Abstract

In this paper is explored a way to reduce the rate of re-clustering and speed up the clustering process on categorical time-evolving data. This method introduces two algorithms, RDE (Replicated Data Elimination) and RCRDE. The RDE algorithm removes the successive surveys of replicated data and considers counters to keep this data. Hence, the number of created windows via the sliding window technique is limited, and this leads to a decrease in the number of implementations of clustering algorithms. The RCRDE algorithm, based on MARDL (MAXimal Resemblance Data Labeling) framework, decides about re-clustering implementation or modification of previous clustering results. The presented method is independent of the clustering algorithm's type and any kind of categorical clustering algorithm can be used. According to the results obtained on different data
sets, this method performs well in practice and facilitates the clustering implementation on categorical data. Also, this method can be utilized to cluster a very large categorical static database with higher quality than previous work.

References

- C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," Proceedings of the 29th International Conference Very Large Data Bases (VLDB), Sep. 2003, pp. 81–92.
- E. H. Han, G. Karypis, V. Kumar, B. Mobasher, "Clustering Based on Association Rule Hyper graphs," Proceedings of ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery (DMKD), 1997.
- Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, 1998, pp. 283-304.
- V. Ganti, J. Gehrke, R. Ramakrishnan, "CACTUS-Clustering Categorical Data Using Summaries," Proceedings of the ACM SIGKDD international conference on
Knowledge discovery and data mining, 1999, pp. 73-83.

Index Terms

Computer Science
Artificial Intelligence
Keywords
Categorical time-evolving data  clustering  data labeling  drifting-concept detecting