Fusions de bases de données

Réunion d'Unité

Chloé Dimeglio

UMR 1027 équipe 5

18/06/2015

Outline

Plan

Projet Big Data - appel d'offre Région (T. Lang et N. Savy)

Problématiques connexes

Projet européen H2020-Lifepath, Recherche Clinique....

Partenaires

IMT-IRIT-IFERISS-University of North Carolina. Intérêts : IUCT-CHU

Tâches

1. Alignement de variables (C.Dimeglio)
2. Gestion des données manquantes (G.Guernec)
3. Intégration des données en langage naturel (IRIT)
4. Sensibilité d'une base de données (C.Dimeglio, B.Lepage)
5. Introduction de variables latentes (C.Dimeglio, B.Lepage)
6. Réflexion autour de la temporalité (équipe 5)
7. Aspects juridiques, éthiques, sociaux (équipe 4)

Au sujet du Big Data

"Le Big Data c'est comme le sexe à l'adolescence :

- Tout le monde en parle,
- Personne ne sait vraiment comment ça marche,
- Tout le monde pense que les autres le font donc tout le monde prétend le faire,
- Les seuls qui n'en parlent pas sont ceux qui l'ont déjà fait car leur première fois ne s'est pas très bien passée"

Dan Ariely

Ce que ça VISE

La question n'est pas celle du volume d'informations mais plutôt de la **quantité d'information utile** à tirer de ces données et des moyens à mettre en oeuvre pour dégager cette information de qualité.

Définition

La fusion de données vise à l'**association**, la **combinaison**, l'**intégration** et le mélange de multiples sources de données représentant des **connaissances et des informations diverses** dans le but de fournir une **meilleure décision** par rapport à l'utilisation séparée des sources de données.

Exemple

Base A		
Sexe	Age	Activité
M	30	1
M	65	0
M	63	1
F	15	0
M	3	0
F	43	1

Base B		
Sexe	Age	Activité
M	32	5
F	28	4
F	46	8
M	68	7
M	8	8
M	11	8

- Deux bases A et B, une même variable Activité codée de deux façons différentes dans chacune des bases.
- Des covariables liées à la variable d'intérêt et communes aux deux bases.

Contexte de l'étude : exemple

Exemple : données longitudinales

- Un changement de questionnaire pour une même étude
- Une même variable recueillie sur les mêmes individus mais pas sur la même échelle au cours du temps –> **fusion de données longitudinales**

Problématique : Comment reconstituer la cohorte ?

Exemple : données transversales

- Un questionnaire différent pour recueillir la même information sur des études différentes
- Une même variable recueillie sur des individus différents à un instant donné du temps –> **fusion de données transversales**

Problématique : Comment considérer l'intégralité de l'information ?

Contexte de l'étude : fusion de bases de données

Méthodes classiques

- Méthodes d'estimation bayésienne
- Cartes d'évidence
- Modèles de Markov cachés
- Modèles graphiques probabilistes
- Technique des moindres carrés

Xu,L., Krzyzak, A. et Suen, C. (1992) : Méthods of combinig multiple classifiers and their application to handwriting recognition

Moravec, H. (1987) : Sensor fusion in certainty grids for mobile robots

Rabiner, L. (1989) :A tutorial on hidden Markov models and selected applications in speech recognition

Pearl, J. (1988) : Probabilistic reasoning in intelligent systems

Abidi, M et Gonzalez, R (1992) : Data fusion in robotics and machine intelligence

Méthode à développer

Fusionner les données sur la base d'une variable commune par du **transport de mesure**.

$$\text{Activité}_{\text{Base A}} \overset{Transport}{\rightarrow} \text{Activité}_{\text{Base B}}$$

Ambrosio, L., Brenier, Y., Buttazzo, G., Caffarelli, L., Evans, L.C., Pratelli, A. et Villani, C. (2001) : Optimal transportation and applications

Villani, C. (2012) : Topics in optimal transportation

Plan

Pré-requis

Cadre

Soit A et B deux bases.
On note X et Y la variable commune d'intérêt codée de deux façons différentes sur les deux bases, telles que :

X	x1	x2	...
P(X=xi)	a1	a2	...
Y	y1	y2	...
P(Y=yj)	b1	b2	...

On note $cov(X)$ et $cov(Y)$ les covariables d'intérêt, communes aux deux bases, associées à X et Y.

