Machine Learning

Association Rules – Examples

Claudio Sartori

DISI
Department of Computer Science and Engineering – University of Bologna, Italy
claudio.sartori@unibo.it

Overview

- Healthcare: clinical decision support
- Cybersecurity: intrusion detection
- Bioinformatics: gene expression patterns
- Public health: outbreak and risk factor analysis
- Manufacturing: fault diagnosis and maintenance
- Education: learning analytics
- Smart cities: traffic and incident analysis

Healthcare: Clinical Decision Support

Context

- Hospitals and research institutes use association rules to find co-occurring diagnoses, treatments, symptoms, and lab results, such as `hypertension` ⇒ `diabetes`
- Supports improved diagnosis, risk prediction, and treatment protocols.

Data structures

- Transactions: individual patient visits.
- Items: diagnoses, medications, lab abnormalities, symptoms, procedures.

Healthcare - Data & Preprocessing

Data acquisition

- Electronic Health Records (EHRs).
- ICD diagnosis codes.
- Medication lists.
- Lab results.

Preprocessing tasks

- Remove personally identifiable information.
- Standardize ICD codes (for example, ICD-10).
- Discretize continuous lab values into categorical bins (for example, Glucose_High).
- Filter noisy or rare events.

Healthcare - Postprocessing & Outcomes

Postprocessing

- Remove medically implausible or coincidental patterns.
- Rank rules by confidence, lift, and leverage.
- Validate rules with clinician review.

Actionable outcomes

- Identification of patient groups at elevated risk.
- Improved triage and screening guidelines.
- Medication interaction warnings.
- Data-driven refinement of clinical pathways.

Healthcare: Typical Questions

- Which diagnoses frequently appear together?
  `{Hypertension}` ⇒ `{Chronic_Kidney_Disease}`
- What medication combinations tend to follow certain diagnoses?
  `{Type_II_Diabetes, Obesity}` ⇒ `{Metformin}`
- Which symptoms strongly predict a future diagnosis?
  `{Night_Sweats, Weight_Loss}` ⇒ `{Tuberculosis}`
- Are there unexpected adverse drug combinations?
  `{Drug_A, Drug_B}` ⇒ `{Abnormal_Liver_Enzymes}`

Healthcare - Patient Record Transactions

```
Transaction T101 (Patient Visit)
--------------------------------
Diagnosis:        [Hypertension]
Diagnosis:        [Obesity]
Medication:       [Metformin]
Symptom:          [Fatigue]
Lab Result:       [Glucose_High]
--------------------------------
T101 = {Hypertension, Obesity, Metformin,
        Fatigue, Glucose_High}
```

```
Transaction T102
--------------------------------
Diagnosis:        [Asthma]
Medication:       [Inhaler]
Symptom:          [Wheezing]
Environmental:    [Pollen_High]
--------------------------------
T102 = {Asthma, Inhaler, Wheezing,
        Pollen_High}
```

Cybersecurity: Intrusion Detection

Context

- Security operations centers use association rules to detect patterns of suspicious behavior.
- Focus on co-occurring events across logs and emerging attack sequences.

Data structures

- Transactions: sessions, sequences of log events, or daily logs.
- Items: event types, IP categories, resources accessed, anomalies.

Cybersecurity - Data & Preprocessing

Data acquisition

- SIEM log events.
- Firewall logs.
- Authentication logs.
- System event feeds.

Preprocessing

- Normalize timestamps.
- Map raw events to categorical labels (for example, port_scan_detected).
- Remove duplicate or irrelevant logs.
- Aggregate logs into session windows (for example, per user per hour).

Meaning of SIEM Log Events

Definition

- SIEM = Security Information and Event Management.
- SIEM log events are security-relevant records aggregated, normalized, and analyzed by a SIEM platform.

Sources

- OS logs (Windows Event Logs, syslog)
- Network devices (routers, firewalls)
- Security tools (IDS, IPS, antivirus, EDR)
- Authentication systems (AD, LDAP)
- Cloud logs (AWS CloudTrail, Azure)

Examples

- Failed logins
- Privilege escalation attempts
- Firewall rule violations
- Malware detection events
- Port scans, unusual traffic

Purpose

- Detect threats via correlation
- Identify anomalies
- Support forensic investigations
- Enable compliance reporting

Cybersecurity - Postprocessing & Outcomes

Postprocessing

- Filter rules with lift > 1.5 to reduce false positives.
- Cross-check with known MITRE ATT&CK techniques.
- Validate rules with security experts.

