

Appunti Database

Brendon Mendicino

December 3, 2022

Contents

1	Introduzione	4
2	Data Warehouse	4
3	Analisi	5
3.1	Finestra di calcolo	6
3.2	Sintassi ORACLE	8
3.3	Esercizi	9
4	Viste materializzate	12
4.1	Documentazione Oracle	12
5	Progettazione fisica	14
6	Alimentazione dei Data Warehouse	15
6.1	Estrazione	15
6.2	Pulitura	15
6.3	Trasformazione	15
6.4	Caricamento	16
7	Data Lake	18
8	Data processing	18
8.1	Data aggregation	19
8.2	Data reduction	19
9	Regole di associazione	22
9.1	Algoritmi Di Estrazione Di Itemsets	23
9.1.1	Apriori	23
9.1.2	FP-Growth	24
9.2	Effetto delle soglie	26
10	Classificazione	27
10.1	Alberi di Decisione	27
10.2	Random Forest	29
10.3	Rule-based Classification	29
10.4	Classificazione Associativa	30
10.5	K-Nearest Neighbor (KNN)	30
10.6	Bayesian Classification	31
10.7	Support Vector Machines	31
10.8	Artificial Neural Network	32
10.9	Model evaluation	33

11 Clustering fundamentals	35
12 Trigger	36
13 Lab 03	39
14 Clustering	41
14.1 K-means	41
14.2 Bisecting K-means	42
15 Introduzione ai DBMS	42
15.1 Buffer Manager	44
15.2 Accesso alla Memoria	45
15.3 Progettazione Fisica	47

1 Introduzione

KDD: Knowledge Discovery from Data

Tecniche di data mining

- Regole di associazione: usate per trovare delle relazioni frequenti all'interno del database. Ad esempio: chi compra pannolini compra anche birra, il 2% degli scontrini contengono entrambe gli oggetti, il 30% degli scontrini che contengono pannolini contengono anche birra. Grazie alle regole di associazione si possono fare dei tipi di analisi come la basket analysis, ma può essere utile anche per le raccomandazioni.
- Classificazione: i classificatori predicono etichette discrete, esempio: nella posta elettronica alcune mail vengono segnate come spam. La classificazione definisce un modello per definire le predizioni, a volte non è sempre possibile creare dei modelli interpretabili ovvero dare una ragione per una determinata scelta.
- Clustering: gli algoritmi creano dei gruppi che raggruppano gli oggetti in esame, senza però dare delle motivazioni delle scelte effettuate.

2 Data Warehouse

Un DW è una base dati di supporto alle decisioni, che è mantenuta separatamente dalla base di dati operativa dell'azienda. I dati al suo interno sono:

- orientati ai soggetti di interesse;
- integrati e consistenti;
- dipendenti dal tempo;
- non volatili;
- utilizzati per il supporto alle decisioni aziendali;

Per la progettazione concettuale di un DW, non esiste un formalismo universale, il modello ER non è adatto ma viene invece utilizzata il modello **Dimensional Fact Model**.

Il DFM è composto da:

- Fatto: modella un insieme di eventi di interesse, che evolvono nel tempo (che può avere diversa granularità).
- Dimensioni: sono gli attributi di un fatto, generalmente sono categorici.

- Misure: descrive una caratteristica numerica di un fatto.

Sulle dimensioni si possono definire delle gerarchie, che definiscono di fatto una dipendenza funzionale tra gli attributi, quindi di 1 a n. Ad esempio: **data** ha un arco **mese**, una data ha uno ed un solo mese (1 a n).

I costrutti avanzati sono:

- archi opzionali;
- dimensioni opzionali;
- attributo descrittivo: sono delle informazioni utili all'utente ma su cui non verteranno le interrogazioni (ad esempio non si farà mai la group by su un indirizzo);
- non-additività: non si può fare la somma sulla metrica, il motivo è che non è modellato in modo tale da fare la somma;
- Fatto: fenomeno di studio;
- Misure: attributi del fatto;
- Dimensioni: tabelle collegate al fatto;

Schema a stella:

Snowflake scheme: si esplicitano le dipendenze funzionali, questo però comporta un aumento delle operazioni di join.

Nella realtà lo snowflake è raramente utilizzato, il motivo è che il costo delle join può diventare oneroso. Un caso di utilizzo dello snowflake è quando si hanno dei dati condivisi.

Archi multipli:

Dimensioni degeneri: sono delle dimensioni con un solo attributo, questo si perchè nello stato attuale non si hanno delle specifiche per quell'attributo ma nel futuro si potrebbe facilmente estendere. Un'altra soluzione potrebbe essere un push down delle dimensioni degeneri nella tabella dei fatti.

Junk Dimension: si può creare una dimensione che contenga tutte le dimensioni degeneri, le informazioni sono collegate semanticamente, è anche possibile unire delle informazioni scorrelate ma non è una scelta poco corretta, una soluzione potrebbe essere avere più junk dimensions.

3 Analisi

Sfruttando solo l'SQL è molto difficile fare delle analisi su un dw, infatti volendo calcolare delle operazioni per due argomenti diversi si devono fare più query. Estendendo il SQL si può, ad esempio, effettuare più operazioni leggendo una sola volta la tabella, ed effettuando il minor numero di join possibile.

Analisi OLAP I tipi di operazione sono:

- roll up: riducendo il livello di dettaglio del dato, ovvero eliminare una o più clausole della groupby o navigare la gerarchia verso l'esterno;
- drill down: si aumenta il livello di dettaglio oppure si aggiunge una dimensione di analisi;
- slice and dice: consentono di ridurre il volume dei dati selezionando un sottogruppo dei dati di partenza;
- tabelle pivot: come viene mostrato il dato;
- ordinamento: ordinamento in base agli attributi;

Queste operazioni possono essere fatte con più o una query.

3.1 Finestra di calcolo

Una finestra di calcolo fa dei calcoli a partire da una query sottostante, la finestra ha 3 operazioni sottostanti:

- partizionamento (**partition by**): partizionamento dei dati, divide i record in gruppi a partire dall'attributo selezionato;
- ordinamento (**order by**): si definisce il criterio di ordinamento delle righe all'interno dei partizionamenti;
- finestra di aggregazione (**over**): porzione di dati, specifica per ogni riga di dato, su cui effettuare dei calcoli;

Example 3.1

Data la tabella Vendite(Città, Mese, Importo), calcolare per ogni città la media delle vendite per il mese corrente ed i due precedenti.

```
1 SELECT Città, Mese, Importo,
2     AVG(Importo) OVER (PARTITION BY Città)
3                       ORDER BY Mese
4                       ROWS 2 PRECEDING)
5     AS MediaMobile
6 FROM Vendite;
```

Quando la finestra è incompleta il calcolo è effettuato sulla parte presente, è possibile specificare che se la riga non è presente il risultato deve essere NULL.

Si può definire un intervallo fisico, superiore o inferiore.

```
1 ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING
```

È possibile definire la tupla corrente e quella che la precedono e che la seguono

```
1 ROWS UNBOUNDED PRECEDING (o FOLLOWING)
```

Il raggruppamento fisico è specifico per quando i dati non hanno delle interruzioni.

Per definire un intervallo logico si utilizza il costrutto **range**.

```
1 SELECT Citta, Mese, Importo,
2        Importo / SUM(Importo) OVER () AS PerOverMax,
3        Importo / SUM(Importo) OVER (PARTITION BY Citta) AS PerOverCity,
4        Importo / SUM(Importo) OVER (PARTITION BY Mese) AS PerOverMonth
5 FROM Vendite
```

Se una **group by** è presente all'interno della query allora, tutte le entry che possono comparire nella finestra di calcolo sono solo quelle che compaiono nella group by.

Funzione di ranking La funzione di ranking serve a creare delle classifiche

- **rank()**: la funzione rank in presenza di più oggetti nella stessa posizione salta al prossimo record;
- **denserank()**: la funzione denserank tiene tutte righe con la stessa posizione;

...

```
1 SELECT Citta, Mese, SUM(Importo) AS TotMese,
2        RANK() OVER (PARTITION BY Citta
3                     ORDER BY SUM(Importo) DESC)
4
5 FROM Vendite, ...
6 WHERE ...
7 GROUP BY Citta, Mese
```

Estensione della group by

- **rollup**: consente di calcolare le aggregazioni su tutti i possibili gruppi, eliminando una colonna alla volta, da destra verso sinistra, esempio: calcola le vendite per: (Citta, Mese, Prodotto), (Citta, Mese), (Citta):

```
1 SELECT Citta, Mese, Prodotto, SUM(Importo) AS TotVendite
2 FROM ...
3 WHERE ...
4 GROUP BY ROLLUP (Citta, Mese, Prodotto)
5
```

- **cube**: consente di calcolare tutte le possibili combinazioni del raggruppamento;
- **grouping sets**: serve a definire degli aggregati su gruppi specifici definiti dall'utente;

3.2 Sintassi ORACLE

Raggruppamento fisico:

Example 3.2

Selezionare, separatamente per ogni città, per ogni data l'importo e la media dell'importo dei due giorni precedenti.

```
1 select citta, data, importo,
2     avg(importo) over (partition by citta
3                        order by data
4                        rows 2 preceding
5     ) as mediaMobile
6 from vendite
7 order by citta, data;
8
```

Raggruppamento logico:

Example 3.3

```
1 select citta, data, importo,
2     avg(importo) over (PARTITION BY citta
3                        ORDER BY data
4                        RANGE BETWEEN INTERVAL '2'
5                        DAY PRECEDING AND CURRENT ROW
6     ) as mediaUltimi3Giorni
7 from vendite
8 order by citta, data;
9
```

Example 3.4

```
1 select COD_A, sum(Q) as sommaPerArticolo,
2     rank() over (order by sum(Q) desc) as graduatoria
3 from FAP
4 group by COD_A
5
```

All'interno di oracle sono preseti delle funzionalità aggiuntive oltre alla funzione di rank, come:

ROW_NUMBER Assegno un numero progressivo ad ogni elemento in una partizione.

```
1 select tipo, peso,
```



```

2     row_number over (partition by tipo
3                       order by tipo)
4 from ...
5 where ...;

```

CUME_DIST Consente di calcolare le distribuzioni cumulative all'interno di una partizione, permette di definire un valore sulla distribuzione dei valori.

