Technical Summary: Graph Mining Healthcare Approach: Analysis and Recommendation

Fatima Dossa and Maira Khan
Group 23
L3

April 8, 2025

1 Problem and Contribution

Problem Addressed

The paper titled "Graph Mining Healthcare Approach: Analysis and Recommendation" addresses the challenge of enhancing predictive accuracy in healthcare applications. Traditional clustering methods often fall short due to issues like data preparation requirements and sensitivity to outliers. The authors propose leveraging graph mining techniques to overcome these limitations and improve prediction outcomes.

Main Contribution

The primary contribution of this study is the application of Frequent Subgraph Mining (FSM) and Discriminative Subgraph Mining (DSM) techniques to analyze patient data. By modeling patient symptoms and medical conditions as directed graphs, the study identifies recurring subgraph patterns associated with patient recovery outcomes. The findings indicate that DSM, in particular, offers higher predictive accuracy compared to FSM in forecasting patient recovery.

2 Algorithmic Description

The core algorithms in the paper are **Frequent Subgraph Mining (FSM)** and **Discriminative Subgraph Mining (DSM)**. Both techniques operate on patient event graphs constructed from the MIMIC dataset, where each node represents a medical event (like a lab test or procedure), and edges reflect the temporal sequence or co-occurrence of those events for a given patient.

Input and Graph Representation

The input is a set of graphs $G = \{G_1, G_2, ..., G_n\}$, where each graph corresponds to a patient record. Nodes are labeled by event types, and edges represent the ordering or co-occurrence of events.

Each graph is labeled as belonging to a particular outcome class (e.g., positive recovery or negative outcome), turning the mining task into a supervised learning problem.

Frequent Subgraph Mining (FSM)

FSM identifies subgraphs that occur frequently across patient graphs, regardless of class label.

- **Step 1:** Convert each patient record into a graph with events as nodes and edges as timestamps or logical flows.
- **Step 2:** Use the `gSpan` algorithm to explore subgraphs through depth-first extensions.
- **Step 3:** Record subgraphs whose frequency exceeds a predefined support threshold.

gSpan, in particular, avoids candidate generation by lexicographically ordering graphs and uses DFS codes to efficiently explore the search space.

Discriminative Subgraph Mining (DSM)

DSM builds upon FSM by identifying subgraphs that best distinguish between outcome classes.

- **Step 1:** Use FSM to generate a candidate list of subgraphs.
- **Step 2:** For each subgraph, compute a *discriminative score* based on its frequency in different classes.
- **Step 3:** Select top-k subgraphs with the highest discrimination, e.g., using g-index or mutual information.

These discriminative subgraphs are then used as features in a classification model, such as a decision tree or support vector machine.

Example

Imagine two patient graphs:

G_1 (Recovered): [A] $\rightarrow$ [B] $\rightarrow$ [C]
G_2 (Not Recovered): [A] $\rightarrow$ [D] $\rightarrow$ [C]

FSM might identify subgraph [A] $\rightarrow$ [C] as frequent. DSM will note that [B] is more common in recovered cases and [D] in unrecovered ones, so [A] $\rightarrow$ [B] is a better discriminative feature.

Intuition and Technical Complexity

- FSM helps detect patterns common across all patient records, which may not be useful for classification alone.
- DSM filters these patterns to find only those predictive of recovery or decline.
- The challenge lies in graph isomorphism checking and subgraph enumeration, which are computationally expensive (NP-complete problems).
- The use of tools like `LEAP` helps to abstract graph construction and label matching from raw medical data, making the system scalable.

Overall, the algorithms are designed to balance coverage (via FSM) with class discrimination (via DSM) to build a powerful healthcare recommendation pipeline.

3 Comparison with Existing Approaches

Traditional clustering techniques in healthcare analytics often require extensive data preprocessing and are sensitive to outliers, which can compromise predictive accuracy. In contrast, the graph mining approach utilized in this study offers a more nuanced analysis by capturing complex relationships within patient data. The application of Discriminative Subgraph Mining (DSM), in particular, enhances predictive performance by focusing on subgraphs that are most indicative of patient outcomes, thereby providing a novel and effective alternative to conventional methods.

Additionally, there are several other existing approaches to medical recommendation systems as well. Collaborative filtering methods—such as those used in systems like CARE—struggle with data sparsity, scalability, and cold-start issues. Decision trees and random forests, while interpretable, often require a large number of trees for reliable predictions and suffer from high latency during inference. Markov models are limited in their ability to capture complex dependencies between actions, while neural networks, despite their predictive power, lack interpretability and require large, well-labeled datasets -making them less suitable for critical healthcare applications.

By avoiding the need for extensive parameter tuning or a priori knowledge, and by directly modeling action sequences through subgraph structures, our algorithm not only addresses these limitations but also offers a more scalable, interpretable, and clinically meaningful solution.

4 Data Structures and Techniques

The study employs several key data structures and techniques:

- **Graphs:** Used to model patient medical data, capturing the intricate relationships between symptoms and conditions.
- **Frequent Subgraph Mining (FSM):** Identifies common subgraph patterns across patient graphs.

- **Discriminative Subgraph Mining (DSM):** Extracts subgraphs that effectively differentiate between varying patient outcomes.

5 Implementation Outlook

Implementing the proposed approach presents several technical challenges:

- **Data Transformation:** Converting unstructured patient records into structured graph representations requires meticulous data preprocessing.
- **Computational Complexity:** Subgraph mining, especially DSM, can be computationally intensive, necessitating efficient algorithms and possibly high-performance computing resources.
- **Scalability:** Handling large-scale patient datasets demands scalable solutions to maintain performance and accuracy.
- **Data Privacy:** Ensuring patient confidentiality and compliance with healthcare data regulations is paramount during data handling and processing.

6 GitHub Repository Update

To align with Checkpoint 2 requirements, the following updates have been made to our project repository:

- **Checkpoint 2 Folder:** Added a new directory containing the LaTeX source files and the compiled PDF of this report.
- **README.md:** Updated documentation to reflect our enhanced understanding of the project, including a summary of the current report and insights gained.
- **Resources:** Referred to the MIMIC dataset and the primary paper for implementing and analyzing FSM and DSM techniques.

The repository remains organized and free from temporary or backup files, adhering to best practices for version control and collaborative development.