L'automazione della rilevazione di outlier nel processo di Data Auditing

Alfredo Roccato
Marketing Strategico, Unicredit Banca
Alfredo.Roccato@unicredit.it

Riassunto: Most of the databases used by companies can include a large number of anomalous values generally termed as "outliers". The detection of such anomalies is important for the Data Quality process and therefore reduces the risk of making wrong decisions. Although many methods for their detection are often based on subjective criteria, or on predefined rules when the data range is known, it is necessary to rely on robust statistical methods when these rules are not applicable. In this paper we present an approach to automating the Data Auditing process of their individuation based on an example of application that uses some of these methods and compare results for a better insight.

Keywords: data quality, data auditing, outlier detection, robust methods

1. L'importanza della rilevazione dei valori anomali

La definizione statistica di outlier dipende dalla distribuzione sottostante le variabili in esame. Una definizione generale di outlier è data da Barnett e Lewis (1994): "Un'osservazione (o un sottoinsieme di osservazioni) che non sembra essere consistente con il resto di quell'insieme".
Possono essere diverse le cause di presenza di outlier e molto spesso sfuggono all'analista stesso: alcune volte gli outlier son dovuti a meri problemi di immissione (*data entry*) oppure ad errori di conversione e di misura.
Gli outlier, trovandosi spesso lontani dal centro della distribuzione, possono influenzare pesantemente le stime statistiche come la media e deviazione standard, rendendo non affidabili i risultati delle analisi eseguite sull'intero insieme dei dati.
Sebbene l'individuazione e l'esclusione degli outlier possano portare indubbi benefici a tutte le analisi che potranno essere eseguite, non è raccomandabile ignorarli del tutto: la loro eliminazione a priori non consentirebbe di giudicare la qualità dei dati stessi poiché, nel caso non fossero effettivamente "errori", si trascurerebbe una parte della variabilità intrinseca del fenomeno (non a caso l'identificazione degli outlier è una parte vitale del controllo di qualità).
Il valore aggiunto in questa fase sarà proprio l'apporto dato dagli esperti del dominio i quali, analizzando le segnalazioni prodotte dal Data Auditing, possono determinare quali anomalie rappresentano veri errori, ed eventualmente da eliminare, e quali sono da tenere, in ogni caso, in considerazione.

2. Metodi per l'individuazione dei valori anomali

Illustriamo qui di seguito alcuni metodi di natura esplorativa per l'individuazione dei valori anomali univariati, che esaminano individualmente ciascuna variabile, e multivariati, che tengono in conto delle relazioni tra le variabili presenti nello stesso insieme di dati.

2.1. Metodi univariati

La rilevazione degli outlier nel caso univariato è relativamente semplice ed esistono diverse tecniche per individuarli (Barnett e Lewis, 1994).

2.1.1. Deviazione standard (STD)

Viene stimato un parametro di posizione (media) e di dispersione (standard deviation). Un outlier è di solito definito tale se il suo valore si colloca al di fuori dell'intervallo *(media- 2 std, media + 2 std).* Alla base di tale tecnica vi è però l'assunzione di una distribuzione normale. Se la distribuzione fosse, in realtà, asimmetrica, considerare outlier i valori che eccedono di 2 o 3 volte la standard deviation, può non essere un buon criterio.

2.1.2. Distanza interquartilica (IQR)

Questo criterio utilizza la distanza interquartilica (IQR), data da *q3 (terzo quartile) – q1 (primo quartile),* ed un suo multiplo, per definire quale valore è da considerare outlier. Si possono identificare come outlier tutti quei valori che, di solito, sono esterni all'intervallo *(q1 - 1.5 iqr, q3 + 1.5 iqr)*.

