论文题目：基于最佳u-shapelets的时间序列聚类算法

**作者：盛锦超 1032012026064 电子信息**

**1.写作背景**

这是一篇关于时间序列shapelet的文章；2009提出shapelet的思想并在2012年提出u\_shapelet；shapelet是时间序列中的最具有特点的子序列，具有区分类和强解释性的作用，主要应用于金融学、医药学、气候学和生物学等领域；u\_shapelet是shapelet在无监督领域的拓展，相较于shapelet，u\_shapelet更适用于无标签或获取标签困难的场景，能够实现时间序列数据的准确划分

强解释性是指在分类或聚类的情况下，具有与shapelet相关的类被划分的时候，可以用shapelet来解释类的特点，比如鸟叫，鸟在某段时间内的叫声模拟波动是不同的，通过截取这段子序列，就可以解释该时间序列属于哪一类

**2.文章介绍**

**引言：**

本文主要介绍的u\_shapelet的加速算法，先介绍了最基础的u\_shapelet算法和一些算法改进

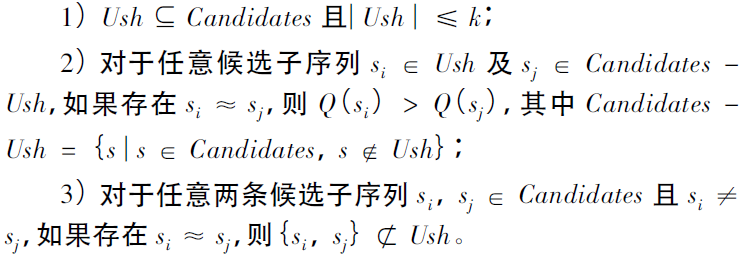
u\_shapelet的最基础算法是ZAKARIA提出来的暴力法，具体见算法1；算法1中首先选择了第一条时间序列，如何通过长度的循环和子序列的选择两重循环计算每一个候选shapelet的gap和dt，dt是将该序列与其他时间序列计算得到的距离序列dis后将其分为左右两类的阈值，通过这个阈值dt就可以计算gap，gap的公式如下：，分别是左右两边的平均值和标准差；然后找最大的gap，将其纳入u\_shapelets集合中，然后将一些与本类类似的时间序列过滤掉；直到出现左边的类只有自己一条时间序列的时候，算法结束；该算法可以自动选择多少条时间序列，复杂度是O(n²m四次方)；对于修改该算法，算法2从SAX即符号聚类近似的角度进行修改，见算法2；其中是对所有相同长度的子序列进行符号化并进行随机掩码，过滤均值和方差不在阈值中的子序列，最后求出最符合的u\_shapelet,

其中选取了1%的子序列作为候选集，通过实验证明了可以在不损失明显精度的情况下提高速度；本文则更多注重精度且保持算法效率与算法2保持一致

提高精度在两方面做调整；第一是对评估方法即gap做出调整，通过实验证明了一种更好的评估标准；第二是引入多元top-k的调整方式，这是一种筛选过程，但这篇文章在这方面给出的内容太少，也是后面我想要思考的重点，稍后再说

**相关定义：**

文章主要介绍了一些定义，如sdist子序列距离：子序列与时间序列之间的距离，通过ED计算，取最小值、距离矩阵Dist：一个N\*m的矩阵，放着N条时间序列和m条shapelet各自的距离，还有一个多元top的概念和基于shapelet1的表达：



实际上就想要表达Ush里面的shapelet彼此不相似且具有更好的质量这件事情

其次介绍了暴力法，这里的暴力法与算法1的唯一区别是这里一开始搜索的是所有时间序列而不是取一条时间序列，可能一开始的复杂度会变高，但其实是有帮助的，即你选择的时间序列可能本质上是一个比较差的时间序列，可能是唯一类，就直接结束搜索了；这样可以保证算法更加稳定运行

