Data fusion e associazione in dati provenienti da crowdmapping

Domenico Vito([a])

([a]) Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, *domenico.vito@polimi.it*

Abstract: L'aggregazione dell'informazione geografica proveniente dall'integrazione di banche dati tradizionali e approcci di crowdsourcing produce una quantità di dati che vanno opportunamente gestiti. L'acquisizione diffusa di questo tipo di informazione implica in particolare due problemi: lo storage di quantità di informazioni da fonti eterogenee e l'attribuzione di senso al dato rispetto all'utenza finale.
Il primo problema rientra nella categoria dei problemi di "data fusion".
Il "data fusion" può essere interpretato come quel processo multilivello che si occupa dell'associazione, della correlazione, della combinazione di dati e informazioni da fonti singole e multiple all'interno di una struttura comune che permetta di fare analisi stime e valutazioni sui dati di partenza.
Le diverse tecniche di fusione dipendono dall'architettura comune che si decide di considerare ed in particolare se essa è centralizzata, distribuita o semidistribuita.
Tuttavia alcuni elementi comuni possono essere identificati nella procedura di aggregazione dell'informazione.
Il processo di data fusion, implica la presenza di dati spuri riidondanza dei dati e problemi di allineamento spaziale e temporale sopratutto se i dati contengono informazioni georiferite.
Una soluzione a questo problema è che alla struttura dati vengano affiancati degli algoritmi di associazione a allineamento che permettano ad esempio di associare a corretta approssimazione della posizione del punto a partire dalle diverse segnalazioni nell'intorno di punto stesso.
Possibili algoritmi di associazione possono essere basati su algoritmi di consenso, teoria dei giochi o potrebbero appartenere alla classe degli algoritmi di ottimizzazione stocastica temporale per un problema in rete
Il lavoro analizzerà l'applicazione di questo tipo di algoritmi su dati georiferiti ottenututi da crowdmapping al fine di determinare come poter effettuare una corretta associazione spazio-temporale.

Il crowdmapping e gli approcci di rilevazione partecipativa delle informazioni georeferenziate producono una grande quantità di dati che devono tuttavia essere gestiti correttamente. La gestione di questa mole di dati è spesso un elemento che viene sottovalutato.
Le questioni chiave da affrontare nel merito della gestione dei dati sono in particolare:

· Come vengono salvati e integrati i dati anche da quelli provenienti?
· Quali sono le informazioni finali da ottenere per l'utente finale?
La prima domanda può essere correlata al problema di *data-fusion* (Durrant-Whyte, 2011).
Il *data-fusion* può essere definito come come "un processo multi-livello che tratta l'associazione, la correlazione, la combinazione di dati e informazioni da singole e multiple fonti per ottenere una definizione accurata della posizione ra, individuare stime e valutazioni complete e tempestive dei dati aggregati, loro significato e le minacce legate alla loro cattiva interpretazione"
A partire da questa definizione, il processo di *data-fusion* può essere scomposto in 3 sub-problemi (Castanedo, 2013):

- Associazione dati
- Stima dello stato
- Decision-fusion: inteso come la combinazione di classificatori per ottenere una migliore precisione di classificazione nel problema del riconoscimento dei pattern.

Questa divisione tipicamente riflette 3 livelli di astrazione dei processi fisici con i quali i dati prodotti sono: misura, associazione di attributi descrittivi, valutazione alla fine di
Le tecniche di *data-fusion* possono essere classificate sulla base dell'architettura di sistema in cui persistono.
Come illustrato nella figura 1 esse possono essere (Durrant-Whyte, 2001):

