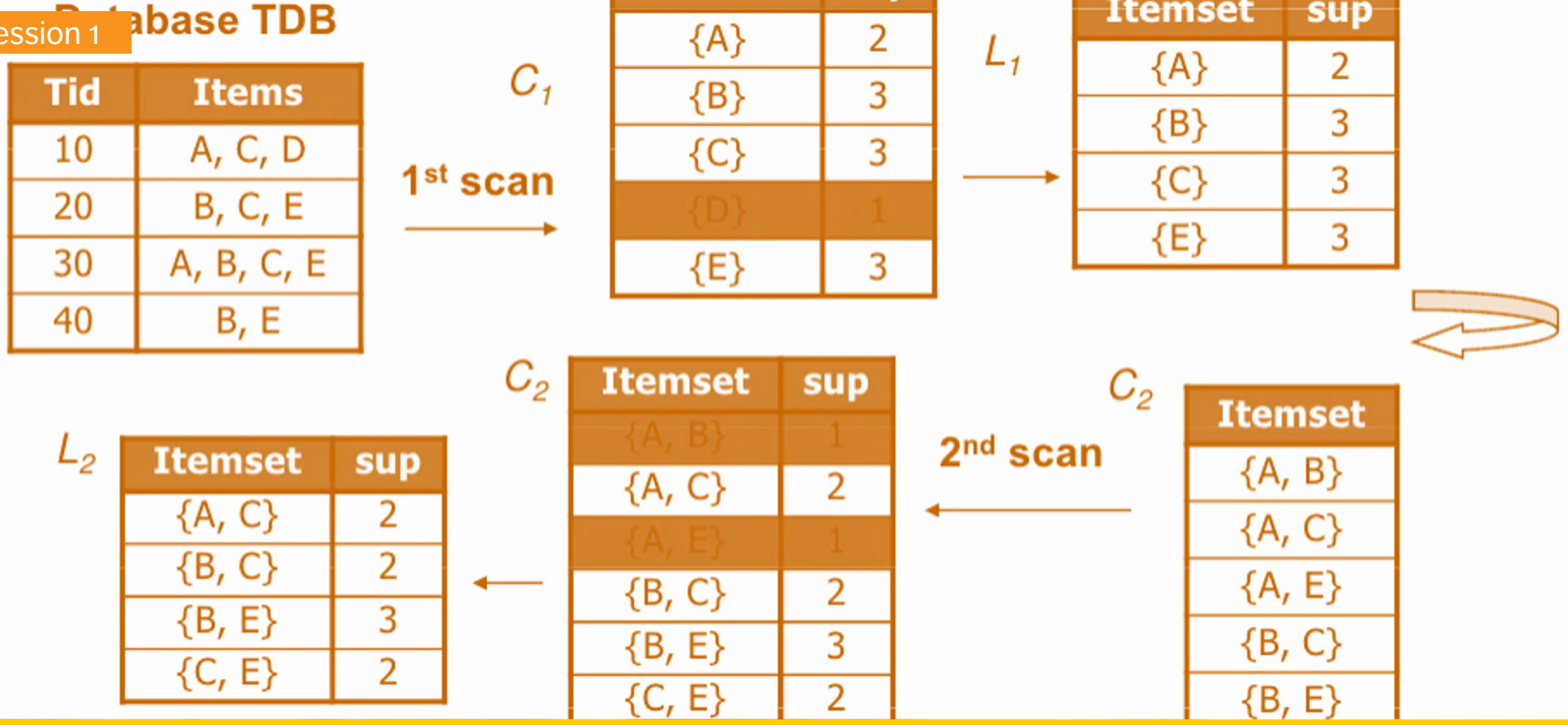
Association Rules Mining

INTRODUCTION TO MACHINE LEARNING

Correlation isn’t causation... but it’s a big hint!

[E. Tufte]



2. Association Rules Overview

Overview

Association rules discovery (ARD) is a type of unsupervised learning that finds **connections** among the attributes and levels of a dataset's observations.