**本文算法：**

第一个调整介绍了四种评价指标，除了gap还有均方根标准差RMSSTD、RS、I指标；具体公式见本文章

其中思想是gap和RS考虑的是子集之间分离度的大小，RMSSTD考虑的是子集内部的紧密性，而I指标是同时考虑了分离度和紧密型，最后通过实验得到了I指标更具聚类的准确度

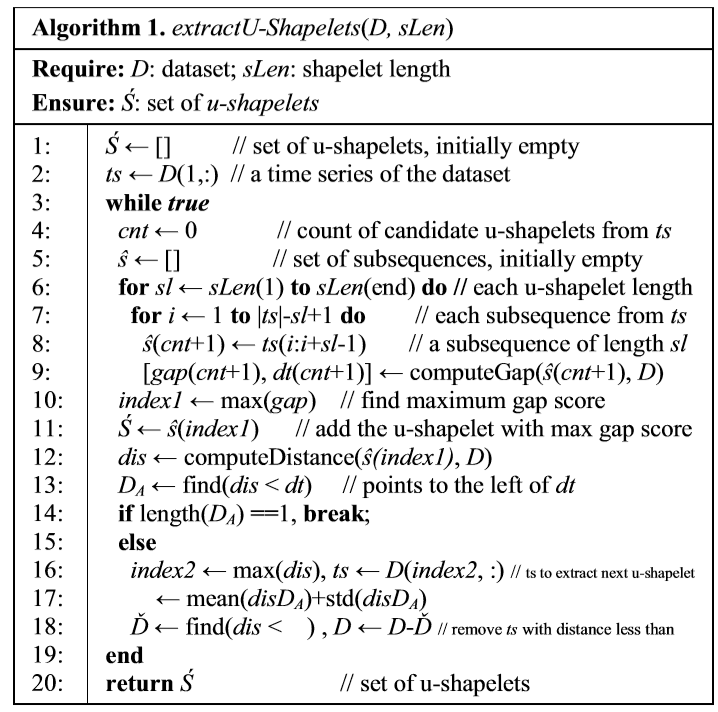
第二个调整是提出新的算法DivUshapCluster算法，该算法见算法3；算法首先通过算法2得到一些具有显著特点的候选集，然后评价候选集，找到最大的候选集，然后将与这个候选集相似的过滤掉，重复多次直到满足循环条件

