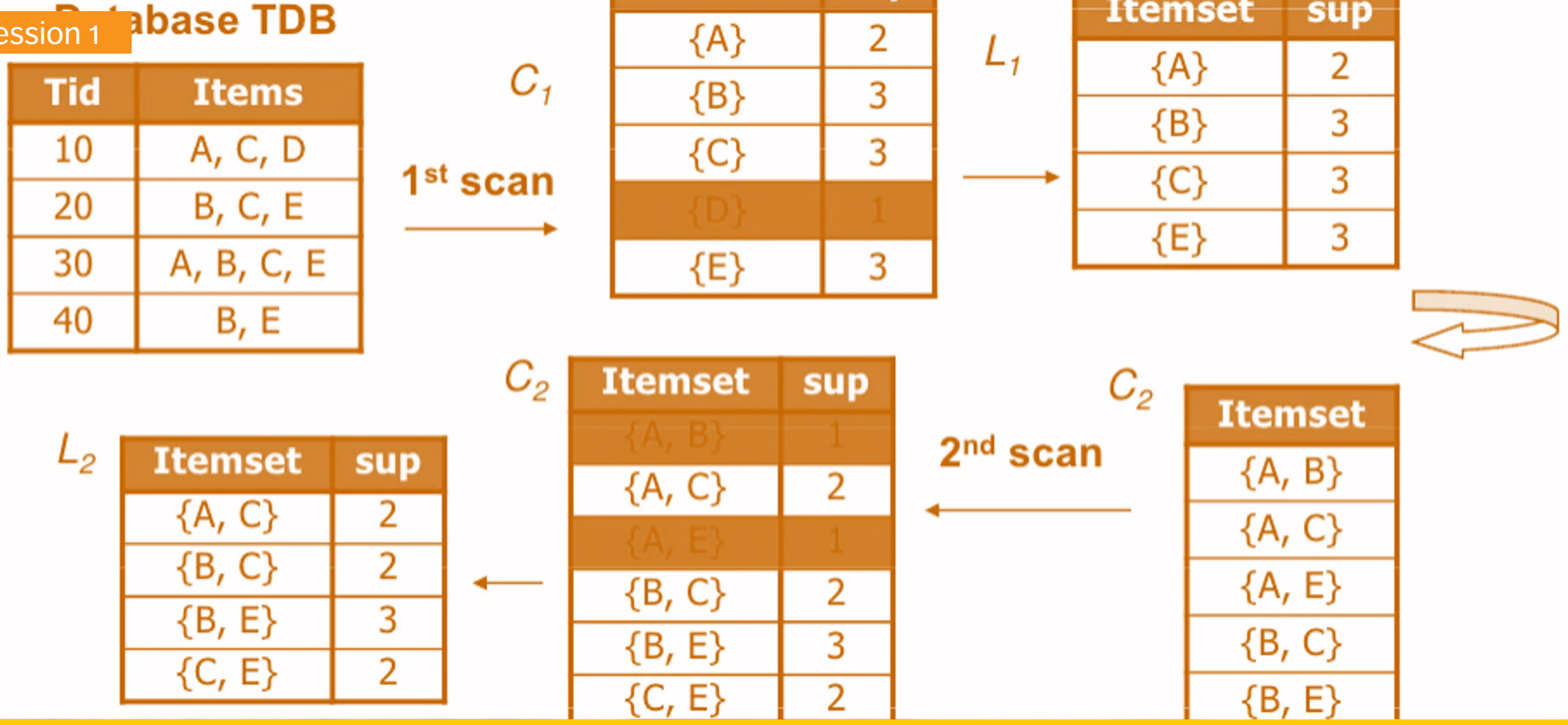
Extraction de règles d'association

INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

Corrélation n'est pas causalité... mais c'est un indice !

[E. Tufte]



2. Aperçu des règles d'association

Vue d'ensemble

La découverte de règles d'association (ARD) est un type d'apprentissage non supervisé qui permet de trouver des **liens** entre les attributs et les valeurs (niveaux) des observations d'un ensemble de données.

Nous pourrions analyser un ensemble de données sur les activités physiques et les habitudes d'achat des Nord-Américains et découvrir, par exemple, que :

- les coureurs qui sont aussi des triathlètes (la **prémisse**) ont tendance à conduire des Subarus, et à consommer des micro-bières (la **conclusion**), ou enore
- les personnes qui ont acheté un équipement de gymnastique à domicile ne l'utilisent pratiquement pas un an plus tard.

