Clustering Streaming Time Series Using CBC

Weimin Li, Liangxu Liu, and Jiajin Le

College of Computer Science and Technology of Donghua University, 1882 West Yan'an Road, Shanghai, China, 200051
108wml@mail.dhu.edu.cn

Abstract. Clustering streaming time series is a difficult problem. Most traditional algorithms are too inefficient for large amounts of data and outliers in them. In this paper, we propose a new clustering method, which clusters Bi-clipped (CBC) stream data. It contains three phrases, namely, dimensionality reduction through piecewise aggregate approximation (PAA), Bi-clipped process that clipped the real valued series through bisecting the value field, and clustering. Through related experiments, we find that CBC gains higher quality solutions in less time compared with M-clipped method that clipped the real value series through the mean of them, and unclipped methods. This situation is especially distinct when streaming time series contain outliers.

1 Introduction

There are extensive studies on efficient algorithms that can manage large data sets [1, 2]. One of the popular data analyzing problems is data clustering. Clustering is an unsupervised learning method, because the data are assigned to the corrected cluster without knowing which cluster they belong to in the learning methods. Efficient and accurate clustering for data stream is a key problem, and many techniques have been proposed for clustering data stream [3, 4, 5].

In this paper, we investigate the clustering time series data streams, which is meaningful [6]. In some data sources, the generation rates of time series data streams have become faster than ever before. The rapid time series data streams have challenged the storage and computation time in our computing systems.

We propose a new approach that clusters on Bi-clipped (CBC) series. We first transform the time series data streams into the Piecewise Aggregate Approximation (PAA) representation, and then clip them into binary series (Bi-clipped) that is different from the process of clipping (M-clipped) in Anthony Bagnall et al. [7]. M-clipped method transforms the original series into a binary one with mean of them. However, mean is influenced by the outliers in original series. This is why we design the Bi-clipped method. Through CBC, we can find that our proposed method not only improves significantly space and time complexity, but also is better to diagnose outliers and have better cluster performance than M-clipped and unclipped series.

The remainder of this paper is organized as follows. In section 2, we briefly discuss related work. Section 3 outlines the proposed method CSC. In section 4, we discuss the advantages of CBC. An experimental evaluation of our method is given in section 5. Section 6 concludes the paper with a summary of the experiment results.

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 629–636, 2007.

2 Related Work

Clustering streaming time series has attracted considerable interest recently. However, main memory and time tend to be a bottleneck. A better approximate representation would be considered for clustering streaming time series. According to this point, many methods including the Symbolic Aggregate Approximation (SAX) [8], M-clipped [7], Extend SAX [9], the Discrete Wavelet Transform (DWT) [10] and the Discrete Fourier Transform (DFT) [11] have been introduced.

Streaming time series is organized along with time axis and processed in a continuous way. However, there are continuous changes of their trends and it is the point to extract useful patterns in data mining research [12]. The Symbolic Aggregate Approximation (SAX) is a symbolic representation through a two-stage process of dimension reduction and discretization [8]. Extended SAX aims to realize efficient and accurate discovering of important patterns, necessary for financial applications [9]. M-clipped [7] transforms the real valued time series into a binary series through mean. Here we give a good representation for streaming time series. Bi-clipped method combined SAX and M-clipped, and improve the mean of M-clipped method influenced by outliers. Bi-clipped data can not only be more compactly represented and efficiently manipulated, but can it be robust to outliers.

3 Clustering on Bi-clipped Stream Data

The streaming time series consist of a set of multidimensional points $S_1 \ldots S_K \ldots$arriving at time stamps $t_1 \ldots t_k \ldots$. Each data S_i is a multidimensional item containing m dimensions, marked by $S_K = (s_i^1 \ldots s_i^m)$. In this section, we combine the reduction technology with the clipped technology, and use it to cluster streaming time series. The clustering on Bi-clipped stream data (CBC) has three phrases, i.e., dimensionality reduction via piecewise aggregate approximation (PAA), Bi-clipped process of stream time series, and clustering the Bi-clipped data.

3.1 Dimensionality Reduction Via PAA

PAA has been proposed by Keogh et al. [13]. It is a simple dimensionality reduction technique to implement compared with more sophisticated techniques like Singular Value Decomposition (SVD), the Discrete Fourier Transform (DFT), and the Discrete Wavelets Transform (DWT).

