GIORNATE DELLA RICERCA IN ISTAT | 10-11 NOVEMBRE 2014

SESSIONE II
PREVENZIONE, VALUTAZIONE E TRATTAMENTO DEGLI ERRORI NON CAMPIONARI

La valutazione dell'errore di *linkage*

Nicoletta Cibella, Niki Stylianidou e Tiziana Tuoto

Istat

Sulla valutazione dell'errore di linkage

Nicoletta Cibella, Niki Stylianidou e Tiziana Tuoto

Istat

cibella@istat.it, styliani@istat.it, tuoto@istat.it

Sommario

L'uso congiunto di dati provenienti da fonti diverse è una opportunità che gli Istituti di Statistica cercano di cogliere sempre più frequentemente. In un contesto in cui grandi patrimoni informativi prodotti da attori diversi possono essere integrati e confrontati, diventa sempre più urgente l'esigenza di corredare i risultati dell'integrazione con valutazioni quantitative della qualità delle operazioni (tecniche e metodologie) che hanno permesso di raggiungere tale risultato. In questo lavoro vengono sperimentati metodi alternativi per migliorare la stima dell'errore di linkage, rispetto a quella fornita dal modello di Fellegi-Sunter, riferimento base per il record linkage probabilistico. Come noto, infatti, il criterio di fellegi e Sunter è molto efficace per la individuazione degli abbinamenti ma generalmente meno affidabile per la valutazione dei risultati ottenuti. Il lavoro propone metodi noti in letteratura e alternative basate su metodologie ancora poco esplorate in questo settore.

Parole chiave: record linkage probabilistico, qualità procedure di integrazione, modelli mistura, modelli a classi latenti, analisi discriminante

Abstract

The combined use of data from different sources is an opportunity that the Institutes of Statistics try to grasp more and more frequently. In a context where huge amount of information produced by different actors can be integrated and compared , it becomes ever more urgent the need of providing integration results with quantitative quality assessments of techniques and methodologies that allowed achieve these resulst. In this work we tested alternative methods to improve the linkage errors estimations, compared to that provided by the Fellegi - Sunter model , that is the reference for the probabilistic record linkage . As well known , in fact , the Fellegi and Sunter criterion is very effective for the identification of links but generally less reliable for the results evaluation. The paper proposes methods well-known in the literature as well as tries to provides alternative ones.

Key words: probabilistic record linkage, linkage quality assessment, mixture models, latent class models, discriminant analysis

1. Introduzione

L'uso congiunto di dati provenienti da fonti diverse è una opportunità che gli Istituti di Statistica cercano di cogliere sempre più frequentemente. In un contesto in cui grandi patrimoni informativi prodotti da attori diversi possono essere integrati e confrontati, diventa sempre più urgente l'esigenza di corredare i risultati dell'integrazione con valutazioni quantitative della qualità delle operazioni (tecniche e metodologie) che hanno permesso di raggiungere tale risultato. Nell'ambito dell'integrazione dei dati a livello micro, le metodologie per il record linkage in Istat sono largamente impiegate e producono generalmente

buoni risultati (quando applicate per mezzo di variabili di abbinamento dal forte potere identificativo) anche se, raramente, le procedure di abbinamento sono accompagnate da indicatori quantitativi di qualità. Eppure, soprattutto nella statistica ufficiale, questi ricoprono un ruolo fondamentale poiché permettono di "certificare" l'accuratezza e la credibilità dei dati, che saranno impiegati in successive analisi statistiche.
In questo lavoro si vogliono esplorare metodi largamente conosciuti in letteratura o metodi ancora poco sfruttati in questo settore per identificare dei criteri per la valutazione degli errori di linkage che siano più performanti rispetto al criterio legato alla regola di decisione di Fellegi-Sunter, che tradizionalmente rappresenta il riferimento di base per il record linkage probabilistico.

