DOI: 10.26314/GARR-Conf18-proceedings-14

Qualità dell'aria: algoritmi di machine learning applicati a dati simulati dal sistema modellistico AMS-MINNI

Angelo Mariano[1], Mario Adani[2], Gino Briganti[3], Mihaela Mircea[2]
in rappresentanza del gruppo MINNI[4]

[1]ENEA laboratorio DTE-ICT-IGEST, Bari, [2]ENEA laboratorio SSPT-MET-INAT, Bologna, [3]ENEA laboratorio SSPT-MET-INAT, Pisa, [4]www.minni.org

Abstract. L'applicazione di algortmi di machine learning nell'ambito della qualità dell'aria richiede la disponibilità di una grande mole di dati per individuare trend stagionali e realizzare algoritmi predittivi che consentano di simulare l'effetto delle molteplici fonti di inquinamento atmosferico. In questo studio sono stati utilizzati il dataset ufficiale nazionale relativo alla simulazione per l'anno 2010, prodotto mediante il sistema modellistico della qualità dell'aria AMS-MINNI e raccolto negli anni sui sistemi di storage distribuito dell'ENEA, e osservazioni estratte dal database BRACE. A questi dati sono stati applicati modelli di deep learning per predire le concentrazioni orarie di alcuni inquinanti. I risultati di questa sperimentazione sono oggetto del presente contributo e suggeriscono una nuova via di ricerca nell'integrazione dei sistemi modellistici della qualità dell'aria con modelli di machine learning.

Keywords. Machine Learning, Deep Learning, Modellistica computazionale, Qualità dell'aria

Introduzione

L'obiettivo di questo studio è verificare quanto il machine learning e le reti neurali "deep" siano in grado di estrarre le caratteristiche ("features") più rilevanti e di dedurre le correlazioni che permettano di predire le concentrazioni di determinati inquinanti per un set di dati simulati con il sistema modellistico AMS-MINNI e le osservazioni estratte dal database BRACE, relativi all'anno 2010. Il sistema modellistico atmosferico (AMS) del modello nazionale MINNI (Modello Integrato Nazionale a supporto della Negoziazione Internazionale sui temi dell'inquinamento atmosferico [1]; http://www.minni.org) è stato sviluppato nell'ambito dell'omonimo progetto avviato da ENEA nel 2002 e finanziato dal Ministero dell'Ambiente e della Tutela del Territorio e del Mare (MATTM). AMS-MINNI è utilizzato per fare simulazioni modellistiche della qualità dell'aria sull'Italia in adempimento del D.Lgs. 155/2010 (art. 22, comma 5) che ha recepito la Direttiva Europea 2008/50/EC.

I dati simulati con AMS-MINNI per l'anno 2010 [2] ad una risoluzione spaziale di 4 km, occupano attualmente circa 20TB di storage. La simulazione della qualità dell'aria richiede una quantità rilevante di risorse computazionali per coprire l'intero territorio nazionale in alta risoluzione ed è stata eseguita con l'ausilio dell'infrastruttura di supercalcolo ENEA CRESCO di Portici [3], inclusa nella griglia computazionale ENEA Grid, distribuita su 6 centri di ricerca connessi dalla rete GARR.

Le reti neurali deep (DNN) e gli algoritmi che le caratterizzano non sono un concetto com-

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

pletamente nuovo, in quanto già teorizzate circa 20 anni fa, ma solo recentemente sono diventate uno strumento molto efficace e in via di forte espansione grazie all'estrema abbondanza di dati digitali, di tecnologie evolute per la gestione dello storage e per il computing, oltre che per ottimizzazioni sopravvenute sugli algoritmi stessi [4]. Ultimamente, il deep learning è la tecnologia alla base di moltissime applicazioni di intelligenza artificiale, come le auto a guida autonoma, la riproduzione del parlato e dei suoni, il riconoscimento facciale e gli assistenti virtuali. Sicuramente si può ritenere che una delle potenzialità più sorprendenti delle reti DNN sia quella di trovare una rappresentazione semplificata delle features iniziali e di stabilire le relazioni tra le grandezze di input tabula rasa da un'ampia mole di dati, grazie ad un'elevata potenza computazionale e alla presenza di big data che anche esperti del dominio sono in grado di interpretare solo parzialmente. In particolare, questi algoritmi con le loro caratteristiche di predittori possono trovare un ampio impiego nel settore relativo all'analisi della qualità dell'aria, grazie alla possibilità di integrare grosse moli di dati provenienti da fonti diverse e di evidenziare le correlazioni tra le condizioni meteo, le condizioni orografiche, le emissioni di inquinanti.