Transporter les mesures

Idée générale

On suppose que deux mesures ν et μ sont associées aux distributions des deux variables X et Y.
On cherche l'application optimale T telle que $\nu = T\mu$

Cas continu

L'application T est **unique** et vérifie donc l'optimalité de la solution.

Cas discret

Toutes les applications T telles que $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ sont solutions, caractérisées par les matrices de transfert de la base A vers la base B

Conséquence

Nécessité de déterminer une fonction de "coût" de passage d'une échelle à l'autre pour définir l'optimalité de l'application.

Plan

Unicité du changement de variable

Si μ et ν ont des densités f et g par rapport à la mesure de Lebesgue, et si T est injective,

$$f(x) = g(T(x))\,|\det(\mathrm{DT}(x))|$$

avec DT l'application jacobienne de T.

Exemple

Soit $X \sim \mathcal{N}(\mu_1, \sigma_1)$ et $Y \sim \mathcal{N}(\mu_2, \sigma_2)$.
Soit $\hat{\mu}_1$ l'estimation de μ_1 sur la base A etc...
On peut facilement vérifier l'égalité de transport suivante :

$$X = (Y - \hat{\mu}_2)\frac{\hat{\sigma}_1}{\hat{\sigma}_2} + \hat{\mu}_1$$

Dans le cas continu, on vérifie **l'existence et l'unicité** de l'application de transport.

Plan

Modélisation

Mesures et matrice de permutation

• Soit $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ la mesure associée à la base A et $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ celle associée à la base B.

• Les plans de transfert sont alors les matrices de permutation γ telles que :

$$\gamma = \sum_{i,j} \gamma_{i,j} \delta_{(x_i, y_j)}$$

Où :

$$\sum_{j} \gamma_{i,j} = a_i$$

et

$$\sum_{i} \gamma_{i,j} = b_j$$

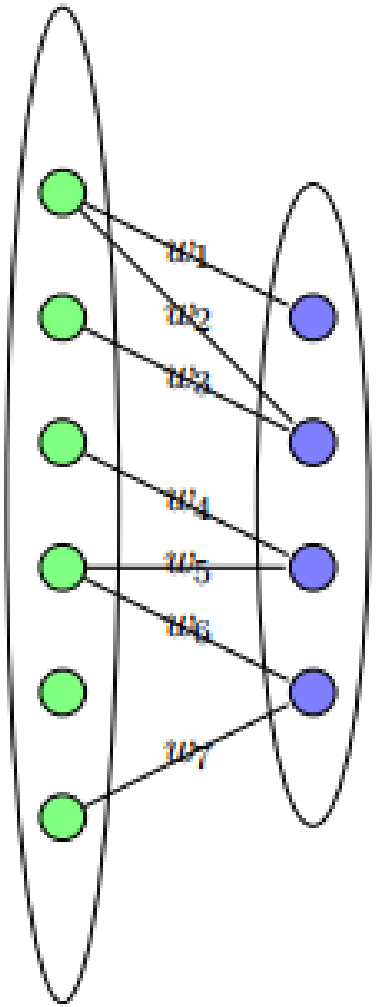
Modélisation

Introduction à la fonction de coût

- Fonction de coût $c(\gamma)$ = **risque de passage** d'une échelle à l'autre.
- Fonction définie à partir d'une distance $c(cov(x_i), cov(y_j))$ associées aux **distributions des covariables.**

$$c(\gamma) = \sum_{i,j} \gamma_{i,j} c(cov(x_i), cov(y_j))$$

Fonction de risque associée au transport



Définition du risque

- Plus la distribution des covariables sur la base A sera éloignée de la distribution des covariables sur la base B plus grand sera le risque.
- Le risque est défini à partir des **écarts d'entropie des distributions de certaines co-variables d'intérêt**.

Il s'agira de minimiser ce risque.

Modélisation

Fonction de coût

Soit K le nombre de covariables associées à la variable d'intérêt.
Soit S le nombre de modalités prises par chaque covariable.
On définit la fonction de coût par :

$$c(\gamma) = \sum_{k=0}^{K} \sum_{i} \sum_{j} \sum_{s=0}^{S} \gamma_{i,j} \left| p_{i,s}^{k} \ln p_{i,s}^{k} - q_{j,s}^{k} \ln q_{j,s}^{k} \right|$$

Où $p_{i,s}^{k} = \mathbb{P}(\mathrm{Cov}_k X = a_s | x_i)$ et $q_{j,s}^{k} = \mathbb{P}(\mathrm{Cov}_k Y = b_s | y_j)$
avec la convention $p \ln(p) = 0$ si $p = 0$.

