Abstract

The data handled in emerging applications like location based services, sensor monitoring systems, and data integration, are often inexact in nature. In this paper, the important problem of extracting frequent item sets from a large uncertain database, interpreted under the Possible World Semantics (PWS) is presented. This issue is technically challenging, since an uncertain database contains an exponential number of possible worlds. By observing that the mining process can be modeled as a Poisson binomial distribution, an algorithm was developed, which can efficiently and accurately discover frequent item sets in a large uncertain database. The important issue of maintaining the mining result for a database that is evolving (e.g., by inserting a tuple) can be presented. Specifically, the proposed mining algorithm can enable Probabilistic Frequent Item set (PFI) results to be refreshed. This reduces the need of re-executing the whole mining algorithm on the new database, which is often more expensive and unnecessary. The proposed algorithm can support incremental mining and provides the accurate results on mining the uncertain database. The extensive evaluation on real data set to validate the approach is performed.
References

- C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent Pattern Mining with Uncertain Data," Proc. 15th ACM SIGKDD Int&apos;i Conf. Knowledge Discovery and Data Mining (KDD), 2009.
- D. Cheung, J. Han, V. Ng, and C. Wong, "Maintenance of Discovered Association Rules in Large Databases: An Updating Technique," Proc. 12th Int&apos;i Conf. Data Eng. (ICDE), 1996.
- C. K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. advances in Knowledge Discovery and Data Mining (PAKDD), 2007.
- A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, &quot;Model-Driven Data Acquisition in Sensor Networks,&quot; Proc. 13th Int'l Conf. Very Large Data Bases (VLDB), 2004.
- J. Han, J. Pei, and Y. Yin, &quot;Mining Frequent Patterns without Candidate Generation,&quot; Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000.
- H. Kriegel and M. Pfeifle, &quot;Density-Based Clustering of Uncertain Data,&quot; Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.
- C. K.-S. Leung, Q. I. Khan, and T. Hoque, &quot;Cantree: A Tree Structure for Efficient Mining of Frequent Patterns,&quot; Proc. IEEE Fifth Int'l Conf. Data Mining (ICDM), 2005.
Information and Knowledge Management (CIKM), 2010.

Index Terms

Computer Science
Artificial Intelligence

Keywords
PFI  PWS  S-PMF  CDF