Vue d'ensemble

La présence d'une **corrélation** entre la prémisse et la conclusion n'implique pas l'existence d'une **relation de cause à effet**, cependant.

Il est difficile de prouver un lien de causalité par le biais d'analyse des données ; dans la pratique, les décideurs se concentrent de manière pragmatique (souvent à tort) sur le fait “qu’il n'y a pas de **fumée sans feu**".

Exemple : être triathlète ne pousse pas à conduire une Subaru, mais Subaru Canada a estimé que le lien était suffisamment fort pour proposer de rembourser les frais d'inscription à une compétition IRONMAN 70.3 (en 2018) !

Analyse des paniers

L'ARD est aussi connue sous le nom d'**analyse des paniers**.

Exemple : l'achat de pain et de lait, mais cela présente un faible intérêt étant donné la fréquence des paniers de marché contenant du lait (**ou du** pain).

Si la présence de lait est **indépendante** de la présence de pain (et *vice-versa),* et si 70 % des paniers contiennent du lait et 90 % du pain, mettons, nous nous attendons à ce qu'au **moins** 90 % × 70 % = 63 % de tous les paniers contiennent les **deux**.

Si nous observons les deux dans 72 % des paniers, disons (facteur: 1.15), on déduit qu'il existe une **faible corrélation** entre les achats de lait et de pain.

Analyse des paniers

Les saucisses et les pains à hot dog ne sont pas achetées aussi fréquemment que le lait et le pain, mais il est possible qu'elles soient achetées ensemble plus souvent qu'on ne le pense.

Si la présence de saucisses est **indépendante** de la présence de pains à hot dog (et *vice-versa),* et si 10 % des paniers contiennent des saucisses et 5 % des pains, mettons, nous nous attendons à ce qu'au **moins** 10 % × 5 % = 0.5 % de tous les paniers contiennent les **deux**.

Si nous observons les deux dans 4% des paniers, disons (facteur: 8), nous en concluons qu'il existe une **forte corrélation** entre les achats de saucisses et des pains à hot dog.

Analyse des paniers

Comment **agir sur** la base de cette idée ? Les supermarchés pourraient annoncer des soldes sur les saucisses tout en augmentant **simultanément** (et discrètement) le prix des pains, ayant pour effet d'attirer plus de clients dans le magasin, dans l'espoir d'augmenter les **volumes de vente** pour les deux articles tout en maintenant le **prix combiné des deux articles**.

Petite histoire : un supermarché a trouvé une règle d'association liant l'achat de bière et de couches et a par conséquent rapproché son étalage de bière de son étalage de couches, ayant confondu corrélation et causalité.

Que pensez-vous qu'il se passe réellement ici ?

Applications

Les utilisations typiques sont les suivantes :

- recherche de **concepts apparentés** dans des documents textuels - recherche de paires (triplets, etc.) de mots représentant un concept commun : {San Jose, Sharks}, {Michelle, Obama}, etc ;
- détecter le **plagiat** – rechercher des phrases spécifiques qui apparaissent dans plusieurs documents, ou des documents qui partagent des phrases spécifiques ;
- l'identification de **biomarqueurs** – la recherche de maladies fréquemment associées à un ensemble de biomarqueurs ;

Applications

Les utilisations typiques sont les suivantes :

- faire des prédictions et prendre des décisions sur la base de règles d'association (il y a des pièges)
- modifier les circonstances pour tirer profit des corrélations (effet causal présumé)
- l'utilisation de connexions pour modifier la probabilité de certains résultats (voir ci-dessus)
- imputation des données manquantes
- remplissage automatique de texte et correction automatique
- etc.

Étude de cas

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

Objectif

En utilisant les données du *Danish National Patient Registry*, les auteurs ont cherché à établir des liens entre différents **diagnostiques :** comment un diagnostique posé à un moment donné permet-il de prédire un autre diagnostique à un moment ultérieur ?

