

Relazione sulla sperimentazione sui Kernel Method

A.A. 2023/2024

Rosso Carlo

Indice

1	Introduzione	2
2	Dataset	2
3	Modello	3
4	Risultati	3
4.1	Binary	3
4.2	Regression	4
4.3	Multiclass	4
4.4	RNN	5
5	Conclusioni	5
5.1	Sentiment vs Syntax vs Subtree	5
5.2	Kernel Methods vs RNN	5

1 Introduzione

La sperimentazione sui Kernel Methods è stata condotta utilizzando il software fornito e descritto nella pagina citata in Moschitti 2006. A causa di alcune incertezze riguardo le modalità operative, ho consultato il professore per chiarimenti. Durante l'incontro, abbiamo deciso di intraprendere le seguenti sperimentazioni:

- Sperimentazione su diversi dataset:
 - Un dataset per la regressione.
 - Un dataset per la classificazione binaria (positivi e negativi).
 - Cinque dataset, ciascuno rappresentante una classe distinta.
 - Dataset sui sottoalberi.
- Sperimentazione esclusiva su Partial Tree Kernels, scelti per le loro buone prestazioni e tempi di allenamento ridotti.

2 Dataset

- È stata ottenuta una versione modificata del treebank, in cui le etichette dei nodi sono state sostituite con i rispettivi POS-tag, utilizzando lo Stanford Parser. In particolare, il parser è stato utilizzato tramite il container Docker descritto nella repository NLPbox 2024.
- I due dataset risultanti sono stati convertiti nel formato svmlight.
- È stato generato un terzo dataset unendo gli alberi dei due dataset precedenti.
- Poiché il software fornito consente di effettuare sia regressione che classificazione binaria, i dataset sono stati ulteriormente elaborati.
- È stato creato un dataset contenente solo esempi positivi o negativi, unendo le classi "negativo" e "un po' negativo", e le classi "positivo" e "un po' positivo". La classe "neutro" è stata eliminata.
- Sono stati creati cinque dataset distinti per la classificazione multiclasse; ciascun dataset è stato progettato per addestrare un modello specializzato su una singola classe.
- Le classi dei modelli sono state normalizzate trattando il valore come un float. In particolare, nel dataset di partenza le classi variavano da 0 a 4, quindi è stata applicata l'operazione $(class - 2)/2$, ottenendo valori compresi tra -1 e 1;
- Sono stati generati dataset contenenti tutti i sottoalberi (togliendo le foglie) dei dataset con etichette di sentimenti.

Il codice utilizzato per generare i dataset è disponibile nella repository Rosso 2024. Non sono stato in grado di generare i sottoalberi per i dataset con etichette di sintassi, poiché gli alberi di sintassi differiscono per poco da

quelli di sentimenti, per cui diventa complesso assegnare il target corretto ai sottoalberi. Per esempio, un un albero di sintassi divide la parola "fully-written" in tre foglie "fully", "-", "written", mentre nell'albero di sentimenti è una sola foglia con etichetta "fully-written". Questa non è l'unica differenza che ho notato, ma anche una differenza così piccola rende complesso da abbinare i sottoalberi alla golden label corretta.

3 Modello

Per l'allenamento dei modelli, è stato sviluppato un programma che utilizza i dataset forniti. L'allenamento complessivo ha avuto una durata media di circa 4 ore per ciascun modello. Tuttavia, il modello allenato per la regressione dei sottoalberi del dataset ha richiesto un tempo di allenamento significativamente maggiore, pari a 20 ore circa.

Il metodo utilizzato per l'allenamento è stato esclusivamente il partial tree kernel, sviluppato da Alessandro Moschitti Moschitti 2006. Questo metodo è stato scelto per le sue buone prestazioni e per i tempi di allenamento ridotti rispetto ad altri kernel.

4 Risultati

Le differenze tra i modelli sono date dai diversi dataset utilizzati nell'allenamento. I test sono stati eseguiti con il medesimo dataset per quanto riguarda la classificazione di ciascun nodo dell'albero; mentre è stata sottoposta la medesima struttura dell'addestramento per valutare i modelli sulla radice dell'albero per ciascun albero.