As defined by Keogh et al. [13], a streaming data S of length n can be represented in a *d*-dimensional space by a vector D, and the i^{th} element of D is calculated by the following equation:

$$d_i = \frac{d}{n} \sum_{j=\frac{n}{d}(i-1)+1}^{\frac{n}{d}i} s_j \tag{1}$$

To transform the streaming data from n dimensions to d dimensions, each sequence of streaming data is divided into *d* "frames" with equal sized and the mean value of each frame is used as a coordinate of a d-dimensional feature vector**.** This

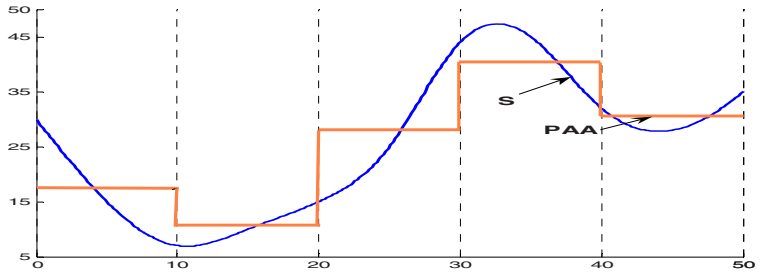

Fig. 1. A streaming data S is represented by PAA

data-reduced representation method is shown in Fig. 1, where the dimensionality is reduced from n=50 to d=5.

3.2 Bi-clipped Data

The dimensionality reduction through PAA is impossible too large, otherwise, more information may be lost. Streaming time series could be dimensionality reduction through PAA within measure. In this section, a new proposed streaming time series representation is described. It is different from the Symbolic Aggregate Approximation (SAX) [8] and M-clipped [7]. We transform PAA series into clipped data through the process of clipping. SAX and M-clipped transformation has similarities to our method. M-clipped transforms the real valued time series into a binary series through mean. However, when the streaming time series contains outliers, they would influence the mean value. Thus, we propose Bi-clipped method to overcome this shortcoming.

In this paper, clipping stream time series is transforming a PAA series D into a binary series Y. D is bisected into [min, a] and (a, max), and we called the whole process Bi-clipped. The elements in [min, a] are mapped to 0, otherwise they are mapped to 1. However, 'a' is not the mean value of D, but the value of the middle situation of the sorted D. In the same way, 'min' and 'max' represent the min and max place of the sorted D. For a time series, we can avoid the uncorrected mean influenced by outliers through using this method. Clipped data Y is produced from the following formula:

$$Y(t) = \begin{cases} 0 & if \quad D(t) \in [\min, a] \\ 1 & otherwise \end{cases} \tag{2}$$

3.3 Clustering the Bi-clipped Data

The problem of clustering is regularly defined as finding a partition of Y into k clusters. The cluster algorithm used in this paper is K-means clustering algorithm that classifies or groups given objects based on attributes into K numbers of group. It is one of the most popular and simple clustering algorithms. All cluster experiments performed in this study involve streaming time series of which the true clusters is not

known. The quality of a clustering model is measured by the sum of squared distances (SSQ) from each point to the cluster where it was assigned [14].

4 Advantages of CBC

4.1 Space and Time Improvements

The aim of using CBC and M-clipped methods is to pack them efficiently. We could gain better performance by analyzing the clipped data. And this reduction technology is effective for very long series like stream data. Because the stream data cannot store on disk, it is possible to analyze the Bi-clipped data of stream data in main memory.

The benefits of using CBC like M-clipped data could be packed efficiently in integer matrix and operated using bit operators, and they could provide a more significant time improvement than using unclipped data in bit operators. A series of doubles of length n using M-clipped method can be stored in an array of n/64 integers [7]. However, a Bi-clipped series of doubles of length n using PAA technology can be stored in an array of d/64. The average time taken to find clusters by different methods can be shown in section 5.2. In this paper, one aim of our study is to focus on improving the performance of clustering stream time series through using CBC.

4.2 Robust Performance to Outliers

When a reduction technology is adopted to deal with stream time series, it is crucial to maintain statistic information of stream time series. In this paper, we propose CBC method to hold the related information. A Bi-clipped series that convert from stream time series have the following properties: through analyzing Bi-clipped series of stream time series, we know that it is approximately a stationary series. We could gain the related information of stream time series from the Bi-clipped of it. In this aspect, CBC method is superior to M-clipped method.

CBC and M-clipped methods could eliminate outliers, but the fashion of them is different. M-clipped method uses the mean of original series to divide it into M-clipped data. However, our method applies PAA to dimensionality reduction and converts it to Bi-clipped through the number region of sorted stream time series, not the values of them. When the probability of outlier is higher, robust of the CBC method is superior to that of M-clipped in theory. Of cause, the robust of both methods absolutely exceed unclipped method.

5 Experiment

We performed a series of experiments to assess the performance of our method. To test the scalability of our method, we generate randomly 50 n-dimensional (n size from 5k to 25k points) datasets S. The data follow uniformly distributions. Through the Bi-clipped process, the data set S is converted to 50 n/d-dimensional points. In section 5.1, we compare the cluster results by using CBC, M-clipped and unclipped data. We find that the CBC method prior to the other methods in stream data that contain outlier. To measure the accuracy and quality of different clustering, we use

the SSQ described in Section 3. Next, another advantage in time complex is described in section 5.2.

