Lazy FCA Model Analysis

A Step-by-Step Approach

By

Mohammad Shazzad Hossain

Why I Select This Dataset

- Dataset Link: https://github.com/aiplanethub/Datasets/blob/master/liver_patient.csv
- Github Link:
- https://github.com/sazzadxy/lazy_Hossain-Mohammad-Shazzad-_report

- The dataset was selected because:
- - It is a well-known dataset for classification problems.
- - Contains a mix of numerical and categorical features.
- - Suitable for evaluating the performance of Lazy FCA and other models.
- - Target variable: Presence of liver disease, which is a binary classification task.

Exploratory Data Analysis (EDA)

- Key observations from EDA:
- - The dataset contains 583 rows and 11 columns, with a mix of numerical and categorical data..
- - Features include age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, liver disease etc.
- - Most columns are fully populated except for 4 missing values detected (Albumin and Globulin Ratio).
- - Target variable is binary, representing presence or absence of liver disease.

Threshold Selection

- Thresholds were selected for binarization based on:
- Defined thresholds for continuous variables based on typical medical reference ranges or domain knowledge.
- Examples:

1. Age: 60 to identify older individuals at risk.
2. Total Bilirubin: 1.2 (upper normal limit).
3. Albumin and Globulin Ratio: 1.0 (healthy balance).

Data Binarization

- The dataset was binarized to transform continuous variables into categories.
- Each feature was binarized to form new binary columns, making it easier for the Lazy FCA model to compute and classify instances.
- Binarization helped standardize features and prepared the data for binary-based comparisons in algorithms like Lazy FCA and Logistic Regression.

Binarize continuous variables

Feature	Threshold	Reason for Threshold
Age	60	Older individuals are at higher risk for liver issues.
Total_Bilirubin	1.2	Upper normal range for bilirubin in healthy individuals.
Direct_Bilirubin	0.3	Normal upper limit for direct bilirubin.
Alkaline_Phosphotase	120	Normal upper limit for alkaline phosphatase..
Aspartate_Aminotransferase	60	Normal upper limit for ALT enzyme.
Aspartate_Aminotransferase	40	Normal upper limit for AST enzyme.
Total_Proteins	6.5	Lower limit of normal total proteins..
Albumin	3.5	Lower limit of normal albumin levels.
Albumin_and_Globulin_Ratio	1.0	alanced healthy reference for the ratio..

Implement Lazy FCA

- Steps to implement Lazy FCA:
- 1. Create a binary context table.
- 2. Define formal concepts dynamically from the dataset.
- 3. Evaluate objects against the generated concepts for classification.

Binary Decision Function, Classifier, Pattern

- Implemented the following:
- - Binary Decision Function: Checks if an object satisfies a concept.
- - Classifier: Matches objects to concepts and predicts class labels.
- - Pattern Extraction: Identifies attribute combinations unique to each target class.

Lazy FCA Accuracy

- Lazy FCA Accuracy:
- - Initial accuracy 70%

Comparison with Main Dataset

- Comparison:
- - Lazy FCA accuracy on the main dataset is 65%.

Comparison with Binarized Dataset

- Comparison:
- - Lazy FCA : 70.%.
- - Machine learning models on binarized dataset:
- - Random Forest: 68%
- - Naive Bayes Classifier: 67%.
- - Logistic Regression: 72%.
- - Decision Tree Classifier: 68%
- - Support Vector Classifier: 69%
- - K-Nearest Neighbors Classifier: 71%

How to Improve Accuracy

- Strategies to improve Lazy FCA accuracy:
- - Refine binarization thresholds based on domain insights.
- - Use partial matching for concepts.
- - Combine Lazy FCA features with machine learning models (hybrid approach).
- - Dynamically adjust concept intents for better alignment.