2.1.3. Median Absolute Deviation (MAD)

E' una tecnica migliore delle precedenti poiché è meno sensibile all'influenza degli outlier: anche per questo motivo tali stimatori sono chiamati "robusti". La mediana è uno stimatore robusto di posizione. Anche per la mediana viene definito uno stimatore robusto di dispersione:

$$MAD = med_i \left\{ \left| x_i - med_j(x_j) \right| \right\}$$

dove med_j (x_j), è la mediana delle n osservazioni e med_i è la mediana degli n valori assoluti degli scarti dalla mediana. Definito un valore M_i per ogni osservazione, tale che:

$$M_i = \frac{0{,}6745(x_i - med_j(x_j))}{MAD} \quad [1]$$

[1], si definisce outlier ogni osservazione per cui $|M_i| > 3.5$ [2]

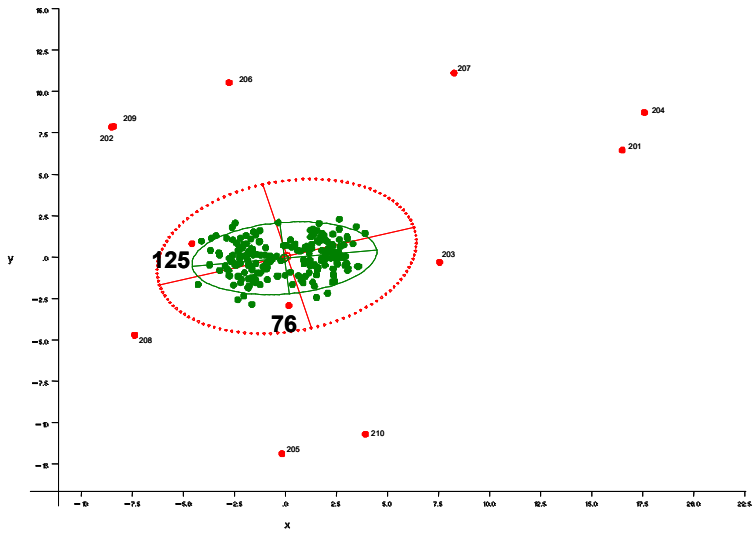
2.2. Metodi multivariati

La rilevazione di outlier multivariati si può fare solo esaminando più fenomeni (variabili) congiuntamente. Si può vedere un esempio in Figura 1 dove le osservazioni, identificate dai punti 76 e 125, possono essere considerate outlier solo con metodi multivariati: né l'una né l'altra possono essere, infatti, considerate outlier se analizzate singolarmente con un test univariato.

Figura 1: *Grafico di dispersione dei punti usati nell'esempio*

[1] Viene usata la costante 0,6745 poiché per una distribuzione normale, E(MAD)=0,6745 σ

[2] Il valore di 3,5 è quello che viene suggerito in letteratura (Iglewicz e Hoaglin, 1993)



La rilevazione degli outlier nel caso multivariato è però più complessa e viene affrontata con metodi robusti che sono, per loro natura, meno sensibili a scostamenti, anche lievi, di posizione e di dispersione.

2.2.1. Minimum Volume Ellipsoid (MVE) e Minimum Covariance Determinant (MCD)

Questi metodi sono stati sviluppati inizialmente da Rousseeuw nel 1984. L'idea di base è di identificare la "forma" principale dei dati, e di identificare come outlier quei punti che giacciono lontani dal loro insieme principale. Il primo metodo, chiamato MVE, perfezionato poi da Rousseeuw e Leroy (1987), cerca di trovare un ellissoide di volume minimo che possa comprendere almeno k osservazioni del suo insieme principale mentre il secondo, chiamato MCD, ricerca, da un insieme n di osservazioni, un suo sottoinsieme k per il quale il determinante della sua matrice di covarianza è minimo. Sebbene il metodo MCD abbia migliori proprietà teoriche, è di solito preferito il metodo MVE perché dispone di algoritmi più veloci.