1. Metodi e strumenti

1.1 Descrizione del modello AMS-MINNI

Il sistema modellistico AMS-MINNI è composto principalmente da un modello meteorologico RAMS, un post-processore diagnostico SURFPRO che stima le variabili micrometeorologiche usando come input i dati meteorologici prodotti da RAMS, un processore per le emissioni antropiche (Emission Manager) e un modello di trasporto chimico FARM.

Le componenti del sistema modellistico (Figura 1) e la simulazione eseguita per il 2010 è descritta in dettaglio in [2].

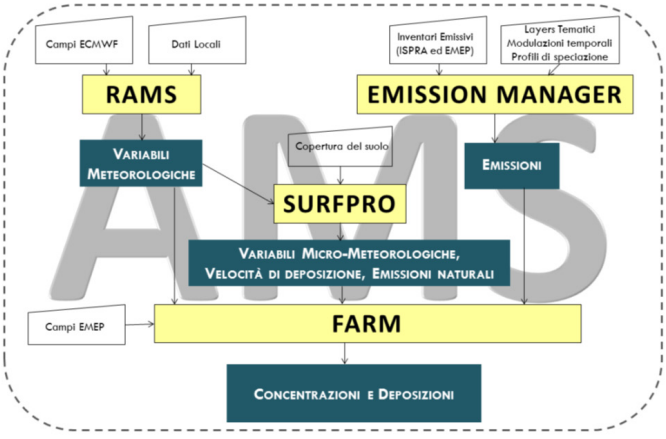


Fig. 1
Rappresentazione schematica di AMS-MINNI

1.2 Descrizione del modello predittivo Machine Learning (ML)

Il modello Machine Learning utilizzato si basa sull'uso estensivo delle celle Long Short-Term Memory (LSTM) [5]. Si tratta di reti neurali "deep" ricorrenti, cioè in grado di analizzare in particolare le dipendenze delle serie temporali o delle sequenze di segnali. Una semplice unità LSTM è composta da una cella, un gate di input, un gate di output ed un

gate di aggiornamento della memoria. La cella è attraversata da una sorta di nastro trasportatore che rappresenta la memoria a breve termine, nella quale le sequenze di segnali vengono scritte o cancellate nel tempo. I compiti nei quali le celle LSTM eccellono sono legati alla classificazione, al processamento e alla formulazione di predizioni nel caso delle serie temporali, dal momento che ci possono essere intervalli di lunghezza non specificata nella ricorrenza di eventi importanti. Le unità LSTM sono state sviluppate per affrontare il problema del gradiente che si azzera durante la correzione a ritroso della rete neurale ricorrente tradizionale (RNN). L'insensibilità rispetto all'intervallo temporale nella ripetizione è uno dei vantaggi delle celle LSTM rispetto alle RNN e più in generale agli altri algoritmi di machine learning applicati alle serie temporali.

La serie temporale di riferimento è la sequenza delle concentrazioni orarie di NO2 misurate da rilevatori disposti in tre punti geografici in Emilia Romagna: sito A non urbano nel comune di Besenzone (PC), sito B urbano nel comune di Piacenza, Giordani-Farnese, sito C urbano nel comune di Piacenza, Parco Montecucco, 24 ore su 24 nell'arco del 2010. Per l'analisi di base del comportamento della serie temporale si è fatto riferimento alle procedure di forecast per le serie temporali Prophet [6], che si basa su un modello additivo in cui si identificano i trend non lineari su base annuale, mensile, settimanale, giornaliera ed oraria.

