위상수학적 질병 지도 및 슈퍼 리스폰더 발굴 시스템 요구사항 명세서

(T-DMAP: Topological Disease Mapping & Analysis Platform)

Yonsei Univ. Dept. of Software
Project Constructor: 황도현 (Lucius)

2026년 1월 27일

Abstract

본 문서는 국민건강보험공단 표본 코호트 데이터(약 244만 건 트랜잭션)를 대상으로 **위상수학적 데이터 분석(TDA, Topological Data Analysis)** 기법을 적용하여 환자의 진료 경로를 시각화하고 분석하는 시스템의 요구사항을 기술한다. 특히 **NVIDIA RTX 4070 Ti Super 기반의 GPU 가속**을 통해 대용량 의료 데이터를 고속 연산하며, 기존 통계가 포착하지 못한 **만성 질환의 악순환 고리(Loop)**와 이를 탈출하여 **완치에 이른 특이 환자군(Super-Responder)**을 식별함으로써 데이터 기반의 최적 치료 경로(Golden Path)를 제시하는 것을 목표로 한다.

1 개요 (Project Overview)

1.1 목표 (Goal)

- **High-Performance TDA:** GPU 가속을 활용해 240만 행 이상의 대용량 의료 데이터에 대한 실시간 위상수학적 구조 분석 수행.
- **Trajectory Analysis:** 환자의 시간적 진료 흐름을 위상 공간상의 궤적(Trajectory)으로 매핑하여 '악화'와 '호전'의 갈림길 규명.
- **Solution Discovery:** 일반적인 악화 패턴에서 벗어나 호전된 '이상치(Outlier)' 그룹의 처방/행동 특성을 역추적하여 해결책 도출.

1.2 범위 (Scope)

- **대상 데이터:** 명세서(ST200), 상병내역(ST400), 진료내역(ST530) 통합 데이터 (Total $\approx$ 2.4M rows).
- **핵심 기술:** KeplerMapper, Giotto-TDA, **RAPIDS (cuML, cuDF)**, Polars.
- **하드웨어:** NVIDIA GeForce RTX 4070 Ti Super (16GB VRAM) Workstation.

2 팀 구성 및 역할 분담 (Team Roles & Responsibilities)

본 프로젝트는 기존 APEDS 프로젝트 팀원 구성을 기반으로 하되, 대용량 처리를 위한 엔지니어링 역량 집중을 위해 역할을 재조정한다.

이름	담당 영역	상세 업무 (R&R)
김택훈	Data Analysis (EDA & Validation)	**[기초 통계 및 비교 검증]**
		- 기존 통계 방식(PCA, K-Means)과의 성능 및 인사이트 비교.
		- TDA로 도출된 특이 그룹(Cluster)에 대한 통계적 유의성 검정.
		- Streamlit 기반 대시보드 UI 기획 및 시각화.
황도현 (Lucius)	Project Constructor (DE & TDA Modeling)	**[데이터 엔지니어링 및 GPU TDA 총괄 & 의료 도메인 리서치 및 데이터 검증]**
		- **Polars** 활용 대용량(2.4M Rows) ETL 파이프라인 구축.
		- **RAPIDS(cuML)** 기반 GPU 가속 UMAP/DBSCAN 구현.
		- Mapper 그래프 생성 및 Loop(H_1)/Flare 구조 탐지 로직 개발.
		- 'Super-Responder' 식별 알고리즘 및 Feature Importance 분석.
		- ICD-10 상병코드 및 ATC/KD 약물 코드 매핑 기준 수립.
		- TDA 분석 결과의 임상적 정합성 검토 및 보고서 작성 지원.
		- 프로젝트 최종 QA 및 시나리오 테스트 수행.

Table 1: 프로젝트 팀 역할 분담표 (수정됨)

3 데이터 명세 (Data Specifications)

3.1 원천 데이터 (Raw Data)

- **규모:** 총 2,441,201 Rows (Big Data Scale).
- **구조:** 명세서(ST200) ↔ 상병(ST400) ↔ 진료(ST530) 간의 1:N:M 관계형 구조.
- **주요 변수:**
  - ST200: SPEC_ID_SNO(키), AGG(연령군), SEX_TP_CD(성별), MSICK_CD(주상병).
  - ST530: GNL_NM_CD(약물코드), TOT_USE_QTY(사용량), AMT(금액), VST_DDCNT(내원일수).

3.2 학습 데이터 (Feature Vectors)

TDA 입력 차원을 구성하기 위해 트랜잭션 데이터를 환자(명세서) 단위로 압축한다.