2.2.2. Robust Mahalanobis Distances (MHL)

I metodi sopraindicati non possono essere utilizzati in pratica quando la numerosità dei dati diventa elevata, perché richiederebbero un dispendio di risorse eccessivo per il calcolo.
In ambito multivariato può essere utile ricorrere alle distanze di Mahalanobis per identificare come outlier le osservazioni che distano maggiormente dal "centro" dei dati. I valori anomali possono, però, alterare non solo il vettore delle medie, ma anche la matrice delle varianze e covarianze rendendo così difficile la loro rilevazione. Un'euristica spesso utilizzata è la procedura di "rifinitura" multivariata di Gnanadesikan e Ketternring (1972). Questa è una procedura iterativa che, ad ogni passo, misura la

distanza di Mahalanobis[3] D^2 tra l'osservazione *i*-esima e il centroide (l'equivalente multidimensionale della media). Le osservazioni più lontane dal centroide hanno un minor valore di probabilità e, quindi, possono essere verosimilmente considerate outliers. In questa procedura vengono tolte, ad ogni interazione, le osservazioni il cui valore di distanza ha una probabilità minore di un valore predefinito e ricalcolati, sulle osservazioni rimanenti, il nuovo vettore delle medie (il centroide) e la nuova matrice delle varianze e covarianze. Questa procedura è ripetuta finché non si trovano nuove osservazioni da togliere oppure è stato raggiunto il numero predefinito di iterazioni.

3. Confronto tra i risultati di varie tecniche

Nel processo di Data Auditing può essere utile disporre di strumenti che possano aiutare a identificare gli outlier utilizzando i metodi sopra illustrati confrontandone i risultati. Al tal scopo è stato sviluppato dall'autore un software specifico che utilizza il linguaggio matriciale SAS/IML.
La macro, OUTLIERS[4], permette all'utente di utilizzare con facilità uno o più dei metodi sopra illustrati tramite una serie di opzioni con le quali può impostare, oltre alla tabella e le variabili da analizzare, anche i criteri per l'individuazione.

Viene qui sotto riportata la sintassi di utilizzo della macro OUTLIERS con le opzioni previste:

```
%outliers (data=,       /* specifica il nome della tabella con i dati da analizzare                */
           vars=,       /* specifica il nome delle variabili della tabella                         */
           method=,     /* specifica il metodo di stima: std, iqr, mad, mve, mcd, mhl              */
           stdmult=,    /* opzione per il metodo std: il numero di deviazioni standard             */
           iqrmult=,    /* opzione per il metodo iqr: il numero distanze interquartiliche       */
           madthrs=,    /* opzione per il metodo mad: il valore di soglia di Hampel                */
           passes=,     /* opzione per il metodo mhl: il numero di iterazioni                      */
           pvalue=,     /* opzione per il metodo mhl: il valore p di χ² di soglia                  */
           test=,       /* identifica il test effettuato e il nome della colonna                   */
                        /* contenente gli outlier di quel test                                     */
           out=);       /* specifica il nome della tabella di uscita contenente, oltre             */
                        /* i dati da analizzare, le colonne contenenti, per ogni test,             */
                        /* gli outlier identificati                                                */
```

e vengono descritte nel seguito, per ogni metodo disponibile, le caratteristiche delle varie opzioni presenti nella macro:

Metodo STD: l'opzione *STDMULT=* permette di impostare la costante moltiplicativa della standard deviation. In quest'esempio sono stati identificati come outlier tutte le osservazioni in cui almeno un valore delle variabili è fuori dall'intervallo di 2 e 3 volte la standard deviation (rispettivamente i test STD1 e STD2):

---

[3] $D^2_{i\,(-)} = (\mathbf{x}_i - \bar{\mathbf{x}}_{(-)})'\mathbf{S}^{-1}_{(-)}(\mathbf{x}_i - \bar{\mathbf{x}}_{(-)})$, dove $\bar{\mathbf{x}}_{(-)}$ è il vettore delle medie e $\mathbf{S}_{(-)}$ è la matrice delle varianze e covarianze calcolati ad ogni iterazione.

