KISDI
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

Asian Journal of Information and Communications

Similarity search for localized patterns in time series data

Ayush Rathi*
Indian Institute of Technology Roorkee, India

Durga Toshniwal**
Indian Institute of Technology Roorkee, India

Abstract

Time series data are of crucial importance as they depict trend of various entities over time. In this paper we give an approach to find similarity in time series based on angular differences and increasing decreasing patterns. The angular differences between two time series are computed by taking differences of their respective angles, where angles are the measured between the consecutive lines connecting the time series values. Differences between corresponding increasing-decreasing patterns are also taken into consideration and a method is proposed to find similarity based on these features. To demonstrate the effectiveness of the proposed approach, it is applied to various different datasets including synthetic, US retail store and Indian rainfall datasets. The comparative analysis of the experimental results of the proposed approach with real life scenarios show that the approach can be efficiently utilized to find similarity in time series for a wide range of applications.

Keywords: time-series, similarity, angular measure, increase-decrease pattern

1. Introduction

Time series have gain a lot of attention in the field of data mining as they carry a lot of information which can be extracted if proper techniques are applied. Various data mining tasks such as clustering, classification, anomaly detection etc. can be applied on time series data to extract meaningful information (Esling and Agon, 2012).

In this paper, we give an approach to find similarity in time series based on the angular differences and the corresponding increasing-decreasing patterns. This approach is robust from factors such as the scaling or amplitude differences which may occur if general methods such as clustering on the basis Euclidean distances over the actual values are applied. The method aims at finding similarities in local patterns exhibited by time series. As we shall show, even techniques such as normalization may not give the desired results, which this method can provide.

The similarity in local patterns is found by considering angular differences and increase decrease patterns.

* arathi@cs.iitr.ac.in,
** Corresponding author: durgatoshniwal@gmail.com

*** The approach presented in this paper was initially formulated in an attempt to find similarities in historic municipal budget data of various cities. We are thankful to https://openbudgetsindia.org/ from where we took various data sets related to municipal budgets. We appreciate the hard work they have done in extracting, documenting and presenting wide range of budget data sets.

Angular differences between two time series are computed by taking differences of their respective angles, where angles are the measured between the consecutive lines connecting the time series values, considering time series is plotted on a 2D graph where X axis represents the time instances and Y axis represents the corresponding values. Differences between corresponding increasing-decreasing patterns are also taken into consideration.

Our method is useful in situations where time series data depicts some local patterns and the time series vary in scale. For example, consider retail data. This method can be used in market basket analysis and association rule mining by finding the products whose sales have shown a similar patterns. Certain product may resemble similar sales pattern over the period of time, but their quantity of sales may vary. For example, say product X follows a pattern that its sales increase in certain period, decreases in certain period or remains constant in certain period. Product Y follows the same pattern, but overall sales of Y may be, say, only 10% of that of X. So here we cannot find similarity in sales of these products if we cluster them on the basis of magnitude of sales. As we shall see, our method can successfully mine these kinds of patterns.

The rest of the paper is organized as follows. In section 2, we provide a brief review of the works done in past for clustering and classifying time series based on patterns and shapes. Section 3 describes our method in detail. Section 4 describes the experiments we performed over various data sets. Section 5 gives the conclusion and future scope.

2. Related work

Similarity search in time series has been a topic of research for a long time and a lot of work has been done for the same. Similarity search in time series can be used for clustering similar time series into groups or classifying them into existing classes. Clustering and classification essentially involve a measure for computing distance between data points and many methods have been proposed for the same. In this section, we first give the review of clustering and classification tasks and then give an overview of work done in the field of mining time series data.

2.1 Clustering and classification

Clustering and classification are two important aspects of data mining. Clustering essentially involves grouping similar data such that objects within same group have high similarity. Classification on the other hand assigns classes to data based on the features exhibited by it. The data is assigned to that class, with which its features resemble the most. A detailed description of clustering techniques can be found in (Xu and Tien, 2015; Han et al., 2011). A study of classification techniques can be found in (Han et al., 2011).

Hierarchical clustering

We applied hierarchical method for clustering using agglomerative clustering with complete linkage (Han et al., 2011) on various data sets by using the distances calculated by the proposed approach. As hierarchical method does not require initializing cluster centers (like K -means) or spatial properties in data (like density-based approaches), it makes former a suitable choice.

