# Sequence Classification: A Regression Based Generalization of Two-stage Clustering

Nusrat Jahan Farin*†, Nafees Mansoor†, Sifat Momen†, Iftekharul Mobin† and Nabeel Mohammed†
*Computer Science and Engineering Department
Jahangirnagar University, Savar, Dhaka
*†Computer Science and Engineering Department
University of Liberal Arts Bangladesh (ULAB), Dhanmondi, Dhaka
Email: nusratfarin89@gmail.com, nafees.mansoor@ulab.edu.bd, sifat.momen@ulab.edu.bd,
iftekharul.mobin@ulab.edu.bd, and nabeel.mohammed@ulab.edu.bd

***Abstract*—Two-stage clustering based approaches has been used for sequence classification. However, certain parameters of these process are either hand picked or found through exhaustive searches. In this paper we propose a simple regression based approach, derived from parameter values found in a previous study, which generalizes the method to find values of three different parameters through proposed equations. We tested the applicability of our method on eight UCR sequence datasets and found our method to be comparable with hand picked approaches and better than single clustering based approaches.**

***Keywords - sliding window size; cluster number; time-series data; kmeans++***

## I. Introduction

Identification and retrieval of appropriate features from time-series data is a challenging [1], and at the same time a rewarding task. These classes of data typically inherent sequence-based partitions, which are not always known apriori. Clustering-based techniques have been successfully applied to this problem domain, mainly due to the structure-discovery mechanism built into such grouping techniques [2].

The approach and results of [3] forms the basis of this study, where a two stage clustering approach is successfully applied in a classification task on some parts of the popular UCR time series data set [4]. Their approach mainly consists of four steps with three manually chosen parameters.

At the first step, the method divides the data into subsequences of a manually selected length $sl$. There exists several techniques to divide the subsequences [5]. These subsequences are then clustered into $k_1$ clusters and each cluster center is used as a bin center to create a Bag-of-Words (BoW) feature vector from the subsequences. These BoW feature vectors are then further grouped into $k_2$ clusters. A classification scheme, which is further detailed in section IV, is used to utilize these $k_2$ clusters to predict the class of time-series data.

In [3], the values of $sl, k_1$ and $k_2$ are all selected through some non-specified methods. It appears that the values reported, the ones which found to give the best performance on each individual data set. This, however, is not an approach that can be repeated under practical constraints.

This paper reports on the finding of some initial work on prediction of the parameter values from information known a priori. Particularly the contribution of this work is a natural extension of [3], where the parameter values found to be useful are used to fit linear models where the dependent variables are the parameters and the independent variables are the number of classes and sequence length.

The results indicate that the proposed models, though extracted from only three data points, are effective and can match the performance of the exhaustively found parameter values and consistently outperforms the single clustering approach.

The rest of the paper is organized as follows. The following section describes the method and evaluation strategy of [3] in details. Section III describes the method used to derive the proposed linear models and how they have used in experiments. In section IV the results are described and compared with the results obtained in [3]. The paper is concluded with the future works in section VI.

## II. Existing Methods

Time series data is sequential data, where the ordering is temporal. There are several popular data sets, which have been widely used in research projects. The UCR time-series is a collection of multiple data sets [6]. It has been used as a benchmark collection of data in multiple studies. This study reports results using 8 data sets from the UCR time-series collection, the details of which can be found in Table I.

In [3], the time-series data is divided into some subsequences. The set of the subsequences is $S = \{S_{nl}|n = 1....,X, l = 1,.....,N\}$, where $X$ is the number of time-series data, $n$ is the number of sequences, $l$ is the length of subseqeuces and $N$ is the number of length.

Each sequence is encoded as a vector and aggregated into a data matrix which is then used in the first of clustering stage. The used clustering algorithm is Kmeans++ [7]. The study reported in [3] uses brute-force technique to determine the number of clusters rather than any particular methodology.

As it is possible for the original sequences to be of variable length, a fixed-length encoding scheme is necessary for feature comparison. The cluster centers obtained from the first clustering stage are treated as 'words' in a Bag-of-Words (BoW) encoding scheme. Briefly, each subsequence is mapped to one of the cluster centers by finding the cluster center that is closest to the subsequence vector in terms of cosine length. For cluster center, the number of subsequences mapped to it is counted, from which it is possible to calculate a frequency histogram, which is the BoW vector.

In [3], numeric vector N is created, such that $V_n = \{x_1, x_2, x_3, .....x_k\}$ where n = 1, 2, 3, ..., N. After the creation of the vector, the numeric value is used for the second cluster. In the second cluster kmeans++ algorithm is used once again [3]. The numeric vector is used for the clustering, where the cluster number is determined randomly for different datasets. In [3], clear and generalized formulas for determining the first and second cluster numbers are absent. Moreover, discussion on the sliding window size is also not presented.

