Alternative Quality Measures for Time Series Shapelets

Jason Lines[1] and Anthony Bagnall[1]

School of Computing Sciences
University of East Anglia
Norwich
UK
{j.lines,anthony.bagnall}@uea.ac.uk
http://www.uea.ac.uk/cmp

Abstract. Classification is a very broad and prevalent topic of research within data mining. Whilst heavily related, time series classification (TSC) offers a more specific challenge. One of the most promising approaches proposed for TSC is time series shapelets. In this paper we assess the current quality measure used for shapelet extraction and introduce two statistical tests into the context of shapelet finding. We show that when compared to information gain, these two quality measures can speed up shapelet extraction whilst still producing classifiers that are not statistically significantly different to the original.

Keywords: time series, shapelets, classification

1 Introduction

Classification is a very broad and prevalent topic of research within the field of data mining. Whilst heavily related, time series classification (TSC) offers a more specific challenge. TSC typically involves problems where the ordering of the data plays a critical role, often where data have been recorded in temporal order at fixed intervals. Many solutions for TSC have been explored, with much of the contribution focused on alternative distance measures for 1-Nearest Neighbour (1-NN) classifiers using either raw time series or transformed representations of the raw data (a comprehensive summary can be found in [5]). In particular, there is strong evidence to support the use of 1-NN classifiers with a Euclidean or Dynamic Time Warping (DTW) distance metric. However, this approach suffers from drawbacks such as poor interpretability of results and relatively slow classification. As a result, many alternatives have been proposed. These include: shapelets [14, 17, 18], weighted DTW [8], support vector machines built on variable intervals [15], tree based ensembles constructed on summary statistics [4], fusion of alternative distance measures [2] and transform-based ensembles [1]. Of these, we feel that shapelets in particular have good potential for TSC due to their interpretability and fast classification of new cases.

Shapelets were first introduced in [18] as time series subsequences that are representative of class membership. The authors construct a decision tree classifier by recursively searching for the most discriminatory shapelet in a data set. They measure the quality of a shapelet by calculating the distance from a shapelet to each instance of data, storing the distances in sorted order, and then finding the point where information gain is maximised. In addition to this implementation, shapelets have also been used in many other applications, such as early classification [19], gesture recognition [6] and as a filter transformation for TSC [11]. For the purpose of this work we do not focus on a specific application of shapelets, but rather we investigate the algorithm used for initially selecting shapelets.

In this paper we investigate the shapelet quality measure used for shapelet extraction by [18]. Whilst it lends itself neatly to a decision tree implementation, we feel that the use of information gain (IF) to assess candidate shapelets involves more computation than is necessary. In response to this, we introduce two new statistical tests into the context of measuring shapelet quality: Kruskal-Wallis (KW) and Mood's Median (MM) tests. We demonstrate the validity of KW and MM for shapelet discrimination in two stages; firstly, we show that there is no significant difference between shapelet tree classifiers built with KW and MM when compared to an IG implementation of [18]. Secondly, we demonstrate that the computation time of the generic shapelet finding algorithm can be reduced by using either KW or MM as the quality measure.

2 Time Series Classification

A time series is a sequence of data that is typically recorded in temporal order at a fixed interval. For the problem of time series classification, suppose we have a set of n time series $T = T_1, T_2, ..., T_n$, where each time series t has m real-value ordered readings $T_i =< t_{i,1}, t_{i,2}, ..., t_{i,m} >$ and a class label c_i. For simplicity, we assume that all time series in T are of length m, but this is not a requirement for TSC. Given a set of data in the form of T, the problem of TSC is to find a function that maps from the space of possible time series to the space of possible class values. Whilst this problem is very similar to the general classification problem, TSC varies from generic approaches as it is often assumed that similarity between time series is to some extent embedded within the autocorrelation structure of the data.

As with all time series data mining, TSC relies to some degree on the use of a similarity measure to compare data. These typically fall into one of three broad categories: similarity in time (correlation-based); similarity in structure (autocorrelation-based); and similarity in change (shape-based). A detailed discussion of time series similarity can be found in [9] and [13].

Many shape-based applications of time series similarity use an elastic measure such as DTW with an instance based classifier (i.e. 1-NN with DTW). However, such an approach risks ignoring discriminatory shapes within a series as they may be masked by noise. This is one of the main strengths of shapelets for TSC; they

allow a mechanism for identifying phase-independent shape-based similarity on a local level, unlike global measures such as DTW that must calculate similarity across entire series.

