Abstract

We consider the problem of discovering frequent item sets and association rules between items in a large database of transactional databases acquired under uncertainty. A probabilistic database considered here is one in which with each transaction associated is a probability, represents the confidence that the transaction will occur with given associated certainty. In this paper, we address the problem of the efficiency of the main phase of most data mining applications: The frequent pattern extraction. This problem is mainly related to the number of operations required for counting pattern supports in the database and we propose a new method, called counting inference probabilistic frequent pattern miner in probabilistic databases, this algorithm allows to perform as few support counts as possible. It is optimized to reduce the number of database scan as well as the number of patterns for which explicit support count is required. Using this method, the support of a pattern is determined without accessing the database whenever possible, using the supports of some of its sub-patterns called key patterns. This method was implemented in the CIPFP, counting inference based probabilistic frequent pattern mining algorithm that is an optimization of the simple and efficient Apriori algorithm. The
goal is to transform all key patterns into non-key patterns as early as possible as for
non-key-patterns database scan is not required at all.

References

1. R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases.
   In Proc. of the Fourth International Conference on Foundations of Data Organization and
database mining applications. In Proc. of the VLDB Conference, pages 560-573, Vancouver,
   IEEE Transactions on Knowledge and Data Engineering, 5(6):914-925, December 1993.
   Special Issue on Learning and Discovery in Knowledge-Based Databases.
4. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in
   large databases. In Proc. of the ACM SIGMOD Conference on Management of Data,
5. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large
databases. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California,
June 1994.
7. R. Brachman et al. Integrated support for data archeology. In AAAI-93 Workshop on
   Knowledge Discovery in Databases, July 1993.
9. P. Cheeseman et al. Autoclass: A Bayesian classification system. In 5th Intl. Conf. on
11. J. Han, Y. Cai, and N. Cercone. Knowledge discovery in databases: An attribute
    oriented approach. In Proc. of the VLDB Conference, pages 547-559, Vancouver, British
    RJ 9567, IBM Almaden Research Center, San Jose, California, October 1993.
    research: Search directions in knowledge discovery. SIGMOD RECORD, 20(3):76-78,
    September 1991.
15. P. Langley, H. Simon, G. Bradshaw, and J. Zytkow. Scientific Discovery: Computational
17. H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering
    association rules. In KDD-94: AAAI Workshop on Knowledge Discovery in Databases, July
    1994.
46. M. Yiu et al. Efficient evaluation of probabilistic advanced spatial queries on existentially uncertain data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 2009.

Index Terms

Computer Science
Algorithms

Keywords

Probabilistic frequent patterns, probabilistic frequent rule, key-patterns, non-key-patterns,