2.2. Time series data mining

Mining time series data is a broad area and a lot of work has been done in this area (Esling and Agon, 2012).

Given the wide scope and importance of mining time series data, it has always remained an area of research. In this section, we give an overview of tasks in mining time series data which forms the background for our work.

Aghabozorgi et al. (2015) have presented a comparison of various time series-based clustering techniques and explored all four components of time series clustering viz. clustering algorithms, prototypes (e.g. averaging etc.), similarity measures (e.g. Euclidean, DTW, LCSS, etc.) and dimensionality reduction (e.g. PAA, SAX, etc.).

Berndt and Clifford (1994) have shown the use of dynamic time warping to find patterns in time series. This technique wraps the time axis of the signals (or time series) to achieve better alignment between them.

Approaches based on derivatives or slopes (similar to angles) have also been proposed. One such approach is derivative dynamic time warping (Keogh and Pazzani, 2001). This approach handles the singularity issues with classic DTW by taking slope as a distance measure instead of actual Euclidean distances. Toshniwal and Joshi (2005) proposed another approach for similarity search by taking cumulative slopes.

Sometimes, a symbolic representation of time series can have various advantages. For example, representing time series in the form of symbols may provide computational advantage of string based algorithms on time series data. For obtaining distance based on increase decrease pattern, we have used one symbolic approach as in section 3. Further, methods such as piecewise aggregate approximation can be applied to reduce the dimensionality of time series data and enhance the computational speed (Keogh and Pazzani, 2000).

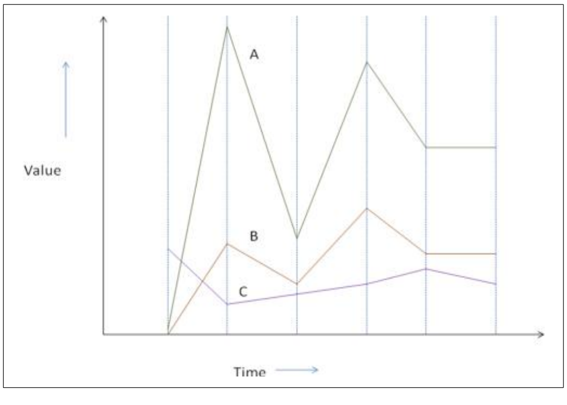
3. Proposed method

In this section we describe our approach for mining localized patterns in time series. Our approach uses both angles and relative increase-decrease pattern as feature. This approach helps in formulating distances between time series which can then be used to cluster similar time series or classify to the most similar class on the basis of local patterns. Since both clustering and classification depend on the distance measure, we explain our method for clustering task, though the distance measure which will be formulated can be used in similar way to classification as well.

Firstly, we explain why contribution of both of these factors as the features is important. After that we explain how we mathematically measure these factors and then finally formulate our method.

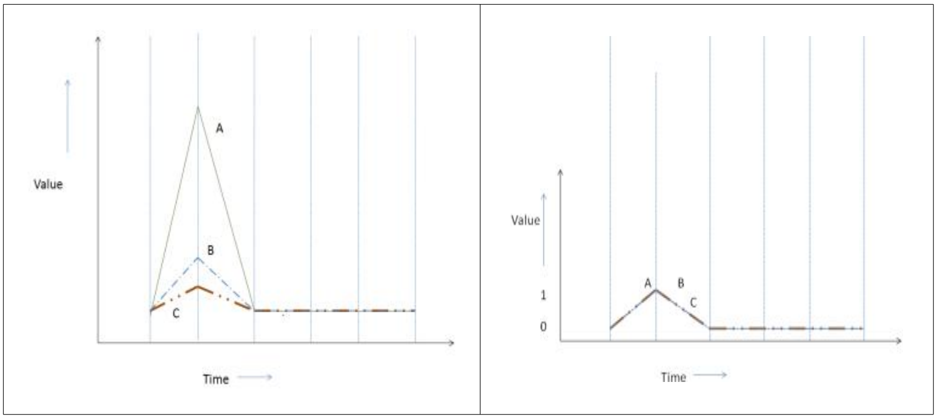
To understand the importance of both of these features, consider the sample time series shown in Figure 1.