- **단위:** Unique SPEC_ID_SNO (예상 5∼10만 건).
- **Feature Engineering:**

- **Medical History:** 주요 약물 계열별 처방 횟수 (Multi-hot Vector).
- **Cost/Intensity:** 총 진료비, 투약 기간, 내원 밀도.
- **Comorbidity:** 동반 상병 개수 및 중증도 가중치.

4 이론적 배경: 위상수학적 분석 방법론 (Theoretical Framework)

본 시스템은 고차원 의료 데이터의 구조적 특징을 보존하기 위해 **Mapper Algorithm**과 **Persistent Homology** 이론을 기반으로 설계된다.

4.1 Mapper 알고리즘의 정의 (Mapper Definition)

데이터 공간 $X \subset R^D$ (환자 벡터 집합)에 대하여, 데이터의 위상적 구조(Topological Shape)를 그래프 $G(V, E)$로 근사하는 과정은 다음과 같이 정의된다.

1. **Filtering (Lens Function):** 고차원 데이터 X를 저차원 공간 Z (Parameter Space)로 사영하는 함수 $f : X \to Z$를 정의한다. 본 프로젝트에서는 $Z = R^2$ 이며, f는 UMAP 알고리즘을 사용한다.
$$f(x_i) = z_i, \quad \text{where } x_i \in X \text{ and } z_i \in Z \tag{1}$$

2. **Covering:** 저차원 공간 Z를 덮는 개집합(Open Sets)들의 유한 집합 $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$를 구성한다. 각 U_α는 일정 비율(p)만큼 중첩(Overlap)된다.
$$Z \subseteq \bigcup_{\alpha \in A} U_\alpha \tag{2}$$

3. **Clustering & Pullback:** 각 U_α의 역상(Pre-image) $f^{-1}(U_\alpha) \subset X$에 대하여, 거리 함수 d_X를 기반으로 클러스터링을 수행한다.
$$f^{-1}(U_\alpha) = \bigcup_j C_{\alpha,j} \tag{3}$$
여기서 $C_{\alpha,j}$는 $f^{-1}(U_\alpha)$ 내의 j번째 클러스터(Node)를 의미한다.

4.2 심플리셜 컴플렉스 및 신경 (Nerve Construction)

데이터의 연결성을 나타내는 1-Skeleton 그래프인 Nerve $\mathcal{N}(\mathcal{C})$는 클러스터 간의 교집합이 존재할 때 엣지(Edge)를 생성함으로써 구성된다.

$$\mathcal{N}(\mathcal{C}) = \{\{\alpha, \beta\} \mid C_{\alpha,i} \cap C_{\beta,j} \neq \emptyset\} \tag{4}$$

즉, 두 클러스터에 공통된 환자(Patient)가 존재하면 두 노드는 연결된 것으로 간주하며, 이를 통해 환자의 전이 경로(Trajectory)를 시각화한다.

4.3 위상적 루프 탐지 (Loop Detection via Homology)

환자의 만성적 악순환 패턴은 위상 공간의 1차원 구멍(1-dimensional hole)으로 나타난다. 이는 k-th Homology Group H_k의 랭크(Rank)인 베티 수(Betti Number, β_k)로 정량화된다.

- β_0 **(Connected Components):** 서로 분리된 환자 군집의 개수 (이질적 집단 식별).
- β_1 **(Cycles/Loops):** 데이터 내에 존재하는 순환 구조의 개수. 본 프로젝트의 핵심 탐지 대상인 '악순환 고리'는 $\beta_1 > 0$ 인 구조체에 해당한다.

$$\text{Target Loop} \iff \exists \gamma \in H_1(\mathcal{N}(\mathcal{C})) \text{ s.t. Persistence}(\gamma) > \tau \tag{5}$$

(여기서 τ는 노이즈를 제거하기 위한 임계값이다.)

4.4 수학적 모델의 기술적 함의 (Technical Implications)

상기 수식들은 단순한 이론적 배경이 아니라, 본 프로젝트의 핵심 기술적 요구사항과 분석 목표에 대한 강력한 근거를 제시한다.

- **GPU 가속의 필요성:** 식 (3)에서 요구되는 수만 개의 부분 클러스터($C_{\alpha,j}$) 생성과, 식 (4)의 조합론적 교집합($\cap$) 연산은 대용량 데이터에서 막대한 계산 비용을 유발한다. 이는 본 시스템이 고성능 GPU(RTX 4070 Ti Super) 기반의 병렬 처리를 채택해야 하는 직접적인 이유이다.
- **악순환 고리의 정의:** 임상적으로 모호할 수 있는 '만성 악순환'의 개념을, 식 (5)의 $\beta_1 > 0$ (1차원 위상 구멍의 존재)이라는 명확한 수학적 조건으로 치환하여 객관적인 탐지 기준을 확립한다.