Actionable outcomes

- Early alerts for suspicious event combinations.
- Improved threat signatures.
- Automated risk scoring.
- Prioritization of machines for forensic review.

Cybersecurity: Typical Questions

- Which event combinations precede a confirmed intrusion?
  `{Multi_Fail_Login, Unusual_Time_Access}` ⇒ `{Unauthorized_Access}`
- Which patterns of behavior distinguish normal from anomalous activity?
- Which attack chains co-occur across different machines?
  `{Port_Scan, SMB_Exploit}` ⇒ `{Ransomware_Deployment}`
- Which user or device profiles correlate with higher breach likelihood?

Cybersecurity - Event Log Transactions

```
Transaction S301 (User Session)
-------------------------------------------
Event:         [Failed_Login]
Event:         [VPN_Login]
Event:         [ privilege_escalation ]
Resource:      [Admin_Panel]
Time:          [Unusual_Time]
-------------------------------------------
S301 = {Failed_Login, VPN_Login,
        Privilege_Escalation,
        Access_Admin_Panel,
        Unusual_Time}
```

```
Transaction S302
-------------------------------------------
Event:         [Port_Scan]
Event:         [SMB_Exploit]
Event:         [File_Encryption]
-------------------------------------------
S302 = {Port_Scan, SMB_Exploit,
        File_Encryption}
```

Bioinformatics: Gene Expression

Context

- Association rules reveal relationships among gene expressions, protein interactions, and pathways.
- Applied to large genomic or transcriptomic datasets.

Data structures

- Transactions: samples, experiments, expression profiles.
- Items: gene up/down states, protein interactions, pathway activations.

Bioinformatics: Typical Questions

- Which sets of genes are co-expressed under certain conditions?
  `{Gene_A_up}` ⇒ `{Gene_B_up}`
- Which expression patterns predict disease phenotypes?
  `{Gene_X_up, Gene_Y_down}` ⇒ `{Tumor_Aggressive}`
- What protein interaction chains commonly appear together?
- Which pathways are co-activated in specific cancers?

Bioinformatics - Gene Expression Profiles

```
Transaction G12 (Tumor Sample)
---------------------------------------
Gene_A:          [Upregulated]
Gene_B:          [Downregulated]
Gene_C:          [Upregulated]
Protein_X:       [Interaction_Active]
---------------------------------------
G12 = {Gene_A_up, Gene_B_down,
       Gene_C_up, Protein_X_interact}
```

```
Transaction G13
---------------------------------------
Gene_D:          [Upregulated]
Pathway_Y:       [Activated]
---------------------------------------
G13 = {Gene_D_up, Pathway_Y_active}
```

Public Health: Outbreak and Risk Factors

Context

- Public health agencies mine co-occurring symptoms, conditions, and social determinants.
- Goal: understand and anticipate disease outbreaks.

Data structures

- Transactions: individual case reports or region-day aggregates.
- Items: symptoms, demographics, exposures, environmental conditions.

Public Health: Typical Questions

- Which symptom clusters strongly indicate a specific disease?
  `{Rash, Fever}` ⇒ `{Measles}`
- Which environmental conditions co-occur with disease spikes?
  `{High_Humidity, Standing_Water}` ⇒ `{Dengue_Outbreak}`
- Which risk factors tend to appear together in severe cases?
  `{Smoking, Air_Pollution}` ⇒ `{Severe_Respiratory_Issues}`
- Which combinations of travel history and symptoms predict imported cases?

Public Health - Epidemiology Case Transactions

```
Transaction C901 (Case Report)
-------------------------------------------------
Symptom:       [Fever]
Symptom:       [Rash]
Exposure:      [Travel_Region_X]
Demographic:   [Child]
-------------------------------------------------
C901 = {Fever, Rash, Travel_Region_X, Child}
```

```
Transaction C902
-------------------------------------------------
Symptom:       [Cough]
Environment:   [Air_Pollution_High]
Behavior:      [Smoking]
-------------------------------------------------
C902 = {Cough, Air_Pollution_High, Smoking}
```

Manufacturing: Fault Diagnosis

Context

- Factories apply association rules to uncover failure patterns.
- Supports predictive maintenance and reliability engineering.

Data structures

- Transactions: machine cycles, daily logs, fault incidents.
- Items: sensor anomalies, part replacements, error codes, vibration spikes.

