Machine Learning per la classificazione di referti anatomopatologici in testo libero

Stefano Martina[1], Leonardo Ventura[2], Paolo Frasconi[1]

[1]Università degli Studi di Firenze, [2]Istituto per lo Studio la Prevenzione e la Rete Oncologica Toscana

stefano.martina@unifi.it, l.ventura@ispro.toscana.it, paolo.frasconi@unifi.it

Abstract

Lo scopo di questo lavoro è di fornire un supporto automatico alla classificazione di casi di cancro usando metodi di *machine learning*. Sono stati realizzati cinque modelli, le cui performance sono state valutate su quattro task di classificazione su dati del Registro Tumori della Toscana (RTT).

1 Background

Il cancro è una delle maggiori preoccupazioni, riduce la qualità di vita e porta ad una mortalità prematura. La misura del carico della malattia è una delle maggiori proccupazioni della sanità pubblica, sono necessarie misure per descrivere lo stato generale della salute della popolazione, per stabilire gli obbiettivi e valutare le performances del sistema sanitario.

I Registri Tumori (RT) sono nati negli ultimi decenni in qualità di strumento strategico per quantificare l'impatto della malattia e per fornire dati analitici. I RT usano fonti cliniche e amministrative con lo scopo di identificare e classificare nuove diagnosi di cancro. Il processo di registrazione è basato parzialmente sulla revisione manuale dei referti di Anatomia Patologica (AP).

I casi di tumore vengono classificati secondo le linee guida del sistema International Classification of Diseases for Oncology Third edition (ICD-O-3). Tale sistema è usato a livello mondiale per informare il controllo dei tumori, l'attivita di ricerca, il planning del trattamento e l'economia della salute.

In questo lavoro abbiamo realizzato cinque modelli di machine learning per mitigare i ritardi dovuti all'intervento umano permettendo di sfruttare i dati prima della loro classificazione manuale. Questo permette ad esempio di monitorare in modo tempestivo l'aderenza del sistema sanitario ai percorsi diagnostico terapeutici oncologici, valorizzando ulteriormente i dati processati dal RTT.

2 Materiali e metodi

2.1 Dataset

abbiamo raccolto un set di oltre $1\ 500\ 000$ record dal registro RTT per cui i referti di AP erano disponibili. Un subset di questi di oltre $90\ 000$ records avevano una diagnosi positiva di tumore con codici topologici e morfologici ICD-O-3 associati.

I record istologici sono costituiti da tre campi di testo libero che riportano *macroscopia*, *diagnosi* e in alcuni casi l'*anamnesi* del paziente. Abbiamo unito i tre campi di testo in uno singolo, in caso contrario abbiamo osservato performances peggiori dovute a rumore nei dati e ad un maggior numero di feature.

2.2 Task di classificazione

Un codice ICD-O-3 topografico è strutturato come *C__._* dove le prime due cifre rappresentano il sito e la terza il sottosito. E.g. *C50.2* rappresenta il quadrante superiore interno (2) del seno (50).

Un codice ICD-O-3 morfologico è strutturato come *____/_* dove le prime quattro cifre rappresentano il tipo di cellula e l'ultima il comportamento. E.g. *8140/3* rappresenta un adenocarcinoma (adeno 8140; carcinoma 3).

In base a tale struttura, abbiamo considerato quattro task di classificazione multiclasse dei referti istologici. Se X è la distribuzione del testo dei referti, Y^s del sito, Y^f del sito completo (sito e sottosito), Y^t del tipo e Y^c del comportamento, allora i quattro task consistono nello stimare le distribuzioni:

- T^s stima $P(Y^s|X)$;
- T^f stima $P(Y^f|X)$;
- T^t stima $P(Y^t|X)$;
- T^c stima $P(Y^c|X)$.

I quattro task hanno un numero variabile di classi, T^s ha 70 classi, T^f 284, T^t 434 e T^c 5. Il dataset è anche fortemente sbilanciato con poche classi di tumore che coprono la maggior parte dei documenti e una lunga coda di documenti corrispondenti a tumori rari.

2.3 Word Vectors

L'approccio classico per le rappresentazioni dei documenti è *bag-of-words* [Manning *et al.*, 2008]. Questo approccio soffre due problemi fondamentali: primo, l'ordine relativo dei termini si perde rendendo impossibile sfruttare la struttura sintattica della frase; secondo, parole distinte hanno rappresentazioni ortogonali anche quando sono semanticamente vicine. Approcci più moderni sfruttano lo spazio vettoriale della rappresentazione delle parole per risolvere tali problemi. I *word vectors* come *word2vec* [Mikolov *et al.*, 2013] o Global Vectors (GloVe) [Pennington *et al.*, 2014] sono costruiti in

modo che analogie tra parole sono direttamente codificate nello spazio vettoriale.

