Big Data Analytics e Predictive Modeling per il Settore Energetico

Roberto Corizzo, Michelangelo Ceci, Donato Malerba
Laboratorio KDDE - Knowledge and Discovery Data Engineering, Dipartimento di Informatica,
Università degli Studi di Bari "Aldo Moro", via Orabona 4, I-70125 Bari, Italy
{roberto.corizzo, michelangelo.ceci, donato.malerba}@uniba.it

Abstract

Questo contributo descrive i recenti risultati conseguiti dal gruppo di ricerca KDDE in termini di nuovi approcci e metodi che operano su dati provenienti da sensori geo-distribuiti, specificatamente progettati per a risolvere problemi di natura predittiva presenti nel settore energetico. La ricerca è stata focalizzata sulla proposta di metodi di Big Data Analytics e di Predictive Modeling in grado di processare grandi quantità di dati provenienti da sensori, mediante l'uso di architetture distribuite, nonché in grado di risolvere il problema della previsione di energia erogata da una rete di impianti, prendendo anche in considerazione il rilevamento e il trattamento di anomalie presenti nei dati. Tale ricerca è coerente con gli obiettivi di progetti di ricerca finanziati dalla Commissione Europea e dal Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica.

1 Introduzione

Il contesto energetico prevede la generazione di grandi quantità di dati provenienti da sensori, presenti presso stazioni meteorologiche e impianti di produzione. Monitorare la produzione e il consumo di energia, sia a livello locale che a livello globale, e disporre di modelli di predizione accurati, è di fondamentale importanza in questo contesto per via delle sempre più pressanti sfide di integrazione dell'energia nella rete, bilanciamento del carico e ottimizzazione dei ricavi nella compravendita di energia. La sfida nasce da una nuova forma di rete che richiede l'integrazione di fonti di energia distribuita e rinnovabile, in quanto tali fonti sono intermittenti e dipendono da fattori incontrollabili, quali le condizioni meteorologiche.

2 Sfide e approcci proposti

La crescente presenza di reti di sensori geo-distribuiti implica la generazione di enormi volumi di dati in più località spaziali ad un ritmo crescente. A causa dell'eterogeneità e dell'elevato volume di dati, è necessario adottare adeguate tecniche di Big Data Analytics che garantiscano accesso e analisi rapida dei dati, caratteristiche, queste, non ottenibili con gli approcci tradizionali alla gestione dei dati.

A tale scopo, è stato realizzato un sistema distribuito per la gestione di grandi quantità di dati raccolti da sensori, basato sul Database management system Apache HBase e sul framework Hadoop, consentendo di disporre di elevate performance computazionali in un ambiente di calcolo distribuito [Ceci *et al.*, 2014] [Ceci *et al.*, 2015]. Tale sistema è stato usato come base per lo sviluppo e l'esecuzione di algoritmi di Predictive Modeling.

Particolare attenzione nel lavoro di ricerca è stata dedicata alle energie rinnovabili e al problema di predizione dell'energia prodotta da reti di impianti ad un orizzonte di previsione di 24 ore. Una delle principali sfide dal punto di vista del Predictive Modeling, è rappresentata dalla collocazione spaziale dei sensori, che comporta la presenza di autocorrelazione spaziale nei dati. Sebbene la considerazione esplicita di queste dipendenze spaziali comporti ulteriore complessità nel processo di apprendimento dei modelli, generalmente consente di ottenere una maggiore accuratezza. La produzione di energia da impianti fotovoltaici risulta fortemente influenzata anche dall'autocorrelazione temporale, poiché: *i)* tende ad assumere valori simili in un dato momento in giorni ravvicinati, *ii)* ha un comportamento ciclico e stagionale. Coerentemente con queste considerazioni, l'attività di ricerca ha investigato diverse strategie per l'embedding di informazioni spazio-temporali nei modelli predittivi. Una di queste ha riguardato l'impiego di indicatori statistici per rappresentare in modo implicito le correlazioni spazio-temporali tra dati prodotti in località spaziali differenti, mediante nuovi attributi descrittivi [Ceci *et al.*, 2017].

Si è inoltre formulato un approccio per modellare l'autocorrelazione spazio-temporale in modo esplicito nel processo di apprendimento di reti neurali. A tal scopo sono stati proposti dei criteri di ottimizzazione basati su entropia [Ceci *et al.*, 2019]. Tali criteri forniscono il vantaggio ulteriore di far fronte ad una distribuzione non gaussiana dell'errore di predizione, tipicamente riscontrata nel contesto di previsione di energia rinnovabile, in cui i tradizionali criteri di apprendimento risultano limitati.

Un altro approccio proposto ha consentito di estrarre l'autocorrelazione spazio-temporale in forma di attributi descrittivi, mediante l'uso di fattorizzazioni tensoriali [Corizzo *et al.*, 2019b]. I modelli tensoriali offrono una più

ricca rappresentazione rispetto a quella matriciale costituita da istanze e attributi, più comunemente adottata. Il modello tensoriale incorpora fenomeni latenti di autocorrelazione spazio-temporale, tipicamente presenti quando i dati sono provenienti da sensori geo-localizzati e distribuiti in un'area geografica. L'approccio proposto ha consentito di ottenere una nuova rappresentazione dei dati che possa essere adottata da qualsiasi algoritmo di Predictive Modeling, per affrontare il problema di predizione di energia.

In tutti i casi considerati, gli approcci proposti hanno dimostrato il beneficio della considerazione dell'autocorrelazione spaziale e temporale nei modelli di predizione e performance di predizione competitive rispetto ad algoritmi molto noti nello stato dell'arte.