We might analyze a dataset on the physical activities and purchasing habits of North Americans and discover that

- runners who are also triathletes (the **premise**) tend to drive Subarus, drink microbrews, and use smart phones (the **conclusion**)
- individuals who have purchased home gym equipment are unlikely to be using it 1 year later

Overview

The presence of a **correlation** between the premise and the conclusion does not imply the existence of a **causal relationship** between them.

It is difficult to prove causation *via* data analysis; in practice, decision-makers pragmatically (often erroneously) focus on "**there's no smoke without fire**."

Example: being a triathlete does not cause one to drive a Subaru, but Subaru Canada thought that the connection was strong enough to offer to reimburse the registration fee at an IRONMAN 70.3 competition (at least in 2018)!

Market Basket Analysis

ARD is aka as **market basket analysis**.

Example: purchase of bread and milk, but that is unlikely to be of interest given the frequency of market baskets containing milk (**or** bread).

If the presence of milk is **independent** of the presence of bread (and *vice-versa*), and if 70% of baskets contain milk and 90% contain bread, say, we would expect **at least** 90% × 70% = 63% of all baskets to contain **both**.

If we observe both in 72% of baskets, say (a 1.15-fold increase), we conclude that there is a **weak correlation** between the milk and bread purchases.

Market Basket Analysis

Sausages and buns are not purchased as frequently as milk and bread, but they might still be purchased as a pair more often than one would expect.

If the presence of sausage is **independent** of the presence of buns (and *vice-versa*), and if 10% of baskets contain sausages and 5% contain buns, say, we would expect **at least** 10% × 5% = 0.5% of all baskets to contain **both**.

If we observe both in 4% of baskets, say (an 8-fold increase), we conclude that there is a **strong correlation** between the sausage and buns purchases.

Market Basket Analysis

How can we **act** on this insight? Supermarkets could advertise a sale on sausages while **simultaneously** (and quietly) raising the price of buns. This could have the effect of bringing in a higher number of customers into the store, hoping to increase the **sale volumes** for both items while keeping the **combined price of the two items constant**.

Little Story: a supermarket found an association rule linking the purchase of beer and diapers and consequently moved its beer display closer to its diapers display, having confused correlation and causation.

What do you think might actually be happening here?

Applications

Typical uses include:

- finding **related concepts** in text documents – looking for pairs (triplets, etc) of words that represent a joint concept: {San Jose, Sharks}, {Michelle, Obama}, etc.;
- detecting **plagiarism** – looking for specific sentences that appear in multiple documents, or for documents that share specific sentences;
- identifying **biomarkers** – finding diseases frequently associated with a set of biomarkers;

Applications

Typical uses include:

- making predictions and decisions based on association rules (there are pitfalls)
- altering circumstances to take advantage of correlations (suspected causal effect)
- using connections to modify the likelihood of certain outcomes (see above)
- imputing missing data
- text autofill and autocorrect
- etc.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Objective

Using data from the *Danish National Patient Registry*, the authors sought connections between different **diagnoses:** how does a diagnosis at some point in time allow for the prediction of another diagnosis at a later time?

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Methodology

1. compute the **strength of correlation** for pairs of diagnoses over a 5 year interval (on a representative subset of the data)
2. test diagnoses pairs for **directionality** (one diagnosis repeatedly occurring before the other)
3. determine reasonable **diagnosis trajectories** (thoroughfares) by combining smaller (but frequent) trajectories with overlapping diagnoses
4. **validate** the trajectories by comparison with non-Danish data
5. **cluster** the thoroughfares to identify a small number of **central medical conditions** (key diagnoses) around which disease progression is organized

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Data

The *Danish National Patient Registry* is an electronic health registry containing administrative information and diagnoses, covering the whole population of Denmark, including private and public hospital visits of all types:

- inpatient (overnight stay)
- outpatient (no overnight stay)
- emergency visits.

