Rapporto Attività di Dottorato
Anno 2002
Menchetti Sauro

Dottorando

Menchetti Sauro
Iscritto al Dottorato di Ricerca in Ingegneria Informatica e dell'Automazione
Dipartimento di Sistemi e Informatica dell'Università degli Studi di Firenze
Ciclo XVII, Anno 2002, Primo Anno di Corso
Tutori: Prof. Paolo Frasconi, Prof. Giovanni Soda
Email: menchett@dsi.unifi.it
Web: http://www.dsi.unifi.it/~menchett

Attività di Ricerca

L'attività di ricerca svolta durante il primo anno di dottorato e che interesserà anche gli altri due anni, riguarda il settore del Machine Learning, una branca dell'Intelligenza Artificiale che si occupa di creare programmi in grado di "imparare" da esempi.

Il lavoro sviluppato tratta due diversi argomenti:

la categorizzazione del testo;

il problema del Parsing nel Natural Language Processing.

Entrambi gli argomenti sono di notevole importanza in un contesto in cui sempre più informazioni vengono mantenute in formato elettronico (basti ad esempio pensare alle Digital Libraries ed al Web). Nasce quindi il bisogno di avere degli strumenti in grado di trattare in modo automatico queste grandi quantità di dati: le tecniche standard dell'informatica come, ad esempio, le basi di dati, non sono più in grado di soddisfare le richieste sempre più complesse degli utenti e servono quindi degli strumenti capaci di eseguire compiti che fino ad oggi erano di competenza solo degli umani. La categorizzazione del testo si prefigge lo scopo di classificare i documenti in un insieme di classi di appartenenza, mentre il parsing cerca di creare l'albero sintattico di una frase. Queste due metodologie combinate insieme posso contribuire alla comprensione del linguaggio naturale da parte delle macchine.

Parallelamente ho realizzato un software generico per il trattamento di strutture dati con algoritmi che utilizzano i kernels come l'SVM ed il Voted Perceptron. Tale software è facilmente estendibile a qualsiasi struttura dati ed indipendente dall'algoritmo utilizzato.

Metodologia

Per quanto riguarda la categorizzazione del testo, si è fatto riferimento ad un dataset disponibile online che raccoglie circa 20.000 articoli battuti dall'agenzia di stampa Reuters. Si tratta di un problema multiclasse in cui ogni documento può appartenere a più classi distinte. Ho impiegato l'SVM, un classificatore molto promettente che si fonda su una solida base teorica. I problemi da risolvere sono stati molteplici: riduzione del problema da multiclasse a binario, bilanciamento tra dati positivi e negativi, vari modelli di documento (binomiale, multinomiale), feature selection, generazione di una probabilità a posteriori, combinazione dei margini tramite alberi di decisione e perceptron, etc..

Il lavoro sul parsing ha impiegato degli algoritmi in grado di trattare strutture dati complesse ed ha utilizzato un treebank che raccoglie circa 40.000 frasi tratte dal Wall Street Journal. Sono stati affrontati il problema del parsing incrementale e di quello completo. Innanzitutto ho implementato un algoritmo in grado di elaborare strutture dati ad albero: si è fatto riferimento al Voted Perceptron con un kernel per gli alberi. Il problema della scelta del miglior albero sintattico è stato risolto con un modello a preferenze che è stato inserito nel precedente algoritmo. A questo è seguita un'attenta sperimentazione ed un confronto con un altro algoritmo implementato per lavorare sullo stesso problema. Il problema del parsing completo è invece ancora in corso d'opera.

Risultati

La prima parte del lavoro è riuscita a replicare i risultati allo stato dell'arte che sono riportati in letteratura, introducendo anche qualche piccolo miglioramento. I miglioramenti prodotti non sono stati sufficientemente rilevanti da essere pubblicati. Il lavoro è comunque servito ad apprendere le metodologie e le tecniche del Machine Learning.

Molto più interessanti invece i risultati ottenuti sul parsing: gli algoritmi con kernel, ritenuti i più promettenti classificatori presenti sul campo, hanno dimostrato invece qualche difficoltà se paragonati ad esempio alle Recurrent Neural Networks (RNNs). Questa difficoltà degli algoritmi con kernel sembra imputabile alla incapacità di creare una rappresentazione che si adatti ai dati in modo automatico, presente invece nelle RNNs. Questo rappresenta un messaggio in controtendenza rispetto alle attese e suggerisce quindi che anche gli algoritmi con i kernels dovrebbero cercare di creare una rappresentazione che automaticamente si adatti ai dati, invece di utilizzare una rappresentazione che viene imposta a priori.

Collaborazioni

Il lavoro sul parsing è stato realizzato in collaborazione con Fabrizio Costa. Per quanto riguarda la parte più algoritmica e quindi meno dipendente dal dominio in esame, ho lavorato con Massimiliano Pontil. Tutto il lavoro è stato supervisionato e guidato dai Prof. Paolo Frasconi e Giovanni Soda.

Pubblicazioni

Una breve descrizione del lavoro svolto sulla categorizzazione del testo è stata riportata sul libro di Pierre Baldi, Paolo Frasconi e Padhraic Smyth intitolato *The Internet and the World Wide Web. Probabilistic Modeling and Algorithms*.

Il lavoro sul parsing è stato presentato a dicembre in un workshop di NIPS 2002.

Partecipazioni a Corsi, Convegni e Seminari

Corsi

Corso *Statistical Learning and Kernel-based Algorithms* tenuto dal Prof. Massimiliano Pontil

Corso *Algoritmi di Ottimizzazione Non Lineare (Locale e Globale)* tenuto dal Prof. Fabio Schoen

Convegni

Partecipazione all'Ottavo Convegno dell'Associazione Italiana per l'Intelligenza Artificiale AI*IA tenutosi a Siena

Seminari

Seminario su algoritmi di apprendimento automatico per la predizione dello stato di legame delle cisteine tenuto dal Prof. Rita Casadio (CIRB, Università di Bologna)

Seminario sul tema delle Digital Libraries tenuto dal Prof. Ian Witten (Università di Waikato, NZ), autore del libro *Managing Gigabytes*

Attività di Ricerca per il Secondo Anno

Durante il secondo anno di dottorato, giungerà a termine il lavoro sul parsing che propone un confronto tra gli algoritmi con kernel e gli algoritmi adattivi come le RNNs. I risultati ottenuti sembrano promettenti e quindi parte del lavoro sarà dedicata all'approfondimento di questi argomenti. Saranno poi approfonditi gli aspetti teorici e algoritmici impiegati nel lavoro. Una probabile direzione di ricerca sarà quella dei kernels adattivi, cioè di algoritmi in grado di creare una rappresentazione automatica dei dati, il tutto applicato sempre alle strutture dati. Durante questo secondo anno verrà definito anche il tema per la tesi di dottorato: probabilmente l'argomento avrà a che fare con gli algoritmi che utilizzano kernel applicati a strutture dati complesse, mentre il dominio applicativo potrà essere legato al parsing o alla categorizzazione del testo.