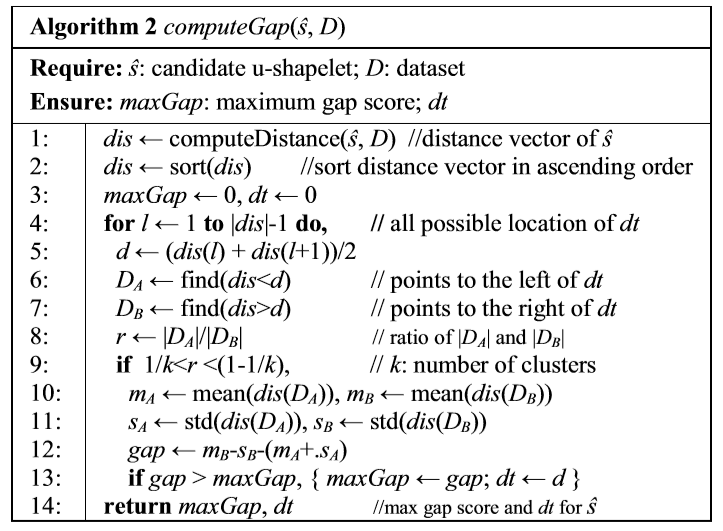
**实验与结论：**

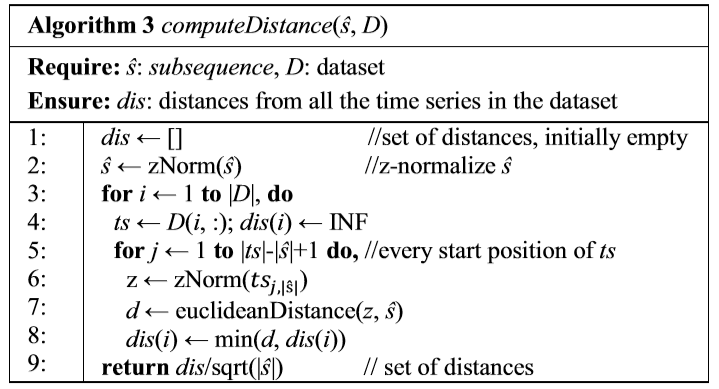
实验是将本算法与暴力法和SUSh进行比较，但作者还与经典的层次聚类，k-means进行聚类，这里不做介绍，最后结论是本算法比其他两个算法整体比较上效果更好(运行时间短且聚类效果好)，聚类的指标用了RI