The data set covers 15 years of such visits, from January ’96 to November ’10, and consists of 68 million records for 6.2 million patients.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Challenges and Pitfalls

- Access to the **patient registry** was protected and could only be granted after approval by *the National Board of Health*.
- There are gender-specific differences in diagnostic trends, but many diagnoses were made predominantly in different sites, suggesting the stratifying by **site** as well as by **gender**.
- In the process of forming small diagnoses chains, they had to compute the correlations using **large groups** for each pair of diagnoses (1 million diagnosis pairs = 80+ million samples) to compensate for **multiple testing** (1000s years' worth of CPU run time) – pre-filtering steps were used to avoid this pitfall.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

The dataset was reduced to **1,171 significant trajectories**.

These thoroughfares were clustered into patterns centred on 5 key diagnoses for disease progression:

- **diabetes**
- **chronic obstructive pulmonary disease** (COPD)
- **cancer**
- **arthritis**
- **cerebrovascular disease**

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

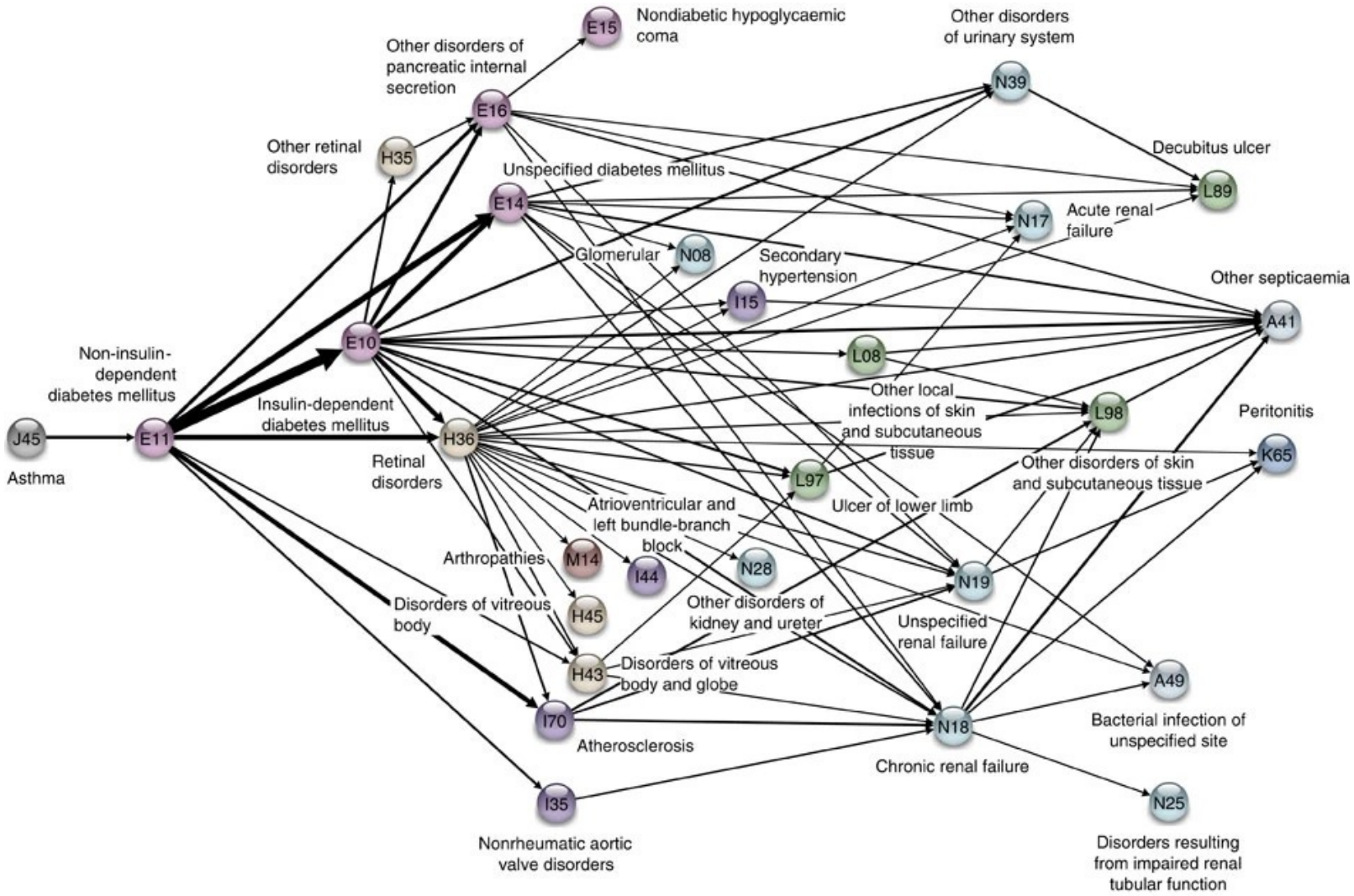
Early diagnoses for these central factors can help reduce the risk of adverse outcome linked to future diagnoses of other conditions.