Figure 1. Sample time series



Suppose we need to cluster time series following a similar pattern. If we use normal Euclidean distance-based measure to find differences in time series, it's more likely that time series 'B' and 'C' will be more similar as pairwise distances between 'B' and 'C' are less even though 'A' and 'B' follow similar pattern. To overcome this issue, normalized Euclidean distance can be used. In that case 'A' and 'B' will have less pairwise distances and it's likely that they will be in the same cluster (class) and 'C' will be allotted a different cluster (class). But an issue can still occur in normalized Euclidean distance as well. For example, consider below the case of Figure 2. If we consider sample time series of Figure 2(a) and normalize the respective time series in a certain range say [0,1] as shown in Figure 2(b), the issue discussed above will be resolved. But in this case, the impact of amplitude factor is lost. Hence, though as per Figure 2(a), 'B' and 'C' are more similar as compared to 'A' and 'B' and 'A' and 'C' respectively, after normalization similarity between 'B' and 'C' will be the same as that between 'A' and 'B'.

Figure 2. (a) Sample time series (b) corresponding normalized time series

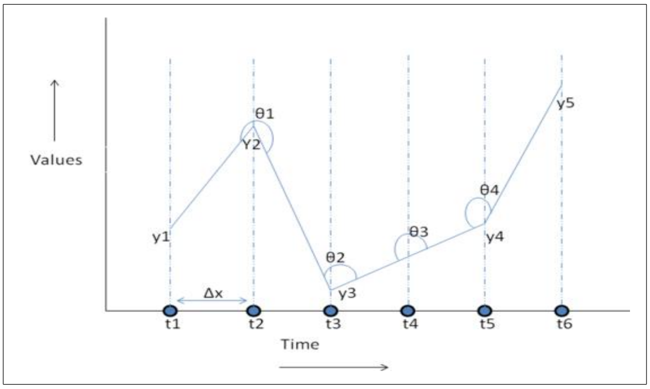


Another measure to capture a similar pattern in time series is to represent time series by a word, where alphabet of word represents the pattern. For example. if 'I' represents an increasing pattern, 'D' represents a decreasing pattern and 'C' represents constant, then the time series 'A' in Figure 1 can be represented by "IDIDC". Then string distance methods like Edit Distance etc. can be used to find similar time series based on pattern. But this method again is not capable of taking magnitude feature into consideration. Hence in cases like in Figure 2, this method will also fail.

3.1 Angles as feature vector

If we take the included angles (see Figure 3) as features instead of actual values of data points, above issues can be resolved. For measuring angles, consider time series is plotted on a graph where x axis represents the time instances and y axis represents the values of time series (at corresponding time instances). The time instances must be of equal gaps. The width between two consecutive time instances on x axis, Δx must be constant as time instances are of equal gaps. However, value of Δx can be suitably taken as per the dataset and depending on how user wants the time series plot. If the time series plot is required with high slopes, Δx can be reduced. If time series should be stretched so that slopes are decreased, Δx can be increased. In general, Δx can be taken as 1.

Figure 3. Included angles as feature vector for time series



The convention which we took to measure angles is clockwise angle between every consecutive line joining the time series values. As can be seen from Figure 3 if time series has ‘n’ data points, the angles will be (n-2). Hence a time series Ti which has values t_{i1}, t_{i2},...., t_{in} can now be represented as θ_{i1}, θ_{i2},...., θ_{in-2} , where θ_{ij} represents the angle between lines joining the points t_j, t_{j+1} and t_{j+1}, t_{j+2} of i^{th} time series.

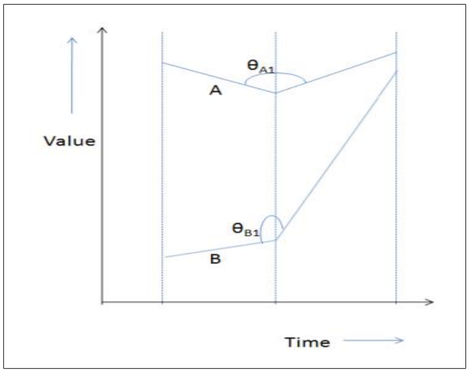
Once we convert the time series points into angular features, we can then find the distances between angle of two time series to find their similarity. Mathematically, if we have two time series ti and tj and there angular features are θ_{i1}, θ_{i2}, ... θ_{in} and θ_{i1}, θ_{j2} ... , θ_{jn} (consider both time series have (n+2) data points so that the angles are ‘n’), the angular distance between two is given by