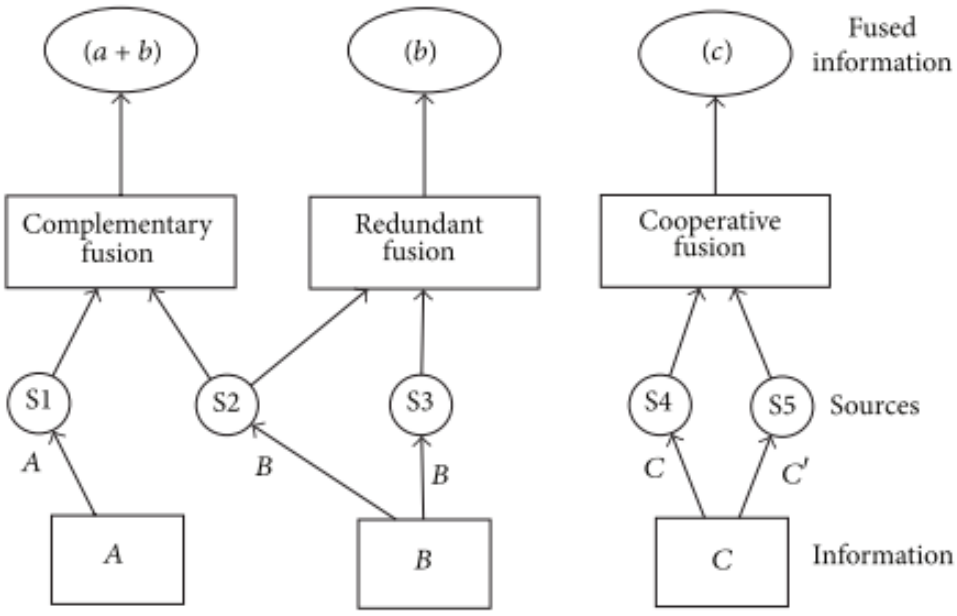

Figura 1 - Strategie di data fusion a) Complementari b) Ridondanti c) Cooperative
(Fonte Castanedo , 2013)

a)**complementari**: quando le informazioni fornite dalle fonti di input rappresentano parti diverse della scena e possono pertanto essere utilizzate per ottenere informazioni globali più complete.
b)**ridondanti**: quando due o più fonti di input forniscono informazioni sullo stesso punto e possono quindi essere fuse per incrementare l'accuratezza della misura
c)**cooperative**: quando le informazioni fornite sono combinate in nuove informazioni tipicamente più complesse delle informazioni originali.
La giusta strategia di fusione dei dati dipende dalle capacità di calcolo del sistema di informazione e a quale livello vengono elaborate le informazioni.
Il livello di elaborazione delle informazioni può essere classificato considerando innanzitutto il modello concettuale di infrastruttura informativa consistente in:

- *risorse*, entità che forniscono e scambiano dati
- *human-computer interaction* (HCI), che è l'interfaccia che si utilizza per interagire con l'origine e l'archiviazione finale e la
- *database management system*, che raccoglie i dati primari e li integra

Sulla base di questo modello si può distinguere un framework di 5 livelli di processamento dell'informazione (Framework JDK).
Questi sono stati originariamente proposti dal Dipartimento Americano di Difesa (DoD) (Castanedo, 2013).

- *Livello 0 - source preprocessing:* questo è il livello più basso del processo di fusione dei dati e comprende la fusione a livello di segnale, ai livelli di pixel e al processo di estrazione delle informazioni. A questo stadio si riduce la quantità di dati e si mantengono le informazioni utili per i processi di alto livello.
- *Livello 1 - raffinamento dell'oggetto:* questo livello impiega i dati elaborati dal livello precedente per effettuareil'allineamento spaziale, l'associazione, la correlazione, le tecniche di raggruppamento o raggruppamento, la stima dello stato, la rimozione di falsi positivi, la fusione di dati unificabili e la combinazione delle caratteristiche estratte dalle immagini. I risultati di output di questa fase sono la discriminazione degli oggetti (classificazione e identificazione) e il monitoraggio degli oggetti (stato dell'oggetto e dell'orientamento).Questa fase trasforma le informazioni di input in strutture di dati coerenti.
- *Livello 2 - situation assessment:* il situation assessment mira a individuare gli scenari probabili in base agli eventi osservati e ai dati ottenuti. Stabilisce relazioni tra gli oggetti. Le relazioni (cioè, la prossimità, la comunicazione) vengono valutate per determinare il significato delle entità o degli oggetti in un ambiente specifico. L'obiettivo di questo livello include l'esecuzione di inferenze di alto livello e l'individuazione di attività e eventi significativi tramite modelli generali. L'output è un insieme di inferenze di alto livello
- *Livello 3 - impact assessment:*valuta l'impatto delle attività rilevate nel livello 2 per ottenere una prospettiva corretta. La situazione attuale viene valutata e viene effettuata una proiezione futura per identificare possibili rischi, vulnerabilità e opportunità operative. Questo livello include una valutazione del rischio o della minaccia e una previsione del risultato logico;