Among the specific results, the following “surprising” insights were found:

- a diagnosis of anemia is typically followed months later by the **discovery of colon cancer**
- a diagnosis of gout was identified as **a step on the path** toward eventual diagnosis of cardiovascular diseases
- COPD is **under-diagnosed** and **under-treated**

Case Study

Danish Medical Data



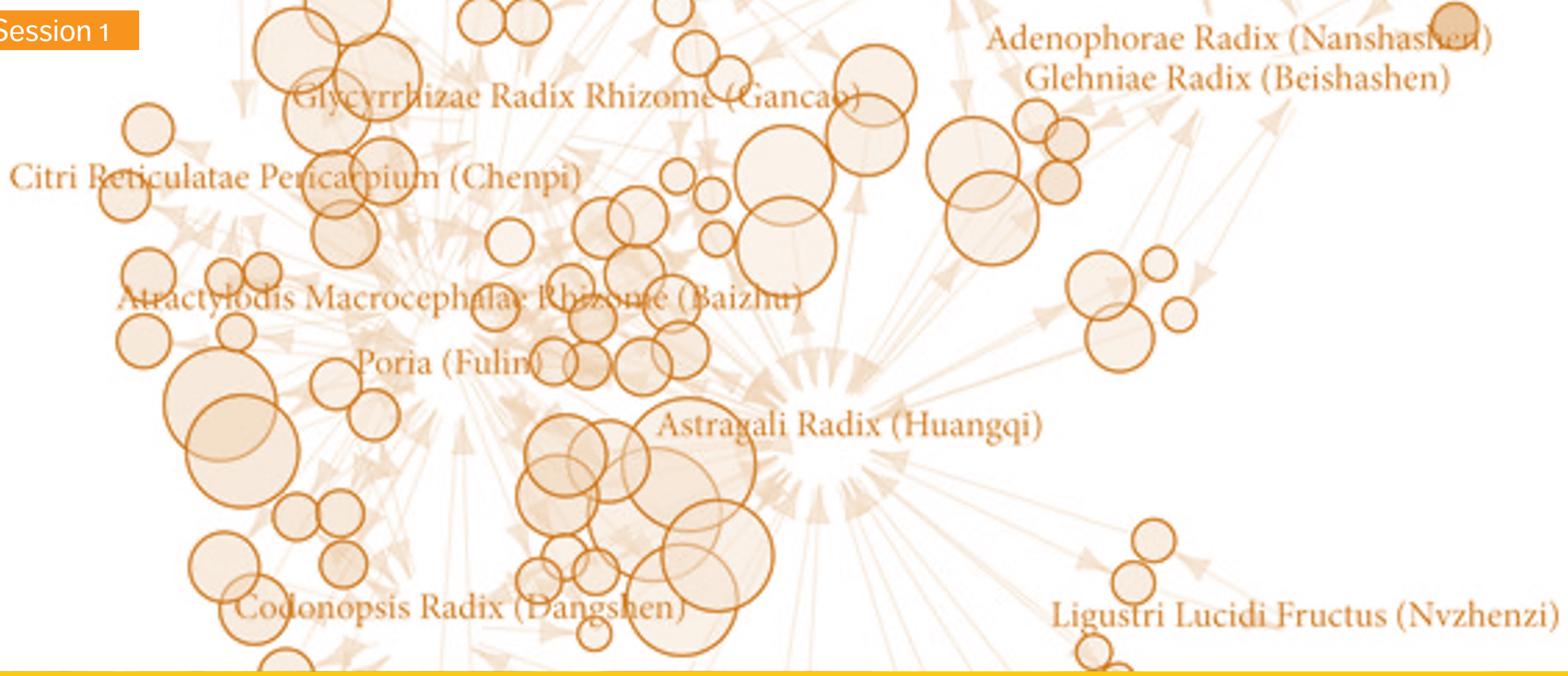
Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Suggested Reading