NTILE(n) Una funzione che dà la possibilità di dividere le partizioni in sottogruppi

```

1 select tipo, perso,
2        ntile(3) over (partition by tipo order by peso) as ntile3peso
3 from ...
4 where ...;

```

3.3 Esercizi

Cliente(CodCliente, Cliente, Provincia, Regione)

Categoria(CodCat, Categoria)

Agente(CodAgente, Agente, Agenzia)

Tempo(CodTempo, Mese, Trimestre, Semestre, Anno)

Fatturato(CodTempo, CodCliente, CodCatArticolo, CodAgente, TotFatturato, NumArticoli, TotSconto)

1. Visualizzare per ogni categoria di articoli

- la categoria
- la quantità totale fatturata per la categoria in esame
- il fatturato totale associato alla categoria in esame
- il rank della categoria in funzione della quantità totale fatturata
- il rank della categoria in funzione del fatturato totale

```

1 select categoria, sum(numArticoli),
2        sum(totFatturato),
3        rank() over (order by sum(numArticoli) desc),
4        rank() over (order by sum(totFatturato) desc)
5 from fatturato f, categoria c
6 where f.codCatArticolo = c.codCat
7 group by categoria;

```

2. Visualizzare per ogni provincia

- la provincia
- la regione della provincia
- il fatturato totale associato alla provincia
- il rank della provincia in funzione del fatturato totale, separato per regione

```
1 select provincia, regione,
2     sum(totFatturato) as fatturatoPerProvincia,
3     rank() over (partition by regione
4                  order by sum(totFatturato) desc
5                  ) as rankFatturatoPerRegione
6 from cliente c, fatturato f
7 where c.codCliente = f.codCliente
8 group by provincia, regione;
```

3. Visualizzare per ogni provincia e mese

- la provincia
- la regione della provincia
- il mese
- il fatturato totale associato alla provincia nel mese in esame
- il rank della provincia in funzione del fatturato totale, separato per mese

```
1 select provincia, regione, mese,
2     sum(totFatturato) as fatturatoPerProvinciaPerMese,
3     rank() over (partition by mese
4                  order by sum(totFatturato)
5                  ) as rankFatturatoPerMese
6 from cliente c, fatturato f, tempo t
7 where c.codCliente = f.codCliente and t.codTempo = f.codTempo
8 group by provincia, regione, mese;
```

4. Visualizzare per ogni regione e mese

- la regione
- il mese
- il fatturato totale associato alla regione nel mese in esame
- l'incasso cumulativo al trascorrere dei mesi, separato per ogni regione
- l'incasso cumulativo al trascorrere dei mesi, separato per ogni anno e regione

```
1 select regione, mese,
2     sum(TotFatturato) as fatturatoPerMese,
3     sum(TotFatturato) over (
4         partition by regione
5         order by mese
6         rows unbounde preceding
7     ) as incassoCumulativoTot,
8     sum(TotFatturato) over (
9         partition by regione, anno
10        order by mese
11        rows unbounded preceding
12    ) as incassoCumulativoPerAnno
13 from cliente c, fatturato f, tempo t
14 where c.CodCliente = f.CodCliente and t.CodTempo = f.CodTempo
15 group by regione, mese, anno;
```

4 Viste materializzate

Le viste materializzate sono necessarie per ridurre la lentezza delle operazioni di group by per grandi moli di dati, le viste materializzate sono dei sommari precalcolati della tabella dei fatti.

Le VM usano con il costrutto principale la group by, quando si crea una VM è conveniente includere anche le dimensioni a granularità superiori, in modo da poter riutilizzare la tabella.

Per rappresentare le dipendenze delle viste materializzate si utilizza un **reticolo multidimensionale**. Più ci si trova in alto al reticolo più ci si avvicina alle dimensioni della tabella dei fatti, più si va in basso più si trova una granularità maggiore.

La scelta delle viste tra tutte le possibili combinazioni è data da:

- si sceglie una sola vista da cui è possibile raggiungere tutti gli attributi;
- creo una vista per ogni query;
- scelgo delle viste intermedie che possono portare a rispondere a più query;

4.1 Documentazione Oracle

Riducono i tempi di esecuzione delle group by e non si eseguono più le join. Nel DBMS Oracle esiste la **query rewriting**, che permette grazie all'ottimizzatore di interpretare le query e se i risultati corrispondono alle condizioni di creazione delle viste, allora la query viene riscritta con la vista.

```
1 create materialized view NAME
2 [build {immediate|deferred}]
3 [refresh {complete|fast|force|never}
4         {on commit|on demand}]
5 [enable query rewrite]
6 as
7     QUERY
```

- immediate: lo schema della tabella viene popolata immediatamente, dato dallo schema di attributi presenti nella select;
- deferred: la vista viene creata, ma viene popolata successivamente;
- complete: i dati vengono presi interamente dal database;
- fast: i dati vengono presi in modo incrementale;
- force: se possibile viene eseguito il refresh in modalità fast, oppure in modalità complete;
- never: la vista non viene mai aggiornata;

- on commit: ogni volta che viene fatto un commit sulla tabella della query anche la vista viene aggiornata;
- on demand: viene definito dall'utente quando aggiornare la vista;
- enable query rewrite: abilita il dbms ad usare la vista per accelerare le query;

Per effettuare il refresh esistono dei tipi di job (a differenza del tipo di prodotto). Quando abbiamo bisogno del fast refresh, la tabella ha bisogno delle informazioni aggiuntive, ovvero dei file di log che ci informano delle nuove informazioni aggiunte al db, la **materialized view log** è associata ad una tabella che ha subito delle variazioni:

```
1 create materialized view log on
2   TABELLA
3 with sequence, rowid
4   (Attributo, ...)
5 including new values;
```

- sequence: istante temporale in cui è avvenuta la modifica;
- rowid: indica la tupla che ha subito una modifica;

Su queste keyword si deve definire una lista di attributi da monitorare, si aggiunge **including new values** per supportare l'inserimento di nuove tuple.

5 Progettazione fisica

Fare una progettazione fisica comporta analizzare il carico di lavoro e definire delle strutture fisiche accessorie per velocizzare le operazioni. Si possono definire delle viste oppure degli indici, ad esempio: indici bitmap, indice di join, ...

La progettazione fisica è dipendente dal carico di lavoro.

La progettazione fisica è caratterizzata da una fase di tuning, utilizzata per testare gli indici e le viste create e decidere se mantenerli o meno.

Gli indici si possono creare sugli attributi che vengono selezionati più frequentemente, se il dominio è ridotto (come quelli categorici dei DW) si utilizza un bitmap, altrimenti un B-tree.

6 Alimentazione dei Data Warehouse

Essendo dei dati derivati, la prima operazione da effettuare è l'ETL, se questo è complesso si va a definire un area di staging in cui il dato viene mantenuto temporaneamente. Il processo di ETL va gestito sia per il popolamento del DW sia per quando verrà aggiornato con dati nuovi.

6.1 Estrazione

L'estrazione statica è la prima estrazione effettuata per popolare il DW. Per fare l'estrazione incrementale si possono:

- creare delle applicazioni ad hoc per i sistemi legacy;
- usando dei log, che non vanno ad interferire con il carico del db;
- usando dei trigger: sono procedure che si attivano quando degli si effettuano delle operazioni specifiche;
- basata su timestamp: dove i recordi hanno il timestamp dell'ultima modifica effettuata su di essi;

6.2 Pulitura

Quando si effettua una estrazione ci si potrebbe trovare di fronte a:

- dati duplicati;
- dati mancanti;
- campo non previsto;
- valori errati o impossibili;
- inconsistenza del valore;

Ogni errore richiede una tecnica specifica per essere risolto, le più comuni sono l'uso di dizionari per controllare errori di battitura, oppure il **join approssimato**, ad esempio: due database non hanno una chiave condivisa per identificare un utente dall'ordine effettuato, allora per fare una join si dovranno prendere i campi comuni, controllandone sempre la consistenza, oppure i problemi di **merge/purge**, ad esempio: facendo il merge di due db le informazioni potrebbero essere duplicate ...