$$\sqrt{\sum_{k=1}^{n}(\theta_{ik}-\theta_{jk})^2} \tag{1}$$

If this distance measure is used for finding similarity, for Figure 1, Time series ‘A’ and ‘B’ will be more similar as compare to ‘B’ and ‘C’ and for Figure 2, Time series ‘B’ and ‘C’ will be more similar to ‘A’ and ‘B’. Hence angular measure can efficiently help in finding similar patterns. Further weights can be assigned to terms such that corresponding to ‘newer’ time instance is more. Higher weights of later time instances will give priority to patterns similar in recent time as compared to older. Using the weights, each θ_k effectively becomes, $\theta_{kt,}$ where, $\square_{kt}=w_{k*}\square_k$ and weights are assigned such that $w_k<w_{k+1}$. To avoid exponential increase, condition such as “sum of all weights should be constant” can be imposed. The use of weights does not impact the method and all further discussion is based on equation (1), though the angles in equation (1) can be transformed by applying weights without impacting the method. Use of weight solely depends on purpose. If data points corresponding to recent time should be given more importance, weights can be applied.

Angular measure alone however cannot suffice to find the similarity in pattern. Figure 4 depicts on such scenario. In this case two time series ‘A’ and ‘B’ are considered, each containing three data points. The angle θ_{A1} and θ_{B1} are equal but time series does not have similar patterns. If a pattern is represented in the form of word (as mentioned earlier in this section) then time series A will be represented as “DI” and time series B will be represented as “II”.

Figure 4. Similar angle but different pattern



So, above scenario depicts that angles alone are not sufficient and increase–decrease pattern should also be taken into consideration.

3.2 Increase–decrease pattern as feature vector

The increase decrease patterns can simply be captured by representing time series through a word as described earlier. Time series containing ‘n’ points, can be represented by word containing (n-1) letters, where alphabet of letter are (I,D,C) and i_{th} letter is ‘I’ if there is an increase in value in time series from t_i to t_{i+1}, ‘D’ if there is decrease in value in time series from t_i to t_{i+1} and ‘C’ if there is increase no change in value in time series from t_i to t_{i+1}. Note that the problem depicted in Figure 2 will again come when we compute differences based on this feature. However as explained in section 3.3, both these features (angles and increase-decrease pattern) combinedly can overcome issues for this type of scenario.

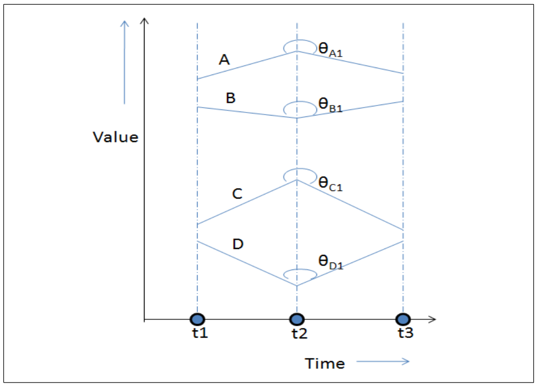
The difference in two time series can be calculated by distance in corresponding strings. Further difference can be given weights such that difference between ‘I’ and ‘D’ is more as compared to difference between ‘I’ and ‘C’ as the earlier depicts a completely opposite pattern. So, if two time series are represented by words S_i and S_j, the difference between ‘k’ the letter of the both words is given by

$$\partial_{S_{ik,}S_{jk}} = \begin{cases} 0, S_{ik} = S_{jk} \\ 1, S_{ik} \neq S_{jk} \text{ and } S_{ik} = 'C' \text{ or } S_{jk} = 'C' \\ \gamma, S_{ik} \neq S_{jk} \text{ and} S_{ik} \neq 'C' \text{ and } S_{jk} \neq 'C' \end{cases} \quad (2)$$

The value when the both letters are opposite, that is, one is ‘I’ and other is ‘D’ is taken as ‘ γ ’ to give it more weight as compared to the case when both letters are not same (same pattern of increase or decrease), but one of them is constant. Optimally, $1 \leqq \gamma \leqq 2$.

One more observation is that the impact of differences in increase decrease pattern should be more when increase or decrease is relatively high. The angular feature can be used as weight to capture this. Consider Figure 5. In this time series ‘A’, ‘B’, ‘C’, ‘D’ respectively depict the pattern “ID”, “DI”, “ID” and “DI”. However as can be seen in Figure 4, the difference in pattern for time series ‘C’ and ‘D’ should be more as compared to that of ‘A’ and ‘B’. One can find by observation that the more linear the plots of time series is, the more the included angle is closer to 180°. As the increase or decrease gets higher, the difference between the included angle and 180° gets higher.