2. L'integrazione dei dati e la misura della qualità dei risultati

Il record linkage è quell'insieme di tecniche e di metodologie finalizzate a riconoscere la stessa entità del mondo reale (individuo, unità economica, evento), anche se diversamente rappresentata. Le tecniche di record linkage, e in particolar modo il record linkage probabilistico, sono molto diffuse nella statistica ufficiale data la loro utilità nel caso in cui un identificativo univoco per le unità è assente e le variabili che congiuntamente permettono di identificare l'unità sono affette da errori di vario, compresi valori mancanti.
Generalmente un problema di record linkage può essere visto come un problema di classificazione in cui tutte le coppie generate dal prodotto cartesiano tra tutti i record dei due (o più file) da integrare devono essere classificate in due insiemi disgiunti, l'insieme degli Abbinati (i Matches) e l'insieme dei Non abbinati (gli Unmatches)
I principali indicatori statistici di qualità delle procedure di integrazione devono rappresentare gli errori che possono risultare da una applicazione di record linkage: si tratta quindi essenzialmente di errori di falso abbinamento e di errori di mancato abbinamento. Si parla di errore di mancato abbinamento quando due unità registrate in due o più file distinti di fatto rappresentano la stessa entità del mondo reale ma la procedura di record linkage non è in grado di identificare questo abbinamento e quindi come risultato le due unità rimangono distintamente assegnate all'insieme degli Unmatches. Si parla invece di errore di falso abbinamento quando la procedura di integrazione assegna all'insieme dei Matches una coppia di unità che di fatto non si riferiscono alla stessa entità del mondo reale.
Le strategie di record linkage spesso cercano di raggiungere un compromesso tra i due tipi di errore: infatti generalmente se si vuole ridurre al minimo il numero di falsi positivi (i falsi abbinamenti) bisogna accettare un numero maggiore di falsi negativi (i mancati abbinamenti) e viceversa. I due tipi di errori possono avere diversa importanza a seconda degli obiettivi dello specifico problema di integrazione.
Indicatori per la misura di questi errori possono essere costruiti sulla base della tabella che identifica gli abbinamenti rispetto al reale stato di abbinamento: la successiva tabella 1 riporta in riga i risultati forniti dalla procedura utilizzata per il record linkage (in termini di Matches - Unmatches) e in colonna il reale stato di abbinamento (Links – No Links) delle coppie considerate:

Tavola 1 – Esito delle procedure di record linkage

	Links	No Links
Matches	a – veri positivi	b – falsi positivi o falsi abbinamenti
Unmatches	b – falsi negativi o mancati abbinamenti	d – veri negativi

Dalla tabella n.1 vengono generalmente ricavati i seguenti rapporti:

- il tasso di falso abbinamento: $f = b/a+b$
- il tasso di mancato abbinamento: $m = c/c+a$

Nelle applicazioni reali, il vero stato di abbinamento (Links/No Links) non è noto e si vuole appunto riuscire a predirlo attraverso la procedura di record linkage.

3. Metodi per la valutazione degli errori di abbinamento

I metodi statistici più noti per la valutazione degli errori di linkage seguono principalmente l'approccio dei modelli mistura con variabili latenti. La variabile (non osservata) che rappresenta il reale stato di abbinamento costituisce appunto la variabile latente che si cerca di predire attraverso l'osservazione dei risultati dei confronti sulle variabili osservabili (negli approcci con modelli non-supervisionati) o per mezzo dei risultati noti in un training set (nei modelli supervisionati). La teoria di Fellegi e Sunter (1969) per risolvere problemi di record linkage rientra appunto in un approccio basato su modelli mistura non-supervisionati.
Secondo il modello Fellegi-Sunter, il criterio classificazione della coppia si basa su due soglie: T_m e T_u ($T_m > T_u$), se il peso di abbinamento $r_{(a,b)}$ della coppia (a,b) è superiore alla soglia T_m la coppia (a,b) viene assegnata all'insieme dei Matches M^*, se il peso $r_{(a,b)}$ è inferiore alla soglia T_u la coppia viene assegnata all'insieme degli Unmatches U^*, se il peso assume un valore compreso tra le due soglie non viene presa alcuna decisione sullo stato di abbinamento della coppia:

$$\begin{aligned} r_{(a,b)} > T_m &\quad \Rightarrow \quad (a,b) \in M^* \\ T_m \geq r_{(a,b)} \geq T_u &\quad \Rightarrow \quad (a,b) \in Q \\ r_{(a,b)} < T_u &\quad \Rightarrow \quad (a,b) \in U^* \end{aligned}$$

