Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2304

Improving Efficiency of High Utility Sequential Pattern Extraction

Vikrant Sharma

Asst. Professor, Department of Computer Science, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number: 234-242
Publication Issue:
Vol. 70 No. 1 (2021)

Article History
Article Received: 25 January 2021
Revised: 24 February 2021
Accepted: 15 March 2021

Abstract
Text mining used on texts and publications in the biomedical and molecular biology fields is referred to as "biomedical text mining." It is a relatively new area of study at the intersection of computational linguistics, bioinformatics, and natural language processing. Superior usefulness the goal of sequential pattern mining is to identify statistically significant patterns among data instances when the values are presented sequentially. Time series mining is typically regarded as a distinct activity even if it is closely linked since it is typically assumed that the values are discrete. Structured data mining has a unique use known as sequential pattern mining. High utility pattern (HUP) mining is one of the most relevant study areas in data mining nowadays since it is capable of taking into consideration the nonbinary frequency values of items in transactions as well as different profit values for each item. The utilization of previous data structures as well as mining outcomes, yet, enables incremental and interactive data mining to eliminate the need for further calculations when a database is updated or the minimum threshold is modified. The method in this study suggests three innovative tree architectures for effective incremental and interactive HUP mining. The high utility sequential pattern mining issue has formalised key ideas and elements.

1. Introduction

Data mining's field of "sequential pattern mining" seeks statistically significant patterns among instances of data when values are presented sequentially. Time series mining is typically regarded as a distinct activity even if it is closely linked since it is typically assumed that the values are discrete. Structured data mining has a unique use known as sequential pattern mining.

In this discipline, a number of significant classical computational issues are addressed. Building effective databases and indexes for sequence information, identifying recurring patterns, assessing the similarity of sequences, and recovering missing sequence elements are a few of these. String mining, which is often based on string processing methods, and item set mining, which is primarily focused upon association rule learning, are two main categories for sequence mining challenges.

Text mining used on texts and publications in the biomedical and molecular biology fields is referred to as "biomedical text mining." It is a relatively new area of study at the intersection of computational linguistics, bioinformatics, and natural language processing. Due to the rise in electronically accessible articles kept in databases like PubMed, there is a growing interest

in text mining as well as information extraction techniques used on the biomedical as well as molecular biology literature.

The main advancements in this field have been in the detection of biological entities (named entity recognition), which involves names of proteins, genes, chemicals, and drugs in free text, association of gene clusters found through microarray experiments with the biological context stipulated by the pertinent literature, automatic extraction of protein interactions, along with association of proteins with functional theories. Information extraction and text mining technologies have even been used to address the subcellular localization of proteins or the extraction of kinetic characteristics from text. To extract information on biological processes and disorders, information extraction including text mining techniques have been investigated.

One of the most crucial elements in a retail setting is the shelf on which items are displayed due to the wide variety of products and consumer purchasing habits. By properly managing shelf space allocation and product presentation, retailers may both boost profits and cut costs. In order to address this issue, George and Binu (2013) have presented a method that uses the PrefixSpan algorithm to mine customer purchase patterns and arranges the goods on shelves in accordance with the patterns.

A sequence database (SDB) may be mined for common sequences using sequential pattern mining. It can extract important information from SDBs by keeping the order of the items in a sequence. For instance, user X purchased a DVD player a week after purchasing a TV. User Y has now reached web page W2 after first navigating web page W1. Due to this, sequential pattern mining is crucial in a variety of real-world application fields, such as market basket analysis, online use mining, biological gene data analysis for illness diagnosis and medicine manufacture, telecommunications data analysis, the stock market, including weather trend prediction.

The current sequential pattern mining algorithms only take into account the binary frequency values of items in sequences plus the equal importance/significance values of different items, despite the fact that sequential pattern mining is a crucial component of data mining applications. Additionally, they employ support measures to determine the frequency of a sequence. The quantity of transaction sequences (TSs) including a sequence in the SDB determines its support/frequency. The goal of sequential pattern mining is to locate all sequences in the SDB that meet a minimal support criterion chosen by the user. However, many situations in real life cannot be accurately modelled by this assumption. A user may purchase several copies of the same item in a retail market, for instance, when each item has a distinct pricing and profit value.