- *Livello 4 - raffinamento del processo:* questo livello migliora il processo dal livello 0 al livello 3 e fornisce la gestione delle risorse e dei sensori. L'obiettivo è quello di ottenere una gestione efficiente delle risorse, tenendo conto delle priorità delle attività, della pianificazione e del controllo delle risorse disponibili.

La fusione ad alto livello inizia normalmente a livello 2 come descritto nella Figura 2.

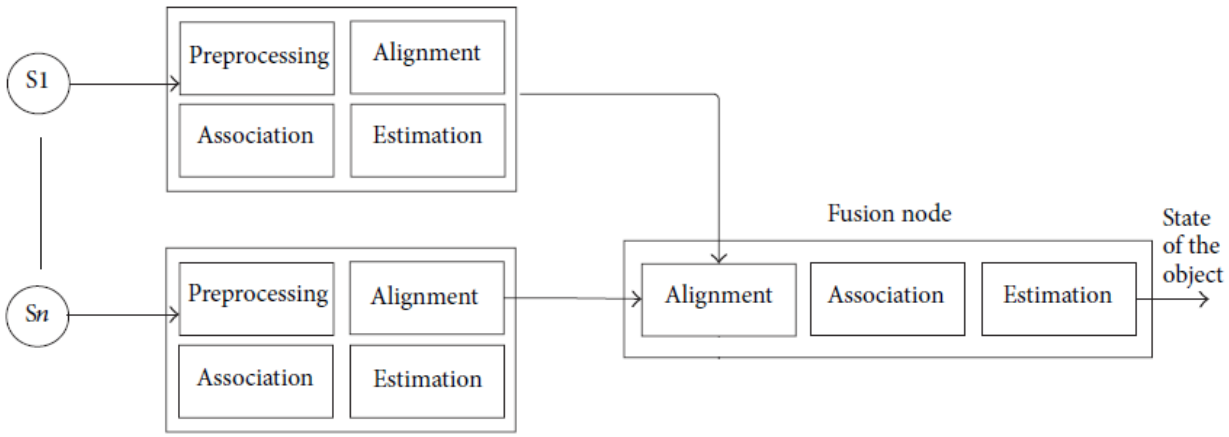

Figura 2 - Architetture di sistema A) Centralizzate B) Distribuite

(Fonte: Castanedo , 2013)

Teoricamente, un'architettura centralizzata sarebbe ottimale con una trasmissione totalmente efficiente e con elevate capacità computazionali.
Considerando il contesto di un paese in via di sviluppo, invece, un'architettura leggera distribuita potrebbe sembrare la soluzione migliore. Sfrutta la moltitudine degli utenti e persino riduce il potenziale di calcolo dei dispositivi remoti.
In questo modo l'integrazione dei dati di carico complessivo sarà distribuita parzialmente (in operazioni primarie come pre-elaborazione e associazioni semplici) su ciascun dispositivo.
L'effetto collaterale di questo approccio risiede nel fatto che il paradigma distribuito introduce problemi legati a dati spuri, ridondanza dei dati e allineamento spaziale e temporale (Castanedo, 2013).
Una possibile soluzione per risolvere questo tipo di problema può essere dato dall'utilizzo di algoritmi di associazione (Fig.3) in cui l'associazione è definita come "il processo di pesi di assegnazione del peso ai dati e alle tracce (set di punti ordinati che seguono un percorso e generati da la stessa destinazione) da un insieme di osservazioni in un altro " (Castanedo, 2013).