La regola di decisione Fellegi e Sunter si basa sui livelli di errore fissati come accettabili: una coppia di fatto costituita da unità differenti può assumere un valore $r_{(a,b)} > T_m$ e la frequenza di tale errore, dovuto alla decisione M^*, è

$$\mu = \sum_{\gamma \in \Gamma} u(\gamma) P(M^* \mid \gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma) \quad \text{dove} \quad \Gamma_{M^*} = \{\gamma : T_m \leq m(\gamma)/u(\gamma)\}$$

dove $r_{(a,b)} = m(\gamma)/u(\gamma)$ è il peso di abbinamento, costruito come rapporto tra la verosimiglianza della coppia di appartenere all'insieme dei Matches $m(\gamma)$ e la verosimiglianza della coppia dei appartenere all'insieme degli Unmatches $u(\gamma)$, dato il risultato osservato del confronto sulle variabili di abbinamento sintetizzato nel vettore dei confronti γ.
Allo stesso modo, una coppia realmente costituita da record che rappresentano la stessa unità può assumere un valore $r_{(a,b)} < T_u$, la frequenza di tale errore, dovuto alla decisione U^*, è

$$\lambda = \sum_{\gamma \in \Gamma} m(\gamma) P(U^* \mid \gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma) \quad \text{where} \quad \Gamma_{U^*} = \{\gamma : T_u \geq m(\gamma)/u(\gamma)\}$$

Fellegi e Sunter suggeriscono di selezionare i valori "accettabili" per i livelli di errore μ e λ, fissati i quali è possibile ottenere i valori delle soglie T_m e T_u risolvendo le formule precedenti. I valori degli errori μ e λ possono essere stimati attraverso la stima delle distribuzioni $m(\gamma)$ e $u(\gamma)$:

$$\hat{\mu} = \sum_{\gamma \in \Gamma_{M^*}} \hat{u}(\gamma) \qquad \hat{\lambda} = \sum_{\gamma \in \Gamma_{U^*}} \hat{m}(\gamma)$$

La stima degli errori così come proposta nell'approccio di Fellgi-Sunter risulta fortemente dipendente dalla precisione delle stime delle distribuzioni m(γ) e u(γ). Errori nella specificazione del modello sottostante tali distribuzioni (ad esempio l'affidabilità della assunzione di indipendenza condizionale), mancanza di informazioni e così via possono causare perdi-

ta di precisione nelle stime delle distribuzioni e di conseguenza forti distorsioni nelle stime degli errori. E' da notare come gli stessi problemi di errata specificazione del modello hanno comunque solo un effetto minore sulla qualità degli abbinamenti individuati, in quanto il valore del peso di abbinamento $r_{(a,b)}$ permette comunque di realizzare una corretta graduatoria per le coppie di record da abbinare.

