University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 10 (Sep 24)

Reminders

- Homework 3: due at 9:30am, Th, Sep 24
- Homework 1&2: graded, check grades in Canvas
- Temporary 2-week remote instruction: till W Oct 7
- No new homework this week
  - work on course project proposal

Course Project

- Hands-on experience with real-world data mining
- Pick your own project of interest
- Suggested group size: 3-4, can mix across class section
- Week 6: proposal; Week 12: checkpoint; Week 16: final

Course Project Proposal

- Introduction: what problem? why this problem?
- Related work: what has been done already?
- Proposed work: what do you plan to do?
- Evaluation: what metrics? how to claim success?
- Milestones: when to accomplish what?

How do I get started?

- What's your interest? Who are on your team?
- Data mining: Different views
  - Application: e.g., sports, health, election, weather, business, …
  - Data: e.g., games/teams/players, COVID-19, Twitter, reviews, …
  - Knowledge: frequent patterns, key factors, trends, anomalies, …
  - Method: understand, preprocess, manage, model, evaluate

Key Points

- Start early, talk to people, and keep it evolving
- Data availability
- Prioritized subtasks
- Existing tools
- Team coordination, individual contributions

Project Proposal Meetings

- Availability survey due at 9:30am, Tu Sep 29
- Meeting schedule: Th Oct 1 to Wed Oct 7
- Meeting with instructor, public to the whole class
- Presentation: 5~10 minutes per team
- Discussion & feedback: ~5 minutes

Project Announcement

- Due at 9:30am, Thursday, Oct 1
- Canvas assignment: Course Project Announcement
- One announcement per team
  - project title, team members (name, CSCI 4502 or 5502)
  - brief project description,
  - dataset(s) to use, potential tool(s) to use

Project Proposal Slides

- Due at 9:30am, Thursday, Oct 1
- Canvas assignment: Course Project Proposal Slides
- Slides to use for the proposal meetings
  - title, team, introduction, related work
  - proposed work (data, subtasks), evaluation, milestones

Project Proposal Report

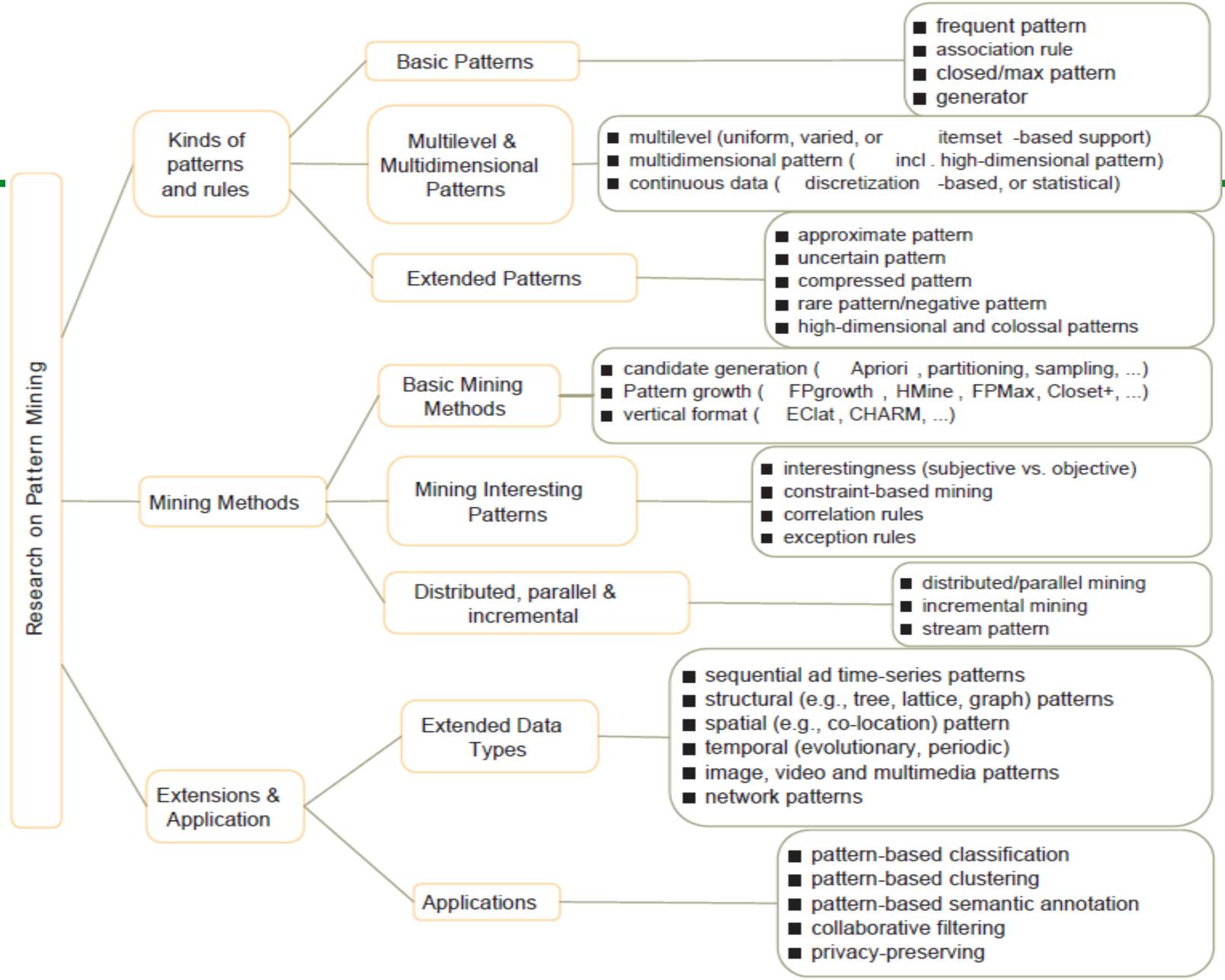
- Due at 9:30am, Th Oct 8
- Canvas assignment: Course Project Proposal Report
- ACM Master Article Template: Word, LaTex, Overleaf
- Page length: ~3 pages
  - title, team, introduction, related work
  - proposed work (data, subtasks), evaluation, milestones

