

Mining changes of patterns from multi-period datasets

Chun-Ta Dai

Institute of Information Management

National Cheng Kung University

Pattern discovery is a common task in data mining. Given the transaction datasets of multi periods, we are concerned with a temporal data mining problem that detects any pattern of interested changes that have been consistent from some period to the last period. Discovering such changes from the transaction database of multi periods will help the managers to detect the tendency of customer needs so that potential customers may be identified. To the best of our knowledge, previous studies in change mining only focus on datasets of two datasets, although the tendency of changes are more meaningful for datasets of multi periods in real-world applications. Conventional data mining techniques that seek frequent patterns could be modified for mining changes from datasets of multi periods, but such approaches would require many pairwise comparisons between datasets of consecutive periods and thus not so efficient.

In this thesis, we propose an algorithm called MCP for mining changes from multi-period datasets. MCP is based on a novel data structure modified from the popular frequent-pattern tree (FP-tree), and seeks the target patterns in a very efficient way. In particular, starting from the last two periods, our algorithm first constructs a candidate-pattern forest (CP-forest) to store those patterns of qualified changes, and then iteratively updates the CP-forest using the dataset of each period. The CP-forest is carefully designed such that useless information will not be stored and qualified patterns can be easily identified by tree traversals. Computational experiments have been conducted to compare MCP and another algorithm called modiFP which is modified from the popular FP-growth algorithm for mining the changes of patterns from multi-period datasets. Several parameters have been used to evaluate the performance of MCP and modiFP, and the results show that MCP is much more efficient than modiFP, especially when the number of periods increases or when the datasets of consecutive periods share more similarities.

Keywords: pattern discovery, change mining, candidate-pattern forest, temporal data mining, algorithm, data structure