Figure 5. Impact of angles in increase-decrease pattern



Hence, to find the difference corresponding to increase-decrease pattern in two time series, following approach can be used.

- Take three consecutive points in both of the time series.
- Find the average of the angle made by these points for both of the time series where angle is measured as explained in section 3.1. Subtract this angle form 180°. Take modulus of this difference. Let this be called as linearity factor (LF).
- For the three consecutive points, there will be a two-letter substring of the word used to represent the time series by the method defined earlier in this section. Find the difference for each corresponding letter of two letter substrings by using equation (2). Multiply this difference by LF.
- Repeat above for every chain of three consecutive points (or time instances) and add all the sub results to get the final distance.

Formally, if the time series 'i' and 'j' have (n+1) data points corresponding to (n+1) data instances, they can be converted to words of 'n' letters by the method described earlier in this section. Let those words be represented by S_i and S_j respectively. Also let the angle θ_{ij} represents the angle between the lines joining the points t_j, t_{j+1} and t_{j+1}, t_{j+2} of ith time series. This angle also will correspond to the j^{th} and $(j+1)^{th}$ letter when time series is transformed to the word representation. (Note that a letter in the word representation depicts the pattern of line joining two data points corresponding to that letter, as increase, decrease or constant). Then the difference on the basis of increase – decrease pattern between two time series is given by

$$\Delta_{S_i,S_j=}\sum_{\mathrm{k}=1}^{\mathrm{n}-1}\left(\left(\partial_{\mathrm{S}_{\mathrm{ik},}\mathrm{S}_{\mathrm{jk}}}+\partial_{\mathrm{S}_{\mathrm{ik}+1,}\mathrm{S}_{\mathrm{jk}+1}}\right)*\mathrm{LF}_{\mathrm{k}}\right) \tag{3}$$

where, $\partial_{s_{ik,}s_{jk}}$ is given by (2) and,

$$LF_K=\left|(180°-(\frac{\theta_{ik}+\theta_{jk}}{2}))\right| \tag{4}$$

and, θ_{ik}, $\theta_{j\mathrm{k}}$ are calculated as explained earlier.

3.3 Combining angular and increase – decrease pattern feature vectors

As noted previously, the similarity search based on increase decrease pattern cannot cover scenarios such as shown in Figure 2, while the scenarios such as shown in Figure 4 cannot be covered by angular features. However, if both the features are combined, all such scenarios can be overcome.

Hence, distance between two time series can be calculated by taking effectively the contributions of both equations (1) and (2). However since (1) contains the angular measures and (2) has string distances which are linear in nature, each angular term in (1) must be converted into linear term. This can be done by multiplying each term of (1) by (π/180°). This is because (2πr) represents the perimeter of circle of radius 'r', so for angle θ, the length of arc will be (θ*r/180°). If we consider the circle of unit radius, this will be (π/180°).

So, (1) can now be re-written as

$$\sqrt{\sum_{k=1}^{n}((\theta_{ik}-\theta_{jk})*(\pi/180°))^2} \tag{5}$$

Now to add the increase-decrease feature based distances, some weights should be multiplied so that linear combination of both the factors is obtained.

So, the effective distance therefore becomes:

$$\sqrt{\sum_{k=1}^{n}((\theta_{ik}-\theta_{jk})*(\pi/180°))^2} + \beta*\left(\Delta_{S_i,S_j}=\sum_{k=1}^{n-1}\left(\left(\partial_{S_{ik},S_{jk}}+\partial_{S_{ik+1},S_{jk+1}}\right)*LF_k\right)\right) \tag{6}$$

Now, since ($\Pi/180^{\circ}$) and β both are constants, they can be effectively changed to a single constant 'α'. So, the effective distance between two time series 'i' and 'j' becomes

$$\sqrt{\sum_{k=1}^{n}(\theta_{ik}-\theta_{jk})^2} + \alpha*\left(\Delta_{S_i,S_j}=\sum_{k=1}^{n-1}\left(\left(\partial_{S_{ik},S_{jk}}+\partial_{S_{ik+1},S_{jk+1}}\right)*LF_k\right)\right) \tag{7}$$

where, all the symbols have their usual meanings as described earlier.

