RELAZIONE PROGETTO ARCHITETTURE DATI

MISURAZIONE E MIGLIORAMENTO DELLA QUALITA' DEI DATI IN AMBITO FINANZIARIO

Beltramelli Fabio 816912

Finati Davide 817508

1) PREMESSA:

Il numero di informazioni presenti sul web sta nell'ultimo periodo diventando sempre più imponente, moltissimi dati riguardanti ambiti completamente diversi e forniti da molteplici fonti sono ormai consultabili da tutti coloro che ne hanno bisogno. Nascono però per tali motivazioni alcune problematiche sia nel gestire un numero così elevato di dati sia nell'ottenere informazioni rilevanti e corrette da essi.

Per prima cosa è stato quindi individuato un ambito di nostro interesse su cui poter misurare la qualità dei dati ed eventualmente migliorarla. L'ambito scelto è stato quello finanziario, siamo perciò partiti da un dataset composto da 1000 titoli su cui era già stato applicato un task di data integration rispetto alle molteplici fonti dalle quali questi dati provenivano.

La definizione di qualità dei dati non è univoca, poiché la sua misurazione avviene tramite diverse "dimensioni" che forniscono delle caratteristiche con le quali poter affermare se dei dati sono buoni o meno e in che misura. Nel nostro contesto sono state individuate diverse dimensioni ritenute rilevanti nel contesto preso in considerazione:

1) Completezza dei dati: un primo fattore importante che è stato analizzato è stata la dimensione della completezza. Per completezza di intende come e quanto le diverse fonti forniscano le stesse tipologie di dati e come queste influiscano nella problematica dei dati mancanti. Questo è un aspetto fondamentale per poi poter valutare in modo migliore anche le altre dimensioni e con il quale si può fornire una prima valutazione della fase precedente di integrazione.
2) Ridondanza: un secondo aspetto analizzato è stata la dimensione della ridondanza, la premessa nel nostro contesto è di sapere di avere un dataset con numerosi valori ripetuti, questo perché uno degli obiettivi è valutare come le diverse fonti si comportano nella descrizione degli stessi titoli. Quello che è stato fatto è quindi un'analisi volta a confermare tale ipotesi.
3) Consistenza: tale dimensione serve a misurare quanto i dati rispettino dei vincoli di dominio e vincoli di integrità tra i dati stessi. L'analisi ha comportato l'applicazione e la misurazione di vari criteri, da quelli più banali fino a controlli più complessi.
4) Precisione: un ultimo aspetto considerato è stata la dimensione della precisione dei dati, per fare ciò è stato preso in considerazione uno standard (da noi scelto come i dati ottenuti da Nasdaq.com) che è stato utilizzato come metro di confronto per valutare dapprima la precisione degli attributi presi singolarmente e in secondo luogo la precisione delle diverse fonti.

2) DOMINIO ED OBIETTIVI:

Dopo una prima valutazione del dataset è stato deciso di lavorare su una porzione più ristretta dei dati che comprendesse i 100 titoli quotati sull'indice Nasdaq per le solo cinque fonti più "autorevoli" durante la sessione di mercato del 01/07/2011, questo ha permesso delle analisi più accurate ed evitato che fonti meno riconosciute andassero a modificare in maniera eccessiva i risultati di tali analisi.

L'obiettivo principale del progetto è stato quindi misurare la bontà delle dimensioni scelte e cercare di ottenere dei miglioramenti su di esse.