3 Shapelets

Shapelets were first introduced in [14] to provide a mechanism for measuring the similarity of time series using subsections that are particularly indicative of class membership. There are three main components of shapelet discovery: candidate generation; a distance measure between a shapelet and a time series; and a measure of shapelet quality.

3.1 Generating Candidates

A shapelet candidate is any contiguous subsequence S of length l within a time series T_i of length m, where $l \leq m$. A series of length m contains $m - l + 1$ unique subsequences of length l. We denote the set of all subsequences of length l for series T_i to be $W_{i,l}$, and the set of all possible subsequences of length l for the data set to be $W_l = W_{1,l}, W_{2,l,}, ..., W_{n,l}$. The set of all candidates in T is $W = W_{min}, W_{min+1}, ..., W_{max}$ where $min \geq 1$ and $max \leq m$. For all possible lengths $l = 1, 2, ..., m$, there are a total of $m\frac{(m+1)}{2}$ shapelets in W. As this number can be very large with long series, [18] specify a minimum and maximum length parameter to constrain the search. The generic shapelet finding algorithm is defined in Algorithm 1.

Algorithm 1 ShapeletSelection (T, min,max)

```
1: bsfQuality = 0;
2: bestShapelet = ∅;
3: C = classLabels(T);
4: W = generateCandidates(T, min, max);
5: for l = min to max do
6:     for all subsequence S in W_l do
7:         D_S = findDistances(S, W_l);
8:         quality = assessCandidate(S, D_S);
9:         if quality > bsfQuality then
10:            bsfQuality = quality;
11:            bestShapelet = S;
12:        end if
13:    end for
14: end for
15: return bestShapelet;
```

Note that our implementation of Algorithm 1 independently normalises each element of W before using the distance function. We justify this as we are searching for local similarity between series, so wish to remove any offset caused by scale. Whilst no mention of this appears in [18], an amortised constant-time normalised distance measure is proposed in [14].

3.2 Shapelet Distance Calculations

The Euclidean distance between two subsequences S and R, where both are of length l, is calculated as:

$$dist(S, R) = \sum_{i=1}^{l} (s_i - r_i)^2. \quad (1)$$

The distance between a time series T_i and a subsequence S of length l is calculated using a sliding window to find the minimum distance between S and all possible subsequences in T_i of length l

$$d_{i,S} = \min_{R \in W_{i,l}} dist(S, R). \quad (2)$$

As $d_{i,S}$ is a minima, an early abandon is used to avoid unnecessary calculations. This calculation is used during shapelet extraction to calculate the distance from a candidate S to each time series in a data set T, $D_S = D_{S,1}, D_{S,2}, ..., D_{S,n}$, where n is the number of series in T. Note that [14] use a more efficient constant-time distance calculation based on maintaining a set of statistics. As the distance metric is incidental to the contribution of this paper, we retain the use of this simpler distance measure to keep the emphasis on shapelet quality measures.

4 Shapelet Quality Measures

The shapelet-finding algorithm defined in 1 requires an objective function for assessing shapelet quality, which is performed in [18] using information gain. This is where the main contribution of this paper lies; we believe that whilst information gain provides a good solution and lends itself neatly to the decision tree classifier implementation of [18], it involves an excessive amount of computation that could be removed using different shapelet quality measures. In response to this, we introduce Kruskal-Wallis and Mood's Median tests into the context of shapelet finding.

To assess the quality of shapelet S for data set T, a prerequisite of each quality measure is the a set of distances D_S must be calculated, where $D_S = D_{S,1}, D_{S,2}, ..., D_{S,n}$ and n is the number of instances in T.

4.1 Information Gain

Information gain [16] (IG) is a non-symmetrical measure of the difference between two probability distributions. The shapelet finding algorithm in [18] uses information gain to assess candidate shapelets. D_S is sorted and the information gain at each possible split point sp is assessed for S, where a valid split point is the average between any two consecutive distances in D_S. For each possible sp, IG is calculated by partitioning all elements of $D_S < sp$ into A_S, and all other elements into B_S. The information gain at sp is calculated as

$$IG(D_S, sp) = H(D_S) - \frac{|A_S|}{|D_S|} H(A_S) + \frac{|B_S|}{|D_S|} H(B_S) \quad (3)$$

where $|A_S|$ is the cardinality of the set A_S, and $H(A_S)$ is the entropy of A_S. Entropy is calculated by

$$H(D_S) = -\sum_{c \in classes\{D_S\}} p_c \log_2 p_c \tag{4}$$

The IG $info_S$ of S is calculated as

$$info_S = \max_{sp \in D_S} IG(D_S, sp). \tag{5}$$

Note that [18] introduce an upper-bound for calculating IG. However, in this paper we do not implement the early abandon. We justify this for two reasons; firstly, in the most pessimistic cases for multi-class problems, the computation involved for implementing a nave approach of the upper-bound would far out way the benefits provided by it. Secondly, the style of upper-bound used by [18] could also be implemented for KW and MM. We wish to directly compare the three quality measures, so by not using an early abandon, any implementation of the three quality measures must evaluate the same number of shapelet candidates.

