MedInfo2019Lyon **Satellite Symposium AIM**

Association Française Informatique Médicale
AIM

DataChallenge Record Linkage

Présentation

Si vous êtes prêt à relever un défi de « Record Linkage », alors AIM vous propose de participer à un DataChallenge ! C'est l'occasion de rejoindre d'autres équipes pour une compétition sur terrain « neutre » et saisir l'opportunité de présenter votre solution pendant le congrès mondial MedInfo2019 en août prochain à Lyon.

Le « DataChallenge Record Linkage » propose de réaliser une tâche de Record Linkage (chaînage ou résolution d'entité) entre 2 fichiers d'enregistrements.

- la particularité de cette tâche tient à l'absence de clé d'appariement, le chaînage ne pouvant s'appuyer que sur la combinaison de traits d'identification,
- sa complexité tient au fait que, comme dans la réalité, ces traits d'identification sont susceptibles d'être imparfaits, dégradés (pouvant par exemple contenir des erreurs typographiques de transcription), voire d'être partiellement manquants.

Le défi consiste à proposer un algorithme complétement automatisé pour optimiser la concaténation de 2 bases (A et B) en proposant d'aller jusqu'au bout du processus, à savoir fournir une solution contenant les décisions définitives de chainage.

En collaboration avec d'autres partenaires, AIM vous offre la possibilité de vous mettre à l'épreuve sur des données synthétiques (libres de droit) mises à la disposition de tous.

Les règles sont les suivantes :

- la participation est autorisée pour une équipe composée de 1 à 4 concurrents maximum,
- après l'inscription (cf. date limite d'inscription sur l'échéancier), chaque équipe est destinataire des données d'entraînement (contenant les 2 bases d'enregistrements à chainer avec une clé globale authentique, une clé spécifique à chaque base, et l'ensemble des traits d'identification, i.e. Nom, NomDeNaissance, Prenom, Sexe, DateDeNaissance),
- dans le but d'améliorer les performances de la solution proposée, l'ajout de données d'entraînement supplémentaires reproduisant les principes détaillés dans la description des méthodes de dégradation des données d'entraînement est autorisé, à la seule condition qu'elles soient libres de droits et/ou sans limitation réglementaire d'accès car elles sont susceptibles d'être exigées par le jury,
- chaque équipe soumet sa solution dans la limite de la période de soumission (cf. date limite de soumission sur l'échéancier) sous la forme d'un abstract précisant le ou les paradigmes méthodologiques utilisés (déterministe, probabiliste, similarités/distances, machine/deep learning, …), d'un fichier de chaînage contenant la solution proposée (cf. modèle infra), du code source reproductible et d'un notebook commenté, l'explicitation de la solution ainsi que son caractère opérationnel étant pris en compte dans l'évaluation du jury.

L'équipe gagnante aura l'opportunité de présenter sa solution pendant le congrès mondial d'informatique médicale MedInfo à Lyon en août 2019 et se verra décerner un prix par le(la) président(e) du jury durant le symposium satellite dédié au « DataChallenge Record Linkage ».

Au plaisir de vous accueillir au « DataChallenge Record Linkage » en vous souhaitant bonne chance !

Détail de la première édition 2019

La présentation du palmarès et de l'équipe gagnante aura lieu en présence du jury durant le congrès MedInfo2019 (du 26 au 30 août 2019 à Lyon).

1. Jeu de données, évaluation des résultats algorithmes/systèmes

Le but de ce challenge est de proposer les décisions de chainage les plus conformes à la réalité des entités contenues dans les données. Le matériel fourni et le principe sont les suivants :

- les données (DRCL_TrainingSet.csv) et leur description ((DRCL_VarDictionnary.txt)
- les principes de dégradation des données :
  - insertion de caractère(s) : ROBERTSON dégradé en ROBERTESON
  - omission de caractère(s) : ROBERTSON dégradé en ROBERTON
  - inversion de caractère(s) : ROBERTSON dégradé en ROBERSTON
  - toute autre erreur typographique, incluant la suppression du champs entie

2. Participation et principes du classement final

Au début de la période de soumission des solutions (cf. échéancier), chaque correspondant d'équipe sera destinataire du fichier test (DRCL_ChallengeSet.csv) à traiter et partir duquel les solutions seront générées par l'algorithme.

Les prérequis à d'évaluation sont les suivants :

- la solution doit obligatoirement aller au bout du processus de décision de chainage aboutissant pour chaque couple d'enregistrement traité à un variable binaire Match = 0 ou 1,
- la solution jointe (DRCL_Solution_[Team].csv) sera limitée à la liste des seuls couples d'enregistrements à chainer (Match = 1),
- la solution doit être accompagnée du code opérationnel qui la génère dans le langage choisi par l'équipe (C, C++, R, Python, SAS,…).

Les critères d'évaluation sont les suivants :

- critère principal : la conformité entre la réalité et la solution de chainage sera mesurée à l'aide du score F1
  - $\text{F1 score} = 2 \cdot \frac{\Pr(\text{Same}=1|\text{Match}=1)\cdot\Pr(\text{Match}=1|\text{Same}=1)}{\Pr(\text{Same}=1|\text{Match}=1)+\Pr(\text{Match}=1|\text{Same}=1)}$
  - où « Same = 1 » désigne un couple d'enregistrements identiques,
  - et « Match = 1 » désigne un couple apparié par la solution proposée.
- critère secondaire : en cas d'ex-aequo sur le score F1, les solutions seront départagées en fonction d'un score Fβ (à la discrétion du jury) et du temps de calcul des solutions de chainage (dans l'idéal, inclure une fonction d'affichage du temps CPU dans le code de la solution).

