DSDC: A Large-scale KPI Clustering Model Based on DWSBD Improved DBSCAN

Haiyan Feng, Mingwei Li*

Department of Statistics, Northeastern University, Shenyang 066004, Liaoning, China
DOI: 10.32629/memf.v5i2.1973

Abstract: In response to the problem of a large amount of noise, anomalies, and phase shifts in KPI time series, which makes it difficult to obtain the correct number of clusters and good clustering accuracy, this paper proposes an improved DBSCAN based clustering algorithm — DSDC algorithm. Firstly, the underlying shape extraction technique of KPI data is proposed based on the existence of a large amount of noise and anomalies in KPI data. Secondly, the DBSCAN clustering algorithm is used to solve the phase shift problem in KPI time series. Finally, for the problem that the DBSCAN algorithm is parameter sensitive and cannot handle multi-density data, a new similarity measure is proposed, i.e., density-weighted shape-based distance (DWSBD). The experimental results show that the DSDC algorithm has higher ACC, NMI, F-score and shorter clustering time compared with K-Medoids and Spectral clustering algorithms which are also based on the underlying shape extraction technique and DWSBD distance.

Keywords: DBSCAN, DBSCAN, DWSBD, KPI, clustering, density-weighted

1. Introduction

Key Performance Indicator (KPI) is a time series data with practical application significance obtained by timing sampling[1]. There are tens of thousands of KPI data in practical applications. if each KPI curve is examined, KPI labeling, model selection, model training, and parameter tuning would consume a lot of human and material resources. Fortunately, due to the implicit correlation and similarity between KPI time series, the aforementioned overhead can be significantly reduced by fast and scalable clustering of KPIs.

For this problem, we propose a new clustering algorithm DSDC for clustering KPI time series, which is based on the similarity of KPI bottom shape and the improved DBSCAN. As shown in Figure 1. The first part is offline KPI clustering, which first samples from a large stream of historical KPIs to improve the clustering efficiency, preprocesses the extracted KPI subsets to extract the underlying shapes of the KPIs, then uses DWSBD as a distance metric to handle the multi-density problem, and clusters these subsequences using the improved BDSCAN algorithm. The cluster centroids of mass are calculated and the remaining KPIs are assigned to the clusters, and finally an anomaly detection model is trained independently for each cluster. The second part is online KPI categorization, which preprocesses the newly generated KPI streams and calculates their distances from each centroids, categorizes them, and selects the corresponding anomaly detection model to detect anomalies, so that similar KPIs share an anomaly detection model.

2. Related Work

2.1 DBSCAN clustering algorithm

The traditional clustering methods mainly include divisional methods, hierarchical methods, density-based methods, grid-based methods and model-based methods. Among them, density-based methods have wide applications in clustering analysis because they do not need to specify cluster number in advance and can find arbitrary number and shape of clusters in the noisy data sets. The distribution of KPI data set has no assumptions, which means that it may have arbitrary shape and different densities. Therefore, density-based clustering is chosen as the clustering algorithm for KPI data.

Density-base Spatial Clustering of Applications with Noise (DBSCAN) is a class of density-based clustering algorithm. The idea of clustering process is to start from any data point in the data set, and expand to a maximum area according to the condition of density reachable. If the initial point is the core point, then the maximized region is a class; If the initial point is a boundary point, it will jump to the next initial point; If the initial point is a noisy point, it will be directly marked as a noisy point.