4.2 Kruskal-Wallis

Kruskal-Wallis [10] (KW) is a non-parametric test to observe whether data originates from a single distribution. The calculated statistic represents the squared-weighted difference between ranks within a class and the global mean rank. For use with shapelets, KW is calculated for S as

$$KW_S = \frac{12}{|D_S| \cdot (|D_S| + 1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(|D_S| + 1) \tag{6}$$

where $|D_S|$ is the cardinality of D_S, k is the number of classes in D_S, R_i is the sum of ranks for class i and n_i is the number of instances of class i in D_S. Note that in order to calculate ranks, D_S must be sorted as it was with IG. However, we believe that KW will be more efficient for shapelet finding than IG because the statistic only needs to be calculated once, rather than for each possible split point in D_S.

4.3 Mood's Median

Mood's Median [12] (MM) is a non-parametric test to determine whether the medians of two samples originate from the same distribution. Unlike IG and KW, MM does not require D_S to be sorted, so therefore should be faster in that respect. Only the median is required for calculating MM, which can be found in $O(n)$ time using quickselect [7]. The median is used to create a contingency table from D_S, where the counts of each class above and below the median are

recorded. The MM statistic is obtained by calculating the Chi-Squared statistic of the table

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{7}$$

where r and c are the rows and columns of the contingency table and o_{ij} and e_{ij} are the observed and expected values of row r, column c respectively.

5 Experimental Procedure

The experiments in this paper are designed to establish the validity and advantages of using KW and MM for shapelet finding. This is demonstrated in two stages; firstly, we use a diverse range of data to build shapelet decision trees akin to [18] using IG, KW and MM as quality measures, and show that the classifiers produced are not statistically significantly different. Secondly, we perform timing experiments to show the relative time performance of KW and MM compared to IG for finding the most discriminatory shapelet in a data set.

5.1 Shapelet Classifier Implementation

We implement four distinct shapelet tree classifiers; the first uses IG as the shapelet quality measure as in [18]; the second uses KW; and the final two use MM. We slightly modify the algorithm for KW and MM classifiers due to the nature of the statistics calculated. In the KW tree, we use the quality measure to find the best shapelet, but the value is calculated from a whole set of distances and no split point is implied. Therefore, once we establish the best shapelet we use a single set of IG calculations to find the best split point. We justify this because the costly IG calculations for each candidate are replaced by KW and we only use IG one on the best shapelet. For MM, we implement two classifiers; the first simply uses the median from the MM calculation of the best shapelet as the split point, whilst the second the same approach as the KW classifier to identify the best split point using IG.

The minimum and maximum shapelet lengths for each data set were computed using the simple cross-validation approach in [11]. The parameters vary across data sets, but are consistent for each classifier to ensure the same number of candidates are evaluated by each for a fair comparison.

6 Results

The results that we report are split into two sections; firstly we wish to demonstrate that Kruskal-Wallis and Mood's median are valid statistics for measuring the quality of shapelets. We demonstrate this through a number of classification experiments and report the error rates of a diverse range of data sets. Secondly, we wish to demonstrate that these new quality measures speed up shapelet discovery; we demonstrate this with a number of timing experiments using the same data sets.

6.1 Classification Performance

The results in Table 1 show that whilst the IG classifier achieves the top rank on more data sets than any other classifier (10 of 26), it is in fact the MM with IG tree that has the best overall rank. The KW tree also has a better overall rank than IG, whilst MM using the median to split has the lowest overall rank. This supports our decision to use IG to find the best split point. To further demonstrate the validity of KW and MM as quality measures, we show that there is no statistically significant difference between the classifiers in Figure 1 using a critical difference diagram (as decribed by [3]). The diagram is derived from the overall test of significance of mean ranks where classifiers are grouped into *cliques*, represented by solid bars. The diagram shows that all classifiers are part of a single clique, and therefore are not statistically significantly different. This supports our claim that MM and KW are valid metrics of shapelet quality.

