ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Effective Data Mining Of High Quality Patterns in Sequence-Level and Element-Level Uncertain Data Models

Sumamadhuri Roy
M.Tech(CSE),
Department of Computer Science and Engineering,
Sarada Institute of Science Technology and Mangement, Srikakulam.

V.Lakshmiprasad
Assistant Professor,
Department of Computer Science and Engineering,
Sarada Institute of Science Technology and Mangement, Srikakulam.

Abstract:

Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.Data mining methodologies have been developed for exploration and analysis of large quantities of data to discover meaningful patterns and rules.

Frequentpattern mining is an important model in data mining.Sequential pattern mining is key task in data mining which detects frequent sub itemsets in the sequential databases. There are many approaches which is based on pattern length and prefix according to projected databases.

In traditions approaches frequent pattern mining taking more processing time and outputs inaccurate frequent items. So we proposed a acyclic top-down directed approach for finding the frequent itemsets and it results best patterns after optimization.

Keywords:

Data mining, Frequent patterns, uncertain databases, Existential probability, Apriori algorithm, Incremental mining.

Introduction:

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices.

These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

Data mining involves six common classes of tasks:

Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression – attempts to find a function which models the data with the least error.
Summarization – providing a more compact representation of the data set, including visualization and report generation.

"Pattern mining" is a data mining method that involves finding existing patterns in data. In this context patterns often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. For example, an association rule "beer potato chips (80%)" states that four out of five customers that bought beer also bought potato chips.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise." Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search methods.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation involves combining data together (possibly from various sources) in a way that facilitates analysis (but that also might make identification of private, individual-level data deducible or otherwise apparent). This is not data mining per se, but a result of the preparation of data before – and for the purposes of – the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when the data were originally anonymous.

FREQUENT PATTERN MINING:

Definition: Let I = {x1,x2 ..... xn} be a set of items. An item set X is a subset of items, ie, X I. For the sake of brevity, an item set X = {x1, x2.... xm}is also denoted as X = x1 x2...xm. A transaction T = (tid, X) is a 2-tuple, where tid is a transaction-id and X is an item set. A transaction T = (tid, X) is said to contain item set Y if andonly if Y⊆ X. A transaction database TDB is a set of transactions. The number of transactions in TDB containing item set X is called the support of X, denoted assup(X). Given a transaction database TDB and a support threshold minsup, an item set X is a frequent pattern, or a pattern in short, if and only if sup (X) >minsup.The problem of frequent pattern mining is to find the complete set of frequent patterns in a given transaction database with respect to a given supportthreshold.

Existing System:

The Apriori-All algorithm is one of the earliest toapriori-based approaches. It first finds all frequent itemsets,transforms the database so that each transaction isreplaced by all frequent itemsets it contains, and thenfinds patterns. The GSP algorithm is an improvementoverApriori-All.

To reduce candidates, GSP only creates anew length k candidate when there are two frequentlength (k-1) sequences with the prefix of one equal to thesuffix of the other. To test whether a candidate is a frequentlength k pattern, the support of each length k candidateis counted by examining all the sequences.

ThePSP algorithm is similar to GSP except that theplacement of candidates is improved through a prefix treearrangement to speed up pattern discovery. The SPIRIT algorithm uses regular expressions as constraints anddeveloped a family of algorithms for pattern mining underconstraints based on a priori rule.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Proposed System:

In our work we proposed an improved sequential pattern mining algorithm which consists of bidirectional pattern mining which means forward and backward pattern mining. Commonprefix and grown uni-directionally along the suffixdirection of detected patterns. At each level of recursionthe length of detected patterns is only grown by 1.

If wecan grow the patterns bi-directionally along both ends ofdetected patterns, we may grow patterns in parallel ateach level of recursion. The motivation of this paper is tofind suitable partitioning, projection, and detection strategiesthat allow for faster pattern growth. To support bidirectional pattern growth, instead ofpartitioning patterns based on common prefix, we canpartition them based on common root items. For a databasewith n different frequent items (without loss of generality,we assume these items are 1, 2, … , n), its patternscan be divided into n disjoint subsets.

The ith subset (1 ≤ i≤ n) is the set of patterns that contain i (the root item ofthe subset) and items smaller than i. Since any pattern insubset i contains i, to detect the ith subset, we need onlycheck the subset of tuples whose sequences contain i indatabase D, i.e., the projected database of i, or iD. In the ithsubset, each pattern can be divided into two parts, prefixand suffix of i. Since all items in the ith subset are no largerthan i, we exclude items that are larger than i in iD.

Algorithms Used:

Algorithm (ATDD based pattern Mining).
Input: A database D and the minimum support
Output: P, the complete set of patterns in D
Method: findP (D, minSup){
P=FISet=D.getAllFI(minSup);
D.transform();
for each FI x in FISet{
UDVertexrootVT=new UDVertex (x)
findP(D.getPreD(x), rootVT, up, minSup)
findP(D.getSurD(x), rootVT, down, minSup)
findPUDDAG(rootVT)
P = P rootVT.getAllPatterns()
}
}

The algorithm then transforms the database. A direct-edacyclic graph is built to represent the containingrelationship of FIs. For each (sorted) itemset, we check allits FIs with children in the DAG, and verify whether theFI corresponding to each child is valid in the itemset. Ifso, we add the id of the child to the itemset and furthercheck the children of that child.Based on the transformed database, for each FI x, thealgorithm creates a root vertex for <x>, detects all the patternsin the prefix projected database and suffix projecteddatabase of x, creates x-ATDD, detects Px using x-ATDD, and add Px to P.