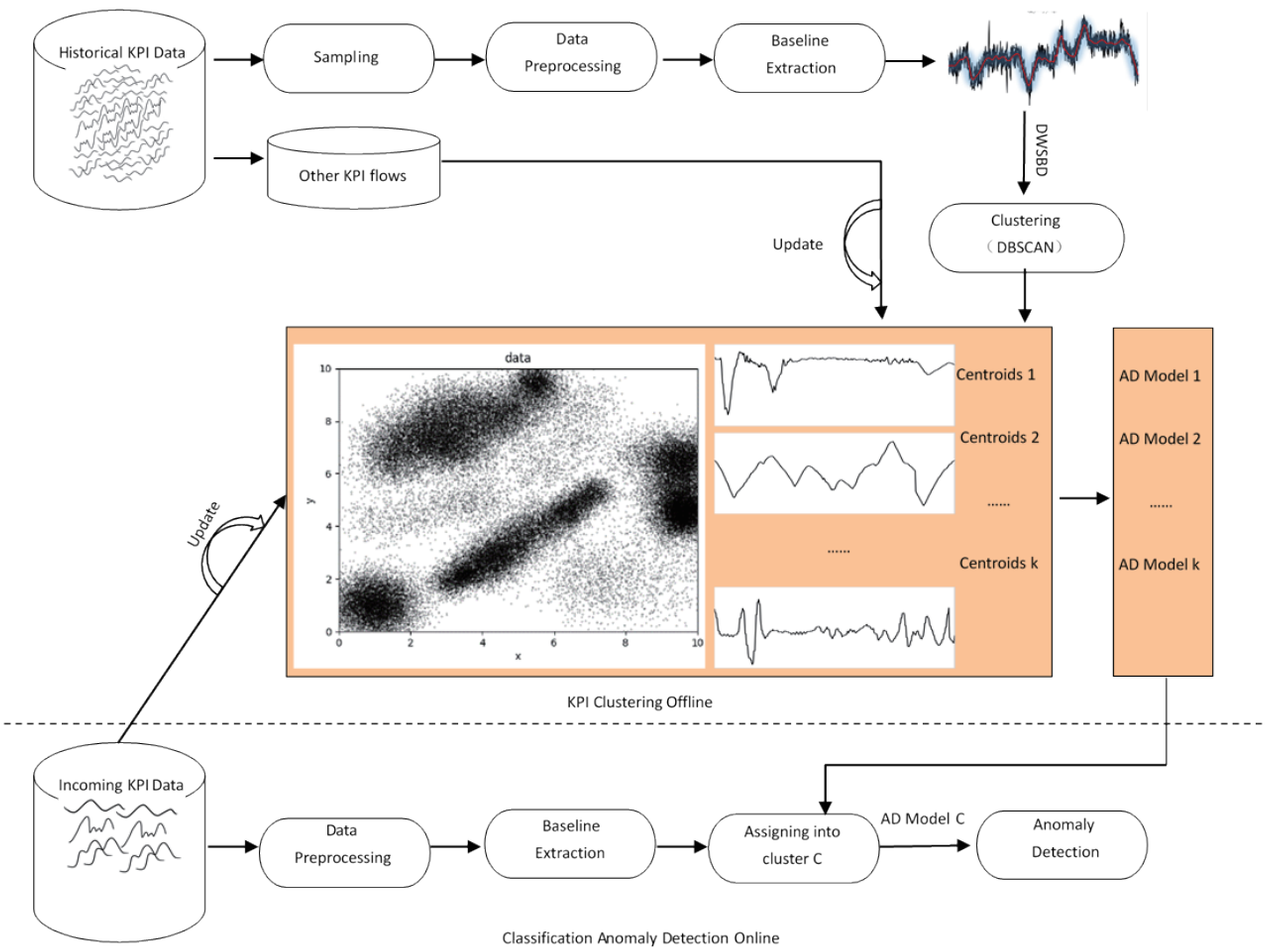


Figure 1. overall framework of DSDC

2.2 K reach-distance

In the process of DBSCAN clustering, because each data object has different distribution in local space region, it is necessary to determine the personalized parameter of data object, i. e. local density, to characterize the distribution of data in a dataset. Introducing the reach-distance proposed by Breunig et al. in 2000[2] to define the local density of each time series. The definition of reach-distance is related to k-distance(p): The reach-dist(p,o) is the maximum of the k-distance(O) and the direct distance between p and o. The k reach-distance of point P to point o is defined as

$$reach\text{-}distance_{k(p,O)} = max\{d_k(O), d(p,O)\} \tag{1}$$

This definition means that the greater of the k-distance(O) and the distance of p from o is chosen as the Reach-dist(p,o).