5.1 Clustering Streaming Time Series

We perform experiments to analyze the ability of using CBC method compared with M-clipped and Unclipped. The Min/Average/Max SSQ value for K-means clustering using CBC, M-clipped and Unclipped data without outliers are shown in Table 1. Fig.2 shows the average value of SSQ. When the streaming time series do not contain outliers, the clustering quality of clipped series from CBC and M-clipped is

Table 1. Quality comparison without outliers

	CBC	M-clipped	Unclipped
N×M	SSQ value (Min/Avg/Max×1.0E+05)		
20×5000	0.593/0.613/0.645	0.489/0.545/0.639	0.636/0.668/0.687
20×10000	1.091/1.180/1.293	0.987/1.103/1.191	1.186/1.263/1.389
20×15000	1.641/1.757/1.950	1.795/1.871/2.099	1.787/2.015/2.093
20×20000	2.191/2.26932.395	2.190/2.341/2.590	2.584/2.635/2.988
20×25000	2.735/2.892/3.236	2.479/2.699/2.731	3.223/3.255/3.475
20×30000	3.290/3.440/3.590	3.287/3.472/3.584	3.870/3.984/4.476

better than that of unclipped data. When the length of data becomes longer, the trend is more distinct. Compared with M-clipped, CBC does not have obvious advantages, but it is faster than the M-clipped method in finding cluster, and the related contents will be discussed in section 5.2.

Table 2 shows the Min/Average/Max SSQ value for K-means clustering using the three methods with outliers. The clustering quality of clipped series including CBC and M-clipped is better than that of unclipped data as above. However, CBC method

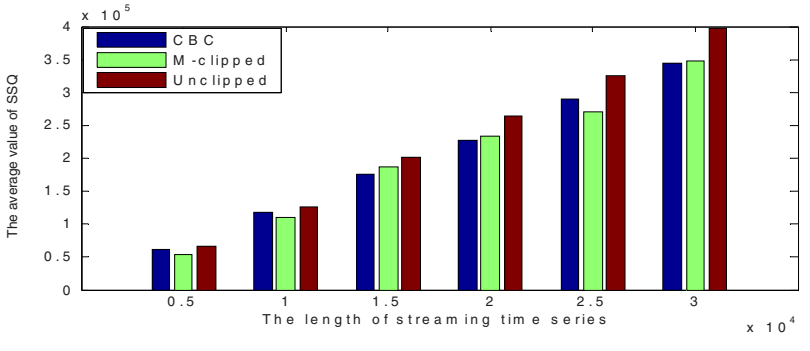

Fig. 2. The average value of SSQ without outliers

exceeds M-clipped method when the outliers exist. Fig.3 shows the average value of SSQ with outliers, which indicates that the Outliers significantly influence the clustering performance of Unclipped data. CBC method has better robust performance to outlier than M-clipped method. Therefore, when the streaming time series have outliers, the clustering quality of clipped series including CBC and M-clipped is better than that of unclipped data. Moreover, there are distinctly differences between CBC and M-clipped. CBC method has better robust performance than the other two methods. When the length of streaming time series is getting longer, CBC method has more distinct advantages than the other two methods in clustering performance. Therefore, CBC method gains better cluster results, which testify the thought about the benefit of CBC.

Table 2. Quality comparison with outliers

	CBC	M-clipped	Unclipped
N×M	(Min/Avg/Max×1.0E+05)		
20×5000	2.453/2.598/2.703	2.245/2.644/2.952	2.630/2.633/2.641
20×10000	4.982/5.164/5.465	5.026/5.381/5.966	5.898/6.481/6.839
20×15000	8.157/8.181/8.208	7.589/8.388/8.993	8.789/9.238/10.940
20×20000	9.059/9.985/11.816	10.021/10.839/11.862	11.743/12.33/13.611
20×25000	12.486/12.540/12.564	12.532/13.731/14.899	14.747/15.190/15.928
20×30000	13.690/14.908/16.506	16.503/16.882/17.949	17.744/19.491/20.544

