data, fails to choose general enough shapelets to accomodate slight variations in the testing data - a kind of over-fitting.

If the shapelet discrimination is poor on the training set, then the first issue is to blame for the poor performance. If the discrimination is good on the training set and poor on the test set, the second issue is to blame. I ran an additional experiment to evaluate the performance of the extracted shapelets on the exact training sets they were extracted from, with an F-Score of

4.2 Missing data

4.3 Multiple shapelets

5 Conclusion

References

A. Mueen, E. Keogh, and N. Young. Logical-shapelets: An expressive primitive for time series classification. 2011.

L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956. ACM, 2009.