6.3 Trasformazione

Conversione dei dati nel formato di quelli presenti nel data warehouse.

6.4 Caricamento

In fase di caricamento i dati si caricano nel seguente ordine:

- dimensioni;
- fatti;
- indici e viste;

Problem 6.1 – Progettazione Magazzini

Tabelle:

Tempo(codT, data, mese, 3m, 4m, 6m, anno)

Magazzino(codMa, magazzino, città, provincia, regione)

Modello(codMo, modello, categoria)

UsoMtqMagazzino(codMa, codT, mtqLiberi, mtqTot)

UsoProdMagazzino(codMa, codMo, codT, numeroProdottiTotale, valoreTotaleProdotti)

Query:

1. Relativamente al primo trimestre dell'anno 2013, considerando solo i magazzini della città di Torino, trovare per ogni coppia (magazzino, data) il valore complessivo di prodotti presenti in tale data nel magazzino e il valore complessivo medio giornaliero di prodotti presenti nel magazzino nel corso della settimana precedente la data in esame (data in esame inclusa):

```

1 select magazzino, data,
2     sum(valoreTotProdotti) as valoreTot,
3     avg(sum(valoreTotProdotti)) over (
4         partition by magazzino
5         order by data
6         range between interval '7'
7         day preceding and current row
8     ) as valoreMedioSuGiornoCorrenteESettimanaPrecedente
9
10 from UsoProdMagazzino u, Tempo t, Magazzino m
11 where u.codT = t.codT and
12        u.codMa = m.codMa and
13        città = 'torino' and
14        anno = 2013 and
15        3m = 1
16 group by magazzino, data;
```

2. Relativamente all'anno 2004, trovare per ogni coppia(città,data) la percentuale di superficie liberagiornaliera nella città. Associare ad ogni coppia un attributo di rank legato alla percentuale disuperficie libera giornaliera nella città (1 per la coppia con la più bassa percentuale di superficie libera giornaliera).


```
1 select citta, data,
2       sum(mtqLiberi) / sum(mtqTot) * 100 as
   percentualeMtqLiberi,
3       rank() over (
4         order by sum(mtqLiberi) / sum(mtqTot) * 100
5       ) as rankLowestPercentuale
6 from Tempo t, Magazzino m, UsoMtqMagazzino u
7 where t.codT = u.codT and
8       m.codMa = u.codMa and
9       anno = 2004
10 group by citta, data;
```

3. Relativamente ai primi sei mesi dell'anno 2014, trovare per ogni coppia (magazzino,data) la percentuale di superficie libera giornaliera.

```
1 select magazzino, data,
2       100 * sum(mtqLiberi) / sum(mtqTot) as
   percentualeMtqLiberi
3 from Tempo t, Magazzino m, UsoMtqMagazzino u
4 where t.codT = u.codT and
5       m.codMa = u.codMa and
6       anno = 2014 and
7       mese <= 6
8 group by magazzino, data;
```

4. Relativamente all'anno 2013, trovare per ogni coppia (magazzino,mese) il valore complessivo medio giornaliero di prodotti presenti.

```
1 select distinct magazzino, mese,
2       avg(sum(valoreTotProdotti)) over (
3         partition by magazzino, mese
4       ) as valoreMedioGiornalieroComplessivo
5 from UsoProdMagazzino u, Tempo t, Magazzino m
6 where u.codMa = m.codMa and
7       u.codT = t.codT and
8       anno = 2013
9 group by magazzino, mese, data;
```

5. Relativamente all'anno 2015, trovare per ogni regione il valore complessivo medio giornaliero di prodotti presenti nella regione.

```
1 select regione, mese,
2       avg(sum(valoreTotProdotti)) over (
3         partition by regione, mese
4       ) as valoreMedioGiornalieroComplessivo
5 from UsoProdMagazzino u, Tempo t, Magazzino m
6 where u.codMa = m.codMa and
7       u.codT = t.codT and
8       anno = 2015
9 group by regione, mese, data;
```

6. Relativamente all'anno 2014, trovare per ogni coppia(mese, regione) la percentuale di superficie libera giornaliera nella regione.

```
1 select regione, mese,
2     avg (100 * sum(mtqLiberi) / sum(mtqTot)) over (
3         partition by regione
4     )
5 from UsoMtqMagazzino u, Tempo t, Magazzino m
6 where u.codMa = m.codMa and
7       u.codT = t.codT and
8       anno = 2014
9 group by regione, mese, data;
```

7 Data Lake

I data lake sono dei repository di dati, storicizzati per utilizzo futuro così come sono disponibili, in qualsiasi formato. Questi raw data potrebbero essere utilizzati in futuro.

I data lake danno la possibilità di storicizzare i dati per un uso futuro, inoltre tutti i dati sono contenuti in un repository comune, infatti è caratterizzato da bassi costi di storage e mantenimento, però può essere difficile estrapolare dei dati.

8 Data processing

Una collezione è costituita da oggetti di dato,

- **Attributo:** è una proprietà dell'oggetto;
- **Tipi di Attributo:** possono essere nominali, ordinali, intervalli, rapporti;
- **Proprietà dei valori degli attributi:** possiamo definire equivalenza, ordine, addizione, moltiplicazione;
- **Attributi discreti e continui:** discreti hanno un numero finito di valori, i continui hanno dei valori reali;

Tipi di dato da analizzare:

- record: sono i dati presenti in una tabella;
- grafi: come la struttura di una pagina web;
- ordinato: dati in cui esiste il concetto di sequenza;

Esistono vari tipi di dato:

- **Document Data** Per ogni riga ho un documento, per ogni colonna ho degli attributi che descrivono delle parole chiavi all'interno del documento, ogni riga è un array che contiene la pesatura degli attributi, la pesatura può essere calcolata con algoritmi specifici;
- **Dato transazionale** Un dato transazionale è formato da un insieme di items all'interno della tabella, ogni transazione è identificata da un ID, nelle tabelle transazionali non esiste il concetto di ordine né tra le righe, né all'interno della transazione;
- **Dato a grafo**: ad esempio le pagine web hanno dei link ad altre pagine, potrei considerare i link come gli archi del grafo, ognuno con un peso specifico, e la singola pagina come un nodo del grafo;
- **Qualità del dato**: nella maggior parte dei casi i dati presentano degli errori, questi possono essere causati da rumori e outliers, dati mancanti o duplica, ...;
- **Outliers**: dati che escono al di fuori del comportamento medio e quindi molto rumorosi, l'obiettivo dell'analisi di outlier detection è l'individuazione di questi dati e dell'eliminazione;

8.1 Data aggregation

Consente di combinare più record o più attributi, si fa questo per effettuare una riduzione della quantità di dati, ed avere dei dati più stabili con una variabilità minore.

8.2 Data reduction

Si possono effettuare due tipi di riduzione: riduzione degli attributi o riduzione dei valori. Per effettuare queste riduzioni esistono delle tecniche specifiche.

Una di esse è il **sampling** è una tecnica di statistica di analisi, ad esempio nella pipeline di data science il sampling serve per trovare delle rappresentazioni adatte al dataset, facendo N esperimenti su un sample preso dal dataset si andranno a fare dei test sull'intero dataset per verificare che le ipotesi siano confermate, in figura si può vedere come il grafico col numero più piccolo di campioni non sia più rappresentativo del dataset originale.

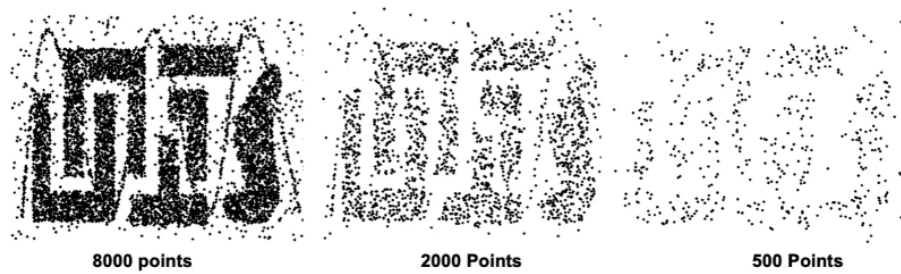


Figure 1: Sampling Example

I tipi di sampling possono essere:

- randomici: un dato estratto non viene reinserito;
- con rimpiazzo: un dato estratto può essere riestratto;
- sampling stratificato (più utilizzato): si può stratificare il dataset a partire da uno o più attributi (le partizioni vengono detti bucket), e poi vengono presi dei valori casuali;

Il problema con questi tipi di approccio è che all'aumentare delle dimensioni i dati diventano più distanti tra loro. Per ridurre gli effetti di avere troppe dimensioni si possono applicare delle tecniche di riduzione dimensionale. Per questo motivo vengono utilizzate tecniche di **dimensionality reduction**, alcune di queste sono:

- **pca, svd, ...**: tecniche statistiche;
- **feature selection**: si vuole avere la proiezione del dato su un nuovo attributo in modo da aumentare la varianza, in qualche modo ridurre le feature ridondanti: le tecniche di feature selection sono: brute force; embedded approach (tramite un albero di selezione si selezionano solo i dati più significativi, può essere fatto quando gli attributi non sono elevati); filter (basati sull'analisi di correlazione, per verificare se esistono delle correlazioni lineari); wrapper (viene utilizzato del data mining come black-box per identificare delle combinazioni di feature); feature creation (consiste nel combinare più variabili in una sola variabile, viene effettuato anche un processo di feature selection);
- **discretizzazione**: per effettuare una discretizzazione si deve mappare un valore continuo in un range di numeri discreti, una tecnica è quella di generare degli intervalli di uguale lunghezza e se la variabile ricade in un intervallo gli viene associato il simbolo corrispondente, questo potrebbe modellare dei valori outliers o rumorosi, si potrebbe anche usare del clustering, ovvero l'aggregazione di dati in base alla distanza tra vari valori, solitamente vengono usate due tecniche per validare i dati dopo la pipeline. Un caso particolare della discretizzazione è

la **binarizzazione** che consiste nel discretizzare una variabile e poi viene fatto il one-hot encoding (i valori vengono mappati su una bitmap);

- **trasformazione**: un attributo va trasformato quando si vogliono riportare i valori in un'altra scala, una delle tecniche più comuni è la normalizzazione, ad esempio negli algoritmi di clustering viene definito uno spazio normato per calcolare la distanza tra i valori, le normalizzazioni più usate sono:

Theorem 8.1 – min-max

$$v' = \frac{v - \min}{\max - \min}(\text{new_max} - \text{new_min}) + \text{new_min}$$

Theorem 8.2 – z-score

$$v' = \frac{v - \text{mean}}{\text{stand_dev}}$$

Theorem 8.3 – decimal scaling

$$v' = \frac{v}{10^j}$$

j intero più piccolo tale che: $\max(|v'|) < 1$

La similarità e la dissimilarità ci permettono di dire quando degli attributi sono simili o dissimili tra di loro, la similarità viene espresso in un intervallo $[0, 1]$, con $1 =$ identici, per definire la similarità si definisce un concetto di distanza, ed una **matrice di similarità**, in cui ogni riga e colonna sono presenti i valori, ogni cella rappresenta le distanze tra i due valori. Le tre distanze utilizzate sono: Manhattan, Euclidea, Mikowski. In caso queste distanze non soddisfino i criteri si definisce una distanza attraverso uno spazio vettoriale.

Una distanza importante è la **distanza di mahalanobis**, che mostra quanto due punti sono distanti in una distribuzione.

Theorem 8.4 – Mahalanobis

$$\text{Maha}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})$$

Σ matrice delle covarianze.

La correlazione è molto importante per trovare relazioni tra i dati. Una tecnica molto usata è trovare le combinazioni lineari attraverso il coefficiente di pearson.

9 Regole di associazione

L'estrazione delle regole di associazione è un modo di trovare delle associazioni tra i valori presenti in un database transazionale. Per decidere queste associazioni solitamente si guardano le ricorrenze statistiche di valori comuni.

Una regola di associazione si definisce come:

Theorem 9.1 – Regola di associazione

$$A, B \implies C$$

Dove degli insiemi di oggetti (**itemset**) possono implicare altri insiemi.

- A, B = corpo della regola;
- C = testa della regola;

La freccia indica la **co-occorrenza**, indica che il corpo è legato alla testa nelle transazioni del db. Per esempio: coca, pannolini \implies latte.

Se si lavora con un db relazione possiamo estrarre una transazione associando ad ogni valore il suo attributo.

Definizioni:

Definition 9.1 – k-itemset

È un itemset che contiene k oggetti.

Definition 9.2 – support count

È la frequenza con cui una transazione appare.

$$\begin{aligned} \#Roma, Colosseo &= 2 \\ sup(Roma, Colosseo) &= 2 \end{aligned}$$

Data una regola di associazione $A \implies B$:

Definition 9.3 – Supporto

$$\frac{\#A, B}{|T|}$$

È la frazione di transazioni che contengono sia A e B. $|T|$ è la cardinalità del db.

Definition 9.4 – Confidenza

$$\frac{sup(A, B)}{sup(A)}$$

È la frequenza delle transazioni B che contengono anche A.

Per creare dei modelli per estrarre le relazioni, vengono definiti due parametri che indicano la frequenza con quale frequenza devono apparire le relazioni:

- supporto > **minsup** threshold;
- confidenza > **minconf** threshold;

Questo viene fatto per limitare il numero di relazioni che vengono estratte, perché nella maggior parte dei casi i db sono molto grandi.

L'estrazione si compone di due fasi:

1. si estraggono gli itemset frequenti, attraverso il vincolo sul supporto;
2. si estraggono le regole di associazione, attraverso il vincolo sulla confidenza;

Il passo più oneroso è l'estrazione degli itemset frequenti.

Il **candidato** è l'oggetto che potrebbe essere estratto, il **frequente** è l'oggetto che supera il valore di minsup. Per migliorare

9.1 Algoritmi Di Estrazione Di Itemsets

9.1.1 Apriori

Apriori si basa sul principio di quanto un itemset è frequente, allora la porzione di lattice che ha come radice l'itemset allora si può esplorare il sottospazio dell'itemset, altrimenti no.

$$A \subset B \implies sup(A) \geq sup(B)$$

L'algoritmo di **Apriori** si basa sulla suddivisione degli itemset in livelli, in ogni livello si prendono dei candidati, facendo il join tra candidati frequenti (che superano la minconf) di livello k si generano candidati di livello k+1, per itemset che non rispettano il criterio di frequenza viene fatto il pruning dell'albero delle scelte.

Sui candidati di lunghezza 2 va applicato il pruning (apriori). Alla fine si troverà l'insieme delle soluzioni che sarà l'unione di tutti gli itemset estratti.

Le limitazioni principali di questo algoritmo sono i costi di scansione del database, infatti dovrà essere letta più volte, inoltre se le transazioni sono molto lunghe l'algoritmo dovrà essere ripetuto n+1 volte (n = lunghezza transazione), per superare questo limite si possono usare degli algoritmi per diminuire i costi legati alla lettura.

9.1.2 FP-Growth

Sono state proposte delle varianti dell'algoritmo per ottimizzare i problemi . Negli anni 2000 è stato proposto un nuovo algoritmo basato sulla memorizzazione. L'algoritmo **FP-growth**, instanzia in memoria un albero dove si trova la proiezione del database originale che considera gli item che soddisfano la soglia di supporto, una volta creata la struttura di supporto in memoria non si accede più in memoria secondaria, ottimizzando la lettura degli item. La struttura in memoria prende il nome di **FP-tree**. L'algoritmo funziona nel seguente modo:

1. vengono contati gli item e vengono scartati quelli con sotto la soglia;
2. viene creata una header table con gli item in ordine decrescente rispetto al supporto;
3. viene scansionata per l'ultima volta il database, le transazioni vengono ordinate in base alla header table;
4. ogni scansione di una transazione ordinata, viene preso l'item ed inserito nell'albero, ogni nodo contiene l'item ed il numero di volte che è stato trovato per ogni transazione letta;
5. l'inserimento nell'albero viene fatto a partire dall'ordine degli item nella transazione (simile agli alberi formati a partire da ogni lettera di parole), ogni volta che si passa da un nodo già inserito il suo contatore aumenta;
6. è importante collegare la header table ai nodi nell'albero, questo viene fatto attraverso una **node link chain**: l'header punta ad un nodo con lo stesso item, quando percorrendo l'albero si trova un altro item il nodo precedente avrà come nodo successivo il nodo corrente;

Viene proposto anche un algoritmo di visita per estrarre gli itemset:

1. viene letta la header table dall'item col supporto più basso;
2. viene creato un **conditional pattern base** di un item, una proiezione dell'fp-tree condizionato ad un item;
3. il CPB viene visitato ricorsivamente per trovare i candidati;
4. ad ogni nodo visitato viene recuperato il path dell'albero fino ad esso, se il supporto del nodo è minore del minsup il path viene scartato, e si passa al prossimo nodo della node-link chain;