It is for this reason that a high-useful sequential pattern mining framework for SDBs is being developed.Due to the fact that candidates are rejected before they are created and assessed, PBCG saves more time and space than PACG. Additionally, the calculation of the utilities accounts Effective data structures should be employed to decrease this complexity because they account for the majority of the computational complexity involved in pattern extraction. We suggest a broad architecture for high utility sequential pattern mining in this paper. This

ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2304

framework presents a strict upper constraint, relying upon the HuspExt algorithm plus Cumulated Rest of Match (CRoM), which is utilised to exclude potential patterns before creation, which makes use of effective data structures when calculating utility. To assess the strategy's efficacy, we run a number of tests on both simulated and actual datasets. The outcomes demonstrate that, at low utility thresholds, the suggested approach efficiently recovers high utility sequential patterns from huge datasets.

2. Literature Survey

In order to extract more useful information from sequence databases, this research suggests a unique framework for mining high utility sequential patterns. UtilityLevel and UtilitySpan, two novel scalable and effective algorithms for mining highly useful sequential patterns, are introduced. Numerous real-world application domains, including market basket analysis, web usage mining, biomedical gene analysis, telecommunications data analysis, stock market, and weather trend prediction, depend on sequential pattern mining. The framework for mining high-utility sequential patterns proposed in this research takes into account both a sequence's internal and external utilities and provides a new metric called sequence utility (SeqUtility) to determine a sequence's utility value. The current HUP mining paradigm is made for databases that are not sequential and do not keep the order in which a pattern's elements appear. To get over the drawbacks of the support measure, which determines the real profit value of a pattern in a transaction database, it employs a measurement called "utility". Very significant and beneficial patterns may be found with this measure, which may not be attainable with the support measure. Other application domains, such as stock tickers, network traffic measures, webserver logs, data feeds from sensor networks, and telecoms call records, can have comparable solutions in addition to the actual retail market [1].

The Intrusion Detection System (IDS), which can identify many forms of assaults, is essential to network security. SVM (Support Vector Machine) is used in this suggested system to categorise data and assess the system's efficacy. 4,900,000 single connection instances with 42 characteristics, including assaults or normal, are included in the NSL-KDD Cup'99 training dataset. In order to make the nominal features sufficient input for classification using SVM, the data set transformation method is used. This study suggests a technique for intrusion detection using SVM that, when combined with a Gaussian RBF kernel, can speed up the classification model building process and improve intrusion detection precision. The empirical findings demonstrate that the disadvantage of SVM, namely the lengthy time necessary to develop a model, may be addressed provided data sets are adequately handled and the appropriate SVM kernel is selected. When the classification was modified to 10-fold cross validation and evaluated again using the provided test set and the same RBF SVM kernel function, the attack detection accuracy rose to 98.5749% [2].

The detection of common subgraphs within graph data sets is the focus of the major study field of graph mining, which falls under the umbrella of data mining. The core of graph mining is known as frequent subgraph mining (FSM), which aims to extract all frequent subgraphs in a given data collection with greater than expected occurrence counts. During the recent years, a large number of significant FSM algorithms have been suggested, indicating the field is

becoming more developed. FSM's wide range of applications demonstrate how important it is. The three application fields of chemistry, web, and biology all employ FSM algorithms. The creation of candidates, navigating the search space, and occurrence counting are the main topics of this paper's "cutting Edge" study of FSM. Candidate generation and support computation are the two components that need the greatest computing, with candidate generation being the most expensive. According on candidate generation, search methodology, and approach to frequency counting, FSM algorithms are classed. Graphs can be used with FTM algorithms, but trees and graphs can both be used with FGM algorithms. The resultant collection of frequent subgraphs has to be reduced, and the mining work needs to be improved, among other things. Combinatorial complexity and the usefulness of frequent subgraphs are trade-offs [3].