Modélisation

Transport optimal

Soit $i \in [1, n]$. Soit $c^A(i)$ la classe de l'élément i sur la base A.
Etant donnée une classe c_1 de la base A et d_1 une classe de la base B, $N(c_1, d_1)$ donne le **nombre de transitions possibles** de la classe c_1 vers la classe d_1 par le transport optimal.

$$N(c_1, d_1) = \sum_{i=1}^{n} \mathbb{1}_{\left[c^A(i)=c_1,\, T_{\text{opt}}(c^A(i))=d_1\right]}$$

Modélisation

Règle d'allocation individuelle

Pour tout i fixé, $c(\hat{i})$ est la classe affectée par estimation à l'individu i. On définit :

$$V^k(i) = \left\{ d_1 \mid \sum_{i=1}^{k} \mathbb{1}_{\left[c^A(i)=c_1, c(\hat{i})=d_1\right]} \leq N(c_1, d_1) \right\}$$

On a finalement :

$$\text{Ind} = \underset{j \mid c^B(j) \in V^k(i)}{\arg\min} \; d(\text{Cov}(j), \text{Cov}(i))$$

Et

$$c(\hat{i}) = c^B(\text{Ind})$$

Plan

Cas d'application

Contexte

- On s'intéresse à la catégorie de revenus d'un individu.
- Dans la base A elle est évaluée par l'individu lui même sur une échelle de 1 à 2.
- Dans la base B elle est évaluée sur une échelle de 1 à 3.

Répartition des données

- Sur la base A, 3 individus se sont évalués comme appartenant à la classe 1, et 5 à la classe 2.
- Concernant la base B, 4 individus se sont évalués en classe 1, 2 en classe 2 et 2 en classe 3.

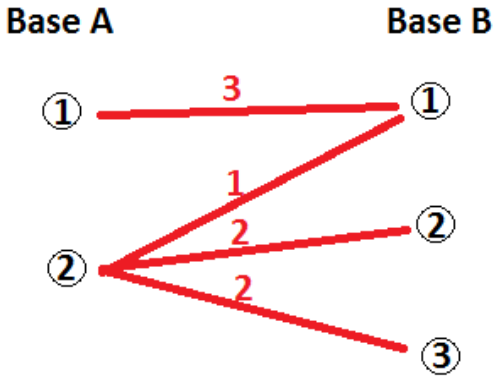
Cas d'application

Solution admissible

Une application permettant de transporter la distribution de la variable de la base A à la base B satisfait la matrice de transfert suivante :

$$\gamma = \begin{pmatrix} 3 & 0 & 0 \\ 1 & 2 & 2 \end{pmatrix}$$

Graphe correspondant



Cas d'application

Solution admissible

La matrice suivante est également solution : $\gamma = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & 1 \end{pmatrix}$

Question

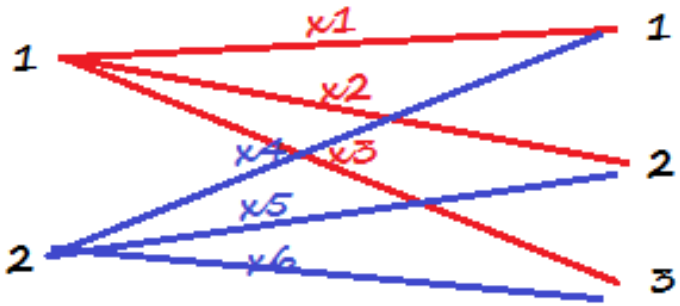
Comment déterminer un transfert optimal ?

Résolution

Flot de coût minimum

- On cherche $\text{argmin}_{i,j} c(\gamma)$ sous les contraintes $Ax = b$
- Avec $A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$, $b = \begin{pmatrix} 3 \\ 5 \\ 4 \\ 2 \\ 2 \end{pmatrix}$ et $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}$

Illustration



Résultats

- Lorsque la variable d'intérêt est **entièrement déterminée par les covariables**, nous obtenons une parfaite adéquation entre la prévision et la "vérité".
- Attention toutefois, cette situation ne reflète pas la réalité clinique.
- Ce premier travail fait l'objet d'un article en cours d'écriture.

