Relazione sulle attività svolte durante il corso del dottorato

May 18, 2021

Titolo della tesi:	*A Tensor Framework for Learning in Structured Domains*
Candidato:	Daniele Castellana
Supervisori:	Prof. Davide Bacciu
Comitato interno:	Prof. Giancarlo Bigi e Prof. Paolo Milazzo
Revisori esterni:	Prof. Barbara Hammer e Prof. Guillaume Rabusseau

1 Sommario della ricerca

Durante il dottorato ho svolto attività di ricerca nell'ambito del Machine Learning (ML) per dati strutturati. I dati strutturati sono composti da una serie di unità atomiche d'informazione (solitamente chiamate *nodi*) che sono collegate tra loro mediante una *struttura*. Lo sviluppo di modelli ML per questi dati è un problema molto studiato nel mondo della ricerca dato che questi sono molto frequenti in ambiti come la chimica, analisi del linguaggio naturale, ecc. Tuttavia, la presenza della struttura complica l'apprendimento in quanto il modello ML deve essere in grado di (1) adattarsi alle diverse strutture che possono comparire nei dati (i.e. essere *adattivo*) e (2) estrarre l'informazione contestuale presente nella struttura. Dunque, le tecniche tipicamente usate per i dati non-strutturati (o "flat") non posso essere applicate direttamente.

Di seguito, discuto brevemente i risultati principali raggiunti durante il dottorato.

Sviluppo di nuovi modelli ML ricorsivi per dati strutturati basati su decomposizioni tensoriali

Molti modelli ricorsivi presenti in letteratura combinano le informazioni atomiche contenute in un nodo e le informazione contestuali ottenute dalla struttura tramite una somma. La somma permette di ottenere modelli "efficienti", in quanti il loro numero di parametri cresce linearmente con la dimensione del contesto (i.e. il numero di nodi vicini). Tuttavia, la somma non è in grado di modellare intreazioni di ordine superiore tra le informazioni considerate. D'altro canto, se le informazioni contestuali sono aggregate mediante un tensore, è possibile modellare interazioni di ordine superiore. Purtroppo, l'approccio tensoriale è spesso irrealizzabile in pratica a causa di una relazione esponenziale tra il numero di parametri del modello e la dimensione del contesto. Curiosamente, questo comportamente è identico per i modelli ricorsivi sia neurali che probabilistci.

Il framework tensoriale proposto permette di creare nuovi modelli ricorsivi (sia neurali che probabilistici) per dati strutturati applicando le decomposizioni tensoriali per ridurre la complessità dei modelli basati su tensori. Il principale beneficio di questa formulazione è la possibilità di regolare il trade-off tra capacità espressiva del modello e numero di parametri dello stesso variando il rank della decomposizione usata. Alti valori di rank, permettono di modellare interazioni complesse e richiedono un numero elevato di parametri. Viceversa, bassi valori di rank modellano interazioni semplici comprimento la complessità del modello.

Anche i modelli basati su somma possono essere inquadrati all'interno del framework tensoriale in quanto la somma è una particolare funzione multi-lineare. Dunque, il secondo vantaggio del framework è quello di poter interpretate ogni modello ricorsivo come un'approssimazione di un modello basato sull'approccio tensoriale. Da questo punto di vista, ogni approssiamzione introduce un bias che può essere sfruttato dal modello.

I risultati della mia ricerca sono stati pubblicati in conferenze e riviste internazionali [8, 4–7, 2]

Realizzazione di modelli Bayesiani non-parametrici per dati strutturati

Determinare gli iper-parametri corretti per un modello ML è spesso dispendioso in quanto questo richiede di ri-eseguire il training diverse volte con varie configurazioni (questo processo è chiamato "model selection"). I modelli non-parametrici offrono un'alternativa alla model selction adattandosi automaticamente sui dati durante il training.

Durante il dottorato ho approfondito le tecniche Byaesiane per sviluppare due modelli non-parametrici per dati strutturati. Il primo è una mistura infinita per dati strutturati basata su Dirichlet Process che è in grado di determinare il numero di clusters presenti nei dati in maniera automatica. Il secondo modello sviluppato è l'estensione Bayesiana di un modello probabilistico basato su decomposizione tensoriale il quale è in grado di determinare il valore del rank in maniera automatica.

I risultati della mia ricerca sono stati pubblicati in conferenze e rivister internazionali [3, 1, 4].

2 Formazione

Esami sostenuti

A.A. 2019/2020

- *Cyber-Physical Systems and Cloud for Smart Industry*: Corso del Dottorato in Informatica, tenuto da Daniele Mazzei (Università di Pisa). Durata: 20 ore;
- *Programming Tools & Techniques in the Pervasive Parallelism Era*: Corso del Dottorato in Informatica, tenuto da Marco Danelutto (Università di Pisa), Patrizio Dazzi (CNR). Durata: 20 ore.