3) DATASET:

Il dataset considerato è composto dai seguenti attributi:

- Source: sorgente dei dati (Bloomberg, GoogleFinance, MSNMoney, Nasdaq, YahooFinance)
- Symbol: ticker dell'azienda
- ChangePerc: variazione percentuale del titolo
- ClosePrice: valore dell'azione al momento della chiusura del mercato
- OpenPrice: valore dell'azione al momento dell'apertura del mercato
- ChangeInDollars: variazione del valore del titolo durante la giornata
- Volume: numero di contratti scambiati durante la sessione di mercato
- HighPrice: valore massimo dell'azione raggiunto durante la sessione di mercato
- LowPrice: valore minimo dell'azione raggiunto durante la sessione di mercato
- PreviousClose: valore dell'azione alla chiusura del mercato il giorno precedente
- YearHigh: valore massimo dell'azione raggiunto durante l'anno
- YearLow: valore minimo dell'azione raggiunto durante l'anno
- NShares: numero di azioni in circolazione
- PE: rapporto prezzo/utili per una singola azione
- MarketCap: valore totale di una azienda (Nshares * ClosePrice)
- Dividend: dividendo distribuito per azione
- DividendYield: rapporto tra dividendo e prezzo di una azione
- EPS: utile per azione (utile netto / Nshares)

4) ANALISI DATASET E DOMINIO APPLICATIVO:

Abbiamo effettuato un'analisi iniziale sul dataset e del dominio finanziario, tali analisi introduttive sono state utilizzate per valutare in maniera migliore le dimensioni di qualità, attraverso tale analisi sono stati notati i seguenti aspetti:

1. L'attributo PE matematicamente può assumere un valore negativo che però non è accettato nella cultura finanziaria, nel dataset quattro fonti su cinque infatti non riportano i valori di PE negativi, al contrario solo msn-money riporta tali valori
2. Come vedremo nell'analisi di consistenza il valore che assume l'attributo PreviousClose può non coincidere con il valore di PriceOpen in quanto sono possibili transazioni a mercato chiuso
3. L'attributo Dividend differisce nelle varie fonti in quanto il suo calcolo può essere fatto utilizzando scale differenti (annuale, semestrale, quarto di anno)

5) METRICHE CON DATASET INIZIALE:

Verranno ora riportate le analisi effettuate sulle metriche individuate in precedenza sul dataset originale, tali metriche sono state misurate normalizzando i dati, cercando perciò di ovviare ai vari problemi di eterogeneità sui dati e gestendo i valori mancanti

5.1) COMPLETEZZA:

I risultati più rilevanti della misurazione di tale dimensione sono stati i seguenti:

- Completezza di attributi:

  40% di NULL nella colonna PreviousClose.

  20% di NULL nella colonna NShares.

  57% di NULL nella colonna Dividend.

  58% di NULL nella colonna DividendYield.

  0.4% di NULL nella colonna EPS.

- Completezza totale:

  10% di NULL nell'intero dataset.

Tale analisi ha mostrato come all'interno del dataset, pur avendo preso in considerazione solo le fonti autorevoli, queste siano composte da set di attributi differenti che generano nel dataset valori mancanti, in particolare due fonti su cinque non riportano i valori di previous close, cioè il valore dell'azione al momento della chiusura del mercato il giorno precedente (40% di nulli) ed una fonte su cinque non riporta invece l'attributi Nshares, cioè il numero di azioni in circolazione per quel titolo (20% di nulli).

Per quanto riguarda invece gli attributi Dividend e Dividend Yield la motivazione dei valori nulli è valida in quanto non è obbligatorio per una azienda distribuire dividenti e perciò tale valore potrebbe risultare mancante, tale assunzione è difatti riscontrato per tutte le fonti.

Globalmente il valore di attributi nulli è abbastanza contenuto nonostante siano presenti le problematiche appena descritte.

L'analisi di tale dimensione è importante nella misurazione delle altre poiché valori mancanti non possono essere valutati in maniera consistente con dati che invece sono valorizzati, per questo primo approccio si è perciò deciso di non prendere in considerazione i valori che risultano mancanti per non influenzare negativamente la misurazione delle altre metriche.

5.2) RIDONDANZA:

Come già preventivato nel nostro campo di applicazione siamo di fronte ad una ridondanza molto elevata.