The value of α depends on dataset features like domain and characteristics of data, user requirements like clustering on basis of certain factors or classification with high accuracy etc. Methods such as gradient descent can be used in the case of classification to find an optimal value of α.

3.4 Similarity search in time series with angular feature and increase – decrease pattern

Suppose there are 'M' time series in data set. The distance between each $\left\langle{M \atop 2}\right\rangle$ pair of time series can be founded by applying the method described in 3.3. Finally, a distance matrix can be created and tasks such as clustering and classification can be done on that to cluster or classify time series following similar pattern.

4. Experiments and results

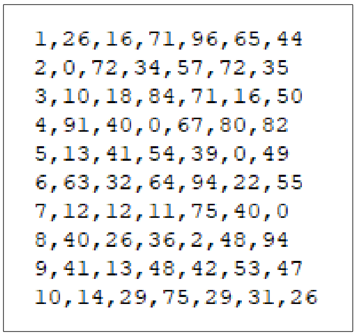
We conducted experiments over variety of data sets including synthetic and real ones. This section covers the details of experiments performed and the results obtained.

4.1 Synthetic dataset

The synthetic data set was designed to explain the working of approach. It consisted of ten time series with each time series comprising of seven data points (hence four angles would be formed for each time series). The value of data points was taken randomly in range [0,100]. The distance between each pair of them was calculated using the discussed approach. Finally, hierarchical agglomerative clustering was applied with complete linkage to cluster them into four groups.

The sample data set in csv form is as follows. Please note that first value in each line represents the time series number (for identification purpose, and then it is followed by data points).

Figure 6. Synthetic time series data set (sample data)



```
1,26,16,71,96,65,44
2,0,72,34,57,72,35
3,10,18,84,71,16,50
4,91,40,0,67,80,82
5,13,41,54,39,0,49
6,63,32,64,94,22,55
7,12,12,11,75,40,0
8,40,26,36,2,48,94
9,41,13,48,42,53,47
10,14,29,75,29,31,26
```

The time series in string representation as explained in section 3.2 is as follows:

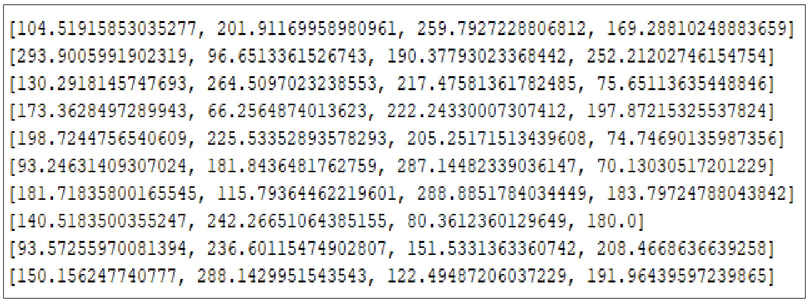
DIIDD,IDIID,IIDDI,DDIII,IIDDI,DIIDI,CDIDD,DIDII,
DIDID,IIDID;

where each comma separated word represents a pattern of single time series.

To plot the above time series on x-y graph with aspect ratio of y:x as 10:20, the value of Δx (that is the distance of two consecutive time instances on x axis) is taken to be 33.33 (since y range is from [0-100] and x range is from [1-7] and since aspect ratio of y:x is 10:20 so Δx will be 33.33 by unitary method.

The value of angles formed for the time series are shown in Figure 7.

Figure 7. Angular representation of synthetic time series data set



```
[104.51915853035277, 201.91169958980961, 259.7927228806812, 169.28810248883659]
[293.9005991902319, 96.6513361526743, 190.37793023368442, 252.21202746154754]
[130.2918145747693, 264.5097023238553, 217.47581361782485, 75.65113635448846]
[173.3628497289943, 66.2564874013623, 222.24330007307412, 197.87215325537824]
[198.7244756540609, 225.53352893578293, 205.25171513439608, 74.74690135987356]
[93.24631409307024, 181.8436481762759, 287.14482339036147, 70.13030517201229]
[181.71835800165545, 115.79364462219601, 288.8851784034449, 183.79724788043842]
[140.5183500355247, 242.26651064385155, 80.3612360129649, 180.0]
[93.57255970081394, 236.60115474902807, 151.5331363360742, 208.4668636639258]
[150.156247740777, 288.1429951543543, 122.49487206037229, 191.96439597239865]
```