Association Rules Overview

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.3 Association Rules Mining
    - Overview
    - Case Study: Danish Medical Data



3. Association Rules Concepts

Correlation and Causation

Association rules can automate **hypothesis discovery**, but one must remain correlation-savvy (less prevalent than one might hope...).

If attributes A and B are correlated in a dataset, there are various possibilities:

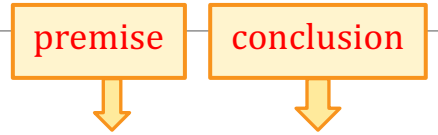
- A and B are correlated entirely by chance in this particular dataset
- A is a re-labeling of B (or *vice-versa*)
- A causes B (or *vice-versa*)
- some other attributes $C_1, \ldots, C_n$ (which may not be available in the data) cause A and B
- etc.

[E. Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die]

Correlation and Causation

Insight	Organization
Pop-Tarts sales shoot up before a hurricane	Walmart
Higher crime, more Uber rides	Uber
Typing with proper capitalization indicates creditworthiness	A financial services startup company
Users of the Chrome and Firefox browsers make better employees	A human resources professional services firm, over employee data from Xerox and other firms
Men who skip breakfast get more coronary heart disease	Harvard University medical researchers
More engaged employees have fewer accidents	Shell
Smart people like curly fries	Researchers at the University of Cambridge and Microsoft Research
Female-named hurricanes are more deadly	University researchers
Higher status, less polite	Researchers examining Wikipedia behavior

Definitions



A **rule** $X \rightarrow Y$ is a statement of the form "if X then Y" built from any logical combinations of a dataset attributes.

A rule **does not need to be true for all observations** in the dataset – there could be instances where the premise is satisfied but the conclusion is not.

Some of the "best" rules are those which are only accurate 10% of the time, as opposed to rules which are only accurate 5% of the time, say.

It depends on the context.

Definitions

To determine a rule's strength, we compute various **rule metrics**, such as the:

- **support** (the frequency at which a rule occurs in a dataset) – low coverage values indicate rules that rarely occur
- **confidence** (the reliability of the rule: how often does the conclusion occur in the data given that the premises have occurred) –high confidence rules are "truer"
- **interest** (the difference between its confidence and the relative frequency of its conclusion) – rules with high absolute interest are more "interesting"
- **lift** (the increase in the frequency of the conclusion which can be explained by the premises) – with a high lift (> 1), the conclusion occurs more frequently than expected
- also **conviction**, **all-confidence**, **leverage**, **collective strength**, etc.

[**Note:** $\text{Freq}(A) \in \{0,1,\ldots,N\} =$ # of observations for which A holds]

Definitions

If N is the number of observations in a dataset, then:

$$\text{Support}(X \to Y) = \frac{\text{Freq}(X \cap Y)}{N} \in [0,1]$$

Proportion of instances where the premise and the conclusion occur together

$$\text{Confidence}(X \to Y) = \frac{\text{Freq}(X \cap Y)}{\text{Freq}(X)} \in [0,1]$$

Proportion of instances where the conclusion occurs when the premise occurs

$$\text{Interest}(X \to Y) = \text{Confidence}(X \to Y) - \frac{\text{Freq}(Y)}{N} \in [-1,1]$$

$$\text{Lift}(X \to Y) = \frac{N^2 \cdot \text{Support}(X \to Y)}{\text{Freq}(X) \cdot \text{Freq}(Y)} \in (0, N^2)$$