Le variabili di input considerate sono state le features non nulle che di solito il modello AMS-MINNI utilizza in input, rilevate in un range di 2 ore indietro nel tempo, in accoppiata con la concentrazione di NO2 misurata dalla centralina nella stessa ora. La topologia del modello ML per la predizione delle concentrazioni orarie di NO2 ha una struttura composta da un layer di 50 celle LSTM e un layer denso che fornisce la predizione sulla concentrazione di biossido di azoto in un determinato sito. Le celle LSTM quindi processano e identificano al loro interno le sequenze temporali estese fino a 2 ore indietro nel tempo e le correlazioni tra le features del modello AMS-MINNI e la concentrazione di biossido di azoto a parità di ora. Per evitare l'overfitting dei dati di input, all'interno della rete neurale deep si è inserita la funzionalità del dropout, ossia la rottura random dei collegamenti fra i nodi ad ogni livello della rete nel 20% dei casi. Il dataset di input, su questi siti specifici è composto da 1GB di dati ed è stato così suddiviso: training il primo 70%, test il secondo 20%, validazione il restante 10%.

Una prima sperimentazione è stata effettuata prendendo in considerazione un singolo sito A (non urbano) corrispondente alla presenza di una centralina e addestrando la rete neurale a replicare i risultati reali. La seconda sperimentazione ha avuto l'obiettivo di replicare con la rete neurale deep i risultati della simulazione di AMS-MINNI, per ottenere uno strumento veloce ed affidabile per verificare l'incidenza delle perturbazioni dei parametri di input sull'output del modello. In quest'ultimo caso le concentrazioni di NO2 a parità di fascia oraria sono state non quelle misurate dalla centralina, ma quelle simulate con il modello.

2. Risultati e discussioni

Fig. 2 e 3 mostrano i "pattern" caratteristici ottenuti dal modello additivo per le concentrazioni di biossido d'azoto su una scala temporale settimanale e, rispettivamente, su scala giornaliera nel sito A. Si evidenziano i tipici fenomeni di accumulo di inquinanti in cor-

rispondenza del traffico veicolare intenso di mattina e pomeriggio e nei giorni lavorativi
La Fig.4 mostra, nell'ambito del set di validazione, il valore atteso della concentrazione di NO2 predetto dal modello ML, il valore misurato dalla centralina ed il valore "airq"

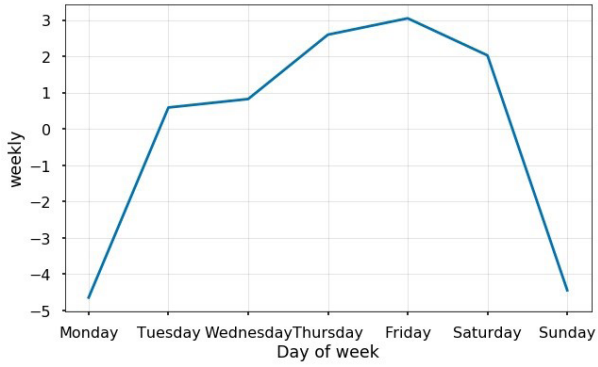


Fig. 2
Illustrazione
del trend a base
settimanale

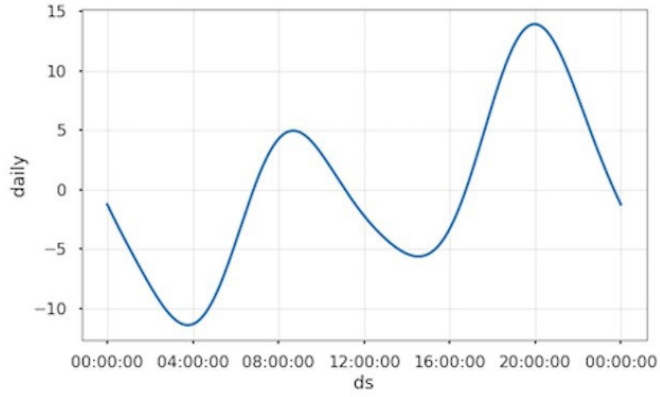


Fig. 3
Illustrazione
del trend a base
giornaliera

corrispondente al valore simulato dal modello AMS-MINNI per il sito A. Come si vede dal grafico, il modello ML prevede con buona approssimazione i valori di concentrazione dell'inquinante nel sito considerato. Il "relative squared error", definito rispetto al valore atteso $\mathbf{y}$, al valore predetto $\hat{\mathbf{y}}$ e alla deviazione standard $\boldsymbol{\sigma}$ come

