Reliably Capture Local Clusters in Noisy Domains From Parallel Universes

F. Höppner, M. Böttcher

– Extended Abstract –

When seeking for small local patterns it is very intricate to distinguish between incidental agglomeration of noisy points and true local patterns. We propose a new algorithm [2] that addresses this problem by exploiting temporal information which is contained in most business data sets. The algorithm enables the detection of local patterns in noisy data sets more reliable compared to the case when the temporal information is ignored. This is achieved by making use of the fact that noise does not reproduce its incidental structure but even small patterns do. In particular, we developed a method to track clusters over time based on an optimal match of data partitions between time periods. Using the terminology of parallel universes in [1], our approach is characterised as follows:

- Assumptions of Existence:
  - The set of all possible objects Ω contains certain objects of interest, say customers, that develop over time (e.g. change opinions or habits, establish new trends, ...).
  - We assume that there are small but interesting groups of objects, possibly embedded into larger but less interesting groups (clusters). The goal is to detect the individual groups, in particular the small ones (local patterns), and to assign the objects to them.
- Input:
  - We periodically gather some data about some of the customers, but we do not recognise a customer if we have seen her before (that is, there is no such thing as a customer ID). The observations of a specific period of time (e.g. observations of one month) constitute a data set (or universe). Our universes therefore do not differ in terms of features but in terms of actually observed customers and in the observation period.
- Output:

Dagstuhl Seminar Proceedings 07181
Parallel Universes and Local Patterns
http://drops.dagstuhl.de/opus/volltexte/2007/1261

  - partial models: Data from a specific period of time (from one universe) is clustered to find groups of similar customers. Given the obtained partition, we could assign each customer to its group, but in this case we would have no interaction between the universes. Therefore, we keep ambiguity in our partial models: Rather than assigning each object to a single cluster we keep a number of possible assignments (and a number of possible small clusters) and postpone the final assignment (and the final decision about the existence of a cluster).
  - model join: Subsequently generated partial models are compared against each other. Since we are not interested in incidental data agglomerations we use the rationale that a true local pattern should repeat itself in the next partial model. So the more often a certain structure is observed, the more we get convinced that this structure is substantial and not just noise. By this comparison of results across universes, we can finally decide about the significance of local structures in each individual universe.

- Goal:
  - In our experimental evaluation it turns out that we are able to detect local patterns more reliable by comparing parallel universes compared to the analysis of a single universe only.

References

[1] M. Berthold, B. Wiswedel: Learning in Parallel Universes (Seminar Summary), Dagstuhl Seminar 07181

[2] F. Höppner, M. Böttcher: Matching Partitions over Time to Reliably Capture Local Clusters in Noisy Domains. (accepted at PKDD 2007)