... ?!?

$$\text{Conviction}(X \to Y) = \frac{1 - \text{Freq(Y)}/N}{1 - \text{Confidence}(X \to Y)} \geq 0$$

Interpretation of the Lift: 70% of those born before 1976 own a copy, whereas 59% of the data's population own a copy.

$$1.18 \approx \frac{0.70}{0.59}$$

Example

RM: if an individual is born before 1976 (X), then they own a copy of the Beatles' *Sergeant Peppers' Lonely Hearts Club Band*, in some format (Y).

Assume that :

- $N = 15{,}356$
- $\text{Freq}(X) = 3888$
- $\text{Freq}(Y) = 9092$
- $\text{Freq}(X \cap Y) = 2720$

$$\text{Support}(\text{RM}) = \frac{2720}{15,536} \approx 18\%$$

$$\text{Confidence}(\text{RM}) = \frac{2720}{3888} \approx 70\%$$

$$\text{Interest}(\text{RM}) = \frac{2720}{3888} - \frac{9092}{15,356} \approx 0.11$$

$$\text{Lift}(\text{RM}) = \frac{15,356^2 \cdot 0.18}{3888 \cdot 9092} \approx 1.2$$

$$\text{Conviction}(\text{RM}) = \frac{1 - 9092/15,356}{1 - 2720/3888} \approx 1.36$$

Interpreting Association Rules

All this seems to point to the rule RM being not entirely devoid of meaning, but to what extent, exactly? **This is a difficult question to answer**.

It is difficult to provide thresholds, but evaluation of a lone rule is **meaningless**.

It is recommended to conduct a **preliminary exploration** of the space of association rules (using domain expertise) in order to determine reasonable threshold ranges for the specific situation; candidate rules would then be discarded or retained depending on these metric thresholds.

This requires the ability to “easily” generate potential candidate rules.

Generating Association Rules

The real challenge of association rules discovery is to **generate** candidate rules without wasting time generating rules which are likely to be discarded.

An **itemset** for a dataset is a list of attributes with values. A set of **rules** can be created from the itemset by adding "**IF** ... **THEN**" blocks to the instances.

From $\{\text{membership} = \text{True}, \text{age} = \text{Youth}, \text{purchasing} = \text{Typical}\}$, we can get:

- **IF** (purchasing = Typical AND membership = True) **THEN** age = Youth
- **IF** age = Youth **THEN** membership = True, etc.

$\boldsymbol{n}$ **items** $\Rightarrow$ $\boldsymbol{2^n - 1}$ **rules** (combinatorial explosion)

Brute Force Algorithm

1. Generate item sets (of size 1, 2, 3, 4, etc.).
2. Create rules from each item set.
3. Calculate the support, confidence, interest, lift, conviction, etc., for each rule.
4. Retain only the rules with “high enough” coverage, accuracy, interest, lift, conviction, or other appropriate metrics.
5. These rules are considered to be **true** for the dataset – they are **new knowledge derived from the data**.

A Priori Algorithm

The combinatorial explosion is a problem – it disqualifies the **brute force** approach for any dataset with a realistic number of attributes.

How can we generate a small number of **promising** candidate rules?

The ***a priori*** algorithm is an early attempt to overcome that difficulty.

Initially, it was developed to work for **transaction data** (i.e. goods as columns, customer purchases as rows); every reasonable dataset can be transformed into a transaction dataset using dummy variables.

