University of Colorado Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 9 (Sep 24)

Announcements (1)

- Homework 3
  - due at 9:30am, Thursday, Sep 26
- Homework 1 & 2
  - will finish grading soon
  - check scores at moodle
  - contact GSS first with grading questions
- No new homework this Thursday

Announcements (2)

- Office hours (regular)
  - Tu 11-12pm, Fr 1-2pm
  - Tu 4-5pm, W 11-12pm, Th 5-6pm
- Midterm exam
  - Week 10: Thursday Oct 31
  - midterm review, sample exam questions
  - in-class exam, closed book
  - distance students: in-class, on campus different time, (online) proctoring

Review

- Chapter 6: Mining Frequent Patterns
  - improve the efficiency of Apriori
    - #scans, #candidates, support counting
  - FP-growth
    - grow patterns w/o generating candidates
    - if c is frequent in DB|ab, then abc is frequent
  - correlation rules

Correlation Rules

- Correlation rule
  - A ⇒ B [support, confidence, correlation]
- Measure of dependent/correlated events

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$$

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

Other Correlation Measures

$$all\_conf(A,B) = \frac{sup(A \cup B)}{max\{sup(A), sup(B)\}} = min\{P(A|B), P(B|A)\}$$

$$max\_conf(A, B) = max\{P(A \mid B), P(B \mid A)\}$$

$$Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$$

$$cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{sup(A \cup B)}{\sqrt{sup(A) \times sup(B)}}$$

$$= \sqrt{P(A|B) \times P(B|A)}.$$

Comparison (1)

Data Set	mc	$\overline{m}c$	$m\overline{c}$	$\overline{m}\overline{c}$	χ^2	lift	all_conf.	max_conf.	Kulc.	cosine
D_1	10,000	1000	1000	100,000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10,000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100,000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100,000	24740	25.75	0.5	0.5	0.5	0.5
D_5	1000	100	10,000	100,000	8173	9.18	0.09	0.91	0.5	0.29
D_6	1000	10	100,000	100,000	965	1.97	0.01	0.99	0.5	0.10

- Null-transaction: e.g., ¬m¬c
- Null-variant: lift and X^2
- Null-invariant: all_conf, max_conf, Kulc, cosine

Comparison (2)

Data Set	mc	$\overline{m}c$	$m\overline{c}$	$\overline{m}\overline{c}$	χ^2	lift	all_conf.	max_conf.	Kulc.	cosine
D_1	10,000	1000	1000	100,000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10,000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100,000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100,000	24740	25.75	0.5	0.5	0.5	0.5
D_5	1000	100	10,000	100,000	8173	9.18	0.09	0.91	0.5	0.29
D_6	1000	10	100,000	100,000	965	1.97	0.01	0.99	0.5	0.10

- Imbalance ratio

$$IR(A,B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

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
  - 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

- Techniques categorized by 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
    - 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 an itemset S violates the constraint
  - so does any of its superset
- Example
  - sum(S.price) ≤ 100: yes
  - sum(S.price) ≥ 100: no
  - range(S.profit) ≤ 15: yes

Monotonicity

- Monotonicity
  - if an itemset S satisfies the constraint
  - so does any of its superset
- Example
  - sum(S.price) ≥ 100: yes
  - min(S.price) ≤ 100: yes
  - range(S.profit) ≥ 15: yes

Succinctness

- Succinctness
  - enumerate all and only those sets that are guaranteed to satisfy the constraint
- Example
  - min(S.price) ≤ v: yes
  - sum(S.price) ≥ v: no
- Pre-counting prunable
  - no need for support counting

Convertible Constraints

- Convert tough constraints into anti-monotonic or monotonic by properly ordering items
- Example
  - avg(S.profit) ≥25
  - ordering items in descending order
    - <a, f, g, d, b, h, c, e>
  - if afb violates C, so does afb*
  - it becomes anti-monotonic

Strongly Convertible

- avg(S.profit) ≥ 25 is convertible anti-monotonic w.r.t. descending order
- avg(S.profit) ≥ 25 is convertible monotonic w.r.t. ascending order
- avg(S.profit) ≥ 25 is strongly convertible

Classification of Constraints