È pratica comune usare librerie pre compilate di word vector addestrate su miliardi di token estratti da sorgenti. Queste librerie sono concepite per testi generici in inglese, i referti istologici oltre ad essere in italiano, impiegano una terminologia specifica. Nel nostro lavoro abbiamo addestrato dei word vectors con GloVe sui testi dei quasi 1 500 000 record non etichettati, in modo da sfruttarli.

2.4 SVM e Neural Networks

Nel nostro lavoro abbiamo realizzato classificatori che usano Support Vector Machine (SVM), Convolutional Neural Network (CNN) [LeCun *et al.*, 2015] e un tipo specifico di Recurrent Neural Network (RNN): Long Short-Term Memory (LSTM) [Hochreiter e Schmidhuber, 1997]. Le RNN sono un tipo di reti neurali specifiche per processare sequenze, queste hanno uno stato interno che dipende dell'input corrente e dal valore dello stato al passo precedente. Il maggior problema delle RNN è il *vanishing gradient* che è affrontato in LSTM con l'uso di *gates*.

3 Esperimenti

Abbiamo realizzato cinque modelli:

T-SVM SVM con rappresentazioni bag of words usando unigrammi;

T2-SVM SVM con bag of words usando unigrammi e bigrammi;

T2-2LSTM due layer LSTM bidirezionali con bag of words usando bigrammi;

G-CLSTM un layer convoluzionale e uno LSTM bidirezionale con rappresentazioni GloVe;

G-2LSTM due layer LSTM bidirezionale con GloVe.

Per sfruttare tutto il dataset abbiamo usato una politica *leave-one-out ten-fold cross validation* in cui il dataset è diviso in dieci parti e, a turno, l'addestramento è eseguito su nove di queste e il test su di una. per ogni modello e task, le metriche sono calcolate per ciascun test e sono espresse in tabella 1 come media e deviazione standard.

Oltre all'*accuratezza*, viene riportato il *kappa* score che misura quanto l'annotatore umano e quello automatico sono in accordo. *MAPc* e *MAPs* sono misure di information retrieval [Manning *et al.*, 2008], rispettivamente valutano come il modello si comporta nel recuperare record di una specifica classe, e la classe corretta per uno specifico sample. Precision, recall, ed f1 score sono indicati come *macro average* tra le diverse classi, dove ogni classe ha lo stesso peso. I risultati nel caso di *micro average*, dove la media viene fatta su tutti i sample, sono uguali a quelli di accuratezza, e quindi sono omessi dalla tabella.

4 Discussione

Considerando i modelli che usano SVM (***T-SVM*** e ***T2-SVM***), non ci sono miglioramenti degni di nota usando bigrammi invece che unigrammi. Le performance di LSTM non sono migliori di quelle di SVM quando vengono usati i bag-of-words (nel modello ***T2-2LSTM***). Quando vengono sfruttati invece i dati non etichettati con GloVe si nota un miglioramento. Le performance dei due modelli che usano glove sono comparabili tra loro, è comunque migliore quello che usa due layer LSTM.

T^s	*T-SVM*	*T2-SVM*	*T2-2LSTM*	*G-CLSTM*	*G-2LSTM*
acc.	89.8 ± 2.0	89.8 ± 2.0	88.6 ± 2.0	90.0 ± 1.6	$\underline{90.5} \pm 1.6$
kappa	88.5 ± 2.2	88.6 ± 2.3	87.2 ± 2.3	88.9 ± 1.8	$\underline{89.3} \pm 1.8$
MAPs	93.0 ± 1.5	93.0 ± 1.5	92.2 ± 1.5	93.5 ± 1.2	$\underline{93.8} \pm 1.1$
MAPc	61.6 ± 3.9	61.3 ± 4.0	55.7 ± 3.7	62.7 ± 3.5	$\underline{64.1} \pm 4.1$
pre.	$\underline{65.5} \pm 4.8$	64.7 ± 3.2	55.0 ± 2.8	61.5 ± 3.4	61.8 ± 3.7
rec.	55.7 ± 4.1	54.7 ± 3.8	51.6 ± 3.2	56.5 ± 3.0	$\underline{58.1} \pm 3.5$
f1	$\underline{58.4} \pm 4.1$	57.5 ± 3.6	52.1 ± 3.1	57.0 ± 2.7	58.2 ± 3.3

