Relazione Scientifica sui Risultati dell'Attività di Ricerca Svolta da Alfredo Cuzzocrea nell'ambito del Programma di Mobilità di Breve Durata CNR 2009

Questa relazione scientifica presenta i risultati dell'attività di ricerca svolta da Alfredo Cuzzocrea presso il Computer Science Department della University of California, Los Angeles (UCLA) dal 26 Ottobre 2009 al 15 Novembre 2009 sul tema "Data Mining Algorithms for Data Streams" in collaborazione con il Prof. Carlo Zaniolo nell'ambito del programma di "Mobilità di Breve Durata" CNR 2009.

Le attività di ricerca sono state focalizzate sulla definizione e la valutazione sperimentale di algoritmi di Data Mining per la scoperta di conoscenza da data streams, con particolare riguardo all'individuazione di tecniche algoritmiche in grado di estendere le attuali capacità dei linguaggi di interrogazione ed estrazione di conoscenza per data streams basati su sintassi SQL. I domini applicativi di interesse hanno riguardato in particolare l'estrazione di patterns e regolarità da data streams e il Data Mining su data streams multi-dimensionali.

Le motivazioni di base di tale attività di ricerca hanno tratto origine dalle limitazioni degli attuali approcci di Data Mining su data streams, che sono principalmente orientati all'applicazione di tecniche tradizionali di Data Mining sviluppate nel contesto delle Basi di Dati Relazionali (ad esempio, Classificazione, Clustering, Frequent Itemset Mining, ecc.), e che, come dimostrato da alcuni recenti risultati scientifici, risultano essere inadeguati rispetto alle peculiari caratteristiche dei data streams. Questa inadeguatezza riguarda non solo l'efficacia ma anche l'efficienza di tali tecniche tradizionali.

In maggior dettaglio, durante tale attività di ricerca sono stati introdotti e caratterizzati, sia dal punto di vista formale sia da quello sintattico, i due seguenti innovativi problemi di Data Stream Mining, per i quali sono state definite alcune soluzioni efficaci ed efficienti:

OLAPing Uncertain, Imprecise and Imperfect Multidimensional Data Streams from Multiple Sources Questo problema focalizza sulla definizione di modelli, tecniche ed algoritmi capaci di calcolare in modo efficace ed efficiente aggregazioni OLAP (On Line Analytical Processing) su data streams multidimensionali incerti, imprecisi ed imperfetti prodotti da sorgenti multiple. Vi sono numerosi

applicazioni reali che possono essere classificate come istanze di questo problema generale. Ad esempio, le reti di sensori di parametri ambientali (temperatura, pressione, umidità relativa, ecc.). In tale scenario applicativo, i dati prodotti dalla rete di sensori sono infatti caratterizzati dalle seguenti proprietà: (*i*) *multidimensionalità*, che si riferisce al fatto che gli attributi caratterizzanti il modello dati associato alle sorgenti (sensori) definiscono intrinsecamente uno spazio multi-dimensionale; (*ii*) *incertezza*, che cattura il fatto che le letture (readings) prodotte dalle sorgenti non possono essere modellate come grandezze esatte ma piuttosto mediante intervalli di confidenza (di valori) cui sono associate delle probabilità; (*iii*) *imprecisione*, che si riferisce al fatto che, in scenari applicativi reali, non è possibile determinare con precisione l'origine delle letture (ad esempio, si pensi ad un sistema GPS con fattori di errore/rumore) ma piuttosto si riscontrano fattori di imprecisione spazio-temporale; (*iv*) *imperfezione*, che modella il fatto che i modelli dati multi-dimensionali associati alle sorgenti possono rappresentare una partizione del modello dati multi-dimensionale associato alle aggregazioni OLAP da calcolare, anziché essere coincidenti ad esso.

Per tale problema, la soluzione definita consiste nell'applicare tecniche probabilistiche al fine di stimare l'incertezza e l'imprecisione delle letture mediante i principi e gli strumenti delle *Funzioni di Distribuzione di Probabilità*, e tecniche basate sulla *Regressione Lineare* al fine di "completare" aggregazioni OLAP multi-dimensionali "imperfette" a partire dai domini degli spazi multi-dimensionali definiti dalle partizioni di dimensioni disponibili nei modelli dati associati alle sorgenti.

Gradient Analysis over Multidimensional Data Streams Questo problema focalizza sulla definizione di modelli, tecniche ed algoritmi capaci di supportare in modo efficace ed efficiente *analisi gradiente* su data streams multi-dimensionali. L'analisi gradiente di uno dominio dati multi-dimensionale consiste nell'individuare variazioni "globali" sui valori dei dati di tale dominio, cioè variazioni che si verificano rispetto tutte le dimensioni dello spazio multi-dimensionale definito dal dominio, oppure ad una loro partizione (che viene assunta, in questo caso specifico, come input dell'analisi gradiente). Sono state sviluppate in letteratura numerose tecniche per implementare efficacemente ed efficientemente analisi gradiente su domini multi-dimensionali statici, ma nessun approccio ha sinora focalizzato l'attenzione sul caso ben più interessante rappresentato da domini multi-dimensionali prodotti da sorgenti dati tempo-variante come, appunto, i data streams. Rispetto a tale problematica di ricerca, le maggiori limitazioni provengono dal fatto che non è possibile applicare algoritmi

multi-step per il calcolo del gradiente su data streams, poiché in genere i motori di interrogazione di data streams dispongono di un buffer limitato che consente loro di processare solo un set finito delle "ultime" letture dello stream di input, mentre tutte le altre letture precedenti vengono perse, a meno che tali letture non siano memorizzate in strutture di memorizzazione secondarie o terziarie (comunque, questa ultima soluzione renderebbe l'accesso a letture "vecchie" dello stream di input molto problematico in eguale misura).

Per tale problema, la soluzione definita consiste nell'applicare tecniche "ottimizzate" degli algoritmi di analisi gradiente per dati statici, al fine di ottenere un trade-off tra efficacia ed efficienza. In particolare, è stata definita sia la possibilità di calcolare il gradiente dello stream di input mediante l'applicazione dei cosiddetti *Algoritmi Approssimati*, algoritmi, cioè, che modificano o arrestano dinamicamente le loro esecuzioni in dipendenza del risultato "attuale" in output rispetto ad una soglia di accuratezza predefinita, sia la possibilità di ottenere delle particolari accezioni di *Gradiente Probabilistico* o *Gradiente Troncato* quali soluzioni "approssimate" del calcolo del gradiente "totale". In particolare, questa seconda alternativa prevede di calcolare il gradiente dello stream di input mediante tecniche di *Inferenza Probabilistica*, oppure a partire da partizioni multi-dimensionali dello stream di input anziché dalla sua versione "globale".

I risultati conseguiti hanno soddisfatto pienamente gli obiettivi del Programma di Ricerca in oggetto e, soprattutto, costituiscono la base per un successivo sviluppo della collaborazione di ricerca tra i soggetti partecipanti al suddetto Programma di Mobilità di Breve Durata.

Firma del Proponente

Firma del Fruitore