On the consequences of multiple instances in temporal sequence mining

Supervision:
- Thomas Guyet, IRISA/LACODAM (thomas.guyet@irisa.fr)

Context:
Sequence mining is a task that consists in extracting interesting subsequences from a collection of sequences. This pattern mining task is widely used to analyze behaviors from longitudinal traces collected on digital, physical or living systems. For instance, traces can be the logs of servers, it can be the purchase sequences of supermarket clients or the care pathways, *ie* the sequences of cares that sick people had.
The temporal sequence mining takes explicitly into account the continuous nature of time and is looking for patterns that both describe the sequentiality of the events and the inter-event durations. This is of particular importance for care pathways analysis to discriminate this two situations: if two events A and B occur closely in time it likely triggers a disease, but not in case the same events are timely distant.
Such kind of patterns, called discriminant temporal patterns, are the most desirable onces: they can be used to prevent from critical consequences of cares.

To discover discriminant patterns, we have to identify in which case A and B trigger the disease and in which way they did not. But, … it is not so simple, especially in care pathways for which events are repeated several times in one sequence (a patient has several times the same medics/cares). This problem corresponds to a multiple instance problem: one people encountered multiple times the A/B association … which association (if any!) witnesses the disease occurrence?

The problem of multiple instance learning is well-known in image classification. One image is made of several elements, but only one of this elements correspond to the label of the image, or sometime it is the combination of elements that witnesses the class.
The same situation occurs with discriminant temporal sequence mining and we believe that this problem has been under-estimated. It can be argue that if multiple instances are not handle, it lowers the accuracy of the patterns.

The topic of this internship is to study multiple-instance problems in discriminant temporal mining. It will propose new approach to handle the multi-instances in this sequence mining task and study its real impact on synthetic and real datasets.

The main steps of this work will be:
- the study of the state of the art of multiple instance learning and sequence mining,
- the proposal of alternative strategies to mine discriminant temporal patterns in case of multiple instance problems,
- the implementation of new algorithms,
- the proposal of an evaluation strategy on synthetic and real datasets, and the conduct of the experiments

References

- Y. Dauxais, D. Gross-Amblard, T. Guyet, and A. Happe, “Discriminant chronicles mining. Application to care pathways analytics,” in *AIME, 2017*.
- Chen, Y., Bi, J., & Wang, J. Z. (2006). MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 1931-1947.

- Carbonneau, M. A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, *77*, 329-353.

Expected profile and skills

- Programming in C/C++
- Programming in R
- Knowledge in machine learning techniques
- Curiosity
- Strong interest for experimental studies
- Scientific English