$$RSE = \frac{\sqrt{(y - \hat{y})^2}}{\sigma}$$

passa da 1.76 nel caso del modello AMS-MINNI a 0.76 nel caso del modello ML, guadagnando quindi una deviazione standard rispetto ai dati presi in esame anche a causa dell'utilizzo preponderante dei dati misurati.
La capacità di questo modello di riprodurre il dato della centralina è però limitato dalla disponibilità di dati di concentrazioni in real time: ML richiede i dati di concentrazione nelle N ore precedenti. Fino a quando questa limitazione di tipo organizzativo non verrà rimossa, sarà necessario far riferimento a modelli di machine learning che prescindano dalle concentrazione dello stesso inquinante nelle ore precedenti, che però al momento danno risultati meno promettenti, con un RSE di 1.13 sullo stesso set di dati di validazione.
Il secondo esperimento utilizzata una rete neurale deep analoga a quella precedente, in

cui però si è cercato di prevedere il dato di concentrazione simulato nel modello AMS-MINNI. In questo caso il training è effettuato su un solo sito B (urbano) e il modello ML "addestrato" è in grado di predire i dati di concentrazione di NO2 sia nel sito originario B

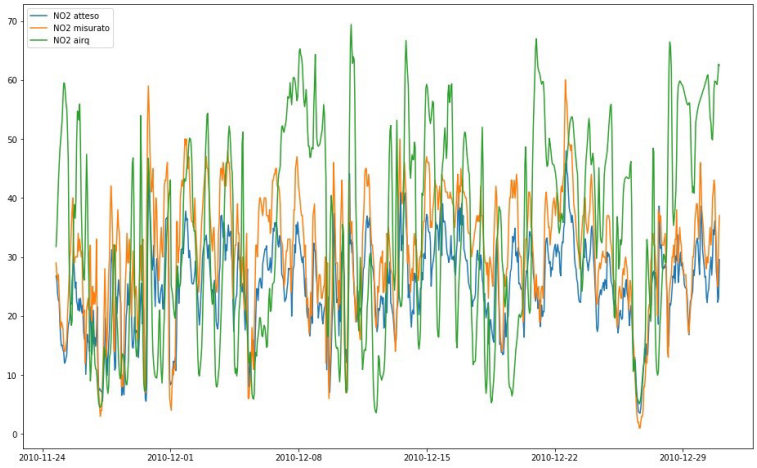


Fig. 4
Confronto tra le concentrazioni orarie di NO2 nel sito A predette dal modello ML (in blu), misurate dalla centralina (in arancio) e simulate dal modello AMS-MINNI (in verde)

(Fig.5) con RSE pari a 0.08 che in altri siti diversi sia urbani (sito C) (Fig.6) che non (sito A) (Fig.7), in quest'ultimo caso con RSE pari a 0.10.
I tempi di calcolo in entrambe le sperimentazioni sono stati dell'ordine di mezzora su un computer portatile. Da questi risultati si evince come l'algoritmo di machine learning sia stato in grado di estrarre da questo grande dataset informazioni utili per stimare in maniera efficace e rapida le concentrazioni a breve termine dell'inquinante considerato.