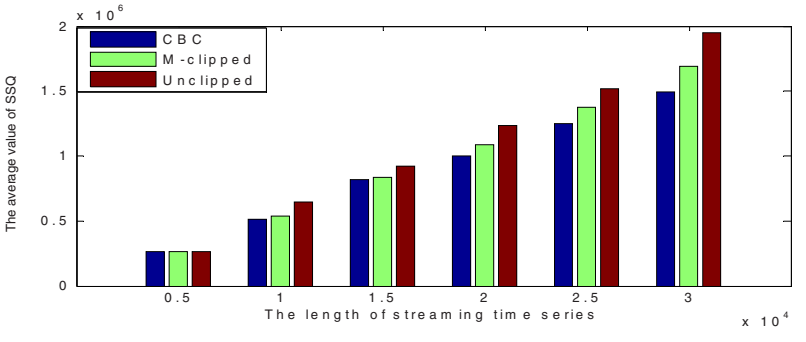

Fig. 3. The average value of SSQ with outliers

5.2 Time Complex

It is distinct that CBC and M-clipped methods could provide a more significant time improvement than using unclipped data in bit operators. In this section, we compare the average time of different means taken to find clusters. Fig.4 shows that with outliers. Each value in Fig.4 gives the time taken to find clusters in different data sets from 5k to 30k in length. The time to find cluster using CBC method is superior to the

other two methods. The time includes the process of clipped. When the length of streaming time series gets longer, it will take more time to find clusters. The ratio of time taken in CBC to that in the other methods is approximately a constant *n*/*d*. These results justify that our CBC method is a good dimensionality reduction and presentation method for clustering purposed.

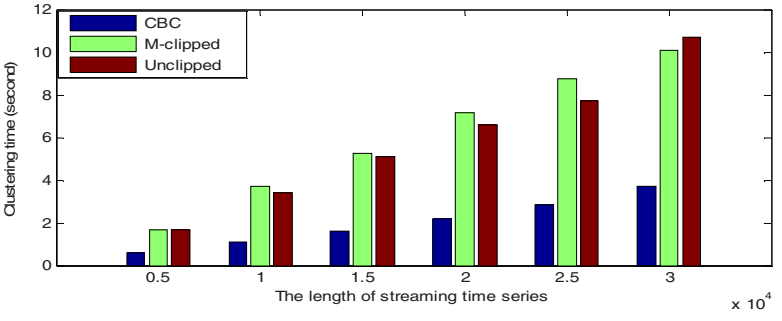

Fig. 4. Average time taken to find clusters with outliers

6 Conclusions

In this paper, considering the problem of clustering streaming time series mentioned above, we developed CBC techniques. Unlike existing methods such as M-clipped and SAX, CBC is based on PAA and Bi-clipped method. It contains dimensionality reduction process of PAA, Bi-clipped process of stream time series, and clustering the Bi-clipped data. The aim of using the CBC is to improve the performance of clustering and improve the time complex. Specific algorithms for binary series would improve time complexity [7]. According to our experiments, CBC performs much better than M-clipped and unclipped data in time complexity. A Bi-clipped series of doubles of length n can be stored in an array of *d*/64. Bi-clipped method allows for longer streaming time series to be stored in main memory. CBC gains better time complexity and space storage. The reduction technology with measure of CBC does not decrease the accuracy. When outliers exist in the streaming time series, CBC has better clustering performance than the other two methods.

References

1. Guha, S., Rastogi, R., Shim, K.: Cure: An efficient clustering algorithm for large databases. In ACM SIGMOD Conference (1998)
2. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. In ACM SIGMOD Conference (1996)
3. Aggarwal, C., Han, J., Wang, J., Yu, P. S.: A Framework for Clustering Evolving Data Streams, Proc. 2003 Int. Conf. on Very Large Data Bases (VLDB'03), Berlin, Germany (2003)

4. Guha, S., Mishra, N., Motwani, R., O'Callaghan, L.: Clustering data streams. In Proceedings of the Annual Symposium on Foundations of Computer Science. IEEE (2000)
5. Guha, S., Meyerson, A., Mishra, N., Motwani, R., O'Callaghan, L.: Clustering Data Streams: Theory and Practice TKDE special issue on clustering, Vol. 15 (2003)
6. Keogh, E., Lin, J., Truppel, W.: Clustering of time series subsequences is meaningless: Implications for previous and future research. In IEEE ICDE (2003) 115-122
7. Bagnall, A., Janacek, G.: Clustering time series with clipped data. Machine Learning 58(2-3) (2005) 151-178
8. --Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In ACM SIGMOD Conference (2003) 2-11
9. Lkhagva, B., Suzuki, Y., Kawagoe, K.: Extended SAX: Extended of Symbolic Aggregate Approximation for Financial Time Series Data Representation. In DEWS2006.
10. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In ACM SIGMOD Conference (1994) 419-429
11. Chan, K., Fu, A. W.: Efficient Time Series Matching by Wavelets. In IEEE ICDE (1999) 126-133
12. Han, J., Kamber, M.: Data Mining- Concepts and Techniques, Morgan Kaufmannn, (2000)
13. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases Journal of Knowledge and Information Systems (2000)
14. O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., Motwani, R.: Streaming-data algorithms for high quality clustering. In IEEE ICDE (2002)