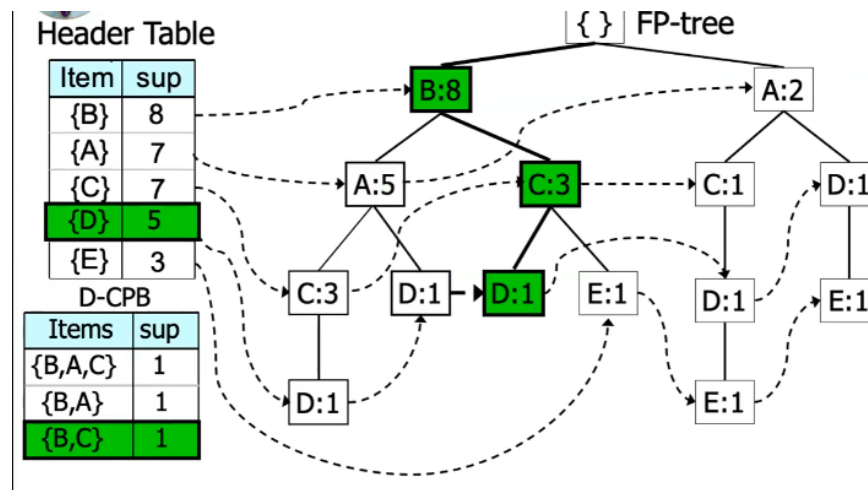


Figure 2: Cpd di D al terzo step

5. data la cpb va creata una Conditional header table, da questa header table, si crea un'altro fp-tree, questo si applica ricorsivamente, ogni chiamata è condizionata (ovvero la cpd sarà preceduta dall'itemset del chiamante: D -, DC -, ...);
6. quando non si riesce a creare una header table si torna al chiamante e si passa alla entry superiore nella header table del chiamante;
7. prima di creare la nuova cpb, si prende l'itemset che arriva dal chiamante e se il valore nell'header tabel del item corrente è maggiore del minsup allora l'itemset concatenato all'item corrente viene insireti negli itemset frequenti;

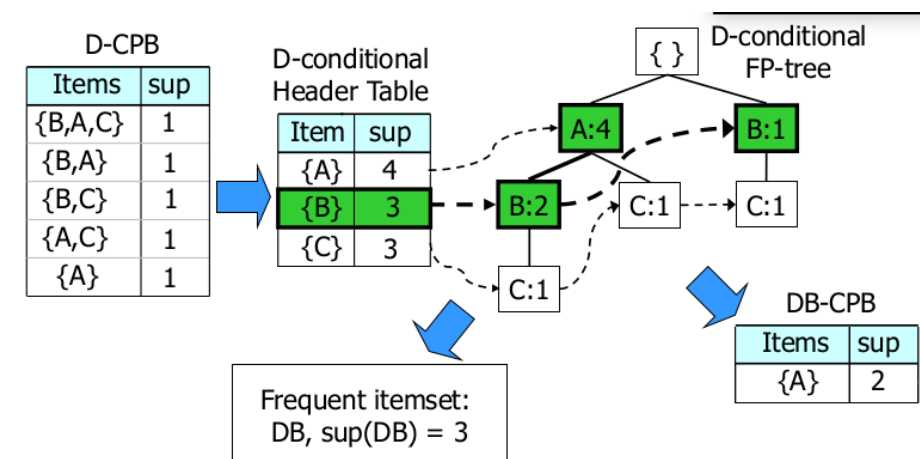


Figure 3: Esempio Aggiunta Itemset Nei Valori Frequenti

Questo algoritmo funziona molto bene se la memoria non viene saturata (bottleneck).

Un altro problema sono l'alto numero di itemset che vengono estratti, infatti anche con db molto piccoli si possono trovare un numero molto elevato di itemset frequenti, per questo si opta per i **itemset massimali frequenti**, IMF per definizione sono gli itemset che hanno il massimo numero di figli, e sono unici nel loro sottolivello, oppure esistono i **closed itemset** che sono gli itemset con nessuno dei suoi immediati superset hanno lo stesso valore.

9.2 Effetto delle soglie

La scelta dei valori di supporto deve essere idoneo, infatti con un valore troppo basso non si identificano delle relazioni e con valori troppo alti emergono relazioni molto deboli. Anche la confidenza comporta dei problemi, se il valore di cui si calcola la confidenza ha un valore molto ampio, si rischiano di ottenere valori errati, per evitare di incrociare queste informazioni si utilizza la regola del lift.

Definition 9.5 – Lift

Dato $r : A \implies B$, allora la Correlazione o lift è:

$$C = \frac{P(A, B)}{P(A)P(B)} = \frac{conf(r)}{sup(B)}$$

Un esempio di regola di associazione potrebbe anche essere l'aggregazione di dati, andano ad accoppiare una tassonomia ai valori, possiamo, aggregare gli attributi, vedere il loro supporto crescere, rappresentando in modo generalizzato un comportamento, per andare a soddisfare un servizio.

10 Classificazione

Le classificazioni cercano, attraverso dei modelli di assegnare dei tag ai dati, attraverso delle tecniche supervised (vuol dire che abbiamo già a disposizione un pool di dati da cui possiamo estrapolare le informazioni per assegnare un tipo di tag).

Per applicare la classificazione si ha bisogno di dei dati di training che hanno già dei tag con il quale si va a generare un modello, per classificare dei nuovi dati si parte dandoli in pasto al modello e partendo dai valori degli attributi si generano delle nuove etichette.

Per poter realizzare un modello di classificazione si ha bisogno di dati di training, usati per generare il modello, e dati di test, usati per validare il modello, ognuno di questi dati ha già associato ad essi una classe di tag. Una volta che si trova un modello adatto, si può insireri in un applicazione per predire i tag.

Gli algoritmi che generano i modelli hanno delle caratteristiche:

- accuratezza;
- interpretabilità;
- incrementalità: il modello può essere aggiornato all'arrivo di nuovi dati;
- efficienza;
- scalabilità: performance dell'algoritmo rispetto al numero di dati;
- robustezza: capacità dell'algoritmo di operare in presenza di dati rumorosi o mancanti;

10.1 Alberi di Decisione

Attraverso un albero di decisione, dati i dati di input è possibile inferire la classe di etichetta.

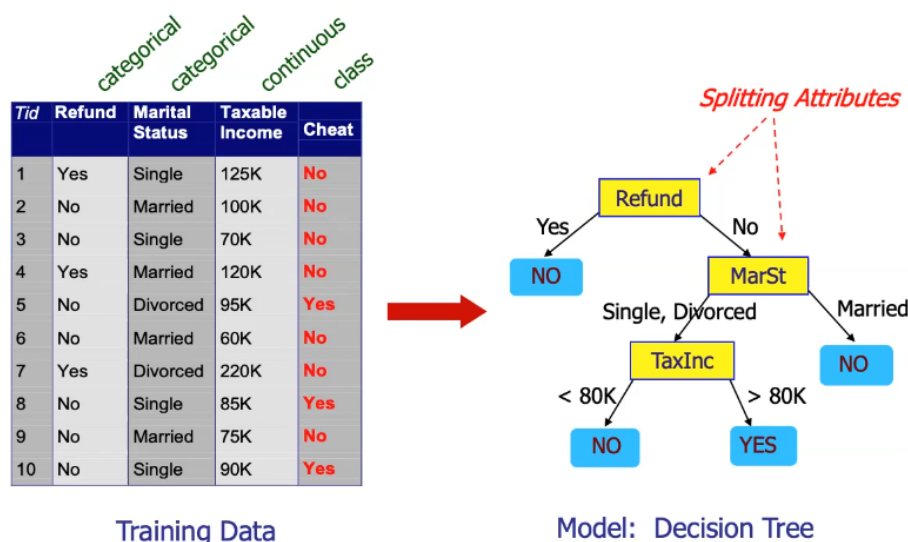


Figure 4: Albero Di Descisione

In un albero di decisione le foglie corrispondono all'etichetta di classe che sarà assegnata all'item in esame. Questo modello dipende molto dai dati, infatti se gli attributi cambiano anche il modello va modificato opportunamente, questo è dovuto al fatto che gli alberi di decisione non sono incrementali. Per generare gli alberi di decisione esistono vari algoritmi, uno di questi è l'**algoritmo di Hunt**, l'algoritmo parte dal leggere il database, il primo passo è quello di individuare l'attributo che riesce a dividere il db in due gruppi più omogenei possibili rispetto al tag. Successivamente si cerca il prossimo attributo che effettua il prossimo miglior partizionamento, fino al raggiungimento delle condizioni di terminazione. Questo albero non è aggiornabile all'arrivo di nuovi dati.

Per effettuare lo split bisogna partire dai tipi di dati con cui si lavora:

- attributo categorico: si possono effettuare n-split diversi per ogni valore, se il tag è binario allora i valori vengono partizionati in due gruppi;
- attributo numerici: si può usare la discretizzazione, oppure si può usare una condizione di test;

Per stimare la purezza (grado di omogeneità) dei nodi che vengono generati, esistono delle metriche per calcolare l'impurità: gini index, entropia, missclassification index. Il primo caso è quello di utilizzo del **Gini index**, con questo metodo viene misurata l'impurità prima e dopo lo split, il gini index si misura con:

$$GINI(t) = 1 - \sum_j (p(j|t))^2$$

$p(j|t)$ = frequenza della classe j al nodo t . Più il gini index si avvicina allo zero, più la classe è **pura**, più il valore si avvicina a $(1 - \frac{1}{\text{numro di classi}})$ più il nodo è impuro, viene

preso l'attributo che genera il gini index col valori più basso. Nelle implementazioni reali vengono fatti dei test utilizzando metriche diverse con successive validazioni.