3.1 Il modello proposto da Belin e Rubin

Per quanto riguarda in particolare la valutazione dell'errore di falso abbinamento, un metodo molto noto in letteratura è quello proposto da Belin e Rubin (1995). Il loro obiettivo è quello di valutare il tasso di falso abbinamento per ogni possibile valore della soglia T_m.
Il metodo di Belin e Rubin richiede la disponibilità di un training set su cui il vero stato di abbinamento delle coppie sia noto. Il modello ipotizza che la distribuzione dei pesi di abbinamento sia una mistura di due distribuzioni: quella dei veri link e quella dei veri non-link. Gli autori considerano i pesi di abbinamento ottenuti secondo la procedura proposta da Fellegi e Sunter, anche se in linea di principio possono essere considerati pesi ottenuti attraverso procedure differenti. In particolare gli autori ipotizzano che le due distribuzioni componenti la mistura siano due Normali, di media e varianza ignote. Per garantire la normalità delle distribuzioni applicano due diverse trasformazioni di Box-Cox, una all'insieme dei pesi appartenenti alla popolazione dei veri link, l'altra per i pesi delle coppie appartenenti alla popolazioni dei veri non-link. I parametri delle trasformazioni di Box-Cox vengono stimati sul training set, in cui si conosce il vero stato di abbinamento delle coppie, e tali stime rappresentano i parametri "globali" della procedura di stima. A questo punti i parametri "globali" delle trasformazioni vengono considerati noti e si calcolano le stime di massima verosimiglianza e i relativi errori standard della mistura di normali sull'insieme dei dati da classificare. Attraverso tali stime, si ottengono stime puntuali del tasso di falso abbinamento, come funzione della soglia T_m e i corrispondenti errori standard per i tassi di falso abbinamento, attraverso una approssimazione di tipo delta.
Il metodo illustrato fornisce buone stime del tasso di falso abbinamento nel caso di abbinamento di tipo 1:1, quando cioè ciascuna unità di un file può essere abbinata al massimo con una sola unità dell'altro file e viceversa; inoltre, è necessario che ci sia una buona separazione tra i pesi di abbinamento associati alla popolazione dei veri link e quelli relativi alla popolazione dei veri non-link.
I limiti legati all'utilizzo di questo metodo dipendono dalla disponibilità di un training set per cui sia noto lo stato di corretto abbinamento e che riproduca la distribuzione dei pesi delle due popolazioni Links e NoLinks così come osservate nell'universo delle coppie da classificare; anche l'effettiva Normalità delle distribuzioni trasformate dei pesi di abbinamento può inficiare l'efficacia del metodo.
Per quanto riguarda la costruzione del training set, in particolare, può essere utile seguire i suggerimenti di Winkler (1995) che consiglia di ridurre l'ampiezza campionaria attraverso la selezione delle coppie più vicine all'area grigia individuata dalle due soglie, anche per mezzo di una strategia di campionamento ponderata; una strategia di campionamento stratificata ottimizzata per la costruzione del training set è stata proposta da Heasman (2014).
Le problematiche esposte riguardo all'applicazione del metodo di Belin e Rubin sono stati riscontrati anche nelle applicazioni proposte nel presente lavoro, così come con altri dati provenienti da processi di integrazione disponibili in Istat, caratterizzati dall'alto potere identificativo delle variabili di match.
Il tentativo di replicare la proposta contenuta nell'articolo di Belin e Rubin è stato molto utile per mettere a fuoco l'esigenza di cercare una regola alternativa che distingua i Links dai No links rispetto alla Regola Principale che distingue i Matched dagli Unmatched. A questo proposito, un metodo poco esplorato nei contesti di record linkage ma che sembra

adatto a gestire questo problema è l'analisi discriminante canonica su due gruppi (introdotta da Fischer nel '36).

3.2 L'analisi discriminte a classi latenti

Il metodo proposto vuole distinguere nell'insieme delle coppie candidate ad essere Matches quelle appartenenti al gruppo (sotto-popolazione) dei Veri link e quelle appartenenti al gruppo (sotto-popolazione) dei Falsi abbinamenti. Anche con questa metodologia, i parametri per la distinzione dei due gruppi saranno stimati ancora una volta sul training set, un sotto-insieme dei dati per cui è noto il vero stato di abbinamento delle coppie. Nel seguito si indichi con Y il vettore dei match: $Y = (Y_1, Y_2)$ dove Y_1 sono le coppie che realmente rappresentano un link, mentre Y_2 è il suo complementare ovvero le coppie corrispondenti a falsi abbinamenti.
La matrice di covariante dei match è $X = (X_1, X_2)$ con $\bar{x}_1$ il vettore delle medie delle covariate dei veri abbinamenti e $\bar{x}_2$ il vettore delle medie delle covariate dei falsi abbinamenti. Sia poi S_1 la matrice della varianza per i veri abbinamenti e S_2 è la matrice della varianza dei falsi abbinamenti. Infine sia n_1 il numero dei veri abbinamenti mentre n_2 è il numero dei falsi abbinamenti.
Il calcolo delle matrici di Varianza e Covarianza:

