CS 4412: Data Mining Project - M1 Proposal

Aman Bhayani

Section: 01
NET ID: abhayani
February 5, 2026

1 Dataset Description

The dataset I will be using for this project comes from the MovieLens (GroupLens) website: MovieLens Dataset.

The dataset contains data from user generated ratings and tags for movies collected by the organization. It represents how users interact with movies over time with different movies by providing ratings, descriptive tags, and genre information. The dataset contains over 32 million ratings with over 2 million tag applications for over 87,000 movies released between January 09, 1995 and October 12, 2023.

- Characteristics of the Dataset:
  - Files: ratings.csv, movies.csv, tags.csv, links.csv
  - Ratings: 32000204 rows
  - Movies and Links: 87585 entries
  - Tags: 2000072 entries
  - File size: 2395 MB

2 Discovery Questions

Here are a few questions about pattern discovery that my project will try to explore.

1. What natural group of movies exist based on user ratings, genres, and tags?

Explore patterns or similarities between movies beyond official genres?

2. Are there distinct types of users based on their rating activity, genre preferences, and tagging behavior?

Identify user behavior such as casual viewers, genre specialists, or highly active users.

3. Are there any seasonal trends or binge-like rating behavior?

Identify seasonal patterns about how engagement changes over time. Identify any high or low activity periods.

3 Planned Techniques

The project will utilize multiple data mining techniques that were discussed throughout the course. The motive outlined for the techniques below demonstrate how they relate to the questions of discovery.

1. Data Preprocessing
- Remove missing or invalid entries
- Convert timestamps to date/time features
- Normalize rating values
- Merge datasets using movieId and userId (if needed)

2. Clustering
Techniques to use: K-Means and DBSCAN.

Motive:
- Cluster movies based on ratings, tags, genres.
- Cluster movies based on user activity.

3. Association Rule
Techniques to use: Apriori and FP-Growth.

Motive:
- Identify any relationships between ratings, tags, genres.

4. Anomaly Detection
Techniques to use: Z-Score and isolation forest.

Motive:
- Identify any unusual patterns in user behavior.

4 Preliminary Timeline

- M2: Initial Implementation (March 5, 2026)
  - Load and clean datasets
  - Implement basic data mining features
- M3: Complete Implementation (April 2, 2026)
  - Implement the algorithms and techniques specified in Planned Techniques
  - Compare results
  - Generate necessary graphs, diagrams, etc.

- M4: Final Deliverable (May 3, 2026)
  - Interpret results
  - Validate patterns
  - Prepare final report, repository, and presentation