4.1 Binary

Modello	Accuracy	Precision	Recall	F1
Subtree sentiment all	0.98	0.99	0.99	0.99
Subtree sentiment root	0.94	0.94	0.94	0.94
Syntax and sentiment all	0.96	0.97	0.95	0.96
Syntax and sentiment root	0.94	0.94	0.94	0.94
Sentiment all	0.95	0.97	0.95	0.96
Sentiment root	0.94	0.95	0.94	0.94
Syntax all	0.45	0.00	0.00	0.00
Syntax root	0.74	0.72	0.78	0.75

Tabella 1: Risultati della classificazione binaria

4.2 Regression

Modello	Mean Absolute Error	Mean Squared Error	R-Squared
Subtree sentiment all	-	-	-
Subtree sentiment root	-	-	-
Syntax and sentiment all	0.47	0.41	-0.94
Syntax and sentiment root	0.94	1.28	-1.93
Sentiment all	0.61	0.66	-2.14
Sentiment root	1.17	1.95	-3.48
Syntax all	0.49	0.35	-0.65
Syntax root	0.56	0.48	-0.11

Tabella 2: Risultati della regressione

In questo caso mancano i risultati per i modelli addestrati sui sottoalberi perché si sta ancora aggiornando, al momento il modello è arrivato a 20 ore di allenamento, stimo che impiegherà 25 ore in totale.

4.3 Multiclass

Modello	Accuracy	Precision	Recall	F1
Subtree sentiment all	0.95	0.92	0.91	0.92
Subtree sentiment root	0.51	0.50	0.48	0.49
Syntax and sentiment all	0.44	0.48	0.59	0.44
Syntax and sentiment root	0.53	0.52	0.49	0.49
Sentiment all	0.40	0.45	0.56	0.40
Sentiment root	0.54	0.54	0.50	0.50
Syntax all	0.18	0.04	0.20	0.06
Syntax root	0.39	0.38	0.34	0.34

Tabella 3: Risultati della classificazione multiclasse

4.4 RNN

Modello	Accuracy	Precision	Recall	F1
RNN 50 unità nascoste all	0.79	0.65	0.53	0.57
RNN 50 unità nascoste root	0.42	0.44	0.39	0.40
Subtree sentiment all	0.95	0.92	0.91	0.92
Subtree sentiment root	0.51	0.50	0.48	0.49
Sentiment all	0.40	0.45	0.56	0.40
Sentiment root	0.54	0.54	0.50	0.50

Tabella 4: Risultati della classificazione multiclasse con RNN rispetto al kernel method

Si noti che l'RNN e il kernel method allenato sul sentiment sono stati allenati con lo stesso dataset. Tutti e tre i modelli qui presentati sono stati valutati sullo stesso dataset di test.

5 Conclusioni

5.1 Sentiment vs Syntax vs Subtree

I modelli focalizzati sulla sintassi hanno mostrato prestazioni inferiori rispetto agli altri due approcci. Questa differenza è particolarmente evidente nell'analisi su tutti i nodi dell'albero. Tale risultato è prevedibile dato che il treebank di test utilizza etichette di riferimento non incluse nella formazione dei modelli, rendendo il confronto meno rilevante.

Nonostante ciò, anche limitando l'analisi alla radice dell'albero, i modelli basati sulla sintassi hanno registrato performance inferiori. Tale tendenza non si verifica nella classificazione dei pattern basata sulla regressione.

Al contrario, i modelli che integrano l'analisi del sentiment con quella sintattica tendono a superare quelli basati esclusivamente sul sentiment, anche se i risultati sono simili quando si focalizzano sulla predizione del target alla radice dell'albero.

In aggiunta, i modelli che operano sui sottoalberi hanno dimostrato un'efficacia notevolmente superiore nella classificazione dei pattern su ogni nodo, pur mantenendo risultati comparabili a quelli dei modelli basati sul sentiment per quanto riguarda la radice.

5.2 Kernel Methods vs RNN

Le tecniche basate su Kernel Methods, quando applicate ai sottoalberi, hanno evidenziato performance superiori rispetto alle RNN con 50 unità nascoste. Le RNN, però, si distinguono nettamente nell'identificazione dei pattern su ciascun nodo dell'albero rispetto agli altri modelli.

Si osserva, inoltre, che nei risultati relativi alla radice dell'albero, i modelli che combinano il sentiment con i Kernel Methods superano significativamente le RNN (0.12), evidenziando un divario maggiore di quello notato quando le RNN confrontano i loro risultati con quelli dei Kernel Methods allenati sulla sintassi (0.3).

Riferimenti bibliografici

- [Mos06] Alessandro Moschitti. *Tree Kernels in SVM-light*. 2006. URL: <https://disi.unitn.it/moschitti/Tree-Kernel.htm> (visitato il 11/07/2024).
- [NLP24] NLPbox. *stanford-corenlp-docker*. 2024. URL: <https://github.com/NLPbox/stanford-corenlp-docker> (visitato il 11/07/2024).
- [Ros24] Carlo Rosso. *bachelor-s_notes*. 2024. URL: https://github.com/danesinoo/bachelor-s_notes (visitato il 11/07/2024).