tra (between)

$$\mathbf{B} = \frac{n_1 * n_2}{(n_1 + n_2)(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^T}$$

e

entro (within)

$$W = \frac{S_1(n_1 - 1) + S_2(n_2 - 1)}{n_1 + n_2 - 2}$$

dei due gruppi permette di costruire l'equazione discriminante di Fisher:

$$Y = XW^{-1}(\bar{x}_1 - \bar{x}_2).$$

Il punto di separazione fra i due gruppi (breaking point) è dato da:

$$break = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

E la regola di assegnazione per ogni unità è data da

$$\min \triangleq \{ |\bar{y}_1 - y_i| \, @ \, |\bar{y}_2 - y_i| \quad \forall y_i \quad .$$

Ovvero ogni unità di *y* andrà assegnata alla popolazione dei veri o dei falsi abbinamenti sulla base della distanza dal baricentro delle popolazioni stesse. Questa metodologia, attraverso la costruzione di opportuni training set, potrebbe permettere di individuare anche i mancati abbinamenti all'interno della popolazione degli Unmaches.

4. Applicazione

I metodi proposti nella sezione precedente sono stati sperimentati con l'ausilio di dati fittizi, per cui è noto il reale stato di abbinamento attraverso la disponibilità di un codice identificativo univoco. In particolare si sono condotte diverse sperimentazioni sui dati perturbati del censimento della popolazione e di una fonte amministrativa che rileva congiuntamente le posizioni contributive e assistenziali (McLeod, Heasman and Forbes, 2011).

4.1 I dati e gli scenari di linkage considerati

I file considerati mimano la presenza di errori e valori mancanti nelle variabili identificatrici che è possibile osservare in contesti reali e che sono la causa principale di errori nei risultati delle procedure di abbinamento: falsi abbinamenti e mancati abbinamenti. Tali dati sono stati creati e resi disponibili da Paula McLeod, Dick Heasman e Ian Forbes, dell'Office of National Statistics (UK), per il progetto Europeo ESSnet Data Integration.
I metodi sono stati testati in due differenti scenari di linkage: il primo presenta un tasso di falso abbinamento di poco inferiore al 2% e per questo verrà indicato nel seguito come Gold Scenario. In questo caso è stato estratto un campione dall'insieme dei dati disponibili di circa 1500 unità, variabili molto discriminanti sono state utilizzate per l'abbinamento secondo il modello di Fellegi-Sunter. In particolare, data la dimensione ridotta dei dati presi in esame non è stato necessario applicare alcuna riduzione dello spazio di ricerca è si è proceduto con la creazione del prodotto cartesiano delle coppie, le variabili di abbinamento selezionate sono state Nome, cognome, giorno e anno di nascita e ciò ha permesso di individuare 1245 abbinamenti di cui 23 falsi abbinamenti. I veri abbinamenti erano in realtà 1301. Gli abbinamenti sono stati assegnati secondo la regola di Fellegi-Sunter, fissando la soglia in corrispondenza dell'errore minimo ottenibile per il tasso di mancato abbinamento, in questo caso inferiore a 2.5%. In questa applicazione, il tasso di errore di falso abbinamento stimato attraverso il modello di Fellegi-Sunter è pari al 3.5% rispetto al tasso osservato di poco inferiore al 2%.
Un secondo scenario di linkage è stato costruito considerando tutti i record che compongono i file, circa 26000 e applicando come metodo per la riduzione dello spazio di ricerca la funzione Simhash applicata dentro opportuni blocchi. Come variabili di abbinamento sono state utilizzate le stesse applicate nello scenario precedente. Questa strategia di abbinamento ha permesso di identificare 21560 abbinamenti, di cui 1336 falsi, con un corrispondente tasso di falso abbinamento di poco superiore al 6%. Gli abbinamenti sono stati assegnati secondo la regola di Fellegi-Sunter, fissando la soglia in corrispondenza dell'errore minimo ottenibile per il tasso di mancato abbinamento, in questo caso inferiore a 3%. In questa applicazione, il tasso di errore di falso abbinamento stimato attraverso il modello di Fellegi-Sunter è pari al 2% mentre al tasso di falso abbinamento osservato è di poco superiore al 6%. I veri abbinamenti sono poco più di 24000. Per questo tale scenario verrà indicato come Silver.