However, the classification performance and the generated cluster quality are checked in [3]. In [3], two methods named as the Pseudo F measure [8] and the Davies-Bouldin (DB) index [9] is compared for measuring the performance. Authors in [3] presented that the high Pseudo F measure indicated the good quality for the clustering, where as the DB index indicated high quality for low measures. Thus the method in [3] performs better than the comparative methods of DB index for three datasets, namely Synthetic, Beef and OliveOil.

In the systems of two-stage clustering process for time-series data, four datasets called Beef, Synthetic, Coffee and OliveOil are used. In [3], kmean++ algorithm is used for two stage cluster, where cluster number and the sliding window of sequence length are predetermined. Authors in [3] use random value for experiments and the best results are explained. Rational explanation on use the particular cluster number and sliding window size are not provided in [3]. Thus, All of the numbers may confuse the users. To avoid the confusion of the user, some generalized formulas are proposed. Thus, using these formulas the sliding window size, number of first and second cluster can be determined.

Two-stage clustering process requires to determine the sliding window size, cluster number etc. The existing systems determine these things arbitrarily for different database. In this paper, we propose a new process to determine the sliding window size, first cluster number and second cluster number for performing the second stage cluster. In the first clustering process the featured data of the time-series data are extracted. Then the featured data is used for the second cluster. Before executing the second cluster, the result of first cluster is converted to numerical vector. The numeric vector is used in the second clustering. Finally, featured data of the time-series data are obtained through the second clustering procedure. The paper also compares the result with the existing methods [3].

### A. *Splitting up The Time-Series Data*

Tomoharu Nakashima uses arbitrary number for different database for sliding window, first cluster number and second cluster number [3]. The series of subsequences are used for the first clustering described below.

### B. *Initial Clustering Procedure*

In the first clustering procedure the extracted subsequences are used to find out the features of time-series data. KMeans++ algorithm is used to cluster the extracted subsequences. KMeans [10] algorithm gives the center of each cluster where kmeans++ the cluster number to be predefined. In these methods, to determine the cluster number formula 3 is used in this paper. Cosine distance [11] of each data from the cluster center is measured by cosine similarity [12] (equation 2).

$$Similarity(A, B) = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

$$Distance_{cosine} = 1 - Similarity(A, B). \quad (2)$$

Cosine gives the similarities between the subsequences. All values of the subsequences are assumed as positive where the maximum value of the similarity is 1 and the minimum is 0. The first cluster result shows the featured data of the time-series [13] data. These data is converted into numeric values. kmeans++ Procedure for clustering is given in Algorithm 1 [7]. The conversion procedure is described in the next section.

**Algorithm 1** KMeans++ algorithm

```
Choose an initial center randomly from the subsequences.
Compute the vector containing the square distances between all points in the dataset.
Choose a second center from subsequences randomly using the probability distribution
Recompute the distance vector until $K_f$ cluster center determined.
Choose a successive center using the average, simple arithmetic mean of all element of subsequences
Recompute the distance vector until convergence.
```

### C. *Time-Series Data to Numeric Vector*

Time-series data in the database are converted into numeric vectors. Numeric vector is implemented from the result of first cluster. This process determines each and every data point of the cluster and determine the repetition frequentness of the cluster in introductory time-series data. We convert the time-series data to numeric vector using the existing method. These methods are described in [3]. Detailed procedure is presented in Algorithm 2:

### D. *Second Clustering*

The kmeans++ algorithm is used for the second cluster. The proposed methods for the second-stage clustering is presented at the next section.

The overview of [3] is presented in section III.

## III. Proposed Method

### A. *Data Analysis*

[3] reports results of their proposed two stage clustering system using different values for the three parameters discussed above. The method used to determine the parameter values are not discussed. Assuming the parameter values were chosen to achieve good classification rates, it is reasonable to attempt to find a pattern in the values. As [3] reported results on four

**Algorithm 2** Time-series data to numeric vector

```
 1: Let num = 1 and len = 1.
 2: Set $V_n$ = ($x_1$,....., $x_n$ )
 3: if $C_{nl}$ the subsequences $Sub_{nl}$ belongs to by Identifying
    its nearest cluster center then
 4:     return Increase the $C_{nl}$-th of $V_n$ by 1.
 5: end if
 6: if len = S then
 7:     return num = num + 1 and len = 1
 8: else
 9:     len = len + 1
10: end if
11: if n > N then
12:     return terminate the process.
13: else
14:     go to step 2.
15: end if
```

different datasets, for each parameter it is possible to obtain four values. Although this is a small number, it is sufficient to fit a straight line through the points.