Subroutine: findP(PD, rootVT, type, minSup){
FISet=PD.getAllFI(minSup);
for each FI x in FISet{
UDVertexcurVT=new UDVertex (x, rootVT)
if(type==up) rooVT.addUpChild(curVT)
elserootVT.addDownChild(curVT)
findP(PD. getPreD(x), curVT, up, minSup)
findP(PD.getSufD(x), curVT, down, minSup)
findPUDDAG(curVT)
}
}

This subroutine detects all the patterns whose ids areno larger than the root of the projected database. The parametersare

1) PD is the projected database;
2) rootVT is
the vertex for the root item of PD;
3) type (up/down) indicatesprefix/suffix PD;
4) minSup is the support threshold.
The subroutine first detects all the FIs in PD. For eachFI x, it creates a new vertex as the Up/Down child (basedon type) of the root vertex. It then recursively detects allthe patterns in PD similar as
findP (D, minSup).
Subroutine: findPUDDAG(rootVT){
upQueue.enQueue(rootVT.upChildren)
while(!upQueue.isEmpty()){
UDVertexupVT=upQueue.deQueue()
if(upVT.upParent == rootVT)
downQueue.enQueue(rootVT.downChildren)
else if (upVT.downParent==null)
downQueue.enQueue(upVT.upParent.VDVS)
elsedownQueue.enQueue(upVT.upParent.VDVS
upVT.downParent.VDVS)

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

```
while(!downQueue.isEmpty()){
UDVertexdownVT=downQueue.deQueue()
if(isValid(upVT, downVT){
UDVertexcurVT=new UDVertex (upVT, downVT)
upVT.addVDVS(downVT)
if(upVT.upParent==rootVT)
downQueue.enQueue(downVT.children)
}
}
if(upVT.VDVS.size>0)upQueue.enQueue(upVT.children)
}
```

Conclusion:

In this paper a novel data structure ADDD is inventedfor efficient pattern mining. The new approach growspatterns from both ends (prefixes and suffixes) of detectedpatterns, which results in faster pattern growth because of less levels of database projection compared totraditional approaches. Extensive experiments on bothcomparative and scalability study have been performed to evaluate the proposed algorithm.

References:

[1] Zhou Zhao, Da Yan, and Wilfred Ng, Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, MAY 2014.

[2] Liang Wang, David Wai-Lok Cheung, Reynold Cheng, Member, IEEE, Sau Dan Lee, and Xuan S. Yang," Efficient Mining of Frequent Item Sets on Large Uncertain Databases", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 24, NO.12, DECEMBER 2012.

[3] Chiu, 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.

[4] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle,"Probabilistic Frequent Itemset Mining in Uncertain Databases," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),2009.

[5] Carson Kai-Sang Leung, Mark Anthony F. Mateo, and Dale A.Brajczuk," A Tree-Based Approach for Frequent Pattern Mining from uncertain Data", T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 653–661, 2008.

[6] C. C. Aggarwal and P. S. Yu. A survey of uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng., 21(5):609-623, 2009.

[7] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In SIGMOD, pages 1-12, 2000.

[8] Q. Zhang, F. Li, and K. Yi, "Finding Frequent Items in Probabilistic Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.

[9] L. Wang, R. Cheng, S.D. Lee, and D. Cheung, "Accelerating Probabilistic Frequent Itemset Mining: A Model-Based Approach," Proc. 19th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2010.

[10] D. Cheung, J. Han, V. Ng, and C. Wong, "Maintenance of Discovered Association Rules in Large Databases:An Incremental Updating Technique," Proc. 12th Int'lConf. Data Eng. (ICDE), 1996 .

[11] C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent Pattern Mining with Uncertain Data," Proc. 15th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (KDD), 2009.

[12] Carson Kai-Sang Leung, Dale A. Brajczuk," Efficient Algorithms for the Mining of Constrained Frequent Patterns from Uncertain Data"ACM SIGKDD Explorations Volume 11, Issue 2.

Author Details :

Sumamadhuri Roy

is student in M.Tech(CSE) in Sarada Institute of Science Technology and Management, Srikakulam. He has received his B.Tech(CSE) from Lenora college of engg, Rampachodavaram, East Godavari Dist. . His interesting areas are Data Mining, Networking.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Mr.V.Laxmiprasad

is working as a Asst.professor in Sarada Institute of Science, Technology And Management, Srikakulam, Andhra Pradesh. He received his M.Tech (CSE) from GMRIT Rajam ,Srikakulam District, JNTU Kakinada Andhra Pradesh. His research areas include Computer networks, Datawarehouse and Datamining.