3. DSDC algorithm

3.1 Underlyin shape extraction

To reduce the influence of noise and anomalies on the clustering results, bottom shape extraction is needed for the KPI series. Firstly, extreme outliers are removed. Intuitively, anomalies have the largest deviation from the mean, and in general, the proportion of outliers in the time series is less than 5% [3]. After standardization, each KPI is normalized to have zero mean and unit variance, so we only need to remove the top 5% of data with the largest deviation from the mean and fill them in using linear interpolation. Then extract the underlying shape and use moving average to denoise the KPI time series, i.e., for the KPI series a moving window of width W is applied and the window is slid along the time series to calculate the average of the new series.

3.2 Distance Metric DWSBD

In time series clustering, the popular methods of similarity measurement are ED, DTW, SBD, etc. Shape-based distance

(SBD) can handle the time shift natively, and its computational complexity can be reduced to O(m log(m)) by convolution theorem and fast Fourier transform[4]. Therefore, the SBD-based distance can be used as a similarity measure for KPI. In order to solve the multi-density and phase shift problems of KPI time series, we define an improved time series distance measure, locally reachable density weighted shape-based distance (DWSBD), which is defined as follows:

For two time series $\vec{x}=(x_1,\ldots\ x_m)$ and $\vec{y}=(y_1,\ldots\ y_m)$, the correlation is held $\vec{x}$ stationary, shift $\vec{y}$ on $\vec{x}$ and calculate the inner product of each shift s.For all possible shifts s ∈ [-m + 1, m- 1], we can compute the inner product $CC_s(\vec{x},\vec{y})$ as the similarity of $\vec{x}$ and $\vec{y}$ with shift S.

$$CC_s(\vec{x},\vec{y})=\begin{cases}\sum_{i=1}^{m-s} x_{s+i}\cdot y_i & ,s\geq 0\\ \sum_{i=1}^{m+s} x_i\cdot y_{i-s} & ,s<0\end{cases} \tag{2}$$

Cross-correlation is the maximum value of $CC_s(\vec{x},\vec{y})$, which means the similarity of A and B at the best phase shift s. In practice, the normalized correlation (NCC) is often used to limit values to [-1,1]. The definition of NCC is as follows:

$$NCC(\vec{x},\vec{y})=\max_s \frac{CC_s(\vec{x},\vec{y})}{\vec{x}_2\cdot\vec{y}_2} \tag{3}$$

Then we can define the SBD distance according to the NCC, which ranges from 0 to 2, where 0 means two time series have exactly the same shape. The smaller the SBD, the higher the shape similarity.

$$SBD(\vec{x},\vec{y})=1-NCC(\vec{x},\vec{y}) \tag{4}$$

Meanwhile, due to the problem that the DBSCAN algorithm cannot cluster multi-density data sets, we define a new distance, DWSBD, based on the SBD distance, which takes the local density between time series into account in the similarity measure of time series. First, define the local density of p as:

$$LRD_{k(p)}=\frac{\left|N_{k(p)}\right|}{\sum_{o\in N_{k(p)}} reach-distance_{k(p,O)}} \tag{5}$$

On this basis, the locally density-weighted shape distance is defined as:

$$DWSBD(\vec{x},\vec{y})=SBD(\vec{x},\vec{y})\cdot\left|LRD(\vec{x})-LRD(\vec{y})\right|^{\alpha} \tag{6}$$

Where, α is the weight index, the larger α is, the smaller the weight of the penalty term, then the greater the attention to the SBD distance of the two time series.