Finding new and practical information from a graph representation of data is known as mining graph data. A crucial component of graph mining is frequent pattern mining (FPM), which aids in finding patterns that theoretically reflect relationships between discrete entities. It is extremely difficult and computationally demanding to develop algorithms that find every frequently occurring subgraph in a big graph dataset since graph and sub graph isomorphism are crucial to the calculations. In order to identify common patterns, this research compares graph mining methods and methodologies. The Apriori-Based Approach, which employs a generate-and-test methodology to create candidate item sets and determine their frequency, is a crucial component in graph mining. The Pattern-Growth Approach is a method for often discovering candidate-less item sets. The method used to determine the support is different in the case of graph mining, which is the primary distinction. Graph isomorphism, an NP-complete issue, is the key obstacle in the creation of the algorithm to achieve high performance to improve the graph mining process. The graph mining algorithm has room for development, for example in terms of speed or sensitivity [4].

This study formalises the concept of aggregate movement behaviour by introducing a unique type of spatio-temporal pattern termed a trajectory pattern. It depicts a collection of independent trajectories that all have in common that they pass through the same order of locations and take around the same amount of time. The areas of interest in the given space and the average travel time for moving things between regions are the two key concepts. A trajectory pattern is a series of regularly visited geographical places in the order that the sequence specifies. Two approaches are put forth to extract trajectory patterns from the source trajectory data: pre-conceived areas of interest, which employ arbitrary prior information to identify a collection of locations of interest. T-patterns, which are utilised to choose the zones in origin-destination matrices, are the fundamental building blocks of spatiotemporal data mining. They are dynamically entwined with the extraction of temporal information from sequences, enabling more accurate trajectory patterns. An empirical study produced encouraging findings [5].

3. Proposed System

This method provides an effective foundation for mining highly useful sequential patterns. It introduces an upper bound based on the Cumulated Rest of Match (CRoM). Prior to generation, it is used to eliminate candidate patterns. A HuspExt algorithm is suggested by the system.

During utility calculations, it makes use of effective data structures. Under low utility thresholds, the suggested solution effectively extracts sequential patterns with high utility from large datasets.

It first begins the mining for biomedical data. Then Obtains information about female cancer institute patients from a source, as well as the patient's mamography test results. A root tree, that is. The report is then analysed, and pruning techniques are used. Get the patient's positive test report data using the extraction algorithm? Using mlo and the image attribute, locate the area where the tissue forms. From x-location, y-location, x_nip, and y_nip, the precise location was discovered. From scanning images, determine the presence of cancerous tissue. The final report on Husp extraction utility mining details where and when the cancer first appeared. The following benefits of the suggested strategy are listed:

- It has excellent performance, is practical given the test setting, and the data's type.
- It permits the solution to reject potential, fruitful patterns even before they emerge.
- With the use of low utility threshold values, it can extract a full set of patterns from large-scale data with less time and memory usage.

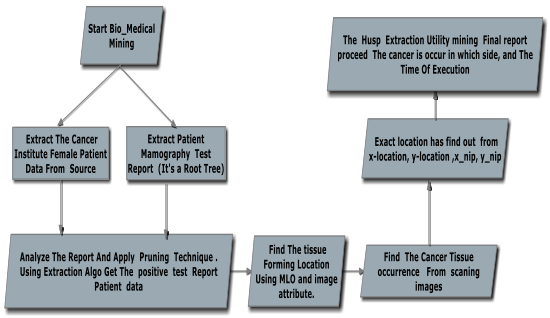

Fig 1: System Architecture

The next section provides an explanation of the many phases that are involved in putting the suggested technique into practice:

1. Data Retrieval

Obtain the data from the data source for the cancer Institute Female Patients Breast Cancer initial test results. The features include patient ID, test results, and scan photos. Tissue Occurrence Area and X and Y Pixel Detail Location.

2. Get Patient Report and Result classification

In this module, we categorise the patient results. To start, we used an extraction technique to extract the patient's ID and data on the tissue results for the female patient. These data form the root of a pattern tree. It includes test positive and negative results, with "10000" denoting a positive report that indicates patients should be affected by breast cancer and "000000" denoting a negative report that indicates patients should not be impacted by cancer. Using the pruning strategy, we remove the reports from normal patients and take the reports from cancer patients to get more data.

3. Ordering Sequence

We examine the patients' scans and MLO reports in this module. Using CRoM(Cumulated Rest of Match) based Upper bound approach, we extract the cancer-affected patients' data and additional checkup reports, but not for all patients. Pick a few patients' scan data at random. The subsequent ordering sequence has taken into account choosing the MLO, Image_id, B_Side Attribute, and choosing the images using the (RMUB) approach. Furthermore from the selective patient (id, result), we randomly choose the images.