Étude de cas

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

Méthodologie

1. calculer la **force de la corrélation** pour des paires de diagnostiques sur un intervalle de 5 ans (sur un sous-ensemble représentatif des données)
2. tester les paires de diagnostiques pour vérifier la **directionnalité** (un diagnostique apparaissant de manière répétée avant l'autre)
3. déterminer des **trajectoires de diagnostique** raisonnables en combinant des trajectoires plus petites (mais fréquentes) avec des diagnostiques qui se chevauchent
4. **valider les** trajectoires par comparaison avec des données non danoises
5. **regrouper les** voies de passage pour identifier un petit nombre de **conditions médicales centrales** (diagnostiques clés) autour desquelles s'organise la progression de la maladie

Étude de cas

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

Données

Le *Danish National Patient Registry* est un registre de santé électronique contenant des informations administratives et des diagnostiques, couvrant l'ensemble de la population du Danemark :

- hospitalisation (nuitée)
- ambulatoire (sans nuitée)
- les visites d'urgence.

L'ensemble des données couvre 15 années de visites de ce type, de janvier 1996 à novembre 2010, et comprend 68 millions d'enregistrements pour 6,2 millions de patients.

Étude de cas

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

Défis et pièges

- L'accès au **registre des patients** est protégé et ne peut être accordé qu'après approbation du *National Board of Health*.
- Il existe des différences entre les sexes dans les tendances diagnostiques, mais de nombreux diagnostiques ont été posés principalement dans différents sites, ce qui suggère une stratification par **site** ainsi que par **sexe**.
- Lors de la formation de petites chaînes de diagnostiques, on a dû calculer les corrélations en utilisant de **grands groupes** pour chaque paire de diagnostiques (1 million de paires de diagnostiques = 80+ millions d'échantillons) pour compenser les **tests multiples** (des milliers d'années de temps d'exécution de l'unité centrale) – des étapes de pré-filtrage ont été utilisées pour éviter cet écueil.

Étude de cas

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

Résumé et résultats du projet

Les données ont été réduites à **1171 trajectoires significatives.**

Ces trajectoires ont été regroupées selon des schémas centrés sur cinq diagnostiques clés pour l'évolution de la maladie :

- le **diabète**
- la **bronchopneumopathie chronique obstructive** (BPCO)
- les **cancers**
- l'**arthrite**
- les **maladies cérébrovasculaires**

Étude de cas

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

Résumé et résultats du projet

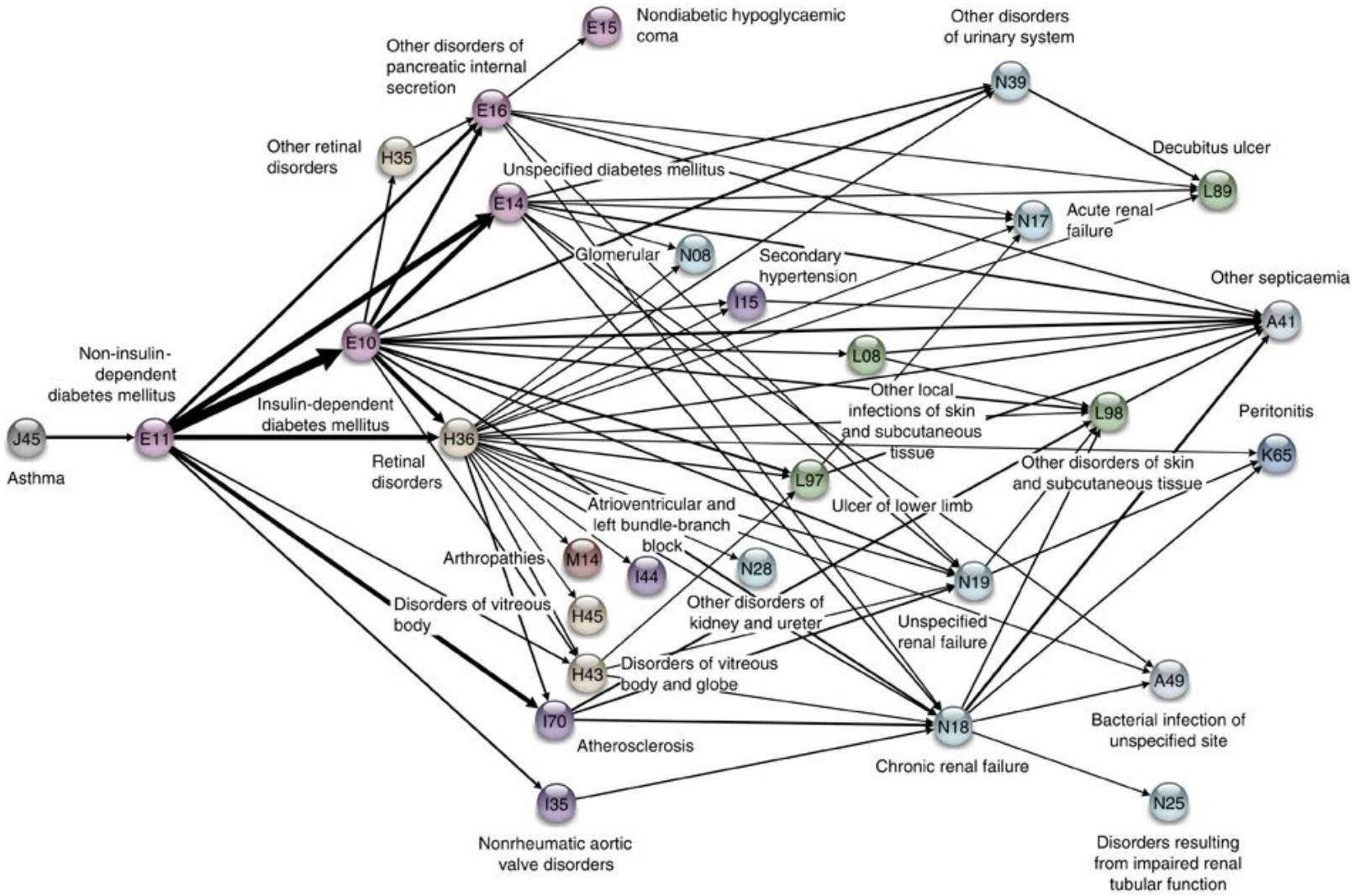
Un diagnostique précoce de ces facteurs centraux peut réduire le risque de résultats défavorables liés à des diagnostiques ultérieurs d'autres pathologies.