Questo tipo di algoritmi potrebbe essere basato su algoritmi di consenso, teoria dei giochi o appartenere alla classe degli algoritmi di ottimizzazione stocastica variabile per un problema in rete (Simonetto et. Al. 2014).
L'applicazione sul crowdmapping e sul sensing partecipativo permette, ad esempio, di integrare le varie fonti correlate ad un allarme per individuare più accuratamente l'evento e diminuire l'incertezza data dalla molteplicità di input.

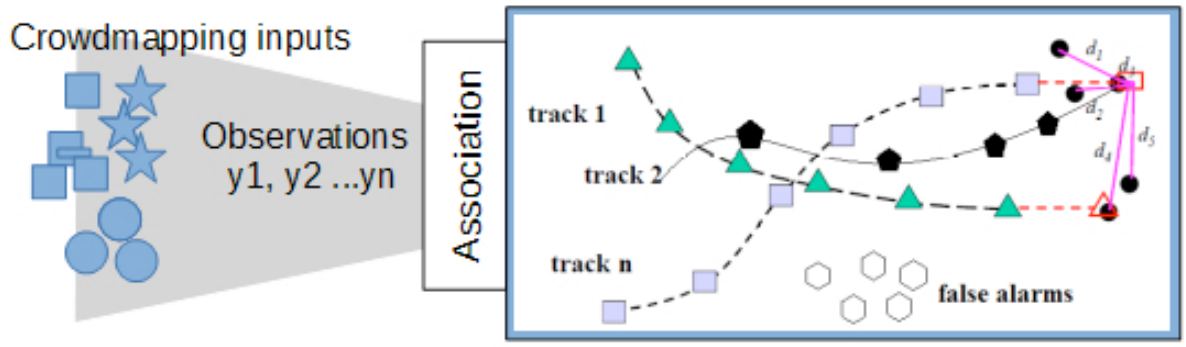

Figura 3 - Problema di associazione per i dati di crowdmapping (Source: Castanedo,2013)

Per una rete connessa, gli algoritmi basati su consenso garantiscono che le stime locali vengono condivise e raffinate tra vicini per raggiungere la stessa media ponderata su tutti i nodi.
La stima dei parametri per i modelli lineari è un problema comune in cui il consenso medio viene regolarmente adottato per simulare un approccio centralizzato senza bisogno di alcun centro di fusione .
Altri esempi di localizzazione delle sorgente si basano su algoritmi stile Ga*ussian Mixture Model* (GMM) (Bui,Huang,2017).
Il *Gaussian Mixture Model* (GMM) è un modello probabilistico che afferma che tutti i punti dati generati sono derivati da una miscela di una distribuzione finita gaussiana che non ha parametri noti.
I parametri per i modelli di miscelazione Gaussiane sono derivati sia dalla stima massima a posteriori, sia da un algoritmo di massimizzazione dell'aspettativa iterativa da un modello precedente ben addestrato.
I GMM sono molto utili quando si tratta di modellare i dati, in particolare i dati provenienti da diverse sorgenti come nel caso dei punti segnalati da una piattaforma di crowdsourcing.

Riferimenti bibliografici

Castanedo F. (2013), "A Review of Data Fusion Techniques", *The Scientific World Journal,* vol.2013: 1-19

Simonetto, A., Kester, L., Leus G. (2014),"Distributed time-varying stochastic optimization and utility-based communication", *arXiv preprint arXiv:*1408-5294

Bolognino, A., Spagnolini U. (2014), "Consensus based distributed estimation with local-accuracy exchange in dense wireless systems", *Communications (ICC), 2014 IEEE International Conference on*.

Bui, D. T., Hoang, N. D. (2017), "A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1. 1) for spatial prediction of floods", *Geoscientific Model Development*, *10*(9).