4. Tissue Location Analysis Using RMUB

Applying pruning techniques, removing the mlo attribute, extracting data using huspextraction, moving forward with the sequence ordering in the x_loc and y_loc attributes, and determining the precise tissue spreading location are all covered in this module. As a result, both the left or right side of the body contains cancer, and both the x_loc and y_loc locations include x-axis tissue.

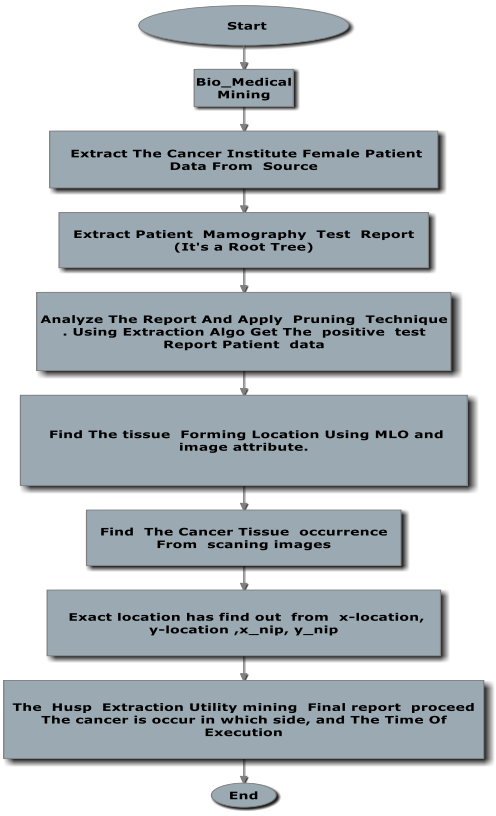

Fig 2: Flow Diagram

5. Mining Report generation

We use this biomedical domain to apply pattern extraction mining, analyse women with breast cancer, locate the no. of cases, the average value, and determine the processing time for this pattern mining.

4. Results

Recent research in the fields of computational linguistics, bioinformatics, medical informatics, and biomedical text mining has emerged. Finding statistically significant patterns across data samples when the values are provided in a sequence is the focus of high utility Sequential Pattern mining, a subfield of data mining. To execute incremental and interactive HUP mining effectively, this study suggests three innovative tree architectures. When a database is updated, it adds a Cumulated Rest of Match to cut down on pointless calculations. HuspExt is an algorithm that has been suggested to extract sequential patterns with high value from big datasets with low utility thresholds.

Prior to generation, candidate patterns are eliminated using a CRoM-based pruning technique, which requires less computational time and storage. In order to efficiently calculate CRoM values, it also makes use of efficient data structures to store sequences and patterns. The pictures below illustrate how HuspExt outperforms cutting-edge techniques in terms of speed and performance, even at very low threshold settings.