Figure 8. Cluster formed for synthetic data set for α = 0

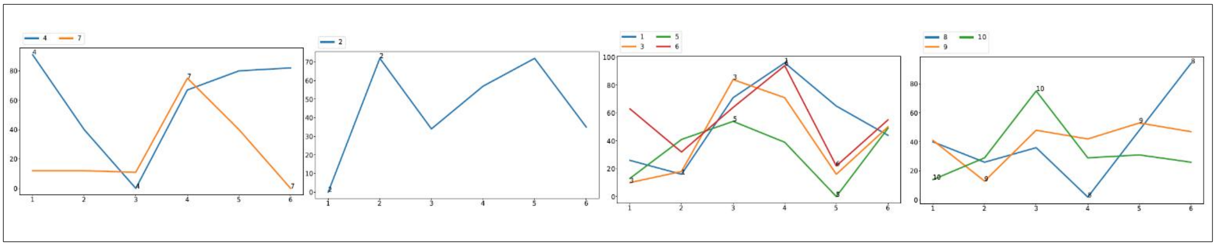
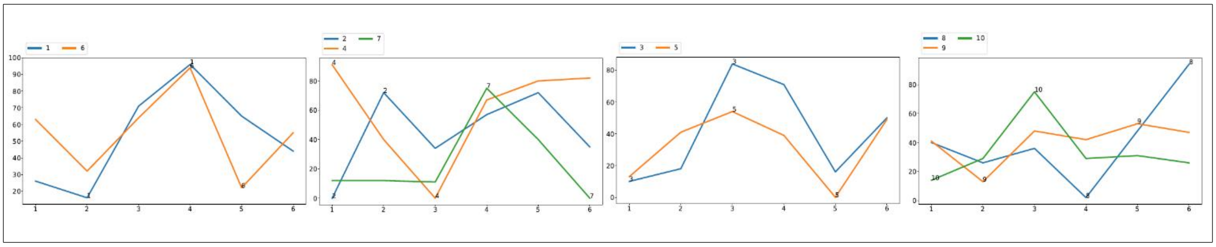


Figure 9. Cluster formed for synthetic data set for α = 0.5



These words and angular presentations were then combined by the method explained in section 3.3 to find similarities in time series and finally they were clustered into four groups.

The experiments were done both with zero value of α(that is not considering increase decrease pattern as feature and solely on basis of angular features) and with non-zero value of α

Results with zero value of ‘α’

Figure 8 shows the cluster formed when ‘α’ value is zero (i.e. only angular features are considered). The legend in the plot depicts the id number of time series. X axis shows the data instance (or time instance) and Y axis represents the corresponding values. As can be seen from the graphs, time series showing similar patterns were grouped in the same cluster.

Results with non-zero value of ‘α’

In the fourth cluster, time series ‘3’ and ‘5’ have the opposite pattern then ‘1’ and ‘6’ for first three data points. The word representation for ‘3’ and ‘5’ is “IIDDI” while for ‘1’ and ‘6’ is “DIIDD” and “DIIDI” respectively. However, the cumulative sum of their angular differences was such that they came in same cluster. Non-zero value of ‘α’ tackled this difference. We separately observed the angular and string-based distances and then set ‘α’=0.5, as this value made both magnitude of differences for both of these features roughly of the same order. As can be seen in Figure 9, ‘3’ and ‘5’ formed a new cluster. But ‘2’, ‘4’ and ‘7’ were also moved to the same cluster to accommodate a new cluster. Changes in ‘α’ value can give better results depending on data characteristics. Further, an optimal number of clusters can be found by applying various techniques to measure goodness of clusters.

4.2 US retail stores data set

This experiment was performed on retail sales data of US stores.[1] The stores considered for the experiment were new car dealers; used car dealers; furniture stores; home furnishings stores; household appliance stores; radio, TV, and other electrical stores; hardware stores; grocery stores; health and personal care stores; pharmacies and drug stores; gasoline stations; mens' clothing stores; women's clothing stores; shoe stores; jewellery stores; sporting goods stores; hobby, toy, and game stores; book stores; office supplies and stationery stores; gift, novelty, and souvenir stores. The data set has sales value for each month and we consider 132 values from January 1991 to December 2017. The weights were adjusted to moderate value so that the both factors have the same order of magnitude of differences (as in 4.2) and the results of clustering were: Cluster 1 (sporting goods stores; book stores), Cluster 2 (furniture stores; household appliance stores; radio, TV, and other electrical stores), Cluster 3 (new car dealers; office supplies and stationery stores), Cluster 4 (used car dealers), Cluster 5 (home furnishings stores; hardware stores; grocery stores; health and personal care stores; pharmacies and drug stores; gasoline stations) and Cluster 6 (mens' clothing stores; women's clothing stores; shoe stores; jewelry stores; hobby, toy, and game stores; gift, novelty, and souvenir stores). As can be inferred from the results, items having similar sales pattern came in the same cluster.

