CHAPTER VII

CONCLUSION

In this thesis, Shape-based Streaming Subsequence Time Series Clustering (3STSC) is proposed to return clustering results in constant time when a new data point arrives in time series data stream. In addition, 3STSC is extended from the proposed Shape-based Subsequence Time Series Clustering (2STSC) which produces more meaningful clustering results in terms of Shape-based Meaningfulness Measurement (SMM). To make 2STSC produce meaningful results, 2STSC utilizes Dynamic Time Warping (DTW) distance measure and Shape-based Averaging function. An intuitive idea is that DTW distance and Shape-based Averaging can handle a set of trivial-matched subsequences which are contiguous subsequences that have small differences because of time shift. Therefore, DTW distance aligns subsequences to find the optimal warping path between two sequences before distance calculation and Shape-based Averaging aligns subsequences to find an optimal warping path between two sequences before averaging. DTW distance and Shape-based Averaging are superior to the Euclidean distance and Amplitude Averaging used in Subsequence Time Series Clustering (STSC) in that Euclidean distance cannot capture the similarity between two subsequences, and Amplitude Averaging cannot preserve characteristics for producing averaged result. In other words, Euclidean distance in clustering algorithm can lead to incorrect grouping of trivial-matched subsequences, and Amplitude Averaging can lead to undesirable smoothing of trivial-matched subsequences. STSC has been proven as meaningless both theoritically and empirically that STSC will always produce sine waves as cluster representatives regardless of input sequences, where these sine waves are unusable. Therefore, in this thesis, 2STSC is proposed to overcome this problem, and then 3STSC is then proposed to support data streams.

This thesis can be extended to improve the performance further in many data mining tasks. Shape-based Averaging and Incremental Shape-based Averaging can be extended to be used in template matching problem and classification. 2STSC and 3STSC can be used as a preprocessing or a subroutine of many data mining tasks such as association rules, classification, pattern discovery, and visualization.

To improve the algorithms proposed in this thesis, a new methodology of sequence alignment and re-sampling technique can be designed for Shape-based Averaging algorithm, and the averaging scheme can be modified to find the optimal averaging result. Incremental Shape-based

Averaging can be improved by adding decremental algorithm so that the characteristics of an averaged result can be removed by a specific sequence. In addition, 2STSC can be improved by speeding up an algorithm and utilizing other clustering algorithm and removing user-defined parameters that are the number of clusters and the length of sliding window. For 3STSC, other than the number of clusters and the length of sliding window that should be removed, the update algorithm of stored subsequences should be improved to reduce distortion of meaningfulness of clustering results.