3.3 DSBD-based DBSCAN clustering algorithm

We decide to adopt DBSCAN, a density-based clustering method in our work for two reasons. One is that it can avoid predetermining the number of clusters. Secondly, the similarity of shapes between KPIs can be used to extend the clusters. Empirically, we set MinPts = 4 [5]. Next we discuss how to determine the key Eps parameters.

1) Plot the k-distance graph. Calculate the DWSBD from each KPI sample to its k-Nearest-Neighbor (KNN), and plot these distances in descending order to form a k-distance curve. The position below the inflection point of the curve, i.e. the flat part of the curve, is the candidate radius value, and the part above the inflection point, i.e., the steep part, corresponds to a drastic change in density and is not suitable as a density radius [6].

2) Take the empirical value 0.05 as the upper bound of the radius [7], and the final density radius is the largest candidate less than this value. In practical, the upper bound can be adjusted to obtain clusters with different granularity. (Figure 2)

After creating a clustering model for the sampled KPIs, we get K natural clusters, calculate the centroid of each cluster, and assign the rest of KPIs to the clusters based on the centroid. Generally, in each cluster, the object with the smallest sum of squared distances from others is considered as the cluster centroid. The centroid is defined as follows:

$$Centroid=\arg\max_{\vec{x}\in cluser_i}\sum_{\vec{y}\in cluser_i} DWSBD(\vec{x},\vec{y})^2 \tag{7}$$

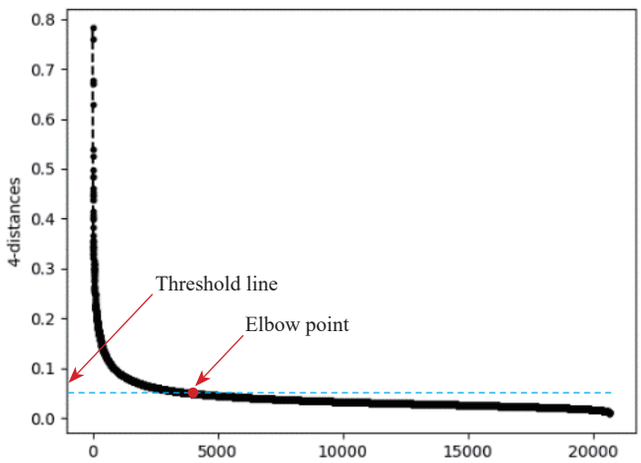


Figure 2. The k-dis curve of the SBD (k = 4). The steep part indicates the drastic density change, and the red dots represent the elbow joint points

4. Experiment and evaluation

4.1 Datasets

(1) Real KPI: Real KPIs collected by the monitoring platform of a domestic Internet company for five consecutive weeks from March 1 to April 5, 2019, and contains data on multiple dimensions including CPU utilization, hard disk temperature, and memory usage.

(2) "Idealized" public time series dataset: the six publicly time series datasets are from the UCR website[8], and the sequence length, number of samples, and number of categories vary significantly. The different datasets come from different scenarios and belong to different domains with good breadth and unknowns.

To quantitatively evaluate the clustering performance of DSDC, we use the following three popular metrics: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), F_1 - *score* .

4.2 Comparison with other distances

In the experiment, we compare DWSBD with other popular distance DWED, DWDTW and SBD, on the real KPI dataset and the "idealized" public datasets to demonstrate the effectiveness of DWSBD. The results of ACC, NMI, F-score, kd and clustering time are shown in Table 1 and Figure 3.