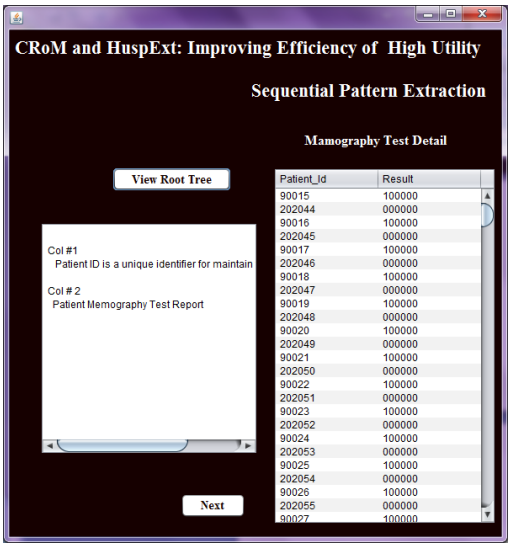

Fig 3: Mamography Test Details

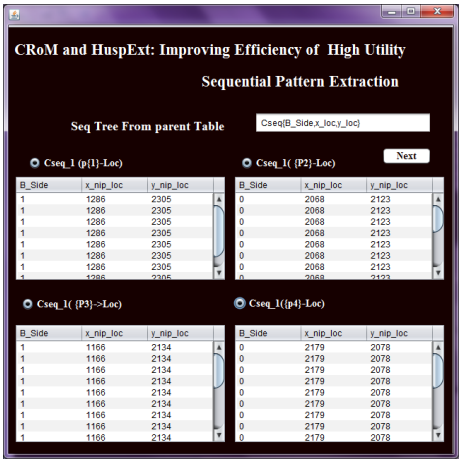

Fig 4: Seq. Tree From Parent Table

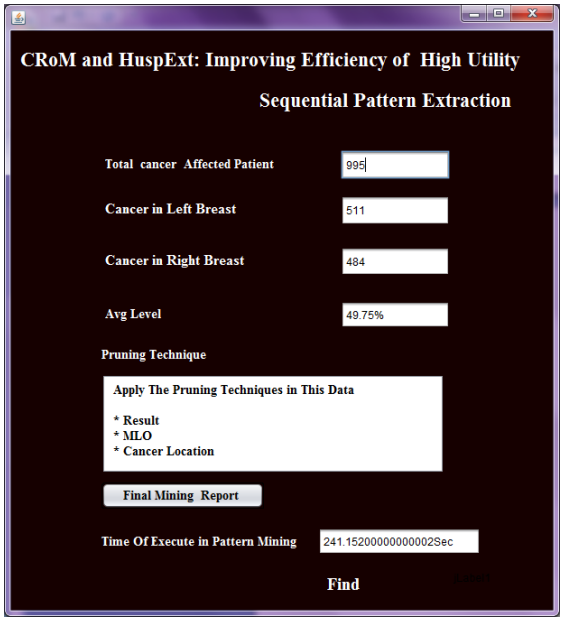

Fig 5: Performance Analysis

5. Conclusion

The system suggested a general framework for this project. Compared to the most recent twu-based upper bound, it is a tighter upper bound on the usefulness of the candidate patterns. Even from vast amounts of data, the Huspext method can extract all of the patterns in less time and memory with the use of low utility threshold settings. Almost often, HuspExt beats cutting-edge algorithms on actual and artificial data sets. In comparison to the prior methods, HuspExt adopts a CRoM-based pruning strategy that places a tighter upper constraint on the utility of the candidate patterns. As a result, the pattern tree's overstated utility of the patterns is reduced. This in turn lowers the amount of time and space needed for computing. HuspExt makes use of effective data structures for storing patterns and sequences in order to compute CRoM values quickly. The method may investigate just the desired itemsets of the pattern-containing sequences thanks to the CSeq structure. HuspExt is quicker than state-of-the-art approaches overall because it eliminates more candidates, and it performs comparably even at extremely low threshold levels.

Reference

1. “A Novel Approach for Mining High-Utility Sequential Patterns in Sequence Databases”, Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, and Byeong-Soo Jeong, 2010.
2. “A Real- Time Intrusion Detection System using Data Mining Technique”, Fang-Yie Leu Kai -Wei Hu, 2009.
3. “A Survey of Frequent Subgraph Mining Algorithms”, Chuntao Jiang, Frans Coenen and Michele Zito, 2013.
4. “cousinPair is that the are not generally applicable”, Harsh J. Patel, Rakesh Prajapati, Prof. Mahesh Panchal, Dr. Monal J. Patel, 2013.
5. “Trajectory Pattern Mining”, Fosca Giannot ti Mirco Nanni1 Dino Pedreschi2 Fabio Pinelli, 2007.

ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2304

6. “Sequential Pattern Mining from Trajectory Data”, Elio Masciari, Gao Shi, 2013.
7. “Mining Frequent Trajectory Patterns for Activity Monitoring Using Radio Frequency Tag Arrays”, Yunhao Liu, 2011.
8. “Utility, Importance and Frequency Dependent Algorithm for Web Path Traversal Using Prefix Tree Data Structure”, L.K.Joshila Grace And dr. V. Maheswari, 2014.
9. “Tracing Efficient Path Using Web Path Tracing”, L.K. Joshila Grace, 2010.
10. “Efficient tree Structures for High Utility pattern Mining in Incremental Databases”, Chowdry Farhan ahmed, Syed Tanbeer, 2010.