Outline

Recommandations

- Ce travail est basé sur **une hypothèse forte** : le comportement des deux populations est similaire pour la variable d'intérêt.
Si on force l'association, on biaise l'information associée.
- Transporter une distribution sur une autre suppose que l'on considère une population comme référence sur la variable d'intérêt. Attention à la représentativité de cette population, peut-on réellement la considérer comme de référence ?

Perspectives

- Tester la validité de l'approche lorsqu'on introduit un alea dans la détermination des variables par les covariables.
- Tester la validité de l'approche lorsqu'on introduit des données manquantes dans la base (cf Gregory)
- Introduire des données en langage naturel (cf partenariat avec l'IRIT)

merci

Lot 1 : Alignements de variables.

- Objectif :
  - Fusion de bases de données
  - Problème : alignement de variables
    variables concernant le même objet soient codées de façon différente dans les 2 bases
  - Lien avec la question du transport de mesures

- Partenaires :
  - IMT (Nicolas SAVY - Sébastien DEJEAN - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO)
  - IRIT (Mohand BENGHANEM - Mathieu SERRURIER - Nathalie AUSSENAC-GILLES)

Lot 2 : Gestion des données manquantes.

- Objectif :
  - Parfaire les connaissances sur l'Imputation multiple
  - Identifier des méthodes alternatives
    notamment dans le cadre complexe des données MNAR
  - Quantifier l'impact des données manquantes sur une réponse

- Partenaires :
  - IMT (Nicolas SAVY - Sébastien DEJEAN - Laurent RISSER - Cécile CHOUQUET - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)
  - IRIT (Mohand BENGHANEM - Mathieu SERRURIER - Nathalie AUSSENAC-GILLES)

Lot 3 : Intégration de données en langage naturel.

- Objectif :
  - Intégrer l'information contenue dans les messages textuels
    - expériences individuelles
    - étude de cas
    - entretiens
  - traitement du langage naturel
  - codage de cette information

- Partenaires :
  - IMT (Nicolas SAVY - Sébastien DEJEAN - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)
  - IRIT (Mohand BENGHANEM - Mathieu SERRURIER - Nathalie AUSSENAC-GILLES)

Lot 4 : Sensibilité d'une base de données.

- Objectif :
  - Validitation des résultats issus de
    - fusion de bases
    - codage de l'information textuelle
    - gestion des données manquantes
  - Approche bayésienne :
    - Simuler des base de données en générant les variables selon leurs lois
    - Analyser la situation
    - Regarder la distribution des résultats

- Partenaires :
  - IMT (Nicolas SAVY - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)

Lot 5 : Introduction de variables latentes.

- Objectif :
  - Un état de santé n'est pas forcément observé directement mais reflété par différentes mesures
  - Meilleure gestion de l'erreur de déclaration (voulue ou pas)
  - Meilleure gestion de l'enchainement des variables observée en passant par la "chaînes" des états latents
  - raisonnement poussé jusqu'à des variables latente comme l'état de santé vrai ou la qualité de vie vraie (si tant est que ces variables aient un sens)
  - Statistique : approche de l'estimation du style EM (markov caché)

- Partenaires :
  - IMT (Nicolas SAVY - Jérome DUPUIS - Laurent RISSER - X X)
  - INSERM Unité 1027 (Chloé DIMEGLIO - Benoit LEPAGE)

Lot 6 : Réflexion autour de la temporalité.

- Objectif :
  - Reconstitution de cohorte biographiques à partir de différentes sources de données
  - Les cohortes de naissance sont rares

- Partenaires :
  - IMT (Nicolas SAVY)
  - INSERM Unité 1027 (Thierry LANG, Cyrille DELPIERRE, Michelle KELLY-IRVING, Sébastien LAMY - Chloé DIMEGLIO - Benoit LEPAGE)
  - Autres equipes de l'INSERM
  - CHU

Lot 7 : Aspects éthique, juridiques et sociaux.

- Objectif :
  - Confidentialité des données
  - Anonymisation des données
  - Conséquences sociales et sanitaires de ces développements du croisement de Bases de Données

- Partenaires :
  - INSERM Unité 1027 (Thierry LANG, Emmanuelle RIAL-SEBAG)
  - IFERISS (laboratoire de Droit d'UT1, Sociologie,..)