Per la creazione di un albero va definito anche un criterio di stop:

- ...
- pre-pruning: se un nodo è quasi puro rimuovo non scendo più nell'albero;
- post-pruning: generato l'albero completo vado a tagliare i rami dell'albero con delle caratteristiche troppo specifiche;

10.2 Random Forest

I random forest sono un'estensione degli alberi di decisione (più alberi di decisione), dato un db si creano n sottoinsiemi dei dati originali, su ogni sottoinsieme si crea un albero di decisione, ognuno di questi modelli per decidere l'etichetta, successivamente si sceglie l'etichetta finale in base ai voti (possono anche essere pesati).

Si parte con il **Bootstrap** che va a estrarre gli n sottoinsiemi (con ripescamento) di oggetti randomici, base alla cardinalità, poi si va a creare un albero di decisione applicando una selezione degli attributi (non tutti gli alberi potrebbero utilizzare le stesse feature), ma si scelgono gli attributi che meglio modellano i dati che si vogliono analizzare.

10.3 Rule-based Classification

Si parte da un database di training. Le classificazioni rule-based sono composte da: (condizione) \implies y, dove la condizione è in insieme di condizioni booleane e y è la classe. Un esempio è:

$$(\text{Blood type} == \text{warm}) \ \&\& \ (\text{can fly} == \text{yes}) \implies \text{Bird}$$

Dato il db di training si genera il modello, che è composto da n regole R.

Ogni regola R che matcha la entry che si sta elaborando, viene associata la classe all'istanza, se esistono delle collisioni, si può utilizzare la mutua esclusione, e soprattutto (idealmente) le regole devono essere esaustive (non devono esistere casi non gestiti).

Per creare le regole si può partire da un albero di decisione, in questo modo si risolve la mutua esclusione ma non l'eshaustività. La strategia che solitamente si segue è il semplificazione di queste regole, le regole possono essere semplificate al completamento dell'albero di decisione, oppure si potrebbe semplificare l'albero con del pruning e poi estrarre le regole, oppure estrarre le regole direttamente dal db. Quando le regole non sono esaustive si usa una classe di default che solitamente è la maggioritaria.

Le caratteristiche di questi modelli sono:

- accuratezza maggiore degli alberi di decisione;
- interpretabili;
- non sono incrementali;
- molto efficienti;
- sono scalabili sia sul training set che sul numero di attributi;
- sono robusti agli outliers;

10.4 Classificazione Associativa

Si utilizzano le regole di associazione per fare delle previsioni, le regole di associazione però vanno ristrette, per la classificazione le uniche regole che interessano sono (condizione) \implies regola di classe (un sottoinsieme delle regole possibili). In questo caso l'estrazione è diversa dalle regole di associazione, infatti la classificazione associativa è composta da regole ordinate con degli indici di qualità, l'ordinamento è fatto sulle regole di associazione (soglia, confidenza, lift). Anche in questo caso non si vuole fare dell'overfitting sul training set, quindi in fase di creazione si dovrà decidere dove fare del pruning.

Le caratteristiche sono:

- ha un'accuratezza maggiore alla rule-based;
- modello interpretabile;
- non incrementale;
- efficienza bassa: dato dall'estrazione delle regole di associazione (dato dal min-sup);
- la scalabilità dipende dai dati: dimensione dell'FP-growth (ad esempio);
- non soffre di missing value e robusto agli outliers;

10.5 K-Nearest Neighbor (KNN)

È un algoritmo di classificazione che non ha un modello, ma si basa sul dataset di training, in questo modo non viene effettuata alcuna fase di training. Quando si vuole classificare un nuovo item, si cercano delle similarità dal dataset di training e poi si assegna la classe. Per decidere quali dati dall'itemset di training sono i più simili al nuovo item si utilizza un concetto di vicinanza, trovati i suoi vicini si assegna la classe, il numero di vicini che vengono scelti è dato dal valore K, la scelta di questo valore

influisce col rumore che possono causare i vicini. Le misure di distanza dipende dal tipo di dato.

Le caratteristiche sono:

- l'accuratezza è simile ai precedenti;
- il modello non è interpretabile;
- il training set può essere incrementale;
- tempi di classificazione lunghi;
- la scalabilità è determinata dalla cardinalità dell'training set;

10.6 Bayesian Classification

Theorem 10.1 – Teorema di Bayes

$$P(C, X) = P(X) \cdot P(C|X)$$

La classificazione bayesiana si basu sul teorema di bayes, questo tipo di classificazione presuppone un indipendenza statistica dei dati tra di loro, che solitamente è l'ipotesi sbagliata. Quando si vuole classificare una nuova tupla X rispetto ad un classe C si calcola $P(X|C)$ per ogni classe, il valore più alto trovato corrisponderà alla classe di appartenenza delle tupla. Per calcolare $P(X|C)$ si calcola:

$$\prod_{i=1}^N P(x_i|C)$$

dove x_i è un attributo della tupla X (è per questo motivo che si richiede l'indipendenza statistica, altrimenti quella moltiplicazione non sarebbe possibile).

10.7 Support Vector Machines

È una tecnica non interpretabile che va identificare un iperpiano che va a separare le classi di interesse. Questo iperpiano deve massimizzare il suo margine (distanza tra l'iperpiano i punti più vicini delle classi). È possibile definire anche delle curve non-lineari modificando il kernel dello spazio vettoriale.

Le caratteristiche sono:

- performance tra le migliori;
- non interpretabile;
- non incrementale;

- molto efficiente con il giusto tuning;
- scalabilità media;
- robusti ad outlier e rumori;

10.8 Artificial Neural Network

L'algoritmo utilizza delle unità di computazioni detti **neuroni**, questi neuroni sono collegati tra di loro attraverso **sinapsi**. Esistono vari modelli di rete neurale:

- Feed Forward NN: la più basilare;
- Convolutional NN, prima viene modellato il dato utilizzando dei filtri, che poi viene dato ad un FFNN per la computazione finale;
- Rcurrent NN,
- auto-encoder: vengono fatte delle elaborazione di dato dove viene fatto del denoising;

FFNN si ha un livello di input ed un livello di output ovvero l'etichetta di classe che deve essere predetta, questo modello è fully connected, ogni nodo genera un output che va in input ad ogni nodo del livello successivo, queste connessioni sono pesate. Ogni nodo è composto da: la somma di tutti gli output pesati, poi questo valore moltiplicato per un coefficiente e poi dato in pasto ad una funzione di attivazione che genera l'output. Le funzioni di attivazione decise come elaborare i valori, le più utilizzate sono la sigmoide e la tangente iperbolica. Altre funzioni sono il binary step o la rampa. La funzione utilizzata per l'output finale è la softmax.

Costruire un algoritmo di rete neurale viene fatto con un approccio iterativo:

1. ai vettori peso vengono assegnati dei valori casuali;
2. l'istanza processata viene processata da tutti i layer della rete, l'output viene confrontato con il vero valore dell'etichetta;
3. ad ogni iterazione viene fatta una backpropagation per aggiustare i valori dei pesi e dei nodi;
4. l'iterazione finisce quando si raggiunge una certa percentuale di guess corretti o quando si finisce il dataset di training;

Le caratteristiche sono:

- performance migliori;
- non interpretabile;

- non incrementale;
- il training è molto lento ma la classificazione è molto veloce;
- richiede grandi moli di dati ma il training diventa più lento;
- robusti in presenza di dati rumorosi ed outliers, questo è ottenuto con un grande pool di dati di test;

Un'architettura utilizzata sono le **Convolutional NN**, la parte di computazione convoluzionale vanno ad effettuare delle feature selection sui dati in input, questi dati vengono poi mandati in input in una rete FFNN. Per estrarre le feature un layer convoluzione fa operazione:

1. convoluzione;
2. funzione di attivazione;
3. pooling;

I dati sono rappresentati in tensori (matrice multidimensionale), ogni matrice di input è un tensore, l'output è anch'esso generato in un matrice tensoriale. La convoluzione viene effettuando applicando una finestra di filtro, che viene fatta scivolare su tutto il tensore, solitamnte viene aggiunto un padding ai bordi, ogni eleborzione viene effettuata considerando il vicinato di un punto che genera un singolo punto in output. Dopo la convoluzione viene applicata la funzione di attivazione, solitamnte nelle CNN viene utilizzata la ReLU (rampa). La fase di pooling è una fase di downsampling, viene fatto applicando dei filtri al tensore in output, il risultato finale è un assegnazione di tag ai vari oggetti presenti nel dato.

Un'altra architettura è quello basato sulle **Recurrent NN** applicate sui tipi di dato con un concetto di tempo, queste architettura hanno un concetto di memoria, infatti all'n-esima elaborazione prendono in considerazione l'elaborazione n-1.

Support Vector Machines

10.9 Model evuation

Nella letteratura esitono delle diverse tecniche per decidere quali dati faranno parte della fase di training di un per la creazione di un modello e quali per la validazione del suddetto. Supponiamo di avera un db con 100 oggetti, si vuole utilizzare una parte di questi oggetti per il training associandogli le caratteristiche dell'input e dell'output, si utilizzano gli oggetti rimanenti per validare il modelo creato, verificando il corretto assegnamento delle etichette.