A.A. 2018/2019

- *Data Stream Processing from the Parallelism Perspective*: Corso del Dottorato in Informatica, tenuto da Gabriele Mencagli (Università di Pisa), Matteo Andreozzi (ARM). Durata:20 ore.
- *Matrix computations: a quick overview and some applications*: Corso del Dottorato in Informatica, tenuto da Gianna Del Corso (Università di Pisa), Federico Poloni (Università di Pisa). Durata: 20 ore.

A.A. 2017/2018

- *An Introduction to Deep Learning*: Corso del Dottorato in Informatica, tenuto da Antonio Gullì (Google). Durata: 20 ore;
- *Skill Boosting for New (Research) Horizons*: Corso del Dottorato in Informatica, tenuto Paolo Ferragina, Michele Padrone (Università di Pisa), Davide Morelli (Biobeats). Durata: 20 ore.

Seminari seguiti

A.A. 2017/2018

- *Academic English* (Joanne Spataro);
- *"Mauriana Pesaresi" Lunch Seminars* (Various speakers);
- *Summer School ACDL 2018*: La scuola "Advanced Course on Data Science and Machine Learning" si e tenuta presso la Certosa di Pontignano, Siena dal 19 al 23 Luglio 2018.

2.1 Scuole di dottorato

- (19–23/07/2018) *Advanced Course on Data Science and Machine Learning Summer School 2018 (ACDL 2018)*, Pontignano (SI), Italia.

3 Pubblicazioni

Pubblicazioni su Riviste Internazionali

[1] Davide Bacciu e Daniele Castellana. "Bayesian mixtures of Hidden Tree Markov Models for structured data clustering". *Neurocomputing* 342 (2019). ISSN: 18728286. DOI: 10.1016/j.neucom.2018.11.091.

[2] Daniele Castellana e Davide Bacciu. "A Tensor Framework for Learning in Structured Domains". *Neurocomputing* (2021). Accepted.

Pubblicazioni su Conference Proceedings

[3] Davide Bacciu e Daniele Castellana. "Mixture of Hidden Markov Models as tree encoder". *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 2018. ISBN: 9782875870476. URL: https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2018-112.pdf.

[4] Daniele Castellana e Davide Bacciu. "Bayesian Tensor Factorisation for Bottom-up Hidden Tree Markov Models". *2019 International Joint Conference on Neural Networks (IJCNN)*. Vol. 2019-July. IEEE, lug. 2019, pp. 1–8. ISBN: 978-1-7281-1985-4. DOI: 10.1109/IJCNN.2019.8851851. URL: https://ieeexplore.ieee.org/document/8851851/.

[5] Daniele Castellana e Davide Bacciu. "Generalising Recursive Neural Models by Tensor Decomposition". *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, lug. 2020, pp. 1–8. ISBN: 978-1-7281-6926-2. DOI: 10.1109/IJCNN48605.2020.9206597. arXiv: 2006.10021. URL: https://ieeexplore.ieee.org/document/9206597/.

[6] Daniele Castellana e Davide Bacciu. "Learning from Non-Binary Constituency Trees via Tensor Decomposition". *28th International Conference on Computational Linguistic (COLING)*. 2020.

[7] Daniele Castellana e Davide Bacciu. "Tensor Decompositions in Recursive NeuralNetworks for Tree-Structured Data". *Proceedings of the the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 2020. ISBN: 9782875870742. URL: http://www.i6doc.com/en/..

Pubblicazioni in Workshops

[8] Davide Bacciu e Daniele Castellana. "Learning Tree Distributions by Hidden Markov Models". *Workshop on Learning and Automata (LearnAut'18)*. 2018.

4 Assistenza alla didattica

Laurea Triennale in Informatica (L-31)

- *Programmazione I e Laboratorio dell'Informatica*, A.A. 2020/2021 (Docente: Prof. Alina Sirbu);
- *Programmazione I e Laboratorio dell'Informatica*, A.A. 2019/2020 (Docente: Prof. Alina Sirbu);
- *Programmazione I e Laboratorio dell'Informatica*, A.A. 2018/2019 (Docente: Prof. Alina Sirbu);

Laurea Triennale in Fisica (L-30)

- *Informatica*, A.A. 2017/2018 (Docente: Prof. Susanna Pelagatti);

5 Tirocini

Hyperborea srl L'obiettivo del tirocinio è stato la sperimentazione di algoritmi e tecniche di classificazione automatica di testi estratti da file audio e processati utilizzando un motore di estrazione semantica. Il tirocinio è stato svolto come esame del corso *Skill Boosting for New (Research) Horizons* dal 11/06/2018 al 30/09/2018.

Pisa, 18 maggio 2021

In fede,
Daniele Castellana