Summarization and Matching of Density-Based Clusters in Streaming Environments
by D. Yang, E. A. Rundensteiner, and M. O. Ward

1 Summary

In this paper authors presented a framework to support clusters detection and matching. Clusters are built in streaming environments and based on density. Given a stream authors designed an on-line strategy to detect and summarize clusters at the same time. All detected clusters are stored together with their summaries in a database. Efficiency of cluster retrieving for given statistics is the most important priority of their framework. Authors implemented and run experiments to confirm efficiency of their approach.

2 Strong Points

- The paper introduces new definitions in a clear and detailed way with examples, which are very helpful to clarify them. Clear explanations and examples make the paper easy to read and understand.
- The main idea of the framework is simple and efficient. It can be easily enhanced (different summarizations, different measure to compare patterns, etc.), described, and implemented.
- The life of each data pattern is described from its detection, through storing in a database, to matching with given patterns.

3 Weak Points

- Authors mentioned the impact of choosing grid cells to the performance of the framework. However, they did not mention its impact to density of points (e.g. Fidelity to Density Distribution). In given example (Figure 5) core points (black dots) are localized inside grid cells. If grid cells would cover core points among many grid cells, densities of such cells would be smaller. Would it impact the overall performance and quality of results?
- Authors did not analyze why Skeletal Grid Summarization (SGS) is better than other summarization measure. In addition, they did not compare SGS attributes to identify their impact while matching patterns. Maybe there is a redundant attribute, which could be dropped or a missing one, which could be added and improve the quality of matching?
- The framework, which is introduced and implemented by authors, bases on simple and efficient approaches of data summarization. Similarity of matched clusters was measured in the experiment that has been run with help of 20 human analysts, and not based on objective quantitative measures. Other experiments (time of generating different summations) show that the new framework requires similar amount of time to generate its summations.

4 Questions and Discussion Points

Defining Initial Grid for Other Dimensions. The framework, which is presented in the paper, clusters points into grid cells based on their location. Location is strongly correlated with density of points, but we can also cluster points based on other attributes. It would be interesting to verify if initial clustering according to other attribute would produce better/worse results. For example, what will be the performance if grid cells are formed based on connectivity of points/nodes, i.e., nodes with similar connectivity would be in the same/close to each other cells.

Privacy. Framework proposed by authors can be enriched by privacy-preserving techniques, which seems to be a natural extension to grid cell partitioning. We could use differential privacy for grid cells or generalizations of point attributes (location, connectivity, etc.). It would be interesting to check the impact of such privacy-preserving techniques to pattern matching.

Paper Presentations – Preferences.

1. Summarization and Matching of Density-Based Clusters in Streaming Environments by D. Yang, E. A. Rundensteiner, and Matthew O. Ward

2. Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data by I. Batal, D. Fradkin, J. Harrison, F. Moerchen, M. Hauskrecht