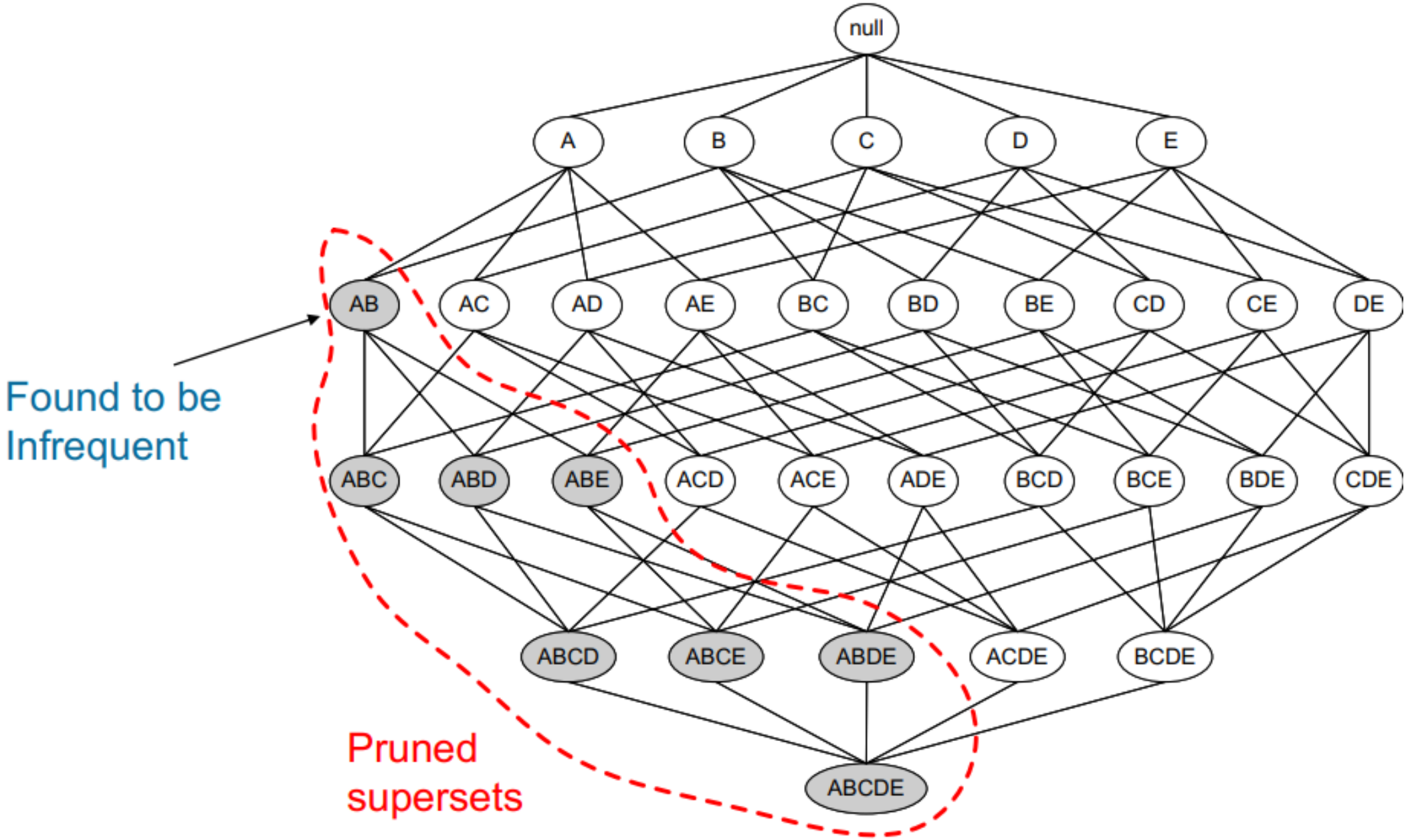
A Priori Algorithm

The a priori algorithm attempts to find **frequent itemsets** from which to build candidate rules, instead of building rules from **all** possible itemsets.

It starts by identifying frequent **individual items** in the database and extends those that are retained into larger and larger **item supersets**, who are themselves retained only if they occur **frequently enough** in the data.

The main idea is that “all non-empty subsets of a frequent itemset must also be frequent”, or equivalently, that all supersets of an infrequent itemset must also be infrequent.