In particolare, avendo preso in considerazione solo le "fonti autorevoli" e solo i titoli presenti nell'indice Nasdaq tutte le fonti forniscono valori su tutti i titoli, quindi a livello di oggetti abbiamo una ridondanza del 100% (poiché in sostanza ogni titolo è ripetuto per le cinque fonti disponibili)

A livello di attributi invece la ridondanza è minore in quanto le diverse fonti forniscono set di attributi diversi (vedi dimensione completezza).

In generale però nel nostro contesto siamo in presenza di una ridondanza generale molto alta, data però dal dominio stesso di tale analisi e perciò tenuta correttamente in considerazione durante la realizzazione del progetto

5.3) CONSISTENZA:

Queste sono state le principali modalità con le quali è stata valutata l'inconsistenza nei dati:

Prima attraverso delle analisi banali come controllare se i valori fossero maggiori di zero, in seguito sono state poi effettuate analisi di consistenza più complesse:

- HighPrice < LowPrice (0%)
- HighPrice < OpenPrice && HighPrice < ClosePrice (0%)
- LowPrice > OpenPrice && LowPrice > ClosePrice (0%)
- YearHigh < YearLow (0%)
- PreviousClose > YearHigh && PreviousClose < YearLow (0%)
- ChangeInDollars != ClosePrice - PreviousClose (58,4%)
- ChangePerc != (ChangeInDollars/PreviousClose) * 100 (59,4%)
- DividendYield != (Dividend/PreviousClose) * 100 (24,6%)
- MarketCap != NShares * ClosePrice (64,4%)

Sono state riportate con il valore in parentesi la percentuale di dati che non soddisfa ognuno di tali criteri di consistenza

Le inconsistenze più elevate sono state riscontrate per gli ultimi tre criteri.

La prima riguarda il valore assunto dall'attributo ChangeInDollars che nel 58,4% dei casi non rispetta la sua semantica, cioè di essere la differenza tra il prezzo del titolo in chiusura di mercato meno il valore al momento della chiusura del mercato il giorno precedente.

La seconda è legata alla prima in quanto indica la percentuale di cambiamento di prezzo al momento della chiusura di mercato del giorno odierno con quello precedente, in questo caso il 59,4% dei dati non rispetta tale vincolo

Il terzo vincolo invece riguarda l'attributo MarketCap che dovrebbe corrispondere al numero di azioni presenti sul mercato (Nshares) per il valore del titolo (ClosePrice), il 64,4% dei dati non rispetta tale vincolo

È stata notata inoltre anche un'inconsistenza nell'attributi di PreviousClose che in molti casi non è uguale all'OpenPrice, questo è stato riportato sottoforma di warning in quanto, come detto in precedenza, è possibile che non sia un errore in quanto sono ammesse anche transazioni a mercato chiuso che non fanno coincidere i due valori.

5.4) PRECISIONE:

Questa dimensione è stata misurata secondo due visioni, una controllando la precisione per ogni attributo mentre la seconda valutando come si comportano le diverse fonti; in entrambi casi è stato utilizzato come metro di paragone uno standard di dati ottenuti di Nasdaq.com

Risultati misurazione precisione sugli attributi:

- ClosePrice: 95,4%
- OpenPrice: 81%
- ChangePrec: 78.4%
- ChangeInDollars: 79.6%
- Volume: 38.6%
- HighPrice: 94.6%
- LowPrice: 95.2%
- PreviousClose: 58,8% *
- YearHigh: 86%
- YearLow: 94%
- NShares: 20%
- PE:34,2%
- MarketCap: 20%
- Dividend: 64,4% *
- DividendYield: 64% *
- EPS: 20,2

Per quanto riguarda la precisione sugli attributi ci sono diverse ragioni per spiegare gli attributi con bassa precisione:

1. Eterogeneità semantica: nella maggioranza casi la bassa precisione è dovuta ad una diversa semantica di rappresentazione rispetto alla groundtruth, soprattutto nel caso di attributi con decimali (diverso arrotondamento e formato) (es. Nshares, MarketCap, PE, EPS)
2. Errori nelle unità di misura: sono presenti alcuni errori di unità di misura, per esempio la maggior parte delle fonti riporta 20M mentre una fonte riporta 20B
3. Errori nei dati: In alcuni casi sono stati riscontrati dei puri errori nei valori dei dati

Risultati misurazione precisione sulle fonti:

- bloomberg: 87%
- google finance: 76%
- msn-money: 77%
- nasdaq: 99% *
- yahoo finance: 23% **

* Tale valore è giustamente elevato poiché la groundtruth considera dati presi da Nasdaq.com

** C'è un'eterogeneità nel calcolare l'attributo che descrive il prezzo di apertura che nella maggioranza dei casi è diverso dalle altre fonti, togliendo quell'attributo la precisione sarebbe del 69%

Sono state calcolate le precisioni rispetto ai quattro attributi fondamentali (OpenPrice, ClosePrice, HighPrice e LowPrice) in quanto le fonti non condividono lo stesso set di attributi.

In generale si può notare come (oltre a nasdaq che è stato preso come groundtruth) la fonte con precisione più alta sia Bloomberg ed al contrario la fonte con precisione più bassa sia Yahoo che però riscontra un problema di eterogeneità su un attributo.

6) MIGLIORAMENTO DIMENSIONI:

Dopo aver ottenuto una misurazione delle dimensioni utilizzando il dataset originale sono state applicate delle modifiche al dataset con l'obiettivo di migliorare tali misurazioni, è stata perciò utilizzata una strategia data-driven, andando cioè a modificare direttamente i dati e non i processi con i quali gli stessi sono ottenuti.

Si è ricavato un nuovo dataset da quello di partenza, applicando delle tecniche di pulizia dei dati, come ad esempio assegnare il valore più frequente tra le varie fonti per i campi nulli. In seguito sono state rimosse le stringhe presenti tra i valori numerici ($,m,mil,b,bil).

I risultati più rilevanti sono i seguenti:

- L'incompletezza delle colonne PreviousClose e NShares ha così raggiunto lo 0%. In generale il dataset ha avuto un miglioramento arrivando al 5% di dati mancanti (partendo dal 10%).
- Il miglioramento nella precisione più rilevante è avvenuto del PreviousClose è passata dal 58% al 97,6%. Anche la precisione di NShares è aumentata fino al 52%.
- L'inconsistenza di ChangeInDollars è diminuita fino al 45%, così come ChangePerc. Anche il valore di inconsistenza di MarketCap è diminuito molto (38%).
- I casi particolari PE < 0 hanno raggiunto lo 0,2%, invece quelli di PreviousClose != OpenClose è decresciuto fino al 42,2%.

7) CONCLUSIONI:

- Le tecniche utilizzate per migliorare le dimensioni hanno apportato un miglioramento visibile dalla misurazione di tali dimensioni prima e dopo la modifica del dataset
- I risultati della valutazione delle varie dimensioni di qualità dei dati hanno mostrato un insieme non indifferente di aspetti importanti, tra cui:
    1. presenza di dati mancanti, che sono stati però parzialmente gestiti modificando il dataset
    2. gestione dell'eterogeneità semantica dei dati che porta ad una diminuzione delle performance (nonostante normalizzazione e gestione valori nulli), anche questo aspetto è stato migliorato se pur in maniera parziale andando a modificare il dataset
    3. precisione globale sulle fonti non da disprezzare, questa però è ancora una volta peggiorata dai problemi di eterogeneità e mancanza dei valori che ne peggiorano le performance per alcuni attributi

8) SVILUPPI FUTURI:

- Migliorare la fase iniziale di normalizzazione dei dati al fine di evitare errori dovuti ad eterogeneità semantica

- Migliorare le tecniche data-driven di pulizia dei dati al fine di migliorare ulteriormente la qualità degli stessi
- Attuare altri tipi di approccio (process-driven) al fine di migliorare la qualità dei dati