L'attività di ricerca ha anche preso in considerazione il problema della predizione di output strutturato e, specificatamente, il problema della regressione multi-target, ovvero dell'apprendimento di modelli di regressione per più attributi target. Ciò ha permesso di apprendere modelli di predizione di serie temporali della produzione energetica. Il modo più semplice per risolvere questo problema consiste nell'adozione di modelli locali per ciascun attributo target, indipendentemente. Approcci più complessi apprendono un modello globale che è in grado di predire il valore di tutti gli attributi target contestualmente. Tali approcci sono stati applicati a differenti modelli, tra cui reti neurali, alberi di regressione e clustering predittivo. I risultati di predizione, nel contesto energetico, hanno evidenziato che gli approcci di output strutturato offrono migliori performance, in quanto in grado di sfruttare possibili dipendenze tra gli attributi target.

Un ulteriore risultato dell'attività di ricerca ha riguardato la sintesi due metodi di clustering distribuito [Corizzo *et al.*, 2019c] [Malondkar *et al.*, 2018]. Il primo, apprende un modello di clustering basato sulla densità, e sfrutta i cluster estratti per risolvere problemi di regressione single-target e multi-target. Il secondo, apprende una gerarchia di modelli Self-Organizing Maps, che si adatta automaticamente ai dati a disposizione, ed è in grado di risolvere problemi di natura descrittiva e predittiva.

Ulteriore aspetto di interesse durante l'attività di ricerca è stato il problema della presenza di anomalie nei dati provenienti da sensori. A tal scopo, è stata proposta una strategia di rilevamento delle anomalie mediante modelli di deep learning non supervisionati [Corizzo *et al.*, 2019a]. Tale strategia è in grado di riparare i dati anomali in una specifica località spaziale, sfruttando i dati non anomali osservati in località spaziali vicine. L'implementazione degli approcci [Malondkar *et al.*, 2018] [Corizzo *et al.*, 2019a] [Corizzo *et al.*, 2019c] è stata condotta adottando le primitive di programmazione distribuite del framework Apache Spark, con il fine di garantire elevate performance di scalabilità in presenza di grandi quantità di dati.

3 Progetti di ricerca

Il gruppo di ricerca KDDE ha affrontato le tematiche descritte in Sezione 2 nell'ambito dei progetti "PON Vi- POC: Virtual Power Operation Center" e "EUFP7 FET MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data". Approcci di clustering preditivo attinenti il settore energetico sono stati proposti in un pilot nell'ambito del progetto EU H2020 "TOREADOR - TrustwOrthy model-awaRE Analytics Data platform". Attualmente, il gruppo di ricerca sta portando avanti l'investigazione di nuovi approcci di Predictive Modeling e Big Data Analytics nell'ambito del progetto PON "ComESto - Community Energy Storage: Gestione Aggregata di Sistemi d'Accumulo dell'Energia in Power Cloud" finanziato dal MIUR. In particolare, l'interesse è quello di modellare funzioni di ottimizzazione multi-obiettivo per la pianificazione della struttura della rete elettrica e, quindi, per poter pianificare interventi migliorativi a medio-lungo termine.

Riferimenti bibliografici

[Ceci *et al.*, 2014] Michelangelo Ceci, Nunziato Cassavia, Roberto Corizzo, Pietro Dicosta, Donato Malerba, Gaspare Maria, Elio Masciari, e Camillo Pastura. Innovative power operating center management exploiting big data techniques. In *18th International Database Engineering & Applications Symposium, IDEAS 2014*, pages 326–329, 2014.

[Ceci *et al.*, 2015] Michelangelo Ceci, Roberto Corizzo, Fabio Fumarola, Michele Ianni, Donato Malerba, Gaspare Maria, Elio Masciari, Marco Oliverio, e Aleksandra Rashkovska. Big data techniques for supporting accurate predictions of energy production from renewable sources. In *Proceedings of the 19th International Database Engineering & Applications Symposium, IDEAS 2015*, pages 62–71, 2015.

[Ceci *et al.*, 2017] Michelangelo Ceci, Roberto Corizzo, Fabio Fumarola, Donato Malerba, e Aleksandra Rashkovska. Predictive modeling of PV energy production: How to set up the learning task for a better prediction? *IEEE Trans. Industrial Informatics*, 13(3):956–966, 2017.

[Ceci *et al.*, 2019] Michelangelo Ceci, Roberto Corizzo, Donato Malerba, e Aleksandra Rashkovska. Spatial autocorrelation and entropy for renewable energy forecasting. *Data Mining and Knowledge Discovery*, 2019.

[Corizzo *et al.*, 2019a] Roberto Corizzo, Michelangelo Ceci, e Nathalie Japkowicz. Anomaly detection and repair for accurate predictions in geo-distributed big data. *Big Data Research (under review)*, 2019.

[Corizzo *et al.*, 2019b] Roberto Corizzo, Michelangelo Ceci, Gama Joao, e Fanaee Hadi. Multi-aspect renewable energy forecasting. *IEEE Trans. Industrial Informatics (under review)*, 2019.

[Corizzo *et al.*, 2019c] Roberto Corizzo, Gianvito Pio, Michelangelo Ceci, e Donato Malerba. Dencast: Distributed density-based clustering for multi-target regression. *IEEE Trans. Parallel Distrib. Syst. (under review)*, 2019.

[Malondkar *et al.*, 2018] Ameya Malondkar, Roberto Corizzo, Iluju Kiringa, Michelangelo Ceci, e Nathalie Japkowicz. Spark-ghsom: Growing hierarchical self-organizing map for large scale mixed attribute datasets. *Information Sciences*, 2018.