[T. Chou, Apriori: Association Rule Mining In-Depth Explanation and Python Implementation]



A Priori Algorithm

The algorithm terminates when no further itemsets extensions are retained, which always occurs given the finite number of levels in categorical datasets:

- **strengths:** easy to implement and to parallelize
- **limitations:** slow, requires frequent scans, not ideal for infrequent and rare itemsets

More efficient algorithms have since displaced it in practice:

- **Max-Miner** tries to identify frequent itemsets without enumerating them – it performs jumps in itemset space instead of using a bottom-up approach
- **Eclat** is faster and uses depth-first search, but requires extensive memory storage

Validation

How **reliable** are association rules?

What is the likelihood that they occur entirely **by chance**?

How **relevant** are they?

Can they be generalized **outside** the dataset, or to **new** data streaming in?

Statistically sound association discovery can help reduce the risk of finding spurious associations to a user-specified significance level.

Validation

We end this section with a few comments:

- frequent rules correspond to instances that occur repeatedly in the dataset, algorithms that generate itemsets often try to **maximize coverage**; when **rare events** are more meaningful we need algorithms that can generate rare itemsets – **this is not a trivial problem**;
- continuous data has to be binned into **categorical** data to generate rules; as there are many ways to accomplish that task, the same dataset can give rise to completely different rules – this could create some **credibility issues** with clients and stakeholders;
- other algorithms: AIS, SETM, aprioriTid, aprioriHybrid, PCY, Multistage, Multihash, etc.

Suggested Reading

Association Rules Concepts

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.3 Association Rules Mining
    - Generating Rules
    - The *A Priori* Algorithm
    - Validation
    - Toy Example: Titanic Dataset

    19.7 R Examples
    - Association Rules Mining: Titanic Dataset

Exercises

Association Rules Concepts

1. Evaluate the following candidate rules in the music dataset:
   - if an individual owns a classical music album (W), they also own a hip-hop album (Z), given that $\text{Freq}(W) = 2010$, $\text{Freq}(Z) = 6855$, $\text{Freq}(W \cap Z) = 132$.
   - if an individual owns both a Beatles and a classical music album, then they were born before 1976, given that $\text{Freq}(Y \cap W) = 1852$, $\text{Freq}(Y \cap W \cap X) = 1778$.

2. Out of the 3 rules that have been established ($X \rightarrow Y$, $W \rightarrow Z$, $Y \,\&\, W \rightarrow X$), which do you think is more useful? Which is more surprising?

Exercises

Association Rules Concepts

3. A store that sells accessories for cellular phones runs a promotion on faceplates. Customers who purchase multiple faceplates from a choice of 6 different colours get a discount. Managers, who would like to know what colours will be purchased together, collected purchases in Transactions.csv.

   Consider the following rules:

   - {**r**ed, **w**hite} ⇒ {**g**reen}
   - {**g**reen} ⇒ {**w**hite}
   - {**r**ed, **g**reen} ⇒ {**w**hite}
   - {**g**reen} ⇒ {**r**ed}
   - {**o**range} ⇒ {**r**ed}
   - {**w**hite, **b**lack} ⇒ {**y**ellow}
   - {**b**lack} ⇒ {**g**reen}

Exercises

Association Rules Concepts

3. (cont.) For each rule, compute the **support**, **confidence**, **interest**, **lift**, and **conviction**. Amongst the rules for which the support is positive (> 0), which one has the highest lift? Confidence? Interest? Conviction? Build an additional 5-10 candidate rules, and evaluate them. Which of the 12-17 candidate rules do you think would be most useful for the store managers? How would one determine reasonable threshold values for the support, coverage, interest, lift, and conviction of rules derived from a given dataset?

Exercises

Association Rules Concepts

4. Go over the titanic association rules example found in DUDADS (see suggested reading). Repeat the process with the UniversalBank.csv dataset (you may need to clean and visualize the dataset first, as well as categorize the numerical variables; can you come up with a reasonable guess as to what each of the variables represent?). Find “true knowledge” about the dataset in the form of reliable and meaningful association rules (use metrics as appropriate).