Table 1: Classification error rates for the shapelet tree classifiers

Data Set	IG	KruskalWallis	MoodMedian	MoodMedIG
Adiac	**0.7008(1)**	0.734(3)	0.7928(4)	0.7289(2)
Beef	**0.5(1)**	0.6667(2.5)	0.6667(2.5)	0.7(4)
ChlorineConcentration	**0.412(1)**	0.474(3)	0.4648(2)	0.4789(4)
Coffee	**0.0357(1)**	0.1429(3)	0.1429(3)	0.1429(3)
DiatomSizeReduction	**0.2778(1)**	0.3889(2)	0.5392(3)	0.5523(4)
DP_ Little	0.3456(4)	0.32(3)	**0.2567(1)**	0.29(2)
DP_ Middle	0.2947(2)	0.3067(3)	0.35(4)	**0.2633(1)**
DP_ Thumb	0.4189(4)	**0.28(1)**	0.3233(3)	0.2967(2)
ECGFiveDays	0.2253(4)	0.1278(2)	0.1568(3)	**0.072(1)**
ElectricDevices	0.451(3)	**0.4416(1)**	0.4492(2)	0.5317(4)
FaceFour	**0.1591(1)**	0.5568(2)	0.5795(3)	0.5909(4)
GunPoint	0.1067(4)	**0.06(1)**	0.1(3)	0.08(2)
ItalyPowerDemand	0.1079(3)	0.0904(2)	0.1322(4)	**0.0894(1)**
Lighting7	**0.5068(1)**	0.5205(2)	0.7671(4)	0.726(3)
MedicalImages	0.5118(3)	0.5289(4)	**0.5(1)**	0.5105(2)
MoteStrain	0.1749(4)	0.1605(2)	0.1605(2)	0.1605(2)
MP_ Little	0.3361(4)	0.3033(3)	**0.2667(1)**	0.2967(2)
MP_ Middle	0.2899(4)	**0.25(1)**	0.2867(3)	0.28(2)
PP_ Little	0.4036(4)	**0.28(1)**	0.3433(3)	0.3267(2)
PP_ Middle	0.3858(4)	0.3167(3)	0.31(2)	**0.3033(1)**
PP_ Thumb	0.3917(4)	0.2867(3)	**0.2667(1)**	0.27(2)
SonyAIBORobotSurface	**0.1547(1)**	0.2729(4)	0.2479(2)	0.2512(3)
Symbols	**0.2201(1)**	0.4432(4)	0.4201(2)	0.4261(3)
SyntheticControl	**0.0567(1)**	0.1(2)	0.1867(4)	0.1433(3)
Trace	0.02(2)	0.06(3)	0.08(4)	**0(1)**
TwoLeadECG	0.1493(3)	0.2362(4)	**0.1343(1)**	0.1466(2)
Mean Rank	2.5385	2.4808	2.5962	2.3846

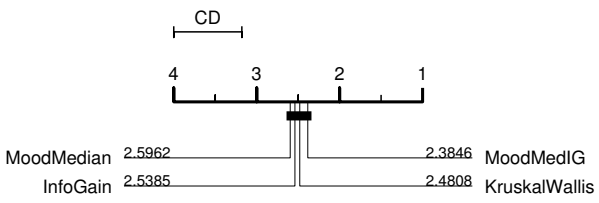

Fig. 1: Critical difference plot for the four different shapelet tree classifiers

6.2 Timing Results

The results in Table 2 were produced using IG, KW and MM to find the best shapelet from each data set. This approach was adopted to ensure fair comparisons could be made between measures, as comparing build times of trees would be biased if they produced classifiers of different depths. Extracting a single shapelet ensures that the same number of candidates are processed for each quality measure.

Table 2: Relative computation times of KW and MM against IG

Data Set	Kruskal-Wallis	Mood's Median
Adiac	0.2723	0.2644
Beef	1.0281	0.9815
ChlorineConcentration	0.6134	0.5735
Coffee	1.0217	0.9716
DiatomSizeReduction	1.0211	0.9803
DP_ Little	0.9319	0.8922
DP_ Middle	0.5339	0.5103
DP_ Thumb	0.9498	0.9034
ECGFiveDays	0.9985	1.0021
ElectricDevices	0.7978	0.7587
FaceFour	1.0471	1.0129
GunPoint	1.0377	1.0120
ItalyPowerDemand	0.5081	0.4903
Lighting7	0.9874	0.9625
MedicalImages	0.5148	0.2355
MoteStrain	1.0149	0.9457
MP_ Little	0.9575	0.9032
MP_ Middle	0.9851	0.9337
PP_ Little	0.9395	0.9008
PP_ Middle	0.9497	0.9000
PP_ Thumb	0.9508	0.8982
SonyAIBORobotSurface	0.9332	0.9715
Symbols	1.0290	1.0132
SyntheticControl	0.4421	0.4131
Trace	0.9753	1.0159
TwoLeadECG	0.9135	0.9099
Average	0.8598	0.8214