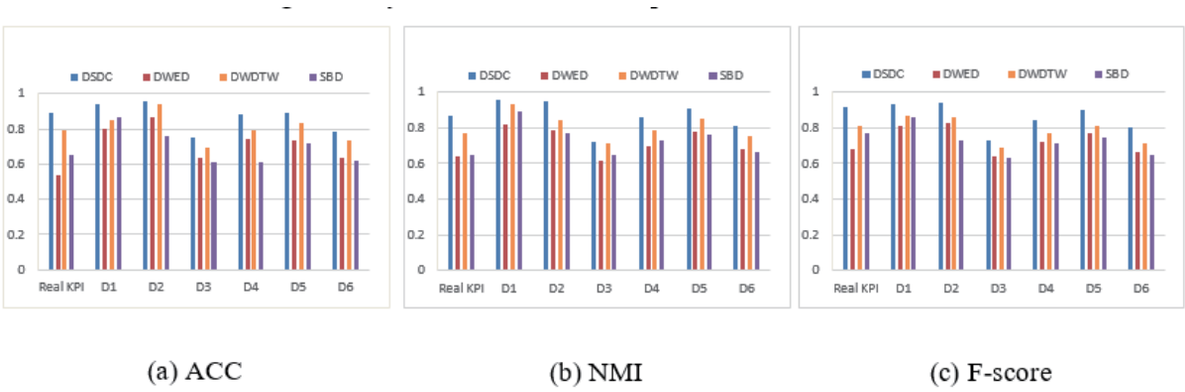


Figure 3. ACC, NMI, Fl-score comparison of different distance measures

As shown in Figure 3, DWSBD distance performs significantly better than the DWED and SBD and slightly better than the DWDTW on all seven datasets, especially in the real KPI dataset. Specifically, the DWED and SBD are confused by the frequent phase shifts and anomalies in the real KPI dataset because it is applied directly on the raw data without the necessary alignment and thus cannot obtain accurate similarities.

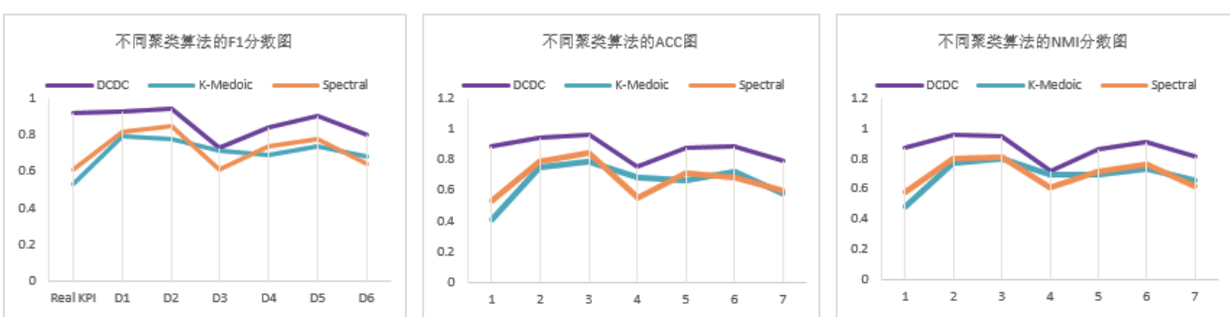
Table 1. Computation time and number of clusters for different distance and clustering algorithms

KPIs	Sample size	Metrics	DCDC	DWED	DWDTW	SBD	K-Medoids	Spectral
Real KPI (true k=6)	371	kd	6	7	5	4	4	5
		Time	490	45	27150	480	650	665
D1 (true k=12)	390	kd	12	15	11	9	10	13
		Time	26	2.1	558	28	40	45
D2 (true k=14)	1690	kd	14	12	14	12	11	12
		Time	22	1.8	486	20	49	44
D3 (true k=4)	88	kd	4	3	4	4	3	4
		Time	14	0.9	322	18	36	38
D4 (true k=7)	175	kd	7	4	7	7	7	6
		Time	15	0.8	472	12	44	39
D5 (true k=2)	150	kd	2	3	2	3	3	3
		Time	17	1.2	490	16	34	29
D6(true k=6)	242	kd	6	5	6	5	5	7
		Time	21	1.1	521	25	46	35

In Table 1 (columns 4-7), bold represents that the correct natural clusters were found for each dataset. Clearly, DSDC is able to find the number of clusters (k = kd) accurately, significantly better than DWED-based DBSCAN. This is because DWED cannot handle the shift problem. When computing distances, DWED assumes that the ith point in one sequence is aligned with the ith point in another sequence, so they are sensitive to distortions in the time dimension. Overall, compared to different similarity measures, DWSBD outperforms ED in most cases and DWSBD outperforms DWDTW in some cases. But DWSBD is faster than DWDTW.

