Progetto Architetture Dati: Record Linkage

Christian Bernasconi 816423
Marco Ripamonti 806785

Appello 23/09/2020

Contesto e obiettivi

Contesto:

- FEIII* Challenge 2016

Obiettivo:

- Implementazione di un processo di Record Linkage
- Valutazione rispetto a diverse strategie e configurazioni

* Financial Entity Identification and Information Integration

Datasets e Ground Truths

Datasets:

- FFIEC: Federal Financial Institution Examination Council
- LEI: Legal Entity Identifiers
- SEC: Securities and Exchange Commission

Ground Truths:

- Matching entities tra FFIEC e LEI
- Matching entities tra FFIEC e SEC

Datasets: FFIEC

Contiene informazioni su banche e istituzioni finanziarie regolate da agenzie affiliate al **F**ederal **F**inancial **I**nstitution **E**xamination **C**ouncil.

Caratteristiche:

- 6,652 records
- 14 attributi
- IDRSSD, Financial Institution Name Cleaned, Financial Institution Zip Code 5, ...

Datasets: LEI

Contiene **L**egal **E**ntity **I**dentifiers per una vasta gamma di istituzioni finanziarie.

Caratteristiche:

- 53,958 records
- 39 attributi
- LEI, LegalNameCleaned, LegalAddress_Line_Cleaned, ...

Datasets: SEC

Contiene informazioni sulle entità registrate con **S**ecurities and **E**xchange **C**ommission.

Caratteristiche:

- 129,312 records
- 24 attributi
- CIK, CONFORMED_NAME, B_STREET, B_CITY, ...

Ground Truths

Costruite partendo dalle coppie individuate da un algoritmo baseline e valutate da esperti.

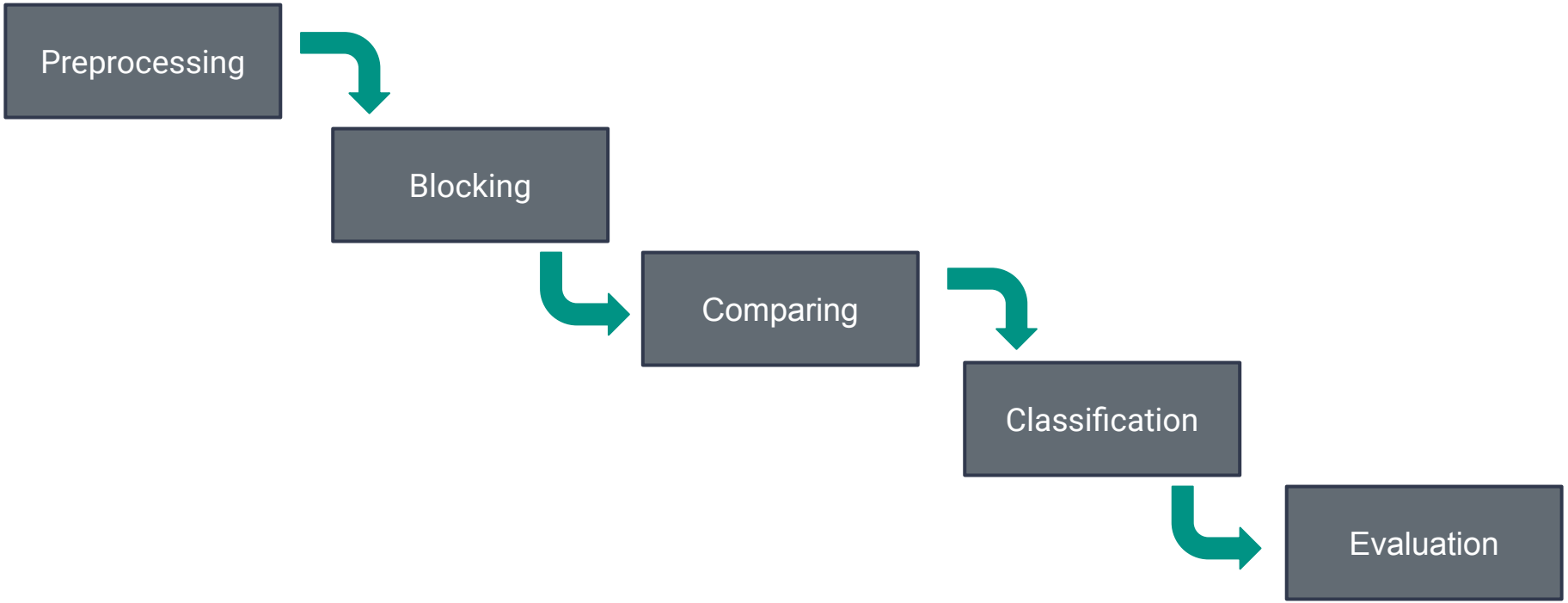
FFIEC - LEI:

- 496 Matching Pairs
- 932 Not Matching Pairs
- 38 Ambiguous Pairs

FFIEC - SEC:

- 230 Matching Pairs
- 560 Not Matching Pairs
- 95 Ambiguous Pairs

Workflow


Preprocessing
Blocking
Comparing
Classification
Evaluation

Preprocessing

Attributi comuni:

- Nome
- Indirizzo
- Città
- Stato
- Zip Code

La funzione di preprocessing è configurabile rispetto agli attributi che si intendono processare

Preprocessing

Attributi comuni:

- **Nome**
- Indirizzo
- Città
- Stato
- Zip Code

Operazioni effettuate:

- Lower-Case
- Rimozione stopwords e punteggiatura
- Sostituzione ‘&’ con ‘and’
- Rimozione modifiers

Full Name	Abbr.
corporation	corp
company	co
federal savings bank	fsb
national association	na
incorporated	inc
trust association	ta

Preprocessing

Attributi comuni:

- Nome
- **Indirizzo**
- Città
- Stato
- Zip Code

Operazioni effettuate:

- Lower-Case
- Rimozione info P.O. BOX
- Uniformazione orientamenti (1)
- Uniformazione tipi strade (2)

(1)

Full Name	Abbr.
north	n
south	s
east	e
west	w
northeast	ne
northwest	nw
southeast	se
southwest	sw

(2)

Full Name	Abbr.
street	st
road	rd
avenue	ave
square	sq
lane	la
suite	su
plaza	pl

Preprocessing

Attributi comuni:

- Nome
- Indirizzo
- **Città**
- **Stato**
- Zip Code

Operazioni effettuate:

- Lower-Case

Preprocessing

Attributi comuni:

- Nome
- Indirizzo
- Città
- Stato
- **Zip Code**

Operazioni effettuate:

- Normalizzazione formato da 5+4 (formato esteso) a 5 cifre
- Normalizzazione da <5 a 5 cifre (zeri iniziali)

Blocking

Full Indexing:

- Considera tutte le possibili coppie
- Aumenta costo fase di compare

- Non è stato applicabile (Out of Memory error)

Exact Blocking:

- Considera solo coppie con attributi selezionati identici
- Efficace quando ci sono pochi errori di spelling

- Fatto sul nome

Sorted Neighbourhood:

- Considera coppie in accordo con una sorting key tenendo conto del vicinato
- Utile quando ci sono errori di spelling

- Fatto sul nome e con vicinato variabile

Comparing

Confronto degli attributi di coppie di record con diverse strategie e metriche:

Esempi: FFIEC → LEI

- Financial Institution Address - LegalAddressLineCleaned

Similarità tra stringhe con metrica Jaro-Winkler e threshold 75%

Comparing

Confronto degli attributi di coppie di record con diverse strategie e metriche:

Esempi: FFIEC → LEI

- Financial Institution Zip Code 5 - LegalAddressPostalCode5

Confronto esatto tra stringhe

Comparing

DET

Address	City	State	Zip Code
1	1	0	0
0	1	1	1
1	0	0	0

ML

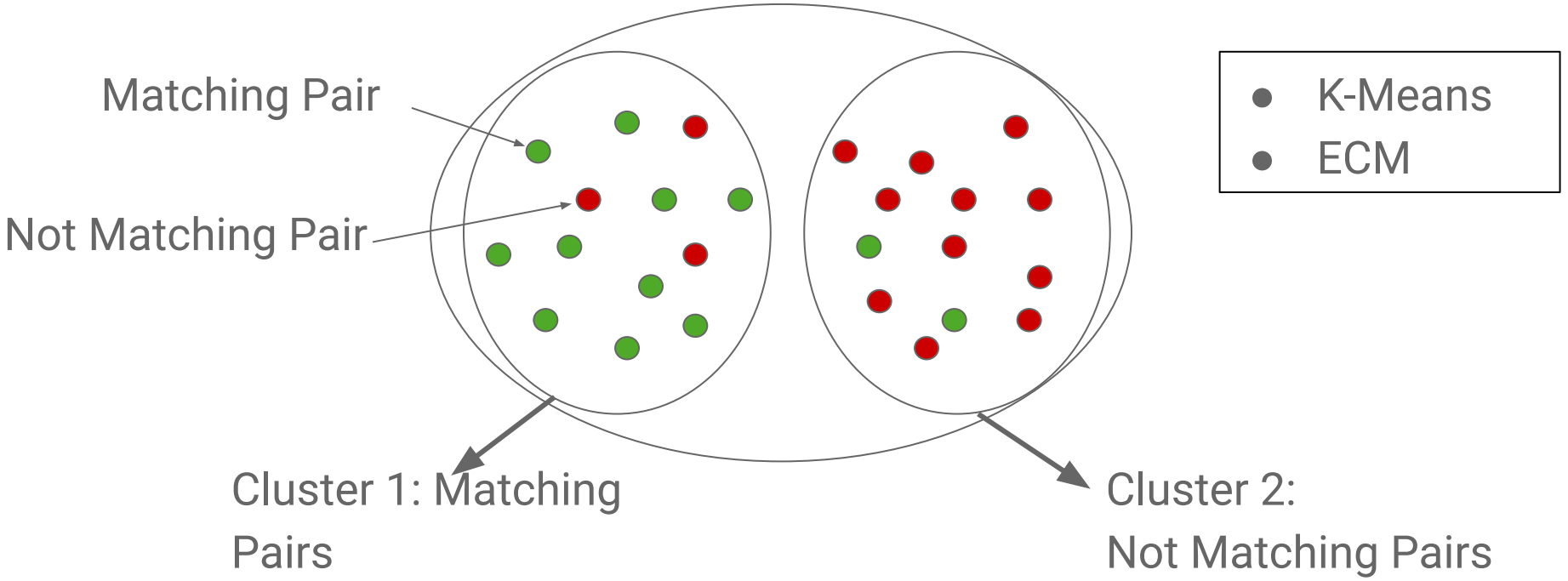
Address	City	State	Zip Code
0.87	1	0	0
0.41	1	1	1
0.77	0	0	0