Parmi les résultats spécifiques, les informations "surprenantes" suivantes ont été trouvées :

- un diagnostique d'anémie est généralement suivi quelques mois plus tard par la **découverte d'un cancer du côlon**
- le diagnostique de la goutte a été identifié comme **une étape sur la voie** d'un éventuel diagnostique des maladies cardiovasculaires
- BPCO est **sous-diagnostiquée** et **insuffisamment traitée**

Étude de cas

Danish Medical Data



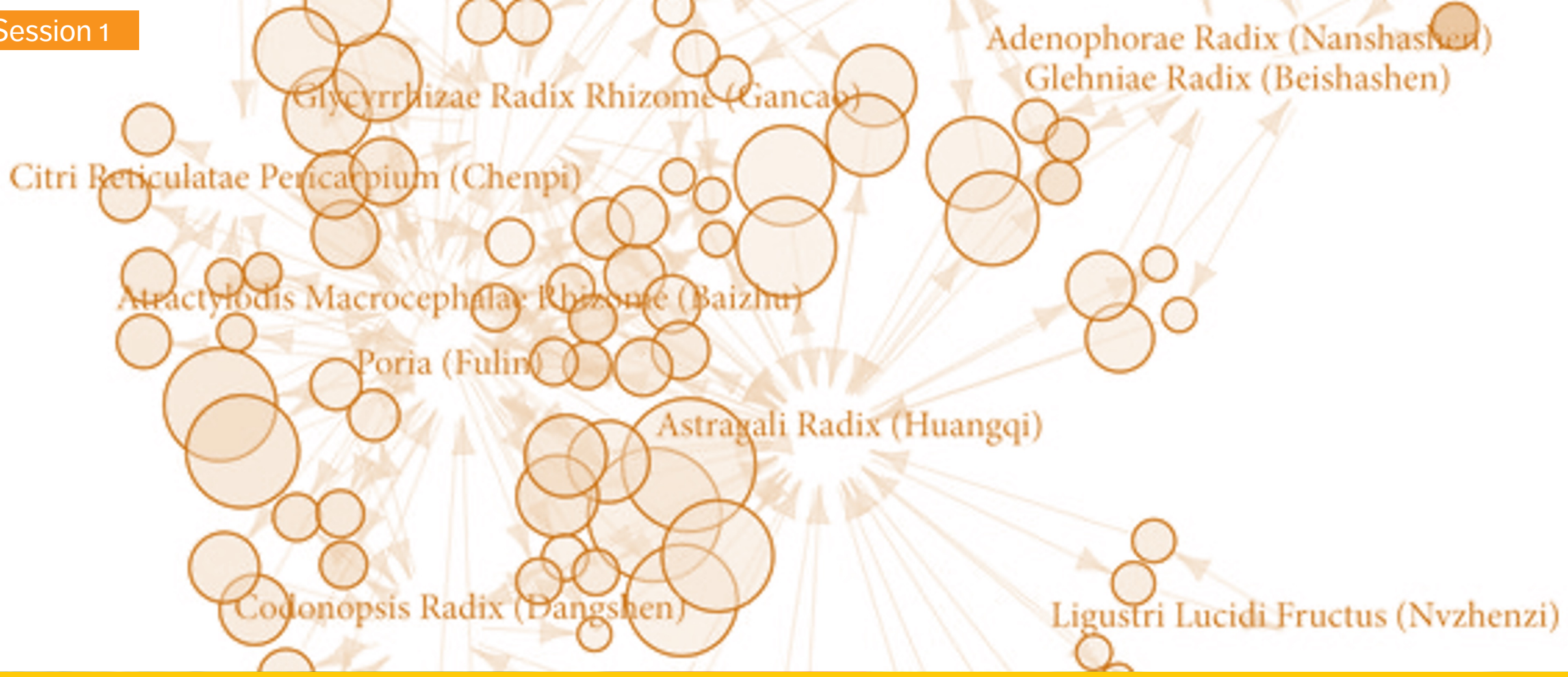
Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Lectures conseillées