[4] Il file di testo contenente la macro OUTLIERS, sviluppata dall'autore, è a disposizione di chiunque ne faccia richiesta

Metodo IQR: l'opzione *IQRMULT=* permette di impostare la costante moltiplicativa delle distanze interquartiliche. Anche in questi caso sono considerati outlier tutte quelle osservazioni in cui almeno un valore delle variabili è fuori dall'intervallo di 1,5 e 3 volte lo scarto interquartilico (rispettivamente i test IQR1 e IQR2):

Metodo MAD: l'opzione *MADTHRS=* permette di impostare il valore di soglia per il valore M_i come indicato in [1]. Gli outlier sono tutte le osservazioni in cui almeno un valore della i-esima variabile $|M_i|$ supera la soglia 2,57 (test MAD1) e 3.5 (test MAD2):

Metodo MHL: questo metodo utilizza la procedura di Gnanadesikan [2] per il calcolo delle distanze di Mahalanobis come proposta da Friendly (1991) in un programma macro SAS[5]. In tale ambito si esaminano iterativamente i dati escludendo di volta in volta le osservazioni per cui dove il valore della distanza di Mahalanobis D^2 ha un valore di probabilità χ^2 minore di *PVALUE=*, il cui valore predefinito è fissato a 0,05. Le osservazioni escluse vengono identificate come outlier e tolte dall'insieme dei dati per l'iterazione successiva. L'opzione *PASSES=* definisce il numero di iterazioni: in questo caso 2 (test MHL).

Metodi MVE e MCD: per questi metodi vengono utilizzate le impostazioni predefinite contenute nella macro OUTLIERS (rispettivamente i test MVE e MCD)[6]: Questi metodi consentono anche una rappresentazione grafica. Il grafico di dispersione riportato in Figura 1 è stato prodotto utilizzando il metodo MVE.

I dati utilizzati per il confronto sono stati creati artificialmente come nell'esempio riportato nel manuale SAS/STAT User's Guide (1989). Quell'esempio illustrava un utilizzo reiterato di cluster analysis finalizzato alla rimozione degli outlier. Tutti i valori delle 210 osservazioni sono numeri casuali ottenuti secondo una funzione normale standardizzata; per poterle meglio evidenziare come outlier, il valore delle ultime 10 è stato moltiplicato per una costante. In Figura 1 è rappresentato il grafico di dispersione di tali dati. In Tabella 1 sono riportati i risultati ottenuti utilizzando la macro OUTLIERS:

[5] La macro OUTLIER di M. Friendly è documentata in rete all'indirizzo http://www.math.yorku.ca/SCS/sssg/outlier.html

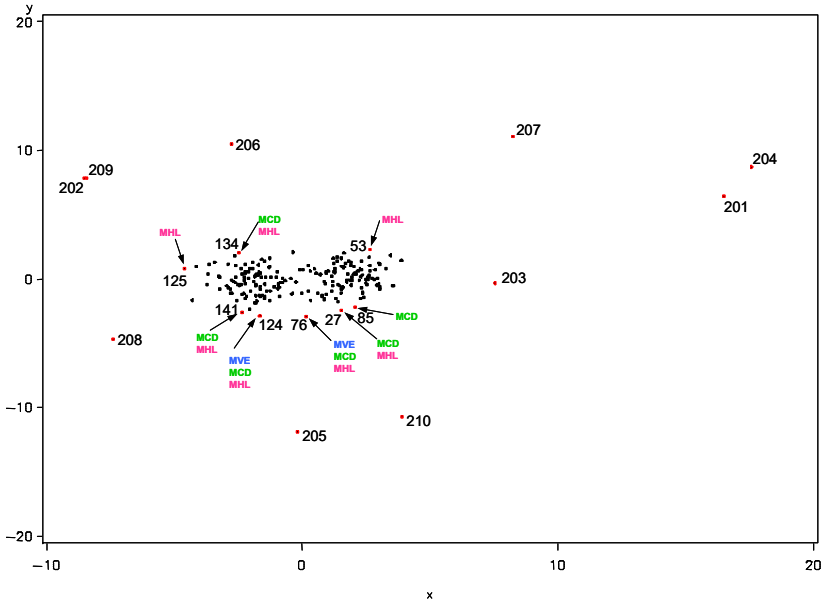
[6] Una descrizione degli argomenti delle routine MVE e MCD di SAS/IML si può trovare in rete all'indirizzo http://ummvsb.missouri.edu/sas8/sashtml/iml/chap17/sect151.htm - idxi170319.html