Figure 2 and Figure 3 shows the number of clusters in first-stage clustering and second-stage clustering respectively plotted against the number of classes.

Figure 4 plots the chosen window size(s) vs the sequence length of the data set.

Each of the figures also shows the best fit straight line. The main focus of this study revolves around these fitted lines, as their equations are treated as general method of determining the values of the three parameters. These equations are shown in Equations 3, 4, 5

$$K_f = Round(-17 * N_{classes} + 130) \qquad (3)$$

Where $k_f$ = First Cluster Number and $N_{classes}$ = Number of Classes

$$K_s = Round(4.34 * N_{classes} + 1.5) \qquad (4)$$

Where $k_s$ = Second Cluster Number and $N_{classes}$ = Number of Classes

$$S_w = Round(0.5 * S_l + 10) \qquad (5)$$

Where $S_w$ = Sliding Window Size and $S_l$ = Sequence length of time series data

## IV. Computational Experiment

In this section the performance of the proposed method is investigated. Experiments are done on 8 datasets from the UCR collection. These datasets are available through the UCR Time-series classification/clustering page [4], [6]. All details of the datasets, used in this experiment, is given in Table I.

In this experiment two types of comparison is made. The first comparison is between the existing proposed methods for the two-stage clustering [3] and the proposed method of the paper. And the second comparison is between the proposed methods and the first clustering process. All the comparison of the result is discussed in section V.

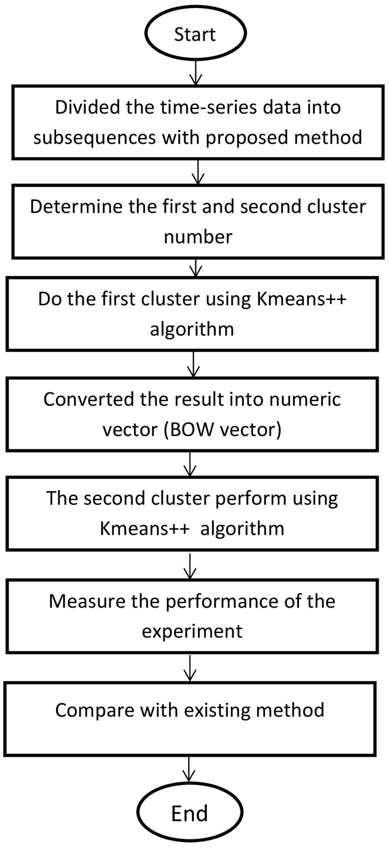
Start
Divided the time-series data into subsequences with proposed method
Determine the first and second cluster number
Do the first cluster using Kmeans++ algorithm
Converted the result into numeric vector (BOW vector)
The second cluster perform using Kmeans++ algorithm
Measure the performance of the experiment
Compare with existing method
End

Fig. 1: System overview of the proposed method

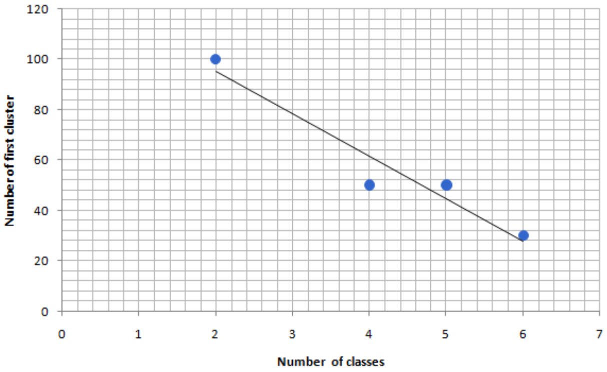

Fig. 2: Number of classes vs number of clusters in first-stage clustering

## V. Results and Discussion

The results obtained using the proposed parameters estimation methods are compared with the results reported in [3].