Manufacturing: Typical Questions

- Which sensor readings in combination precede a specific failure?
  {Temp_High, Vibration_High} ⇒ {Bearing_Failure}
- What component co-failures frequently occur together?
  {Pump_Failure} ⇒ {Valve_Replacement}
- Which maintenance actions tend to resolve related anomalies?
- Which operational conditions correlate with reduced lifespan?

Manufacturing - Machine Cycle Snapshot

```
Transaction M203 (Machine Cycle)
--------------------------------------------
[Sensor]          Temp_High
[Sensor]          Vib_Spike
[Error Code]      E17
[Maintenance]     None
--------------------------------------------
M203 = {Temp_High, Vib_Spike, Error_E17}
```

```
Transaction M204
-----------------------------------------
[Sensor]          Noise_High
[Repair]          Bearing_Replace
-----------------------------------------
M204 = {Noise_High, Bearing_Replace}
```

Education: Learning Analytics

Context

- Universities analyze behavior patterns, resource usage, and outcomes using association rules.
- Aim: improve learning design and early-warning systems.

Data structures

- Transactions: a student term, course session, or weekly activity.
- Items: actions (viewed lectures, submissions), skills, quiz scores.

Education: Typical Questions

- Which behaviors predict high performance?
  {Early_Assignment_Submission, Forum_Participation} ⇒ {High_Grade}
- Which behaviors signal dropout risk?
  {No_Logins_Week2, Missed_Quiz_1} ⇒ {Dropout}
- Which learning resources are often used together?
- How do specific misconceptions co-occur across assessments?

Education - Student Interaction Log

```
Transaction A012 (Week 2 Behavior)
-------------------------------------------------
[Action]          View_Lecture_3
[Action]          View_Lecture_4
[Quiz]            Quiz_1_Attempted
[Performance]     Low_Score
-------------------------------------------------
A012 = {View_Lecture_3, View_Lecture_4,
        Quiz_1_Attempted, Low_Score}
```

```
Transaction A013
-------------------------------------------------
[Action]          Forum_Post
[Action]          Early_Submission
[Outcome]         High_Grade
-------------------------------------------------
A013 = {Forum_Post, Early_Submission, High_Grade}
```

Smart Cities: Traffic Patterns

Context

- Urban planners use association rules on IoT and traffic data to find patterns leading to congestion or accidents.

Data structures

- Transactions: road-segment time windows (for example, 5 minutes).
- Items: traffic volume, weather, accidents, low speed, congestion.

Smart Transportation - Data & Preprocessing

Data acquisition

- Roadside IoT sensors.
- Weather stations.
- GPS and speed detectors.
- Incident reports.

Preprocessing

- Synchronize time windows across sensors.
- Convert continuous values to categories (for example, Speed_Low, Density_High).
- Handle missing sensor data.
- Aggregate into fixed windows (for example, 5-minute intervals).

Pipeline 3: Smart Transportation - Postprocessing & Outcomes

Postprocessing

- Rank rules by lift to detect strongest indicators.
- Validate with historical crash data.
- Cluster rules by road type (for example, highway vs urban).

Actionable outcomes

- Dynamic speed-limit recommendations.
- Accident probability dashboards.
- Better placement of signage and sensors.
- Predictive alerts for drivers.

Smart Cities: Typical Questions

- Which conditions co-occur immediately before accidents?
  `{Rain, Low_Speed, High_Density}` ⇒ `{Collision}`
- What combinations of road conditions cause predictable congestion?
  `{Construction, Lane_Closure}` ⇒ `{Severe_Delay}`
- How do weather patterns influence traffic flow?
- Which intersections exhibit recurring joint anomalies?

Smart Cities - Road Segment Events

```
Transaction T550 (Segment: Highway-12)
-------------------------------------------------
Weather:      [Rain]
Traffic:      [High_Density]
Speed:        [Below_30]
Event:        [Accident]
-------------------------------------------------
T550 = {Rain, High_Density, Speed_Low, Accident}
```

```
Transaction T551
-------------------------------------------------
Weather:      [Clear]
Traffic:      [Moderate]
Event:        [No_Accident]
-------------------------------------------------
T551 = {Clear, Moderate_Traffic, No_Accident}
```

Wrap-up

- Association rules apply far beyond commerce and marketing.
- Common pattern: transactions of events + items as discrete attributes.
- Output: interpretable rules that support decision making and policy.