Aiming at real applications: Subsequence Outlier Detection on Mixed-Type Attributes Data in RDBMS

ABSTRACT

For many real applications, ranging from traditional management information systems (MIS) to e-commerce and socials web, produce large volumes of sequences of multivariate time-stamped observations. The classical data representation for these applications is an information table stored in RDBMS [1], in which rows stand for objects and columns denote attributes. Based on the above scenarios, unsupervised subsequence outlier detection on time series data is a valuable problem in practice which helps in saving the cost of labeling and providing interpretability in real applications. This problem is called subsequence outlier detection tasks [2]. In this abstract, we study this issue and show our preliminary analysis and ongoing work.

1 MOTIVATIONS

Although many classic subsequence outlier detection methods have demonstrated their effectiveness in various scenarios, they may not work well in real applications. Three characteristics of time series data increase the difficulties for the task:

- **M1: Attribute-Level Semantic Structure.** Attributes from information tables, which were designed by domain experts and database administrators carefully, exist an explicit or implicit semantic structure in a specific application scenario. Therefore, to detect the inner structure of attributes would be conducive to measure the isolated degree of a subsequence.
- **M2: Mixed-Type based Non-IID Similarity Metric.** A better subsequence outlier detection model should need to develop a meaningful and useful similarity metric in mixed-type attributes space and capture the dependencies between attributes of different types under Non-IID assumption.
- **M3: Local Characteristics Modeling.** Due to the effect of time drift, the isolated degree of a subsequence built by a series of comparisons between the subsequence and the set of other subsequences should find the different levels of restriction of the set to compare with, i.e., to utilize the characteristics of "locality" to model the isolated degree of a subsequence.

2 METHOD DESIGN

To cope with the challenges like the above, we believe that the following three design ideas are critical for the task.

- **Attributes Structure Learning.** We capture the semantic causal structure by using the structure learning methods of Bayesian network like [4], whose nodes represent attributes and edges represent the direct dependent relationship between attributes. (**M1**)
- **Non-IID Coupled Similarity Metric.** Inspired by the idea of Non-IID Learning [1] and based on the above attribute causal structure, we define a Non-IID coupled similarity metric in mixed-type attributes space consisting of intrinsic intra-attribute and inter-attribute similarity to capture the relationship between attributes and attribute values. (**M2**)
- **Locality Modeling.** According to the definition of traditional generalized local outlier detection framework (TraLOD) [3], we extend TraLOD to handle mixed-type attributes data for subsequence outlier detection tasks by incorporating the above similarity metric to pick out the surrounding neighborhood of subsequences and using the local information to measure the isolated degree of a subsequence. (**M3**)

3 CONCLUSION AND FUTURE WORK

In the above sections, we present motivations and ongoing work for subsequence outlier detection tasks. In summary, it would include the following three major potential contributions:

- **Novel Framework.** Our framework extends TraLOD to handle mixed-type attributes data for subsequence outlier detection tasks by utilizing the local characteristics of a subsequence to measure the isolated degree of being an outlier (called Locality Modeling).
- **Novel Non-IID Similarity Metric.** We propose a mixed-type based Non-IID coupled similarity metric considering the intrinsic intra-attribute and inter-attribute coupling between subsequences and utilize it to pick out the surrounding neighborhood of a subsequence.
- **Flexibility.** The novel framework can flexibly instantiate different classic local outlier detection methods to cope with subsequence outlier detection tasks and is easy to implement.

If this proposal is accepted for CIDR2020, we look forward to showcasing the details of the proposed framework and preliminary results at the conference.

REFERENCES

[1] Songlei Jian, Liang Hu, Longbing Cao, and Kai Lu. 2018. Metric-Based Auto-Instructor for Learning Mixed Data Representation. In *Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th {AAAI} Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New.* 3318–3325.

[2] Eamonn J Keogh, Jessica Lin, and Ada Wai-Chee Fu. 2005. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA.* 226–233.

[3] Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. 2014. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery* 28, 1 (2014), 190–237.

[4] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65, 1 (2006), 31–78.

[1] RDBMS refers to a relational database management system.

, ,
XXXX. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$XX
https://doi.org/10.1145/nnnnnnn.nnnnnnn