Intelligenza Artificiale e Statistica Ufficiale: Esperienze in Istat

Roberta Radini, Monica Scannapieco, Laura Tosco, Diego Zardetto
Istat
radini@istat.it, scannapi@istat.it, tosco@istat.it, zardetto@istat.it

Abstract

Istat utilizza tecniche di IA in settori sperimentali e di produzione dell'informazione statistica. Nel presente contributo si illustra una sintesi (i) delle attività legate all'uso delle ontologie per l'integrazione dei dati nell'ambito del Sistema Integrato dei Registri e (ii) delle attività che prevedono il ricorso a tecniche di *Machine Learning* in progetti di innovazione e per la realizzazione di statistiche sperimentali.

1 L'Uso di ontologie per la produzione di statistica ufficiale

Nell'ambito dei processi statistici, la gestione dei dati e dei metadati ad essi relativi ha da sempre ricoperto un ruolo importante. Negli ultimi anni, l'Istat, così come altri Istituti Nazionali di Statistica europei ed internazionali, ha investito sulla modellazione concettuale dei dati tramite *ontologie*.
I principali motivi per cui l'Istat si è avvicinato a tale paradigma sono:

1. Accesso ai dati trasparente rispetto alla loro memorizzazione fisica.
2. Accesso ai dati facilitato dalla separazione della conoscenza di dominio dalla conoscenza degli aspetti tecnologici connessi alla gestione dei dati.
3. Modellazione formale ed esplicita dei concetti che consente, da una parte, una maggiore formalizzazione dei concetti stessi a vantaggio della correttezza della rappresentazione dei dati, dall'altra, una maggiore facilità di modifica nel caso in cui il dominio vari.
4. Associazione in maniera *machine-actionable* dei metadati ai dati, con conseguente superamento della tradizionale logica di gestione "documentale" dei metadati.
5. Possibilità di utilizzo di strumenti automatici di ragionamento che consentono di scoprire nei dati nuovi pattern di conoscenza, nonché di effettuare ricerche più complesse ed efficienti sui dati.
6. Integrazione dei dati basata sulla definizione di una semantica unica e condivisa dei concetti.
7. Controllo di qualità, in termini accuratezza, completezza, e consistenza [Aracri *et al.*, 2018].
8. Migliore accessibilità dei dati pubblicati attraverso la compatibilità con gli standard del *Semantic Web*.

Istat sta utilizzando le ontologie nell'ambito del progetto di costruzione del Sistema Integrato dei Registri (SIR).
Il SIR rappresenta un'innovazione fondamentale nel sistema di produzione dei dati dell'Istat, centralizzando ed integrando i dati derivati dalle fonti amministrative e dalle rilevazioni statistiche condotte dall'Istituto. Il SIR garantirà una gestione unitaria delle diverse tematiche (statistiche sociali, economiche, territoriali etc.) ed una integrazione concettuale e statistica, oltre che fisica, tra le unità statistiche che lo compongono (individui, famiglie, imprese, luoghi etc.).
La realizzazione del SIR prevede l'utilizzo del paradigma Ontology-based Data Management (OBDM) [Lenzerini, 2011] che include tre layer:

- **Ontologia**: specifica concettuale del dominio di interesse espressa in un linguaggio formale sul quale è possibile fare interrogazione.
- **Basi dati**: rappresentano i dati dei singoli registri.
- **Mapping:** regole di corrispondenza tra i dati dei registri e i concetti del dominio di interesse (definiti nell'ontologia).

Questo approccio consente, non solo di documentare concettualmente il dominio di interesse, ma anche di realizzare e/o gestire l'integrazione fisica dei dati dell'intero sistema SIR tramite l'ontologia.

Ad oggi sono stati realizzate, a diversi livelli di consolidamento, le seguenti componenti del SIR:

- Ontologie degli individui e delle famiglie e Registro degli Individui;
- Ontologie degli indirizzi e Registro dei Luoghi – componente indirizzi;
- Ontologia del Lavoro e Registro del Lavoro.

2 L'uso di tecniche di Machine Learning per la produzione di statistiche sperimentali ed in progetti di innovazione

In linea con il percorso intrapreso da Eurostat e da altri istituti di statistica internazionali, l'Istat sperimenta l'utilizzo di nuove fonti di tipo Big Data e l'applicazione di metodi innovativi nella produzione di informazione statistica.

Le fonti Big Data non sono, in genere, direttamente trattabili con tecniche statistiche tradizionali (si pensi – a titolo di esempio – a specifiche tipologie di dati, come le immagini ed i testi in linguaggio naturale). Ciò motiva e giustifica il crescente interesse degli Istituti Nazionali di Statistica verso le tecniche di *Machine Learning* [UNECE, 2018].

