
BOLLETTINO UNIONE MATEMATICA ITALIANA

Sezione A – La Matematica nella Società e nella Cultura

RENATO TORELLI

Preprocessamento e generazione di piani di taglio per problemi di controllo e correzione dati

*Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. 8-A—La
Matematica nella Società e nella Cultura (2005), n.3-1, p. 645–647.*

Unione Matematica Italiana

[<http://www.bdim.eu/item?id=BUMI_2005_8_8A_3-1_645_0>](http://www.bdim.eu/item?id=BUMI_2005_8_8A_3-1_645_0)

L'utilizzo e la stampa di questo documento digitale è consentito liberamente per motivi di ricerca e studio. Non è consentito l'utilizzo dello stesso per motivi commerciali. Tutte le copie di questo documento devono riportare questo avvertimento.

*Articolo digitalizzato nel quadro del programma
bdim (Biblioteca Digitale Italiana di Matematica)
SIMAI & UMI*

<http://www.bdim.eu/>

Preprocessamento e generazione di piani di taglio per problemi di controllo e correzione dati

RENATO TORELLI

1. – Il problema della correzione degli errori.

Oggigiorno, gli utenti domandano informazioni statistiche più dettagliate e con elevati standard di qualità. Di conseguenza, diviene più difficile per gli istituti di statistica nazionali e analoghi organismi, produrre dati di elevata qualità in breve tempo. Un ulteriore fattore di complicazione è costituito dal fatto che i dati rilevati di norma contengono errori. Infatti, ogni stadio nella collezione dei dati è una potenziale sorgente di errori.

2. – Il modello matematico proposto.

Il problema dell'individuazione degli errori è spesso affrontato formulando un insieme di regole che ogni record deve soddisfare affinché sia dichiarato corretto. Dopo tale fase di individuazione, i record errati dovrebbero essere corretti e usati. Ciò dovrebbe essere fatto seguendo il criterio del minimo cambiamento possibile e, nel contempo, rispettando le distribuzioni di frequenza originali dei dati. Quanto esposto viene generalmente ottenuto per mezzo di procedimenti detti di data editing and imputation automatici.

Nell'ambito di un processo di correzione automatico, le attività svolte dall'essere umano sono essenzialmente di due tipi: fornire metadati e quindi le regole; progettare il processo di correzione e quindi individuare il modello matematico e l'algoritmo di soluzione da utilizzare nel processo stesso.

In particolare, per effettuare una correzione automatica è necessaria una prima fase di localizzazione dell'errore all'interno dei record errati ed una seconda fase di imputazione del valore corretto per i campi del record individuati come errati. Nell'ambito di un'attività di ricerca immediatamente precedente a questa tesi [3], è stato individuato il seguente modello matematico di programmazione lineare mista [4] per rappresentare i problemi descritti sopra:

$$\begin{aligned}
 \min \quad & c'x + d'\bar{x} \\
 & Ax + B\bar{x} \geq 1 \\
 & Dx + E\bar{x} + Fu + Gz \geq h \\
 (1) \quad & x_i + \bar{x}_i = 1 \\
 & z_i \in \{0, \dots, U\}, \quad x_i, \bar{x}_i, u_i \in \{0, 1\} \\
 & A, B \in \{0, 1\}^{m \times n}, \quad D, E, F, G \in \mathbb{R}^{m \times n} \\
 & c', d' \in \mathbb{R}_+^n, \quad h' \in \mathbb{R}^n
 \end{aligned}$$

Tali modelli hanno una forte componente di variabili binarie e di vincoli di set covering [4]. Inoltre, le variabili z possono, a seconda dei casi, essere reali anziché intere.