La prima operazione da effettuare è quella di misurare la distribuzione dei tag nei diversi oggetti, calcolare i costi di missclassification e misurare la cardinalità del training set.

Una tecnica di partizione dei dati è l'**hold out**, dove si definiscono delle percentuali fisse tra training e test, generalmente l'80% dei dati vengono usati per il training ed il 20% per la validazione. Queste percentuali vanno mantenute per tutte le classi presenti, serve dunque un sampling di tipo stratificato attraverso tutte le classi, questo vuol dire che la percentuale con cui appaiono le diverse classi deve essere mantenuta nei partizionamenti. Questa tecnica è utilizzata con db di grosse dimensioni.

La tecnica della **cross validation** presenta nativamente la ripetizione della fase di training. I dati vengono divisi in k **fold**, su $k-1$ fold viene fatto il training mentre sul rimanente viene fatta la validation, questo viene ripetuto per ogni fold. Quando un db è piccolo (e.g. nei casi medici), si prendono il numero di fold uguali al numero di entry nel database, quindi prendendo un singolo oggetto viene fatto il training su tutti gli altri, questa tecnica viene detta **leave-one-out**.

Per poter stimare correttamente un modello deve essere validato in base ai parametri di input (analisi delle sensitività), si devono selezionare i parametri di input del modello e fare delle validazioni, il motivo per il quale il dataset iniziale viene diviso in tre parti:

- parte di training 60%;
- parte di validation (fatto sulla sensitività dei parametri) 20%;
- parte di test 20%;

L'hold-out viene utilizzato per dividere il db in training-validation e test, mentre il cross validation divide il db in training e validation.

Esistono delle metriche per poter stimare in modo oggettivo le predizioni fatte dal modello, per far questo viene utilizzata la **matrice di confusione**. La metrica dell'accuratezza non è sempre affidabile, consideriamo un esempio:

classe	record
0	9900
1	100

assegnando un classificatore di default che assegna sempre la classe 0 la sua accuratezza sarà del 99%, mentre la classe 1 non sarà mai predetta, questo è un esempio di caso in cui le classi sono sbilanciate, il nostro interesse in questi casi è predire con precisione la classe minoritaria, come nella vita reale solo l'1% della popolazione contrae una malattia, ma quell'1% deve essere predetto con la maggiore precisione. Generalmente si danno delle metriche di valore per gli attributi di interesse, queste metriche calcolate per una classe di interesse vengono dette, **recall** (numero di oggetti correttamente assegnati alla classe / numero di oggetti che appartengono alla classe) e **precision** (numero di oggetti correttamente assegnati alla classe / numero di oggetti assegnati alla classe). Queste metriche vanno massimizzate attraverso l'**F-measure** $= \frac{2rp}{r+p}$.

L'ultima tecnica di classificazione dei modelli è la **curva ROC**, ...

11 Clustering fundamentals

... da rivedere

Bisecting K-means

12 Trigger

Consideriamo un esempio di tema di esame.

a:

```
1 misure: sum(incasso), sum(#consulenze)
2 tabelle: INCASSO, TEMPO, SERVIZIO, SEDE-CONSULENTI
3 gb: semestre, tipologia-servizio
4 selezione: regione = 'Lombardia'
```

b:

```
1 misure: sum(incasso), sum(#consulenza)
2 tabella: INCASSO, SEDE-CONSULENTI, SERVIZIO, TEMPO, AZIENDA
3 gb: regione, servizio, anno
4 selezione: Nazionalita = 'Italia' or
5           Nazionalita = 'Germania'
```

c:

```
1 misure: sum(incasso), sum(#consulenze)
2 tabelle: INCASSO, SEDE-CONSULENTI, SERVIZIO, TEMPO
3 gb: tipologia-servizio, regione, semestre
4 selezione: anno >= 2017 and anno <= 2019
```

Date le query precedenti si crei una vista materializzata:

```
1 -- Query blocco A
2 select servizio,
3         tipologia-servizio,
4         semestre,
5         anno,
6         regione,
7         nazionalita,
8         sum(incasso),
9         sum(#consulenze)
10 from incasso i, tempo t, azienda a, servizio s, sede-consulenze sc
11 where condizioni di join
12 group by servizio,
13         tipologia-servizio,
14         semestre,
15         anno,
16         regione,
17         nazionalita
```

L'identificare minimale sarà: (servizio, semestre, regione, nazionalita)

Punto 2: ...

Punto 3:

```
1 insert into viewIncassi(servizio,
2                          tipologia-servizio,
3                          nazionalista,
4                          semestre,
5                          anno,
```

```
6     regione ,
7     incassotot ,
8     numconsulenzetot)
9 (blocco A)
```

Punto 4:

```
1 create trigger refreshViewIncassi
2 after insert on incasso
3 for each row
4 declare
5     varServizio varchar(20);
6     varTipologiaServizio varchar(20);
7     varSemestre varchar(20);
8     varAnno varchar(20);
9     varNazionalita varchar(20);
10    varRegione varchar(20);
11    n int;
12 begin
13 -- leggere le tabelle dimensionale x recuperare
14 -- i valori dell'identificatore della vista materializzata:
15 -- servizio, nazionalita, semestre, regione
16 select servizio, tipologia-servizio into varServizio,
17     varTipologiaServizio
18 from servizio
19 where idServizio = :NEW.idServizio;
20
21 select semestre, anno
22 from tempo
23 where idTempo = :NEW.idTempo;
24
25 select nazionalita into varNazionalita
26 from azienda
27 where idCategoriaAzienda = :NEW.idCategoriaAzienda
28
29 select regione into varRegione
30 from sede-consulenti
31 where idSede = :NEW.idSede;
32
33 -- verifico se esiste una tupla in viewIncassi con
34 -- associati i valori estratti
35
36 select count(*) into n
37 from viewIncassi
38 where nazionalita = varNazionalita and
39     semestre = varServizio and
40     servizio = varServizio and
41     regione = varRegione;
42
43 if (n > 0) then
44     update viewIncassi
45     set incassoTot = incassoTot + :NEW.incasso
```

```
46         numConsulenze += :NEW.#consulenze
47     where servio = varServizio and
48         nazionalita = varNazionalita and
49         semestre = varSemestre and
50         regiono = varRegione;
51 else
52     insert into viewIncassi ( ... , )
53     values (varServizio,
54         varTipoServizio,
55         varNazionalita,
56         varSemestre,
57         varAnno,
58         varRegione,
59         :NEW.incasso,
60         :NEW.#consulenze);
61 end if;
62 end;
```

Punto 5:

```
1 create trigger updateViewIncassi
2 after update of tipologiaServizio on servizio
3 for each row
4 declare
5     typeServizio varchar(20);
6 begin
7
8
9 end;
```

Punto 6:

```
1 create materialized view log on incasso
2 with sequence, row id
3 (...)
4 including new values;
5
6 create materialized view log on servizio
7 with sequence, row id
8 ( ... )
9 including new values;
10
11 create materialized view log on tempo
12 with sequence, row id
13 (. ...)
14 including new values;
```

13 Lab 03

1.

```

1 select mese, tariffa,
2     sum(prezzo),
3     sum(sum(prezzo)) over () as prezzo_comp,
4     sum(sum(prezzo)) over (partition by mese) as prezzo_per_mese,
5     sum(sum(prezzo)) over (partition by tipo_tariffa) as
6     prezzo_per_tariffa
7 from tempo te, fatti f, tariffa t
8 where te.id_tempo = f.id_tempo and
9       f.id_tar = t.id_tar and
10      anno = 2003
11 group by mese, tipo_tariffa;

```

2.

```

1 select mese
2     sum(chiamate) as chiamate_tot,
3     sum(sum(prezzo)) as incasso_tot,
4     rank() over (
5         order by sum(prezzo) desc
6     ) as rank_piu_chiamate
7 from fatti f, tempo t
8 where f.id_tempo = t.id_tempo
9 group by mese;

```

3.

```

1 select mese
2     sum(chiamate) as chiamate_tot,
3     sum(prezzo) as incasso_tot,
4     rank() over (
5         order by sum(chiamate) desc
6     ) as rank_piu_chiamate
7 from fatti f, tempo t
8 where f.id_tempo = t.id_tempo and
9       anno = 2003
10 group by mese;

```

4.

```

1 select tipo_tariffa
2     sum(prezzo)
3 from fatti f, tariffa t
4 where f.id_tar = t.id_tar and
5       mese = 'luglio' and
6       anno = 2003
7 group by tipo_tariffa;

```

5.

```

1 select mese
2     sum(prezzo) as incasso_tot,

```

```
3      sum(sum(prezzo)) over (
4          order by mese
5          rows unbounded preceding
6      ) as incasso_da_inizio_anno
7 from fatti f, tempo t
8 where f.id_tempo = t.id_tempo
9 group by mese;
```