Istat ha recentemente pubblicato sul sito istituzionale due *statistiche sperimentali* derivate da fonti Big Data:

- Modalità di utilizzo dei siti Web da parte delle imprese
- Social Mood on Economy Index

La prima statistica sperimentale è relativa a stime di caratteristiche d'impresa derivate dall'elaborazione statistica dei testi ricavati dallo *scraping* massivo dei siti web delle imprese stesse. Tecniche di *Machine Learning* supervisionato, come ad esempio *random forest*, hanno consentito di predire il valore di variabili quali l'attività di *e-commerce* e la presenza sul sito di offerte di lavoro a partire dal solo contenuto testuale del sito web. L'accuratezza delle statistiche risultante si è rivelata soddisfacente rispetto alle analoghe stime ottenute da indagini tradizionali.

La seconda statistica sperimentale fornisce misure giornaliere del sentiment italiano sull'economia, derivate da campioni di *tweet* pubblici in lingua italiana, catturati in streaming. In questo caso l'utilizzo delle tecniche di *Machine Learning* è stato duplice. Da una parte, tecniche non supervisionate sono state utilizzate per realizzare il *clustering* dei *tweet* negli insiemi di *tweet* 'positivi', 'negativi' e 'neutri'. Dall'altra, i moderni modelli di rappresentazione del linguaggio che vanno sotto il nome di *word embeddings* [Mikolov *et al.*, 2013] [Pennington *et al.*, 2014] hanno permesso di valutare la pertinenza dei *tweet* raccolti rispetto al fenomeno d'interesse (in questo caso la percezione sullo stato dell'economia italiana) [De Fausti *et al.*, 2018].

In aggiunta, Istat sta conducendo un progetto sperimentale finalizzato alla produzione di statistiche di *Land Cover* a partire da immagini satellitari. Per statistiche di *Land Cover* (si veda ad esempio l'indagine europea LUCAS) si intende, da una parte, la classificazione del territorio secondo possibili tipologie di 'copertura' (costruito, campi coltivati, foreste, vegetazione erbacea, etc.), dall'altra, la stima della misura della superficie associata a ciascuna di esse. Il progetto prevede l'uso di tecniche di *Deep Learning*, in particolare *Convolutional Neural Networks*, per prevedere la classe di 'copertura' associata a porzioni di territorio a partire da immagini satellitari Sentinel-2. L'infrastruttura IT di supporto a questo progetto prevede il ricorso ad una *farm* di GPU *on premise* recentemente acquisita dall'Istat.

Un ulteriore progetto sperimentale, ancora in fase di avvio, ha l'obiettivo di realizzare un sistema di *virtual assistance*, ispirato a sistemi come *Siri*. Tale sistema cambierebbe radicalmente il paradigma di interazione degli utenti con l'Istituto, e faciliterebbe loro la ricerca delle informazioni statistiche. L'idea di base è che l'utente possa formulare una query in linguaggio naturale e che il sistema sia in grado di interpretarla, tradurla in una query eseguibile in automatico su una base di dati integrati e di proporre il risultato all'utente. Le tecniche in fase di sperimentazione sono: modelli *Deep Learning* di tipo *sequence-to-sequence* per la componente conversazionale del sistema, algoritmi di *topic modeling* per la componente di *information retrieval* e *ontology modelling* per la componente di *query answering*.

Riferimenti bibliografici

[Aracri et al., 2018] Raffaella Aracri, Adele Bianco, Roberta Radini, Monica Scannapieco, Laura Tosco, Federico Croce, Maurizio Lenzerini, and Domenico Fabio Savo. On the Experimental Usage of Ontology-based data access for the Italian Integrated System of Statistical Registers: Quality Issues. In *Proceedings of the European Conference on Quality in Official Statistics (Q2018)*, Krakov, Poland, May 2018.

[De Fausti et al., 2018] Fabrizio De Fausti, Massimo De Cubellis, and Diego Zardetto. Word Embeddings: a Powerful Tool for Innovative Statistics at Istat. In *Proceedings of 14th International Conference on Statistical Analysis of Textual Data (JADT)*, pages 174–182, Rome, Italy, June 2018. UniversItalia.

[Lenzerini, 2011] Maurizio Lenzerini. Ontology-based data management. In *Proceedings of the 20th International Conference on Information and Knowledge Management (CIKM)*, pages 5–6, Glasgow, Scotland, UK, October 2011.

[Mikolov et al., 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 27th Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, Lake Tahoe, Nevada, USA, 2013.

[Pennington et al., 2014] J. Pennington, R. Socher, C.D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 19th Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014.

[UNECE, 2018] UNECE Blue Sky Thinking Network. The Use of Machine Learning in Official Statistics. *Final report of the UNECE Machine Learning Project*, Geneva, October 2018. Available at: https://statswiki.unece.org/download/attachments/223150364/The%20use%20of%20machine%20learning%20in%20official%20statistics.pdf?version=1&modificationDate=1542811360675&api=v2