There are few cases where IG is fastest, and even in these cases the difference is marginal. It is clear that KW and MM perform much better on some data sets whilst providing at least a modest speedup on the majority of cases. MM is the fastest overall and provides almost an 18% speed-up over IG, whilst KW also provides a marked improvement of approximately 14%. On first glance this may not seem significant, but shapelet extraction can be time consuming and can potentially take hours in some cases, so an improvement of almost 20% is important.

7 Conclusions and Future Work

In this paper we have introduced two new quality measures for shapelets in TSC. We demonstrated the effectiveness of the Kruskal-Wallis and Mood's Median statistics as discriminatory measures by using them to build shapelet decision tree classifiers in the style of [18]. We used these classifiers to illustrate two points; firstly, using these alternatives to information gain does not degrade the discriminatory power of the shapelets that are extracted. This is demonstrated

by producing classifiers that are shown not to be statistically significantly different over 25 data set. Secondly, we limit the shapelet finding algorithm to extract only the best shapelet from each data set, allowing us to directly compare the computation times of the three statistics. Our results show an average improvement in computation time across the 25 data sets of approximately 14% and 18% for Kruskal-Wallis and Mood's Median respectively. With a view to the future, we can investigate the potential of these alternative quality measures in further applications of shapelets, such as extending a shapelet filter for TSC [11].

References

1. A. Bagnall, L. Davis, J. Hills, and J. Lines, *Transformation based ensembles for time series classification*, Proc. 12th SDM, 2012.
2. K. Buza, *Fusion methods for time-series classification*, Ph.D. thesis, University of Hildesheim, Germany, 2011.
3. J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, JMLR **7** (2006), 1–30.
4. H. Deng, G Runger, E. Tuv, and M. Vladimir, *A time series forest for classification and feature extraction*, Tech. report, Arizona State University, 2011.
5. H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, *Querying and mining of time series data: Experimental comparison of representations and distance measures*, Proc. 34th VLDB, 2008.
6. B. Hartmann and N. Link, *Gesture recognition with inertial sensors and optimized DTW prototypes*, Proc. IEEE SMC), 2010.
7. C.A.R. Hoare, *Quicksort*, The Computer Journal **5** (1962), no. 1, 10–16.
8. Y. Jeong, M. Jeong, and O. Omitaomu, *Weighted dynamic time warping for time series classification*, Pattern Recognition **44** (2010), 2231–2240.
9. E. Keogh and S. Kasetty, *On the need for time series data mining benchmarks: A survey and empirical demonstration*, Data Mining and Knowledge Discovery **7** (2003), no. 4, 349–371.
10. W.H. Kruskal, *A nonparametric test for the several sample problem*, The Annals of Mathematical Statistics **23** (1952), no. 4, 525–540.
11. J. Lines, L. Davis, J. Hills, and A. Bagnall, *A shapelet transform for time series classification*, Tech. report, University of East Anglia, UK, 2012.
12. A.M.F. Mood, *Introduction to the theory of statistics.*, (1950).
13. F. Mörchen, I. Mierswa, and A. Ultsch, *Understandable models of music collections based on exhaustive feature generation with temporal statistics*, Proc. 12th ACM SIGKDD, 2006, pp. 882–891.
14. Abdullah Mueen, Eamonn J. Keogh, Qiang Zhu, Sydney Cash, and M. Brandon Westover, *Exact discovery of time series motifs*, SDM, 2009, pp. 473–484.
15. J. Rodriguez and C. Alonso, *Support vector machines of interval-based features for time series classification*, Knowledge-Based Systems **18** (2005).
16. C. E. Shannon, *A mathematical theory of communication*, Bell System Technichal Journal **27** (1948), 379–423, 623–656.
17. L. Ye and E. Keogh, *Time series shapelets: A new primitive for data mining*, Proc. 15th ACM SIGKDD, 2009.
18. ———, *Time series shapelets: a novel technique that allows accurate, interpretable and fast classification*, Data Min. Knowl. Discov. **22** (2011), no. 1-2, 149–182.
19. P. Yu K. Wang Z. Xing, J. Pei, *Extracting interpretable features for early classification on time series*, Proc. 11th SDM, 2011.