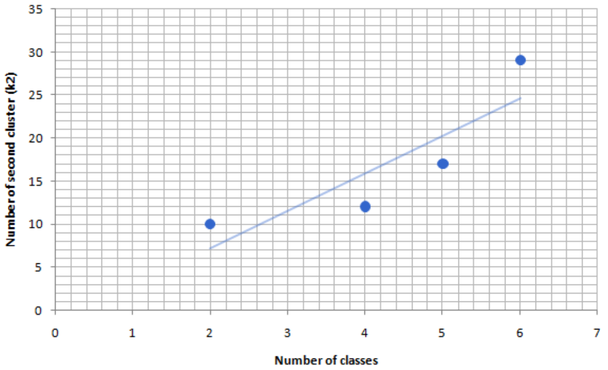

Fig. 3: Number of classes vs number of clusters in second-stage clustering

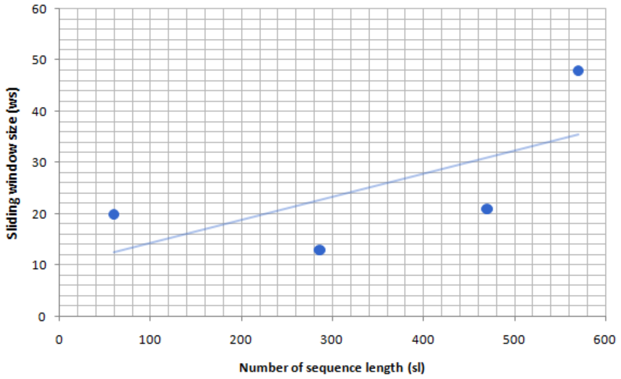

Fig. 4: Sequence length vs window size

The comparison is presented in Table II, which shows that ther results obtained in this study are comparable in our accuracy with those reported in [3]. In some of situation terms of the proposed method give better classification accuracy [3]. Figure 5 shows a bar-chart for visual comparison of the results.

In Table III, the comparison between the proposed methods and the single clustering methods for the proposed formula is given. In most of the cases the proposed methods give the better result than the single clustering process. In Figure 6 the overview of the comparison is given.

TABLE I: The Eight Benchmark Datsets Used in the Experiment

| Datasets | classes | training seq. | test seq. | seq. length |
|---|---|---|---|---|
| Coffee | 2 | 28 | 28 | 286 |
| OliveOil | 4 | 30 | 30 | 570 |
| synthetic | 6 | 300 | 3 | 60 |
| Beef | 5 | 30 | 30 | 470 |
| BeetleFly | 2 | 20 | 20 | 512 |
| BirdChicken | 2 | 20 | 20 | 512 |
| Car | 4 | 60 | 60 | 577 |
| Earthquakes | 2 | 139 | 322 | 512 |

TABLE II: Comparison Performance with Existing Method [3]

| Datasets | Comparison Method | Proposed Method |
|---|---|---|
| Coffee | 0.954±0.032 | 0.964±0.03 |
| OliveOil | 0.850±0.043 | 0.8±0.03 |
| synthetic | 0.977±0.007 | 0.977±0.005 |
| Beef | 0.603±0.057 | 0.47±0.05 |

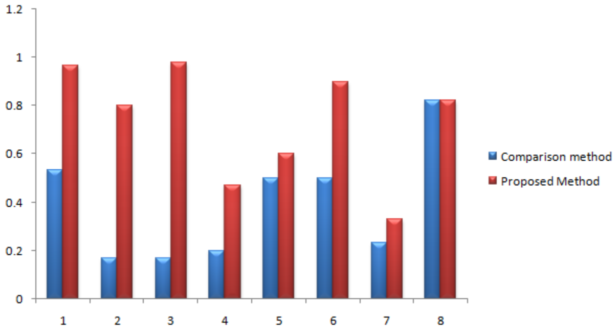

Fig. 5: Comparison between the existing method of single clustering and proposed method

In Table III the classification performance of eight benchmark datasets are given. The average classification rate is given here over ten trials along with standard deviation. The existing system [3] shows the four datasets classification performance. The results of the experiment is compared with the existing results in Table II.

## VI. Conclusion

Two-stage clustering based approach to a time-series data classification task is proposed in [3]. The process involves setting/choosing values of three different parameters. The original study does not report on the method used to choose the parameters. The contribution of this paper is in proposing three equations which can be used as a generalized method of choosing the parameter values based on known data. Experiments were conducted using the proposed parameter estimation equations and the classification accuracy obtained from these experiments were comparable. And most of the time results produced by proposed method, are better than those reported by [3]. The comparison was done on the four datasets used in the previous study. Results were reported on a

TABLE III: Classification Performance on benchmark Datasets with Single-Stage clustering

| Datasets | Comparison Method | Proposed Method |
|---|---|---|
| Coffee | 0.535 | 0.964±0.03 |
| OliveOil | 0.167 | 0.8±0.03 |
| synthetic | 0.167 | 0.977±0.005 |
| Beef | 0.2 | 0.47±0.05 |
| BeetleFly | 0.5 | 0.6±0.001 |
| BirdChicken | 0.5 | 0.9±0.01 |
| Car | 0.233 | 0.33±0.03 |
| Earthquakes | 0.82 | 0.82±0.057 |