Tabella 1: *Confronto dei risultati ottenuti utilizzando i metodi disponibili nella macro OUTLIERS*

Outlier	X	Y	STD1	STD2	IQR1	IQR2	MAD1	MAD2	MVE	MCD	MHL
27	1,5198	-2,4317								√	√
53	2,6456	2,2957									√
76	0,1622	-2,9284					√		√	√	√
82	-0,9075	-0,4825									
85	2,0684	-2,1802								√	
104	-1,7987	-0,4298									
120	-0,0870	0,6941									
124	-1,6508	-2,8541							√	√	√
125	-4,5766	0,8138									√
134	-2,4626	2,0641								√	√
141	-2,3458	-2,5776								√	√
169	-2,6072	-0,2944									
170	-1,2368	-0,9273									
175	-0,4046	-1,1466									
177	-4,2869	-1,6436									
201	16,4976	6,4640	√	√	√	√	√	√	√	√	√
202	-8,5131	7,8566	√	√	√	√	√	√	√	√	√
203	7,5461	-0,3111	√		√		√		√	√	√
204	17,5795	8,7282	√	√	√	√	√	√	√	√	√
205	-0,1681	-11,8663	√	√	√	√	√	√	√	√	√
206	-2,7676	10,5338	√	√	√	√	√	√	√	√	√
207	8,2638	11,1083	√	√	√	√	√	√	√	√	√
208	-7,3873	-4,6949	√		√		√	√	√	√	√
209	-8,4105	7,8918	√	√	√	√	√	√	√	√	√
210	3,9173	-10,6898	√	√	√	√	√	√	√	√	√

Come si può notare i test univariati riescono sì ad individuare gli outlier estremi, ma non quelli più interni (per esempio le osservazioni 76 e 125), che possono perdere addirittura di efficacia se diventano troppo selettivi. Il test MAD1 riesce ad invidividuare l'outlier 76 ma perde anch'esso di efficacia se la soglia è troppo alta, come dimostra il test MAD2. I test MVE e MCD riescono a rilevare entrambi le osservazioni 76 e 124 mentre il solo test MCD ne rileva altre come outlier. I risultati del test MHL si avvicinano molto a quelli del test MCD. Nel grafico di Figura 2 sono rappresentati gli outlier ottenuti con i test MCD e MHL tranne quelli relativi le osservazioni dalla 201 alla 210 poiché presenti in tutti i test.

Figura 2: *Grafico di dispersione dei punti usati nell'esempio con l'evidenziazione degli outlier.*



4. Un caso di studio

Il caso di studio che qui si vuole presentare è relativo ad un'analisi interna all'Unità Operativa Marketing Strategico di Unicredit Banca il cui obiettivo era l'individuazione di zone cittadine interessanti per l'insediamento ottimale di nuovi siti commerciali (agenzie) compatibilmente con la disposizione della rete attuale. A tal fine si è scelto di utilizzare un precedente studio finalizzato alla valutazione della distribuzione della ricchezza delle famiglie bancarizzate (quelle che hanno almeno un rapporto bancario) per fasce di patrimonio presso la loro banca principale.
La tabella seguente riassume la stima del numero medio delle famiglie bancarizzate di Milano per fascia di patrimonio presso la loro banca principale:

Fascia di patrimonio (€) presso la banca principale	Numero medio di famiglie bancarizzate delle 5770 sezioni di censimento di Milano
0-25.000	85.4
25.000-75.000	14.0
75.000-250.000	10.8
250.000-500.000	3.1
500.000-1.000.000	1.8
>1.000.000	1.0

Lo studio permette di associare agli asset[7] finanziari calcolati per ogni singolo item (sezione di censimento) un dato di profittabilità economica (margine bancario generabile) che, sommato secondo criteri di adiacenza geografica, permette di evidenziare zone ad elevata concentrazione di potenziale economico e, come tali, particolarmente appetibili per localizzazioni commerciali (agenzie).
E' stata usata, per scopi esplorativi, la macro OUTLIERS per ognuno dei sei tipi diversi di metodi disponibili: l'interesse si è concentrato su quelle sezioni rilevate come outlier dai metodi multivariati.