3. **A titre d'exemple**

- fichier « d'entrainement » (DCRL_TrainingSet)

GlobalKey	Data	DataKey	LastName	BirthName	FirstName	Gender	BirthDate
1	A	A1	Couture	Boivin	Flore	F	15/01/1931
1	A	A2	Couture	Boivin	Fleur	F	15/01/1931
2	A	A3	Mailhot		Lionel	M	14/08/1976
3	A	A4	Gagné		Orson	M	05/03/1946
4	A	A5	Lefèbvre		Roger	M	07/05/1983
5	A	A6	Lessard		Fanchon	F	16/09/2018
5	A	A7	Lesard		Fanchon	F	17/09/2018
6	A	A8	Marseau		Gil.	M	01/04/1966
7	A	A9	Renaud	Morneau	Claude	F	03/10/1965
8	A	A10	Rousseau	Coulombe	Jessamine	F	06/05/1963
4	B	B1	Lefevre		Roger	M	07/05/1983
5	B	B2	Lessard		Fanchon	F	17/09/2018
5	B	B3	Lessard		Fanchon		17/09/2018
6	B	B4	Marceau		Gilberte	M	01/04/1966
7	B	B5	Renaud	Mornot	Claude	M	03/10/1965
8	B	B6	Coulombe	Rousseau	Jasmine	F	06/05/1963
9	B	B7	Rousseau		Timothée	M	07/09/1960
10	B	B8	Veronneau		Albracca	F	20/09/1980
11	B	B9	Tisserand	Allain	Estelle	F	16/02/1949
12	B	B10	Dastous		André	M	02/12/1982
13	B	B11	Pouchard		Paulette	F	22/09/2011
14	B	B12	Rochefort		Moore	M	20/03/1991
14	B	B13	Rochefort		Moure	M	20/03/1991
15	B	B14	Lussier		Jeoffroi	M	15/09/1935

Tableau 1 : Format du fichier de données d'entraînement

La variable GlobalKey identifie la réalité des entités contenues dans les fichiers. Cette variable sera absente du fichier test.

- format de la solution (DRCL_Solution_TeamAIM)

Solution		
Key1	Key2	Match
A5	B1	1
A6	B2	1
A6	B3	1
A8	B4	1
A9	B5	1
A10	B6	1
B2	B3	1
A10	B7	1
B6	B7	1

Tableau 2 : Format d'une solution du TeamAIM

La variable Match pourra éventuellement être occultée, le fichier ne devant contenir que les couples pour lesquels la solution proposée par l'équipe aboutit à la décision de chainage.

- calcul du F1 Score

Le score F1 est calculé sur la base de la confrontation entre la réalité et la solution proposée, comme suit :

Key1	Key2	Same	Match
A1	A2	1	0
A5	B1	1	1
A6	A7	1	0
A6	B2	1	1
A6	B3	1	1
A7	B2	1	0
A7	B3	1	0
A8	B4	1	1
A9	B5	1	1
A10	B6	1	1
B2	B3	1	1
B12	B13	1	0
A10	B7	0	1
B6	B7	0	1

		Same		
		1	0	
Match	1	7	2	9
	0	5	262	267
		12	264	276

Tableau 3 : Confrontation entre réalité et solution

Le score F1 serait ici environ égal à 0.667.

4. Règlement et inscription

Pour chaque équipe, l'inscription de l'un au moins de ses membres au congrès MedInfo2019Lyon valide sa participation au « DataChallenge Record Linkage ». Le palmarès du « DataChallenge Record Linkage » sera établi parmi les équipes représentées à cet évènement satellite.

Chaque équipe choisit un nom et un représentant (membre correspondant) et renseigne le formulaire d'inscription (cf. formulaire d'inscription). L'envoi des données est conditionné par le renseignement complet du formulaire d'inscription.

5. Dates importantes

Le « DataChallenge Record Linkage » se déroulera selon l'échéancier suivant :

01/03/2019	Début des inscriptions
01/03/2019	Envoi du jeu de données d'entraînement
30/04/2019	Clôture des inscriptions
02/05/2019	Envoi du jeu de données « DataChallenge Record Linkage »
02/05/2019	Début des soumissions
30/06/2019	Fin des soumissions
31/07/2019	Publication des résultats du « DataChallenge Record Linkage »
26--30/08/2019	Présentation pendant MedInfo2019Lyon

6. Jury

Le jury sera composé :

- de membres du CA AIM présents à MedInfo,
- de membres de la communauté Biostatistique,
- d'experts internationaux en Record Linkage

7. Formulaire d'inscription

Le formulaire d'inscription est à renvoyer complétement renseigné (avec un nom d'équipe et au moins un membre correspondant) avant la fin de la période d'inscription (cf. échéancier) à l'adresse lemlih.ouchchane@uca.fr.

"DataChallenge Record Linkage" Registration Form

Team Name	:	
Corresponding Member		Member 1
FirstName	:	
LastName	:	
Email	:	
Country	:	
Institutions	:	
Attending MedInfo2019Lyon (Y/N)	:	Y
TeamMate 1		Member 2
FirstName	:	
LastName	:	
Email	:	
Country	:	
Institutions	:	
Attending MedInfo2019Lyon (Y/N)	:	
TeamMate 2		Member 3
FirstName	:	
LastName	:	
Email	:	
Country	:	
Institutions	:	
Attending MedInfo2019Lyon (Y/N)	:	
TeamMate 3		Member 4
FirstName	:	
LastName	:	
Email	:	
Country	:	
Institutions	:	
Attending MedInfo2019Lyon (Y/N)	:	

Comments :