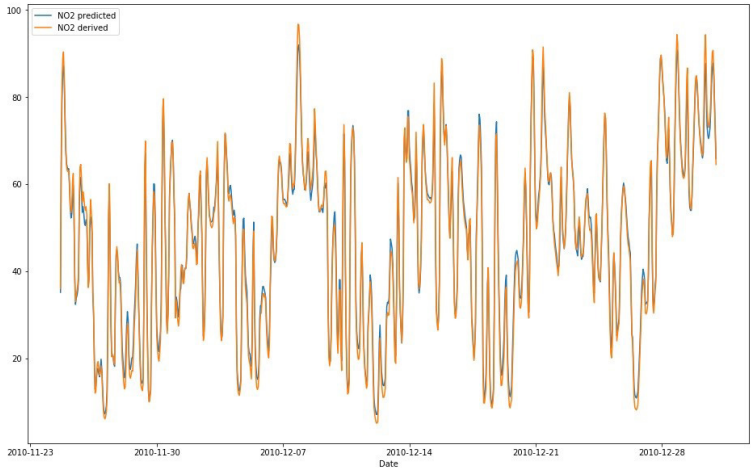


Fig. 5
Confronto tra le concentrazioni orarie di NO2 nel sito B predette dal modello ML (in blu) e simulate da AMS-MINNI (in arancio)

3. Conclusioni

Questo studio dimostra le potenzialità di questo grande dataset distribuito ospitato sull'infrastruttura ENEAGrid – CRESCO e degli algoritmi di machine learning nella costruzione di predittori che supportino la ricerca nel campo della qualità dell'aria. I risultati

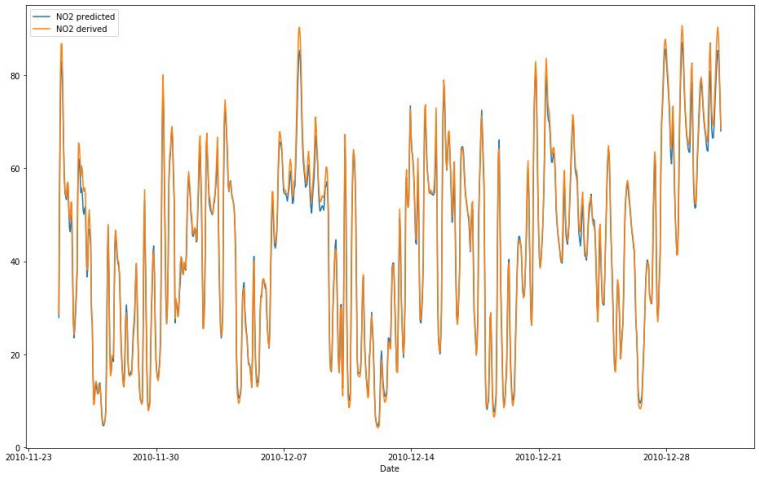


Fig. 6
Confronto tra le concentrazioni orarie di NO_2 nel sito C predette dal modello ML (in blu) e simulate da AMS-MINNI (in arancio) con training sul sito B.

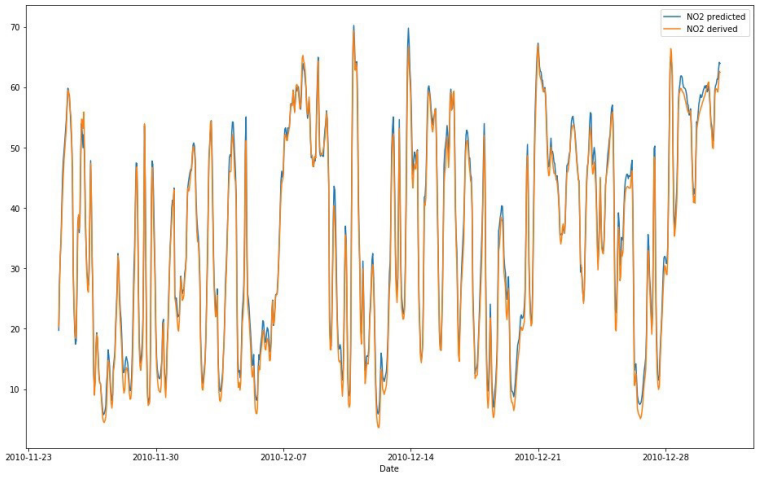


Fig. 7
Confronto tra le concentrazioni orarie di NO_2 nel sito A predette dal modello ML (in blu) e simulate da AMS-MINNI (in arancio) con training sul sito B.

aiuteranno a sviluppare approfondite analisi su altri inquinanti dannosi per la salute pubblica, come l'ozono e il particolato PM10 e PM2.5.
Si ringrazia per il supporto il gruppo di lavoro ENEAGrid [7].