T^f	*T-SVM*	*T2-SVM*	*T2-2LSTM*	*G-CLSTM*	*G-2LSTM*
acc.	68.4 ± 2.3	68.7 ± 2.0	67.4 ± 1.7	70.1 ± 2.1	$\underline{70.9} \pm 2.0$
kappa	66.5 ± 2.4	66.8 ± 2.1	65.6 ± 1.7	68.4 ± 2.2	$\underline{69.3} \pm 2.1$
MAPs	78.4 ± 1.9	78.4 ± 1.7	78.5 ± 1.3	80.6 ± 1.4	$\underline{81.3} \pm 1.4$
MAPc	43.1 ± 2.2	43.4 ± 2.2	36.8 ± 2.3	42.9 ± 2.6	$\underline{45.0} \pm 2.0$
pre.	41.4 ± 1.6	$\underline{41.6} \pm 1.5$	33.0 ± 2.8	38.7 ± 3.1	39.8 ± 2.3
rec.	35.7 ± 1.9	35.1 ± 2.1	32.0 ± 2.5	36.6 ± 3.0	$\underline{38.0} \pm 2.2$
f1	36.6 ± 1.5	36.4 ± 1.7	31.2 ± 2.3	35.9 ± 2.9	$\underline{37.3} \pm 2.1$

T^t	*T-SVM*	*T2-SVM*	*T2-2LSTM*	*G-CLSTM*	*G-2LSTM*
acc.	81.9 ± 1.9	82.9 ± 2.0	82.8 ± 1.4	84.6 ± 1.4	$\underline{84.9} \pm 1.5$
kappa	79.5 ± 2.2	80.7 ± 2.3	80.6 ± 1.6	82.7 ± 1.6	$\underline{83.0} \pm 1.7$
MAPs	87.8 ± 1.3	88.6 ± 1.4	88.7 ± 1.0	90.3 ± 0.9	$\underline{90.6} \pm 1.0$
MAPc	62.4 ± 1.6	64.4 ± 1.8	55.1 ± 3.1	64.2 ± 1.9	$\underline{65.9} \pm 1.9$
pre.	56.1 ± 2.4	$\underline{58.3} \pm 1.9$	47.0 ± 3.3	56.5 ± 1.8	57.0 ± 2.6
rec.	51.1 ± 2.6	52.2 ± 2.2	47.0 ± 2.6	56.8 ± 2.2	$\underline{58.6} \pm 2.0$
f1	51.4 ± 2.5	52.9 ± 1.9	45.0 ± 2.9	54.6 ± 1.9	$\underline{55.5} \pm 2.3$

T^c	*T-SVM*	*T2-SVM*	*T2-2LSTM*	*G-CLSTM*	*G-2LSTM*
acc.	95.9 ± 1.0	96.0 ± 1.1	94.1 ± 3.0	94.4 ± 4.2	$\underline{96.5} \pm 0.8$
kappa	82.3 ± 4.6	82.8 ± 5.0	70.4 ± 25.5	67.6 ± 35.9	$\underline{85.6} \pm 3.4$
MAPs	97.7 ± 0.6	97.8 ± 0.6	96.6 ± 1.8	96.8 ± 2.5	$\underline{98.1} \pm 0.5$
MAPc	85.4 ± 5.9	85.9 ± 5.7	71.4 ± 18.4	75.5 ± 26.4	$\underline{89.5} \pm 4.2$
pre.	87.0 ± 5.0	$\underline{87.9} \pm 4.8$	69.9 ± 19.9	72.7 ± 27.1	85.5 ± 4.0
rec.	78.6 ± 7.3	78.6 ± 7.4	67.6 ± 17.4	72.1 ± 25.4	$\underline{85.9} \pm 4.9$
f1	81.7 ± 6.3	82.0 ± 6.3	68.0 ± 18.5	72.1 ± 26.0	$\underline{85.5} \pm 4.2$

Tabella 1: Risultati per i quattro task dei cinque modelli.

I modelli possono anche essere usati in un contesto di information retrieval, recuperando casi di tumore con valori variabili di recall e precision. Possono essere costruite delle curve recall-precision (non riportate) per calcolare la soglia di classificazione.

Questo lavoro mostra le pontenzialità del machine learning in un contesto non ancora esplorato in Italia.

Riferimenti bibliografici

[Hochreiter e Schmidhuber, 1997] Sepp Hochreiter e Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, e Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Manning *et al.*, 2008] Christopher D. Manning, Prabhakar Raghavan, e Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Mikolov *et al.*, 2013] Tomas Mikolov, Wen-tau Yih, e Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Hlt-naacl*, volume 13, pages 746–751, 2013.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, e Christopher D. Manning. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 00365.