La fase di verifica è stata effettuata analizzando i dati di tipo socio-demografico associati alle sezioni individuate come outlier. Studiando, ad esempio, la sezione di censimento 4521, e considerando i dati sintesi derivati dal censimento 1991, si è potuto scoprire il motivo dell'"anomalia": dal confronto della media dei dati Istat di questa sezione e la media cittadina si è notato un gran numero di individui, ma un numero molto esiguo di famiglie bancarizzate.
Ciò era dovuto al fatto che tale sezione comprendeva quasi esclusivamente la più famosa casa di riposo per anziani di Milano: in questo caso particolare una di quelle sparute "famiglie" era la casa di riposo stessa. La sezione è stata perciò considerata un vero e proprio outlier e, quindi, funzionale all'analisi specifica, che era finalizzata all'individuazione di zone ad alta concentrazione di ricchezza, è stato deciso di rimuoverla dall'insieme dalle osservazioni che sono servite per la costruzione del modello di stima dei margini generabili utili all'individuazione di vie candidate per l'apertura di nuovi sportelli bancari.

[7] Le disponibilità bancarie comprensive di depositi bancari, reddito fisso, gestioni patrimoniali, azioni, fondi e premi assicurativi.

5. Conclusione

La rilevazione e l'individuazione delle anomalie presenti nei dati, uno dei compiti principali del Data Auditing, è parte integrante del processo di Data Quality che sta ormai acquisendo un'importanza sempre più rilevante nell'ambito dell'Information Technology: le società commerciali si stanno infatti sempre più rendendo conto che informazioni non corrette possono portare anche a decisioni sbagliate che possono avere serie ripercussioni sul business.
Un possibile approccio, orientato all'automazione di tale rilevazione, consiste nell'utilizzo di programmi ad hoc, basati su robusti metodi statistici, che possano permettere all'analista di analizzare velocemente i dati confrontare i risultati ottenuti per rendere più efficace la loro individuazione.

Ringraziamenti

L'autore desidera ringraziare Salvatore Ingrassia e Maria Rosaria Ferrante per i loro commenti e suggerimenti durante la preparazione del presente lavoro; si ringrazia inoltre Michael Friendly per aver gentilmente concesso il permesso di utilizzare il codice presente nella sua macro come riferimento per lo sviluppo del metodo MHL nella macro OUTLIERS.

Riferimenti

Inmon,W. H.(1992) Building the DataWarehouse,Wiley, New York
M. Hernandez, S. Stolfo (1997) Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 2(1)
English, Larry P. (1999) *Improving Data Warehouse and Business Information Quality,* Wiley, New York
Luzi, O (2001) Analisi e verifica della qualità dei dati, Individuazione e correzione degli errori statistici, ISTAT
Barnett, V., Lewis, T. (1994) *Outliers in Statistical Data*. 3rd Edition, Wiley, New York
B. Iglewicz, D. Hoaglin (1993) *How to detect and handle outliers,* The ASQC Basic References in Quality Control: Statistical Techniques, Vol. 16, ASQC, pp.10-13.
Rousseeuw, P.J., Leroy, A.M. (1987) *Robust Regression and Outlier Detection*, Wiley, New York
Gnanadesikan, R., Kettenring J.R. (1972) Robust Estimates, Residual, and Outlier Detection with Multiresponse Data, *Biometrics*, 28-81
Friendly, M. (1991) SAS® System for Statistical Graphics. Cary, NC, SAS Institute Inc.
SAS/STAT® User's Guide (1989) Version 6, Fourth Edition, Volume 1, Cary, NC, SAS Institute Inc., 842-850
SAS/IML® Software (1989) Usage and Reference, Version 6, First Edition, Cary, NC, SAS Institute, Inc.
SAS/IML® Software (2001) Changes and Enhancements, Release 8.2, Cary, NC, SAS Institute Inc.