4.3 Comparison with other clustering algorithms

In the experiments, DSDC is compared with K-Medoids, and Spectra [9] clustering algorithms, which are also based on the underlying shape extraction technique with DWSBD, to demonstrate the effectiveness of the DBSCAN algorithm in DSDC. The comparison results are shown in Figure 4 and Table 1.



(a) F-score (b) ACC (c) NMI

Figure 4. Performance comparison different clustering algorithms

As shown in Figure 4, on D3, all three methods perform poorly; on D1 and D2, all three methods perform well. In the real KPI dataset, the DSDC differs the most significantly from the other two algorithms. The purple lines in the three subplots are at the highest position on any dataset, indicating that DSDC has the best clustering results in all seven datasets. From Table 1 (columns 4, 8, 9), we can see that DSDC was able to find the correct number of clusters (k = kd) accurately on all seven datasets. Especially on the real KPI data set, only DSDC found the correct cluster number.

5. Conclusion

In this paper, a new robust density-based spatial clustering algorithm DSDC is proposed to solve the problem that the existing KPI time series are difficult to obtain robust and efficient clustering results due to the large amount of noise,

anomalies, phase shifts, etc. Firstly, the underlying shape extraction technique of KPI data is proposed for the problem of large amount of noise, anomaly and amplitude non-uniformity in KPI time series, and secondly, the DBSCAN algorithm is selected for clustering based on the problem of large amount of phase shift in KPI time series. Finally, the DWSBD distance is defined for the parameter-sensitive problem of DBSCAN algorithm. The DWSBD distance is combined with the DBSCAN clustering algorithm to cluster the KPI time series and extract the clustering center of mass. The evaluation between a real KPI dataset and six public UCR datasets shows that the DSDC algorithm proposed in this paper, significantly outperforms other clustering algorithms in t clustering KPI time series. The average ACC, NMI, and F-score of DSDC are 20% higher than those of K-Medoids and Spectral algorithms, with higher clustering performance with shorter clustering time.

References

[1] A. Fahim, "An extended DBSCAN clustering algorithm", International Journal of Advanced Computer Science and Applications (IJACSA), vol.13, no. 3, 2022.

[2] Nakagawa K, Imamura M, Yoshida K. Stock Price Prediction using k-Medoids Clustering with Indexing Dynamic Time Warping[J]. IEEJ Transactions on Electronics Information and Systems, 2018, 138(8):986-991.

[3] J. Zhao, L. Itti, shapedtw: Shape dynamic time warping, Pattern Recognition 74 (2018) 171–184.

[4] X. Xi, E. Keogh, L. Wei, A. Mafra-Neto, Finding motifs in a database of shapes, in: Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, 2007, pp. 249–260.

[5] M. F. Hassanin, M. Hassan, and A. Shoeb, "DDBSCAN: Different densities-based spatial clustering of applications with noise", International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 401–404, 2015.

[6] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," Radiology, vol. 227, no. 3, pp. 617–628, 2003.

[7] R. Hyde, and P. Angelov, "A fully autonomous data density based clustering technique", IEEE Symposium on Evolving and Autonomous Learning Systems (EALS), pp. 116–123, 2014.

[8] Dau H,Bagnall A,Kamgar K,et al.The UCR time series archive[J].IEEE-CAA Journal of Automatica Sinica,2019,6(6):1293-1305.

[9] Souto M, Coelho A, Faceli K, et al. A Comparison of External Clustering Evaluation Indices in the Context of Imbalanced Data Sets[C]// Neural Networks. IEEE, 2012.