Classification: Deterministic

Address + City + State + Zip Code > 2 → Match

ID_1	ID_2	Address	City	State	Zip Code
4635743	321563	1	1	0	0
632132	135476	0	1	1	1
3231365	123546	1	0	0	0
953213	654899	1	0	1	1

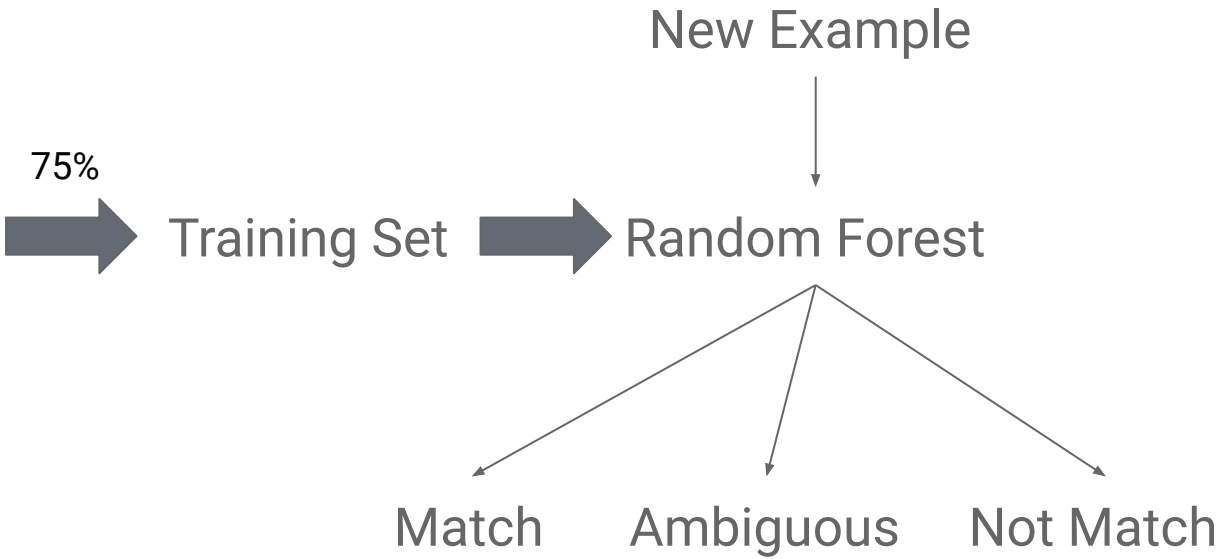
Classification: ML Unsupervised



Classification: ML Supervised

Ground Truth bilanciata

14579	987451	Match
...	...	...
15646	354964	Ambiguous
...	...	...
67989	975522	Not Match
...	...	...



Evaluation: Risultati

Risultati per la configurazione migliore (preprocessing completo)

Datasets	Approccio	Precision	Recall	F-Score
FFIEC-LEI	DET	0,975	0,955	0,965
FFIEC-LEI	ML Unsup.	0,969	0,963	0,966
FFIEC-SEC	DET	0,774	0,700	0,735
FFIEC-SEC	ML Unsup.	0,620	0,752	0,679
FFIEC-SEC	ML Sup.	0,591	0,591	0,580

Evaluation: Considerazioni Ground Truths

Per individuare ulteriori possibili matches sono stati selezionati i candidati attraverso la strategia di Sorted Neighbourhood Blocking.

FFIEC-LEI

Name	Address	City	State	Zip Code
OLD NATIONAL BANK	1 MAIN STREET	EVANSVILLE	IN	47708
Old National Bancorp	1 Main Street	Evansville	IN	47708
NORTHERN TRUST COMPANY	50 SOUTH LA SALLE STREET	CHICAGO	IL	60603
Northern Trust Collective Canada Index Fund - Non-Lending	50 South Lasalle Street	Chicago	IL	60603

FFIEC-SEC

Name	Address	City	State	Zip Code
SKAGIT BANK	301 EAST FAIRHAVEN AVENUE	BURLINGTON	WA	98233
SKAGIT STATE BANCORP, INC.	301 E. FAIRHAVEN AVENUE	BURLINGTON	WA	98233
SUNNYSIDE FEDERAL SAVINGS AND LOAN ASSOCIATION OF IRVINGTON	56 MAIN ST	IRVINGTON	NY	10533
SUNNYSIDE BANCORP, INC.	56 MAIN STREET	IRVINGTON	NY	10533

Fine