4.2 Risultati

Nei due scenari di linkage individuati sono stati testati i metodi per la stima dell'errore di falso abbinamento proposti da Belin e Rubin e quello basato sull'analisi discriminante.
Come universo delle coppie considerate, in cui individuare le due sottopopolazioni di vere coppie e falsi abbinamenti, è stato scelto l'insieme dei Matches, costruito, come descritto nel sottoparagrafo precedente, che secondo la regola ottimale di Fellegi-Sunter. Lo scopo è quello di verificare se i metodi proposti riescono a migliorare la valutazione dei risultati del processo di linkage rispetto al metodo di Fellegi-Sunter, che, come detto, è molto efficace nell'individuazione dell'insieme degli abbinamenti ma ha un suo punto di debolezza nella valutazione dell'errore associato a tale risultato. Tutte le sperimentazioni presentate sono state condotte con il software open-source RELAIS, nella versione 3.0. In teoria, a rigore dai dati individuati come Matches andrebbe estratto un training set rappresentativo del universo dove attraverso revisione manuale bisogna verificare la correttezza dei record individuati e il reale stato di abbinamento delle coppie. Come detto la selezione del training set è un'attività molto delicata ed è molto importante che esso sia rappresentativo perché su esso

verranno calcolati i parametri globali della trasformazione nel caso del metodo di Belin e Rubin o la regola rappresentata dal break-point nel caso dell'analisi discriminante. In questo esperimento, per evitare che l'efficacia dei metodi sia compromessa da estrazioni non ottimali del training set e sfruttando i dati fittizi per cui il reale stato di abbinamento è noto a priori per tutte le coppie, i parametri globali e il valore del break-point sono stati calcolati direttamente sull'intero insieme dei Matches.
In entrambi gli scenari descritti, Gold e Silver, il metodo di Belin e Rubin non permette di pervenire a valutazioni sul tasso di falso abbinamento. Infatti, in entrambi gli scenari, le trasformazioni Box-Cox delle due sottopopolazioni non corrispondono a distribuzioni normali, dato che, come noto, questo tipo di trasformazioni non corregge, ad esempio, la multi-modalità delle distribuzioni osservate. Anche in conseguenza di ciò, le due popolazioni componenti la mistura non risultano identificabili.
Riguardo alla procedura che coinvolge la metodologia di analisi discriminante, lo scopo è quello di stimare il break-point che separa la categoria di true matches dai false matches. Per ottenere questo, il training (nel nostro caso l'intero insieme delle coppie risultate abbinate secondo il criterio di Fellegi-Sunter) viene suddiviso in due sotto insiemi: quello dei true matches e quello dei false matches che di solido è il più piccolo. Formalmente:

$$Y_{training_{all}} = B * X_{training_{all}} + e_{training_{all}}$$

$$\begin{bmatrix} 1 = True\ linked \\ 0 = False\ linked \end{bmatrix} = B * P_{POST_{training_{all}}} + e_{training_{all}}$$

Diventa:

$$Y_{training_{true}} = [1 = True\ linked] = B * P_{POST_{training_{true}}} + e_{training_{true}}$$

$$Y_{training_{false}} = [0 = False\ linked] = B * P_{POST_{training_{false}}} + e_{training_{false}}$$

Nella prima applicazione del metodo è stata usata una sola covariata (P.POST), anch'essa risultante dal processo di integrazione di Fellegi-Sunter, ossia la probabilità a posteriori che dato un determinato profilo di corrispondenze nelle variabili di linkage, una coppia presa a caso tra quelle con quel profilo sia un vero matches. Naturalmente, più alto è il valore di questa probabilità più è probabile che la coppia considerata sia un vero link.