Aperçu des règles de l'association

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. Concepts de règles d'association

Corrélation et causalité

Les règles d'association peuvent automatiser la **découverte d'hypothèses**, mais il ne faut pas tomber dans le paneau vis-à-vis les **corrélations**.

Si les attributs A et B sont corrélés dans un ensemble de données, il y a plusieurs possibilités :

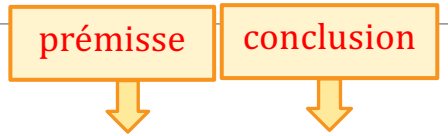
- A et B sont corrélés de manière tout à fait fortuite dans cet ensemble de données particulier
- A est un ré-étiquetage de A (ou *vice-versa*)
- A provoque B (ou *vice-versa*)
- certains autres attributs $C_1, \ldots, C_n$ (qui peuvent ne pas être disponibles dans les données) entraînent A et B
- etc.

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

Corrélation et causalité

Renseignement	Organisation
Les ventes de Pop-Tarts augmentent avant un ouragan	Walmart
Plus de criminalité, plus de trajets Uber	Uber
Le fait de taper avec des majuscules correctes indique la solvabilité.	Une jeune entreprise de services financiers
Les utilisateurs des navigateurs Chrome et Firefox sont de meilleurs employés	Une société de services professionnels dans le domaine des ressources humaines, sur les données des employés de Xerox et d'autres entreprises.
Les hommes qui sautent le petit-déjeuner souffrent davantage de maladies coronariennes	Chercheurs médicaux de l'Université de Harvard
Des employés plus engagés ont moins d'accidents	Coquille
Les personnes intelligentes aiment les frites frisées	Des chercheurs de l'Université de Cambridge et de Microsoft Research
Les ouragans à nom féminin sont plus meurtriers	Chercheurs universitaires
Statut plus élevé, moins poli	Des chercheurs étudient le comportement de Wikipédia

Définitions



Une **règle** est une déclaration de la forme "si X alors Y" construite à partir de n'importe quelle combinaison logique des attributs d'un ensemble de données.

Une règle **n'a pas besoin d'être vraie pour toutes les observations** – il peut y avoir des cas où la prémisse est satisfaite mais la conclusion ne l'est pas.

Certaines des "meilleures" règles sont celles qui ne sont exactes que 10 % du temps, vs. des règles qui ne sont exactes que 5 % du temps, mettons.

Cela dépend du contexte.

Définitions

Pour déterminer la force d'une règle, nous calculons diverses **mesures, t.q :**

- **le support** (la fréquence à laquelle une règle apparaît dans un ensemble de données) – une couverture faible indique une règle qui n'apparait que rarement
- **la confiance** (la fiabilité de la règle : quelle est la fréquence de la conclusion dans les données étant donné que les prémisses se sont produites) – les règles à haut niveau de confiance sont plus "vraies"
- **l'intérêt** (la différence entre sa confiance et la fréquence relative de sa conclusion) – les règles ayant un intérêt absolu élevé sont plus "intéressantes"
- **"lift"** (augmentation de la fréquence de la conclusion qui peut être expliquée par les prémisses) - avec un lift élevé ($\gg 1$), la conclusion est plus fréquente que prévue
- **conviction**, **confiance totale**, **effet de levier**, **force collective,** etc.