5 시스템 아키텍처 (System Architecture)

단계	도구 (Tool)	주요 역할
1. Ingestion	**Polars**	- Pandas 대비 10배 이상 빠른 속도로 2.4M CSV 로드. - 메모리 효율적인 데이터 타입(Float32) 변환 및 전처리.
2. Projection	**RAPIDS (cuML)**	- **GPU 가속 UMAP/PCA**를 사용하여 고차원 벡터 투영. - RTX 4070 Ti Super 활용 시 수초 내 연산 완료.
3. Mapping	**KeplerMapper**	- Lens 공간을 Hypercube로 분할하고 **GPU DBSCAN** 수행. - 노드(Cluster)와 엣지(Link)로 구성된 Simplicial Complex 생성.
4. Analysis	**Scikit-learn**	- 식별된 특이 그룹(Node)에 대한 특성 분석 및 중요 변수 추출.

6 핵심 기능 요구사항 (Functional Requirements)

6.1 위상적 루프(H_1) 및 악순환 고리 탐지

- **기능:** Mapper 그래프 상에서 닫힌 경로(Cycle)를 형성하는 환자 군집을 자동 식별한다.
- **임상적 의미:** 표준 치료에 반응하지 않고 '입원 → 퇴원 → 재발'을 반복하는 '의료 쇼핑' 또는 '만성 난치' 환자군 정의.

6.2 분기점(Bifurcation) 및 플레어(Flare) 분석

- **기능:** 데이터 구조가 'Y'자 형태로 갈라지는 분기점(Critical Point)을 포착한다.
- **임상적 의미:** 경증 환자가 중증 합병증으로 진행되거나, 반대로 회복세로 돌아서는 **'골든타임'** 시점을 특정한다.

6.3 슈퍼 리스폰더(Super-Responder) 및 탈출 경로(Escape Path) 발굴

- **Trajectory Tracking:** 악순환 루프(H_1) 내에 머물다가 궤도를 이탈하여 '호전 클러스터'로 이동한 환자를 추적한다.
- **Outcome Coloring:** 유사한 중증도(Input)를 가졌으나, 진료 결과(Output)가 현저히 좋은 **'성공한 이상치(Lucky Outliers)'**를 시각적으로 필터링한다.
- **Factor Analysis:** 해당 그룹이 공통적으로 경험한 **'Game Changer' 요인(특정 약물 변경, 병원 이동 등)**을 역추적하여 솔루션으로 제시한다.

7 머신러닝 모델링 및 하드웨어 최적화

7.1 GPU Acceleration Strategy

- **라이브러리:** NVIDIA RAPIDS Suite (`cuDF`, `cuML`).
- **성능 목표:** 240만 건 데이터 전처리 및 10만 건 환자 벡터 TDA 연산을 **5분 이내** 완료 (CPU 대비 50배 가속).
- **메모리 관리:** 16GB VRAM 활용을 극대화하기 위해 Batch Processing 및 Data Type 최적화 적용.

7.2 사후 검증 (Post-Hoc Analysis)

- **유의성 검증:** TDA로 발견한 '슈퍼 리스폰더' 그룹과 '일반 환자' 그룹 간의 약물 처방 패턴 차이가 통계적으로 유의미한지($P < 0.05$) 검증한다.

8 개발 일정 (Development Schedule)

기간: 2026년 1월 26일 ~ 1월 30일 (5일 스프린트)

Day 1 (1/26 월): Environment & ETL
CUDA Toolkit 및 RAPIDS 환경 설정. Polars 기반 대용량 데이터 로드 및 Aggregation 파이프라인 구축.

Day 2 (1/27 화): Feature Engineering & Lens Search
환자 임베딩 벡터 생성. GPU UMAP 파라미터(n_neighbors, min_dist) Grid Search 수행.

Day 3 (1/28 수): TDA Construction & Loop Detection
KeplerMapper 파이프라인 가동 및 위상적 질병 지도 생성. 악순환 루프(H_1) 및 분기점 구조 식별.

Day 4 (1/29 목): Insight Mining (The Solution Finder)
'Escape Path' 분석 및 슈퍼 리스폰더 그룹 추출. 그룹별 핵심 변수 비교 분석.

Day 5 (1/30 금): Final Review & Demo

TDA 시각화 결과물(HTML) 및 분석 리포트 작성. 프로젝트 최종 발표 및 시연.