12ª conferenza nazionale di statistica

PIÙ FORZA AI DATI
UN VALORE PER IL PAESE

22 GIUGNO 2016
11.15 | 12.45

INNOVAZIONI E SPERIMENTAZIONI

Nuove prospettive per la statistica ufficiale: metodi bayesiani per la stima dei flussi demografici

Brunero Liseo | Sapienza Università di Roma

Metodi bayesiani vs. metodi classici

Non esistono *quantità incognite* ma solo diversi gradi di conoscenza di un fenomeno.
L'uso di dati – di qualunque tipo - permette di aggiornare le valutazioni complessive intorno ad un fenomeno di interesse mediante specifiche regole di "*updating*"
Se disponibili, è possibile utilizzare *informazioni a priori* – extra sperimentali – all'interno del processo di apprendimento.

Metodi bayesiani in statistica ufficiale

non sono più soggettivi dei metodi classici. …"*Nonetheless, the principal challenge to Bayesian methods that remains is the need to constantly rebut the notion that frequentist methods are "objective" and thus are more appropriate for use in the public domain"*
S. Fienberg Stat. Sci, 2011, 26, 212–226.
più adatti a estrarre valida informazione dai dati a nostra disposizione.
Possono fornire una quantificazione dell'incertezza più precisa (ad esempio, nei modelli per piccole aree, a proposito dell'errore standard delle stime)

Metodi bayesiani in statistica ufficiale

più adatti all'uso combinato di diverse fonti informative

- fonti amministrative e indagini
- più indagini
- più liste
- …

particolarmente utili a fini previsivi

I progetti in corso

Insieme a

- colleghi di Sapienza,
- studenti di dottorato in economia e in statistica
- ricercatori Istat

- Record Linkage
- Small Area Estimation
- Popolazioni dimoranti
- Previsioni demografiche

Record Linkage

- Integrazione di K fonti informative per la individuazione di entità comuni
- Dai confronti deterministici al metodo statistico fino al machine learning
- Cambiamento di prospettiva: dal confronto specifico alla individuazione di entità latenti nella popolazione che possono fornire più segnali in diverse occasioni di rilevazione

Il metodo bayesiano consente di

- Calibrare l'affidabilità delle diverse occasioni di rilevazione.
- Valutare diversamente la plausibilità che una certa entità sia presente o meno in una determinata lista
- Differenziare il contenuto informativo delle diverse variabili
- Tecniche di machine learning consentono ormai di condurre procedure di linkage su strutture più complesse delle consuete matrici unità-variabili (stringhe, alberi di risposta, etc.)

.

Small Area Estimation

- Utilizzo di tecniche RL per la determinazione di ulteriori covariate
- Gestione di diverse fonti di incertezza, dal linkage all'errore di misurazione nelle variabili
- Formalizzazione gerarchica naturale (scambiabilità parziale)

Popolazione dimorante

Problema: Individuazione di grandezze non direttamente osservabili

- Utilizzo di strutture latenti
- Aumento del numero di parametri da stimare
- Inserimento di *vincoli naturali* mediante utilizzo di distribuzioni a priori "*condivise*"
-

Dati amministrativi

Possiamo utilizzare dati amministrativi per produrre statistica ufficiale?

- informazioni non ottenute mediante schemi campionari. Questo rende gli approcci analitici basati sulla teoria dei campioni di valore discutibile.
- le fonti di errore "dominanti" sono in genere di natura non campionaria
  - Over-coverage
  - Under-coverage
  - Mis-classification
  - Missing data
  - Linkage error

Approcci bayesiani per dati amministrativi

- dati osservati: $X^{(O)}$
- "vero" valore del dato $X^{(T)}$ (latente)

Tre ingredienti:

- Concepire un modello che possa generare dati corretti da fonti osservabili

$$p(X^{(T)}|X^{(O)})$$

- Utilizzo di dati di supporto, studi di "validazione" e/0 modelli che spieghino come $X^{(O)}$ possa essere stato generato dal vero $X^{(T)}$.
- Idee che generalizzano gli approcci usualmente utilizzati nel caso di dati mancanti.

Notazione

- S: Dati da indagine di validazione (survey)
- $G = g(X^{(T)})$ (quantità di interesse primario)
- $p(X^{(T)}|\theta)$ Modello di sistema
- $p(X^{(O)}|X^{(T)}, \gamma)$ Modello osservazionale

Dunque,

$$p(X^{(O)}|S) = \prod_{i \in S} p(X_i^{(O)}|X_i^{(T)} = S_i, \gamma) \prod_{i \notin S} p(X_i^{(O)})$$