**3.算法描述**

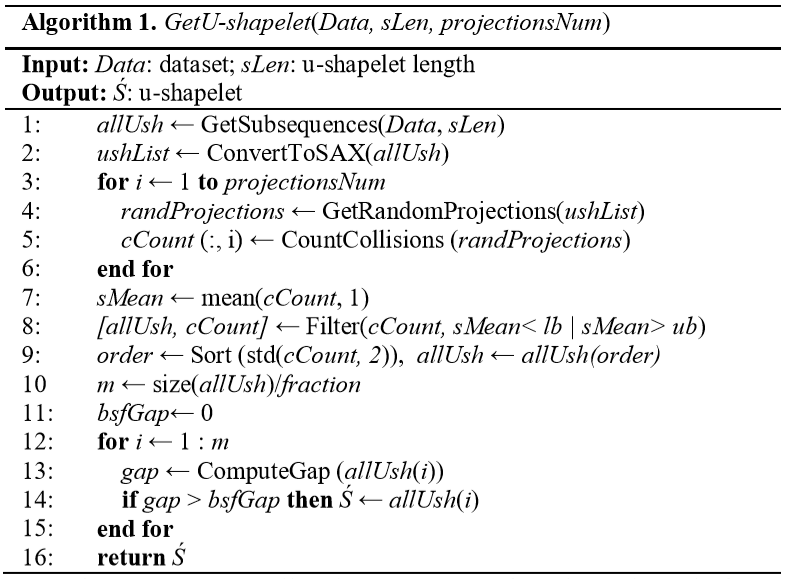
算法1暴力法

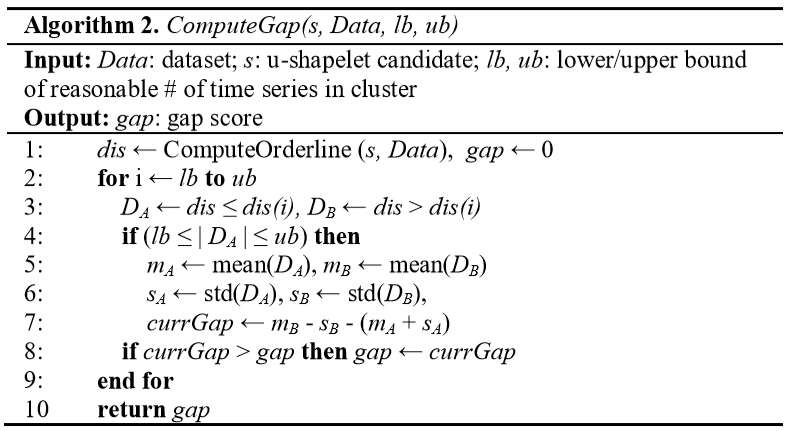




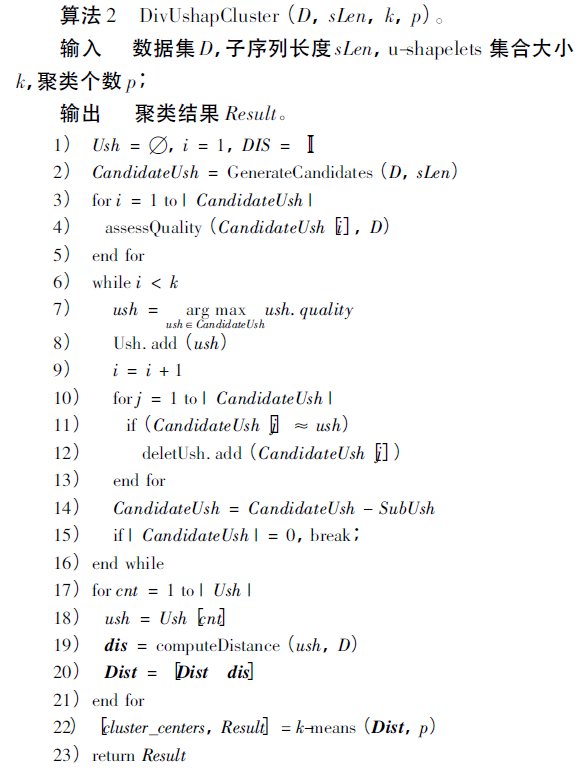


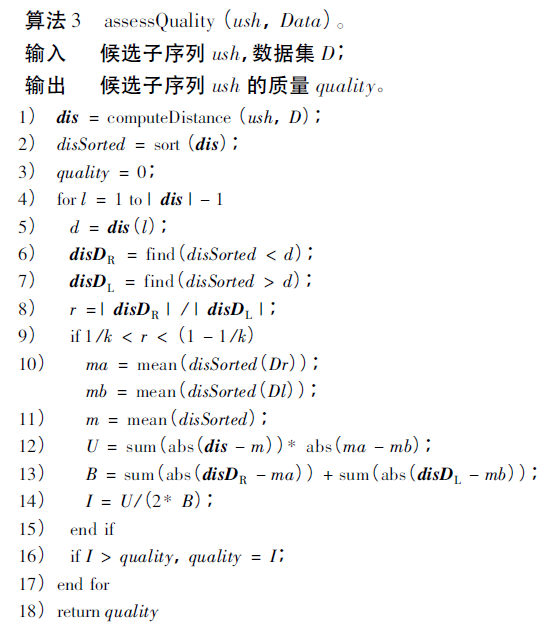
算法2暴力法SAX算法优化





算法3DivUshapCluster





**4.建议与问题思考**

笔者认为文章有两大问题，下面给出笔者的思考和笔者认为的解决方案

①文章对于top-k查询介绍的非常浅显，笔者虽然知道是什么意思，但是在算法中仍旧给出了非常模糊的介绍，即算法3中的11行，笔者以为，若你找的是最大的质量了，上述对top-k描述中若Q(i)>Q(j)才相似这个限制会产生矛盾，既然这是相似那么，第一条最大质量的子序列就和所有序列相似，这是必然的，那么算法没有意义，笔者以为可以更多解释一下；笔者猜测应该需要计算这条候选集和其他候选集之间的距离，取个阈值，当阈值大于一个数量，那么不过滤，反之过滤删除

②笔者提到实验部分的时候，作者和传统的聚类算法也进行了比较，笔者认为是没必要的，shapelet的另一个好处就是可以将其集合作为新的特征，看成原始特征的降维算法，算法1的后续就是通过这个思想来进行k-means算法的，如果相反使用与传统的聚类算法进行比较，反而失去了shapelet的特点

**5.附录说明**

文章作者：余思琴

参考文献：

ZAKARIA J， MUEEN A， KEOGH E． Clustering time series using unsupervised-shapelets ［C］ / / ICDM 2012: Proceedings of the IEEE 12th International Conference on Data Mining． Washington， DC:IEEE Computer Society， 2012: 785 － 794．

ULANOVA L， BEGUM N， KEOGH E． Scalable clustering of time series with u-shapelets ［C］ / / SDM 2015: Proceedings of the 2015 SIAM International Conference on Data Mining． Philadelphia， PA:

SIAM， 2015: 900 － 908