In questo caso, il metodo basato sull'analisi discriminante mostra un break point pari a b= 0.9014709 a partire dal quale i clerkes dovranno andare a verificare i records, sul database di riferimento, con p-post inferiore. Il tasso del fallimento di tale regola è pari a 19.7% (=256/1300).
Applicando lo stesso modello, con la stessa unica covariata, allo scenario Silver si ottiene un break point pari a b=0.9571004. In questo caso il tasso del errore/ fallimento è pari a 12,5%.
E' importante sottolineare che i risultati illustrati sono ottenuti utilizzando una sola covariata e questo può rendere instabile il risultato, mentre la disponibilità di almeno tre covariate permetterebbe di definire meglio le caratteristiche idiosincratiche per ciascun cluster.
Per questo motivo, come miglioramento del metodo proposto, sono state costruite nuove covariate. Procedendo a successivi passi di integrazione degli stessi file considerati originariamente e sfruttando variabili alternative, o combinazioni di diverse variabili derivate dallo scenario Gold, ed effettuando il merge fra tre modelli elaborati in Relais, si ottiene un dataset di pair matches di 1652 records con 327 Falsi e 1325 Veri maches. Il tasso del fal-

limento della regola del break point è nettamente migliorato scendendo all'8%. Tale regola, che è frutto di combinazioni di vari P.Post, è però di uso non immediato per i clerks.

5. Conclusioni

In questo lavoro abbiamo cercato di identificare dei metodi per migliorare la stima dell'errore di linkage, rispetto a quella forntia dal modello di Fellegi-Sunter, che come ben noto è un metodo molto efficace per la individuazione degli abbinamenti ma generalmente meno affidabile per la valutazione dei risultati ottenuti. Negli scenari individuati per le sperimentazioni, il metodo proposto da Belin e Rubin si è rivelato poco utile dati i suoi limiti legati alla mancata verifica delle ipotesi sulle distribuzioni. Per questo abbiamo proposto e testato un metodo che sfrutta i principi dell'analisi discriminante, che non presenta i limiti dell'approccio parametrico di Belin e Rubin ma che comunque non migliora le stime dell'errore fornite da Fellegi e Sunter. Futuri approfondimenti riguarderanno l'applicazione di modelli di predizione di tipo logistico e/o modelli non parametrici come gli alberi di classificazione e di regressione (Breiman et al,1984). In generale, l'uso degli alberi di classificazione può essere finalizzato sia a produrre un'accurata partizione della popolazione rispetto alla variabile target e, quindi, a ricostruire l'informazione sulle unità che appartengono allo stesso nodo di quelle per cui essa è nota sia a rivelare legami nascosti tra la variabile target e altre variabili esplicative. In questo contesto, la metodologia viene usata per identificare dei predittori dei true links rispetto ai false links, la variabile target diviene quindi l'essere un vero abbinamento rispetto all'insieme di tutti i matched dichiarati dalla procedura di abbinamento. E' evidente che, anche in questo contesto, l'uso di un training set su cui sia stato determinato il reale stato di abbinamento delle coppie è richiesto al fine di individuare i predittori del reale stato di abbinamento.

Bibliografia

Belin Thomas R., Rubin Donald B, A Method for Calibrating False-Match Rates in Record Linkage, Journal of American Statistical Association, June 1995, vol.90, no 430, pp.81-94.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. (1984), Classification and Regression Trees, Wadsworth International Group, Belmont, California, USA.

Fellegi, I.P., Sunter, A.B. (1969), *A Theory for Record Linkage*, Journal of the American Statistical Association, 64, pp. 1183-1210.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics 7 (2): 179–188.

McLeod, Heasman and Forbes, (2011) Simulated data for the on the job training, Essnet DI http://www.cros-portal.eu/content/job-training