[**Note:** $\text{Freq}(A) \in \{0,1,\ldots,N\} =$ # d'observations pour lesquelles A est valide]

Définitions

Si N est le nombre d'observations dans un ensemble de données, alors :

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

Proportion de cas où la prémisse et la conclusion se retrouvent ensemble

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

Proportion de cas où la conclusion se produit lorsque la prémisse se produit

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

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

... ? !?

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

Interprétation du "lift" : 70% des personnes nées avant 1976 possèdent un exemplaire et 56% des personnes nées après 1976 en possèdent un, alors que 59% des individus en possèdent une copie.

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

Exemple

RM : si une personne est née avant 1976 (X), elle possède une copie du *Sergeant Peppers' Lonely Hearts Club Band* des Beatles, sous une forme ou une autre (Y).

Supposons que :

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

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

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

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

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

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

Interprétation des règles d'association

Tout cela semble indiquer que la règle RM n'est pas dépourvue de sens, mais dans quelle mesure? Il est **difficile de répondre à cette question** sans fournir des **seuils** ; mais l'évaluation d'une règle isolée **n'est pas utile**.

Il est recommandé de procéder à une **exploration préliminaire** de l'espace des règles d'association (avec expertise du domaine) afin de déterminer des seuils raisonnables pour la situation spécifique ; les règles candidates seraient alors écartées ou conservées en fonction de ces seuils métriques.

Pour ce faire, il faut pouvoir générer "facilement" des règles candidates potentielles.

Générer des règles d'association

Le véritable défi est de **générer** des règles candidates sans perdre de temps à générer des règles qui seront probablement rejetées.

Un **ensemble d'éléments** pour un ensemble de données est une liste d'attributs avec des valeurs. Un ensemble de **règles** peut être créé à partir de l'ensemble de données en ajoutant des blocs "**SI** ... **ALORS**" aux instances.

De $\{\text{membership} = \text{True}, \text{age} = \text{Youth}, \text{purchasing} = \text{Typical}\}$, nous obtenons :

- **SI** (purchasing = Typical AND membership = True) **ALORS** age = Youth
- **SI** age = Youth **ALORS** membership = True, etc.

n **éléments** $\Rightarrow$ $2^n - 1$ **règles** (explosion combinatoire)

Algorithme de force brute

1. Générer des ensembles d'éléments (de taille 1, 2, 3, 4, etc.).
2. Créer des règles à partir de chaque ensemble d'éléments.
3. Calculer le soutien, la confiance, l'intérêt, etc. pour chaque règle.
4. Ne conserver que les règles dont la couverture, la précision, l'intérêt, ou d'autres paramètres appropriés sont "suffisamment élevés".
5. Ces règles sont considérées comme **vraies** pour l'ensemble des données – il s'agit de **nouvelles connaissances dérivées des données**.

Algorithme a priori

L'explosion combinatoire est un problème - elle disqualifie l'approche par **force brute** pour tout ensemble de données réaliste.

Comment générer un petit nombre de règles candidates **prometteuses** ?

L'algorithme ***a priori*** est une tentative pour surmonter cette difficulté.

Initialement, il a été développé pour des **données transactionnelles** (c'est-à-dire les marchandises dans les colonnes, les achats de clients dans les lignes) ; tout ensemble de données raisonnable peut être transformé en un ensemble de données transactionnelles à l'aide de variables nominales.