Review

- Chapter 6: Mining Frequent Patterns
  - basic concepts, Apriori algorithm, correlation: lift
  - improve the efficiency of Apriori
  - FP-growth: grow patterns w/o generating candidates
  - Correlation metrics: null transaction, imbalance ratio

University of Colorado
Boulder

Chapter 7:
Advanced Pattern Mining


Research on Pattern Mining
Kinds of patterns and rules
Basic Patterns
frequent pattern
association rule
closed/max pattern
generator
Multilevel & Multidimensional Patterns
multilevel (uniform, varied, or itemset -based support)
multidimensional pattern (incl . high-dimensional pattern)
continuous data (discretization -based, or statistical)
Extended Patterns
approximate pattern
uncertain pattern
compressed pattern
rare pattern/negative pattern
high-dimensional and colossal patterns
Mining Methods
Basic Mining Methods
candidate generation (Apriori , partitioning, sampling, ...)
Pattern growth (FPgrowth , HMine , FPMax, Closet+, ...)
vertical format (EClat , CHARM, ...)
Mining Interesting Patterns
interestingness (subjective vs. objective)
constraint-based mining
correlation rules
exception rules
Distributed, parallel & incremental
distributed/parallel mining
incremental mining
stream pattern
Extensions & Application
Extended Data Types
sequential ad time-series patterns
structural (e.g., tree, lattice, graph) patterns
spatial (e.g., co-location) pattern
temporal (evolutionary, periodic)
image, video and multimedia patterns
network patterns
Applications
pattern-based classification
pattern-based clustering
pattern-based semantic annotation
collaborative filtering
privacy-preserving

Road Map (1)

- Kinds of patterns
  - set, sequential, structural
- Completeness
  - all, closed, maximal, constrained, approximate, near-match, top-k
- Levels of abstraction
  - e.g., computer ⇒ printer; laptop ⇒ HP_printer

Road Map (2)

- Number of data dimensions
  - computer ⇒ printer; (age:30-39, income:42K-48K) ⇒ HDTV
- Types of value
  - Boolean: presence or absence; quantitative: e.g., age, income
- Types of rules
  - association, correlation, gradient

Various Association Rules

- Single-level, single-dimensional, Boolean value
- Multi-level association rules
  - support: uniform, reduced, group-based
  - redundancy filtering: milk ⇒ wheat bread [8%, 70%]; 2% milk ⇒ wheat bread [2%, 72%]
- Multi-dimensional association rules
- Quantitative association rules

Multi-dimensional Association

- Single-dimensional (intra-dimensional) rules:
  - buys(X, "milk") ⇒ buys(X, "bread")
- Multi-dimensional rules: ≥2 predicates
  - inter-dimensional (no repeated predicates)
    - age(X, "19-25") ∧ occupation(X, "student") ⇒ buys (X, "coke")
  - hybrid-dimensional (repeated predicates)
    - age(X, "19-25") ∧ buys(X, "popcorn") ⇒ buys(X, "coke")

Categorical vs. Quantitative

- Categorical attributes
  - nominal, finite number of possible values, no ordering among values
  - e.g., occupation, brand, color
- Quantitative attributes
  - numeric, implicit ordering among values
  - e.g., age, income, price

Mining Quantitative Association

- How numerical attributes (e.g., age, salary) are treated
  - static discretization: predefined concepts
  - dynamic discretization: data distribution
  - clustering: distance-based association
  - deviation: from normal data
    - e.g., sex = female ⇒ wage: mean=$7/hr (overall mean = $9/hr)

Constraint-Based Mining

- Automatically find all patterns in a data set
  - Unrealistic! Too many patterns, not focused
- Data mining should be an interactive process
  - user directs what to be mined
- Constraint-based mining
  - user flexibility: provides constraints on what to be mined
  - system optimization: more efficient mining

Constraints in Data Mining

- Knowledge type constraint
- Data constraint
- Dimension/level constraint
- Interestingness constraint
- Rule (or pattern) constraint
  - metarules (rule templates)
  - #attributes, attribute values, etc.

Metarule-Guided Mining

- $P_1(X,Y) \wedge P_2(X,W) \Rightarrow$ buys(X, "office sw")
- age(X, "30-39") $\wedge$ income(X, "41K-60K") $\Rightarrow$ buys(X, "office sw")
- $P_1 \wedge P_2 \wedge \ldots \wedge P_a \Rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_b$
  - n = a + b, find all n-predicate sets L_n
  - compute the support of all a-predicate subsets of L_n
  - compute the confidence of rules

Anti-Monotonicity

- Anti-monotonicity
  - if itemset S violates the constraint, so does any of its superset
- Example
  - sum(S.price) $\leq$ 100: yes
  - sum(S.price) $\geq$ 100: no
  - range(S.profit) $\leq$ 15: yes