Riferimenti bibliografici

[1] Mircea, M., Ciancarella, L., Briganti, G., Calori, G., Cappelletti, A., Cionni, I., Costa, M., Cremona, G., D'Isidoro, M., Finardi, S., Pace, G., Piersanti, A., Righini, G., Silibello, C., Vitali, L., Zanini, G., (2014). "Assessment of the AMS-MINNI system capabilities to predict air quality over Italy for the calendar year 2005. Atmospheric Environment", 84, 178–188, ISSN 1352-2310, http://dx.doi.org/10.1016/j.atmosenv.2013.11.006.

[2] Ciancarella L. et al. (2016), "La simulazione nazionale di AMS-MINNI relativa all'anno 2010 - Simulazione annuale del SistemaModellistico Atmosferico di MINNI e validazione dei risultati tramite confronto con i dati osservati", Rapporto Tecnico ENEA, RT/2016/12/ENEA

[3] Ponti G. et al (2014), "The role of medium size facilities in the HPC ecosystem: the case

of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure" In Proceedings of the International Conference on High Performance Computing and Simulation, HPCS 2014, Bologna, Italy, 21-25 July, 2014

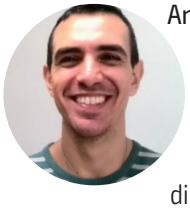
[4] Bengio Y. (2009) "Learning Deep Architectures for AI", Foundations and Trends in Machine Learning, 2 (1), pp. 1-127.

[5] Hochreiter S., Schmidhuber J. (1997). "Long short-term memory". Neural Computation. 9 (8): 1735– 1780.

[6] Taylor SJ, Letham B. (2017). "Forecasting at scale." PeerJ Preprints 5:e3190v2

[7] http://www.eneagrid.enea.it/people/2018EneaGridPeople.html

Autori



Angelo Mariano - angelo.mariano@enea.it

Dottore di ricerca in fisica teorica dal 2004, con una tesi nel campo della decoerenza in meccanica quantistica, ricercatore ICT in ENEA dal 2010, con competenze nel campo della progettazione informatica, dell'integrazione tra i sistemi ICT, del calcolo scientifico ad alte prestazioni, del cloud computing, dell'analisi di sistemi complessi. Attualmente si occupa di informatica gestionale, gestione dei big data, machine learning e intelligenza artificiale.

Mario Adani - mario.adani@enea.it

Ha lavorato nel campo della modellistica e assimilazione dati in oceanografia, partecipando allo sviluppo del modello NEMO. Ha contribuito a diversi progetti europei finalizzati allo sviluppo della capacità previsionali e di reanalisi. Attualmente, sta lavorando con modelli di dispersione inquinanti in atmosfera a fini previsionali e alle relazioni tra il cambiamento climatico e l'inquinamento atmosferico.

Gino Briganti - gino.briganti@enea.it

Gino Briganti si è laureato in fisica nel 1990, con una tesi concernente simulazioni su CRAY della QCD non perturbativa. Si occupa di sviluppo di codici per la simulazione della micrometeorologia dello strato limite e della dispersione e trasformazione chimica degli inquinanti in atmosfera. Una considerevole parte di lavoro riguarda set-up e run di modelli di simulazione su GRID ENEA. La sua attività ha come obiettivo la valutazione di impatto sulla qualità dell'aria di scenari emissivi, a supporto della negoziazione internazionale sull'inquinamento atmosferico.

Mihaela Mircea - mihaela.mircea@enea.it

Mihaela Mircea lavora come ricercatrice nel campo delle scienze atmosferiche, essendo particolarmente interessata alla conoscenza della composizione chimica della troposfera e ai suoi effetti sulla meteorologia, clima, ambiente e salute. I suoi studi sono dedicati allo sviluppo di modelli in modo integrato con le osservazioni per riprodurre in modo realistico i processi atmosferici e la loro interazione con il resto dell'ambiente con lo scopo di poter prevedere con precisione l'impatto delle future attività antropiche sulla qualità dell'aria e sul clima.