Questo lavoro riguarda gli aspetti algoritmici della soluzione di tali problemi. Sebbene la correzione di ogni singolo record possa essere risolta in decine di secondi da una procedura di branch-and-cut stato dell'arte, come ILOG Cplex, è però necessario correggere insiemi di dati molto grandi. Perciò, sono necessarie procedure computazionalmente efficienti. Inoltre, esistono applicazioni per cui la dimensione dei modelli citati cresce in maniera esponenziale, causando grossi problemi alle menzionate procedure di branch-and-cut. Un esempio sono le applicazioni che richiedono il processamento simultaneo di insiemi di record di grandi dimensioni. Una serie di operazioni volte a superare i limiti di tale tipo di procedura per ottenere la soluzione di tali problemi lineari interi misti sono qui studiati e proposti. Queste operazioni vengono tutte effettuate prima dell'inizio dell'esplorazione dell'albero di branching, condotta da una generica procedura di branch-and-bound. In particolare, sono qui proposte una serie di procedure di preprocessamento e di generazione di piani di taglio, finalizzate alla riduzione della dimensione di un dato modello ed a produrre una più robusta formulazione del modello stesso.

3. – L'algoritmo proposto.

Per quanto riguarda le operazioni di preprocessamento [7], tese alla riduzione delle dimensioni del modello matematico, vengono proposte le seguenti operazioni:

- *Rafforzamento dei vincoli*, condotto per migliorare gli upper e lower bound correnti per ogni variabile.
- *Individuazione di Ridondanze o Inammissibilità*, attuata per eliminare i vincoli ridondanti o per provarne l'infeasibility.
- *Fissaggio di Variabili*, allo scopo di eliminare variabili dal modello.

Per quanto riguarda invece le procedure di generazione di piani di taglio, vengono proposte le seguenti operazioni:

- *Probing* [6], ottenuto facendo uso dell'Algoritmo del Volume [1, 2] e di un grafo

dei conflitti aggiornato dinamicamente. Semplificando, esso consiste nel provare a porre un insieme di variabili binarie a 0 oppure ad 1 e ad esplorare le conseguenze. Questa procedura deriva disuguaglianze che non sono necessariamente valide per l'intero insieme dei punti interi ammissibili, ma che sono valide per i punti interi ottimi.

- *Separazione* per problemi di set covering con vincoli di uguaglianza, che permette di derivare nuovi vincoli di set covering in modo da rafforzare la formulazione [5].

4. – Implementazione e Risultati.

L'intera procedura è implementata in C++. Il software sviluppato offre una libreria di classi C++ che abilita l'utente a definire automaticamente modelli per problemi MILP, con uno stile molto simile alla tecnologia Concert di ILOG. La formulazione specificata è quindi ridotta e rafforzata come illustrato e poi risolta per mezzo dell'Algoritmo del Volume. Sono infine riportate comparazioni sui tempi di soluzione ottenuti sia utilizzando il sistema software proposto che facendo uso del pacchetto commerciale ILOG Cplex. I risultati sono molto incoraggianti.

BIBLIOGRAFIA

- [1] BARAHONA F. e ANBIL R., *The Volume Algorithm: producing primal solutions with a subgradient method*, IBM Research Report (1998).
- [2] BARAHONA F. e CHUDAK F., *Near-optimal solutions to large scale facility location problems*, IBM Research Report (1999).
- [3] BRUNI R., REALE A. e TORELLI R., *DIESIS: a New Software System for Editing and Imputation*, Proceedings SIS2002 (2002).
- [4] NEMHAUSER G.L. e WOLSEY L.A., *Integer and Combinatorial Optimization*, J. Wiley, New York (1988).
- [5] NOBILI P. e SASSANO A., *A Separation Routine for the Set Covering Polytope*, Proceedings of the 2nd IPCO Conference, Carnagie-Mellon University Press (1992).
- [6] SAVELSBERGH M.W.P., *Preprocessing and probing techniques for mixed integer programming problems*, Orsa Journal on Computing (1994), 445-454.
- [7] WOLSEY L.A., *Integer Programming*, John Wiley & Sons, New York (1998).

ISTAT - Istituto Nazionale di Statistica, Roma
e-mail: torelli@istat.it

Dottorato in Ricerca Operativa (sede amministrativa: Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma «La Sapienza», Roma) - Ciclo XVI
Direttore di ricerca: Prof. Antonio Sassano, Università degli Studi di Roma «La Sapienza»