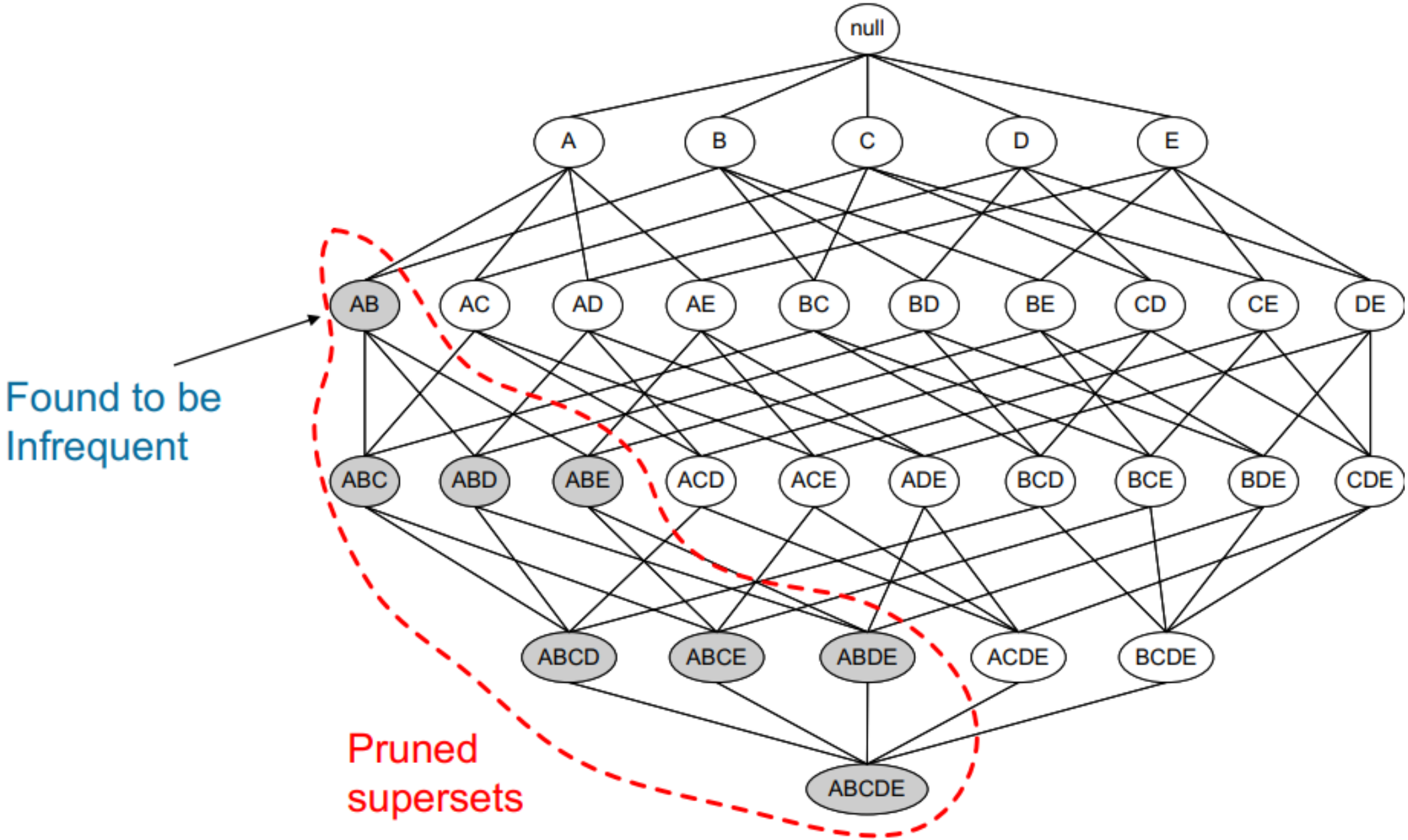
Algorithme a priori

L'algorithme a priori tente de trouver des **ensembles d'éléments fréquents** à partir desquels construire des règles candidates, au lieu de construire des règles à partir de **tous les** ensembles d'éléments possibles.

Il commence par identifier les **éléments individuels** fréquents dans la base de données et étend ceux qui sont retenus à des **supersets d'éléments** de plus en plus larges, qui ne sont eux-mêmes retenus que s'ils apparaissent **suffisamment souvent** dans les données.

Idée : "tous les sous-ensembles non vides d'un ensemble fréquent doivent également être fréquents" ou, “tous les sous-ensembles d'un ensemble peu fréquent doivent également être peu frequents”.

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



Algorithme a priori

L'algorithme termine lorsque plus aucune extension d'ensembles d'éléments n'est retenue, ce qui se produit toujours étant donné le nombre fini de niveaux dans les ensembles de données catégorielles

- **Points forts :** facile à mettre en œuvre et à paralléliser
- **Limitations :** lent, nécessite des analyses fréquentes, n'est pas idéal pour les itemsets peu fréquents et rares

Des algorithmes plus efficaces l'ont depuis supplanté dans la pratique :

- **Max-Miner** tente d'identifier les ensembles d'éléments fréquents sans les énumérer – il effectue des sauts dans l'espace de ces ensembles au lieu d'utiliser une approche ↑.
- **Eclat** est plus rapide et utilise la recherche en profondeur, mais nécessite une grande quantité de mémoire.