4.3 Indian rainfall data

This experiment was performed on rainfall data of India[2]. Rainfall data of 30 subdivisions for period of 146 years was taken. The time series for each subdivision was prepared with the data points in chronological order (month-year wise). The results when subdivisions were divided into 5 clusters by the proposed approach were: Cluster1 (COASTAL ANDHRA PRADESH, RAYALASEEMA, TAMIL NADU, SOUTH INT. KARNATAKA, KERALA); Cluster 2 (GANGETIC W. B, ORISSA, JHARKHAND, BIHAR, EAST UTTAR PR), Cluster 3 (WEST U.P. PLA, HARYANA, PUNJAB, WEST RAJASTHAN, EAST RAJASTHAN, WEST MADHYA P, EAST MADHYA P, GUJARAT, SAURASHTRA & KUTCH, VIDARBHA, CHATTISGARH), Cluster 4 (ASSAM, NAGA.MANI. MIZO. &TRIP, SUB-HIMA. W. B) and Cluster 5 (KONKAN AND GOA, MADHYA MAHARASHTRA, MARATHWADA, TELANGANA, COASTAL KARNATAKA, NORTH INT. KARNATAKA).

The results show that each cluster contains neighboring subdivisions which is supported by the fact that adjacent regions exhibit a similar climatic and rainfall pattern.

5. Conclusion

Euclidean distance and normalized Euclidean distance are the most common distance measures used to determine similarity in the time series. However, they suffer from various issues as discussed in Section 3. So, there is a need for some more robust method to compute similarity in time series. In this paper, we proposed a method to mine similarity in time series based on two features – angular measures and increase decrease patterns. This method is capable of mining local patterns and robust from factors such as amplitude scaling. Instead of considering the global trend, this method considers local shapes to find the similarity and mine the patterns from the time series data.

To prove the correctness of the approach, we applied the method on three data sets. First, a synthetic data set was generated. As observed from its plots, the time series having similar plots (patterns) were grouped in the same cluster. Second, we used the data set of US Retail Stores. The result of clustering of stores aligned with

[1] http://www.economagic.com/

[2] ftp://www.tropmet.res.in/pub/data/rain/iitm-subdivrf.txt; http://www.tropmet.res.in/Home

the real-life shopping behavior of people. Third, we used data set of India rainfall statistics. The result of clustering process was that adjacent subdivisions were grouped in the same cluster, confirming with the fact that nearby regions have similar rainfall patterns across the year.

Mining of local pattern has wide range of applications. It can be used for market basket analysis, to find which items follow similar sales pattern; to group stocks based on their performance and create portfolio and much more.

The proposed approach can be applied in all such scenarios to mine meaningful information. The proposed method performs one to one comparison by considering time series of same length. The support for finding similarity between time series of varying lengths is an interesting problem which needs to be studied further.

References

Aghabozorgi, S., Shirkhorshidi, A. S. and Wah, T. Y. (2015). Time-series clustering-A decade review. *Information Systems, 53*(C), 16-38.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *AAAIWS'94 Proceedings of the 3rd Internat ional Conference on Knowledge Discovery and Data Mining*, 359-370 Seattle, Washington.

Esling, P. and Agon, C. (2012). Time series data mining. *ACM Computing Surveys (CSUR) Surveys, 45*(1), USA.

Han, J., Pei, J. and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Keogh, E. J. and Pazzani, M. J. (2000). Scaling up dynamic time warping for datamining applications. *In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and data Mining*, 285-289, Boston, Massachusetts, USA.

__________________________ (2001). Derivative dynamic time warping. In *Proceedings of the First SIAM International Conference on Data Mining.*

Toshniwal, D. and Joshi, R. C. (2005). Using cumulative weighted slopes for clustering time series data. *International Transactions on Computer Science and Engineering, 20*(1), 29-40.

Xu, D and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193.