

Profilo:

Progettazione ed implementazione di algoritmi di record linkage per il miglioramento della qualità dei dati, mediante approcci di apprendimento automatico

Contesto di lavoro:

Direzione Banche Dati, Sistemi Decisionali, Atenei – Area Dati in Rete

Caratteristiche:

- Conoscenze di basi di dati, tematiche connesse alla qualità dei dati, algoritmi di classificazione

Attività prevalenti inerenti il ruolo:

- Approfondimento delle metodologie di data quality
- Utilizzo di tools di data quality e data mining
- Progettazione ed implementazione di algoritmi di classificazione da utilizzare per la deduplicazione di record

Candidature richieste: n° 1**Sede svolgimento stage:**

CSI-PIEMONTE C.so Unione Sovietica, 216 – 10134 Torino

Durata dello stage:

4-6 mesi full-time.

Facilitazioni previste:

Utilizzo gratuito della mensa.

Riferimenti aziendali:

Dott. Andrea Muraca – Area Dati in Rete
Dott.ssa Laura Caimotto

Tel. 011 316.85.30
Tel. 011 316.83.42

Torino, 23/03/2005

Descrizione Stage

Nel campo del miglioramento della qualità dei dati, il noto problema del record linkage è uno dei più studiati dal punto di vista algoritmico. Il problema consiste nel determinare, per una coppia di record omogenei in un database, se essi rappresentino la stessa entità del mondo reale. Questa determinazione avviene valutando la similarità di campi corrispondenti del record, ed applicando varie tecniche per determinare se gli interi record sono da considerarsi sufficientemente simili da essere dichiarati un *match*.

Tradizionalmente, l'output degli algoritmi di record linkage consiste nell'assegnazione di ogni coppia di records nel database che si intende de-duplicare, ad una fra tre possibili classi: *match*, *non-match* e *don't know*. Quest'ultima classe è riservata ai casi per i quali non esiste sufficiente evidenza (statistica o di altra natura) che consenta di concludere se i due record siano o meno un *match*. Le coppie in questo insieme dovranno essere esaminate da esperti, operazione normalmente costosa. L'obiettivo degli algoritmi presentati in letteratura è quindi, generalmente, di controllare il trade-off tra precisione delle assegnazioni alle classi *match* e *non-match*, e ampiezza della classe degli incerti. La precisione della classificazione si misura in percentuale di falsi positivi e falsi negativi prodotti dal classificatore su un test set adeguato. Questo problema è stato affrontato nella letteratura scientifica utilizzando approcci diversi, di tipo sia deterministico che probabilistico [2,3,4]. Di recente, l'utilizzo di tecniche di machine learning (apprendimento automatico) è stato sperimentato a livello prima di ricerca e poi commerciale [1].

Nel contesto CSI Piemonte, questo problema assume particolare rilevanza per quanto riguarda la de-duplicazione di record in database di forte interesse per la Pubblica Amministrazione regionale. In questo senso, CSI Piemonte è attualmente impegnato in diversi progetti di re-ingegnerizzazione che comprendono, tra l'altro, la de-duplicazione di record in database di dimensioni significative. Nell'affrontare il problema in questo ambito, sarà quindi necessario tenere presente i vincoli tecnologici di progetto esistenti, in particolare prevedendo l'utilizzo di strumenti di data quality analysis and cleaning che fanno parte della piattaforma di analisi SAS.

L'obiettivo dello stage è la sperimentazione dell'applicazione di tecniche di classificazione automatica al problema del record linkage, in modo da sfruttare ed estendere le funzionalità offerte dalla piattaforma SAS. Questo obiettivo diverge in modo significativo dalla attuali proposte di ricerca che prevedono l'utilizzo di tecniche di machine learning.

La sperimentazione sarà condotta su database reali gestiti dal CSI, e comprenderà le seguenti fasi:

- progettazione di una codifica dell'informazione di similarità tra record in termini di vettori di attributi, che utilizzi gli strumenti di verifica di matching tra campi offerte dal tool di data quality della SAS;
- progettazione ed implementazione di un algoritmo per la generazione semi-automatica di training set e test set di tali vettori a partire da una popolazione significativa del database (circa 200,000 record);
- progettazione ed implementazione di un algoritmo di supervised classification per la de-duplicazione dei records, nel contesto CSI descritto sopra, che operi sui training set dati, e verifica della sua performance;
- studio di fattibilità per l'integrazione dell'approccio all'interno delle procedure di analisi di SAS attualmente utilizzate nell'ambito dei progetti CSI riguardanti la qualità dei dati.
- eventuale sperimentazione utilizzo del tool di data mining della SAS a supporto delle fasi sopraelencate.

Riferimenti

[1] Andrew Borthwick , Maggie Soffer, Business Requirements of A Record Matching System, ICIQ04, Cambridge, MA, Nov. 2004.

[2] I.P. Fellegi and A.B. Sunter., A theory for record linkage. Journal of the American Statistical Association, 64, 1969.

[3] W.E.Winkler, Exact matching lists of businesses: Blocking, subfield identification, information theory. In Alvey and Kills, editors, Record Linkage Techniques. US Internal Revenue Service, 1985.

[4] W. Winkler. Methods for record linkage and bayesian networks. Technical report, U.S. Census Bureau, Statistical Research Division, 2002.