Validation

Quelle est la **fiabilité** des règles d'association ?

Quelle est la probabilité qu'ils se produisent entièrement **par hasard** ?

Quelle est leur **pertinence** ?

Peuvent-elles être généralisées au delà de l'ensemble de données, ou à de **nouvelles** données ?

La **découverte d'associations statistiquement valables** peut contribuer à réduire le risque de trouver des associations fallacieuses.

Validation

Nous terminons cette section par quelques commentaires :

- les règles fréquentes correspondent à des instances qui se répètent dans l'ensemble de données, les algorithmes qui génèrent des ensembles d'éléments tentent souvent de **maximiser la couverture** ; lorsque les **événements rares** sont plus significatifs, nous avons besoin d'algorithmes qui génèrent des cas rares – **ce n'est pas un problème trivial** ;
- les données continues doivent être **catégorisées** pour générer des règles ; comme il existe de nombreuses façons d'accomplir cette tâche, le même ensemble de données peut donner lieu à des règles complètement différentes – ce qui pourrait créer des **problèmes de crédibilité** auprès des clients et des parties prenantes ;
- autres algorithmes : AIS, SETM, aprioriTid, aprioriHybrid, PCY, Multistage, Multihash, etc.

Lectures conseillées

Concepts de règles d'association

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

Exercices

Concepts de règles d'association

1. Évaluez les règles candidates suivantes dans l'ensemble de données musicales :
   - si un individu possède un album de musique classique (W), il possède également un album de hip-hop (Z), étant donné que $\text{Freq}(W) = 2010$, $\text{Freq}(Z) = 6855$, $\text{Freq}(W \cap Z) = 132$.
   - si un individu possède à la fois un album des Beatles et un album de musique classique, il est né avant 1976, étant donné que $\text{Freq}(Y \cap W) = 1852$, $\text{Freq}(Y \cap W \cap X) = 1778$.

2. Parmi les 3 règles établies ($X \rightarrow Y$, $W \rightarrow Z$, $Y \,\&\, W \rightarrow X$), laquelle vous semble la plus utile ? Laquelle est la plus surprenante ?

Exercices

Concepts de règles d'association

3. Un magasin qui vend des accessoires pour téléphones cellulaires organise une promotion. Les clients qui achètent plusieurs produits parmi un choix de 6 couleurs différentes bénéficient d'une réduction. Les responsables, qui souhaitent savoir quelles couleurs seront achetées ensemble, ont collecté les achats dans le fichier Transactions.csv.

   Considérons les règles suivantes :

   - {**r**ouge, **b**lanc} ⇒ {**v**ert}
   - {**v**ert} ⇒ {**b**lanc}
   - {**r**ouge, **v**ert} ⇒ {**b**lanc}
   - {**v**ert} ⇒ {**r**ouge}
   - {**o**range} ⇒ {**r**ouge}
   - **{b**lanc, **n**oir} ⇒ {**j**aune}
   - {**n**oir} ⇒ {**v**ert}

Exercices

Concepts de règles d'association

3. (suite) Pour chaque règle, calculez le **soutien**, la **confiance**, l'**intérêt**, le **“lift”** et la **conviction**. Parmi les règles pour lesquelles le soutien est positif (> 0), quelle est celle qui a le plus grand “lift” ? La confiance ? L'intérêt ? Conviction ? Élaborez une dizaine de règles candidates supplémentaires et évaluez-les. Parmi les règles candidates, quelles sont celles qui, selon vous, seraient les plus utiles aux directeurs de magasin ? Comment déterminer des valeurs seuils raisonnables pour le soutien, la couverture, l'intérêt, la portée et la conviction des règles dérivées d'un ensemble de données donné ?

Exercices

Concepts de règles d'association

4. Passez en revue l'exemple de règles d'association pour l'ensemble de données du Titanic dans DUDADS (voir les lectures conseillées). Répétez le processus avec l'ensemble de données UniversalBank.csv (il se peut que vous deviez d'abord nettoyer et visualiser l'ensemble de données, ainsi que catégoriser les variables numériques ; pouvez-vous donner une idée raisonnable de ce que représente chacune des variables ?). Trouvez la "vraie connaissance" de l'ensemble de données sous la forme de règles d'association fiables et significatives (utilisez des métriques (mesures) si nécessaire).