6.

```
1 select tipo_tariffa, mese,
2      sum(sum(prezzo)) over (partition by mese) as incasso_per_mese,
3      sum(sum(prezzo)) over (partition by tipo_tariffa) as
4      incasso_per_tariffa,
5      100 * sum(sum(prezzo)) over (partition by mese) /
6      sum(sum(prezzo)) over (partition by tariffa)
7 from fatti f, tempo t
8 where f.id_tempo = t.id_tempo and
9      anno = 2003
10 group by tipo_tariffa, mese;
```

vista:

```
1 create materialized view view_fatti
2 build immediate
3 as
4 select mese,
5      tipo_tariffa,
6      anno,
7      sum(prezzo) as prezzo,
8      sum(chiamate) as chiamate
9 from fatti f, tempo te, tariffa t
10 where f.id_tempo = te.id_tempo and
11      f.id_tar = t.id_tar and
12      anno = 2003
13 group by mese,
14      tipo_tariffa,
15      anno,
16      prezzo,
17      chiamate;
```


14 Clustering

Il clustering consiste nel raggruppamento di oggetti similari tra di loro, questa tecnica viene detta di **unsupervised learnig** perchè per effettuare questa suddivisione parte solo da informazioni presenti nei dati. Le tecniche del cluster analysis si pone l'obbiettivo di partizionare il db in sottogruppi, dove gli oggetti in un sotto gruppo sono vicini tra di loro mentre sono distanti con gli oggetti degli altri sottogruppi. Col termine clustering si definiscono dei cluster, ovvero degli insiemi di dati, il clustering si suddivide in: partizionale (i dati appartengono ad uno ed us solo gruppo), gerarchico (elementi rappresentati da un albero gerarchico, che identifica n partizionamenti detto dendogramma). Gli algoritmi di clustering possono suddivisi per i gruppi di cluster che si vengono a formare, uno di questi è il partizionamento esclusivo vs non-escusivo un punto può appartenere a più sottogruppi, fuzzy vs non-fuzzy dove un punto è associato ad ogni gruppo ma con un peso per ognuno di essi, parziale quando il partizionamento viene assegnato solo ad sottogruppo esculdendo i dati rumorosi, completa quando ad elemento viene associato un tag. I gruppi identificati possono essere caratterizzati da: cardinalità, densità, forma; esistono degli algoritmi in grado di indentificarà gruppi a densità omogenea ed eterogenea, per ogni caratteristica esiste un tipo di algoritmo. I gruppi che si ottengono possono essere:

- ben separati: distanza massimizzate tra gruppi diversi e minimizzate tra elementi di un gruppo;
- center-based: gruppi rappresentati da un punto medio, **centroide** media dei punti in cluster, **medoide** media dei punti più rappresentativi del cluster;
- cluster continui;
- density-based: i cluster hanno dansità uguale;
- cluster concettuali;

14.1 K-means

Ogni cluster è associato con un centroide, l'algoritmo prende in input un parametro K che corrisponde al numero di cluster. L'algoritmo è formato da:

```

1 seleziona K centroidi casuali
2 do
3   si formano K cluster a partire dai centroidi assegnando i valori
   pi\'u vicini
4   si ricalcolano i centroidi a partire dai cluster
5 while (i centroidi non cambiano)
```

Questo algoritmo converge abbastanza velocemente infatti ha $O(\text{num-punti} * K * \text{iterazioni} * \text{num-attributi})$, per questo motivo l'algoritmo viene fatto eseguire molteplici

volte per eliminare il problema dell'assegnazione casuale dei centroidi che potrebbe portare anche a cluster vuoti. Per valutare quanto è buono un partizionamento si usa l'SSE, Sum of Squared Error.

Definition 14.1 – SSE

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

Questa metrica è molto buona per calcolare dove si deve eseguire l'algoritmo su stessi K (o anche comparare le prestazioni di diversi algoritmi), se però si decide di aumentare K allora l'SSE è sempre confrontabile ma i valori saranno più piccoli al crescere di K, perché la base dati rimane la stessa, dunque con più partizionamento i dati sono molto più coesi tra di loro.

Oltre ad utilizzare l'approccio delle **run multiple** si possono utilizzare come centroidi utilizzando punti del database stesso usando altre tecniche. Si possono anche utilizzare centroidi maggiori di K e poi al termine delle run scegliere tra questi i centroidi iniziali, si può anche utilizzare del postprocessing o un bisect K-means.

Esiste il problema dei cluster vuoti quando si fanno delle iterazioni, per evitare che vengano creati cluster vuoti si prende un punto dal cluster con l'SSE più grande oppure scegliere il punto con l'SSE più grande, riassegnando il centroide col un cluster vuoto al punto estratto si risolve questo problema, se più cluster sono vuoti si ripete fino ad averli tutti con almeno un elemento.

Per ridurre il rumore dei dati eliminando gli outliers si possono applicare operazioni di pre-processing. Una volta finito l'algoritmo si possono applicare operazioni di post-processing come: eliminare cluster piccoli, eliminare cluster con SSE molto grande, oppure fare il merge di cluster vicini tra di loro, per le operazioni di post-processing non esistono delle linee guida, ma solitamente dipende dal contesto.

14.2 Bisecting K-means

Il bisecting

...

15 Introduzione ai DBMS

Il DBMS permette di avere una gestione concorrente delle informazioni, oltre ad avere delle routine che operano sui dati, la sua struttura è molto complessa ed è strutturata in moduli, l'entry point è una query.

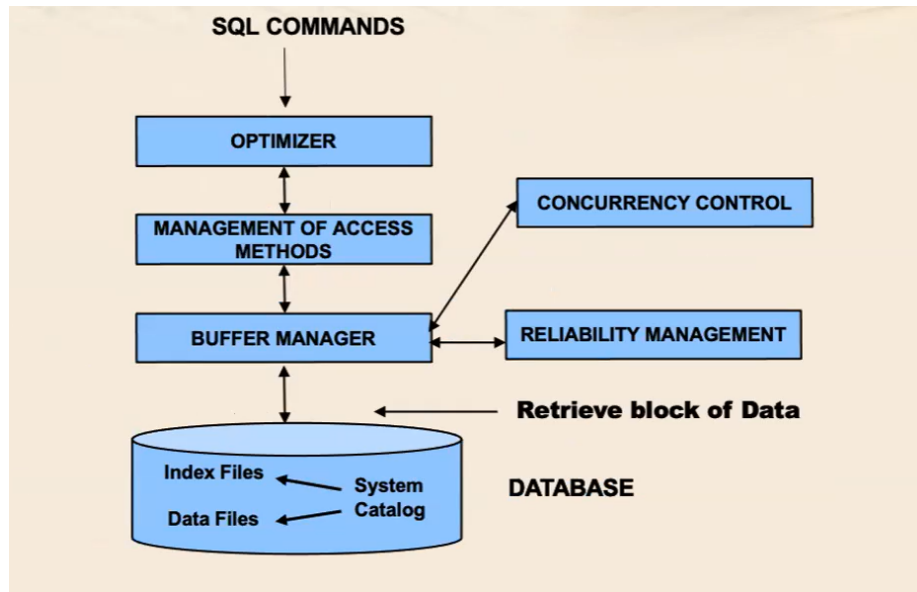


Figure 5: Architettura Dbms

Il DBMS è composto da:

- ottimizzatore: prende delle decisioni in base alla distribuzione dei dati, oltre a fare il parsing della stringa;
- access method manager: viene eseguito un metodo specifico per leggere la memoria;
- buffer manager: buffer nella memoria principale che ottimizza le operazioni di I/O;
- concurrency control: gestisce l'accesso concorrente ai dati;
- reliability manager: garantisce la correttezza del contenuto dei dati del db, utilizzando dei file di log in cui vengono scritte le operazioni effettuate;

Il concetto fondamentale nel contesto dei db sono le **transazioni**: ovvero di una serie di operazioni che rappresentano una singola unità di lavoro. Una transazione termina con COMMIT o ROLLBACK. Le transazioni sono caratterizzate da ACID:

- **atomic**: tutte le operazioni devono andare a buon fine o nessuna, le operazioni che controllano lo stato del sistema sono **UNDO** o **REDO**;
- **consistency**: non vanno violati i vincoli di consistenza del DBMS, vincoli di chiavi, chiavi esterne, ...;
- **isolation**: le transazioni operano in modo indipendente tra di loro, i dati intermedi non sono visibili all'esterno;
- **durability**: i dati di una transazione non possono essere perse, questa proprietà è garantita dal reliability manager;

15.1 Buffer Manager

Il buffer manager ha accesso ad una parte della memoria principale a cui gli applicativi hanno accesso. Il buffer è organizzato in **parole** di dati, che è l'unità minima trasferibile dalla memoria secondaria, per essere efficace nelle operazioni di IO le pagine devono essere presenti nel buffer per velocizzare le operazioni, per questo motivo si utilizza il principio di località. Ogni pagina deve avere:

- l'ID del file;
- l'ID del blocco;

Per ogni pagina esistono anche due stati:

- **count**: numero di transazioni che stanno utilizzando la pagina;
- **dirty bit**: settato quando la pagina è stata modificata e non è stata ancora caricata nella memoria secondaria;

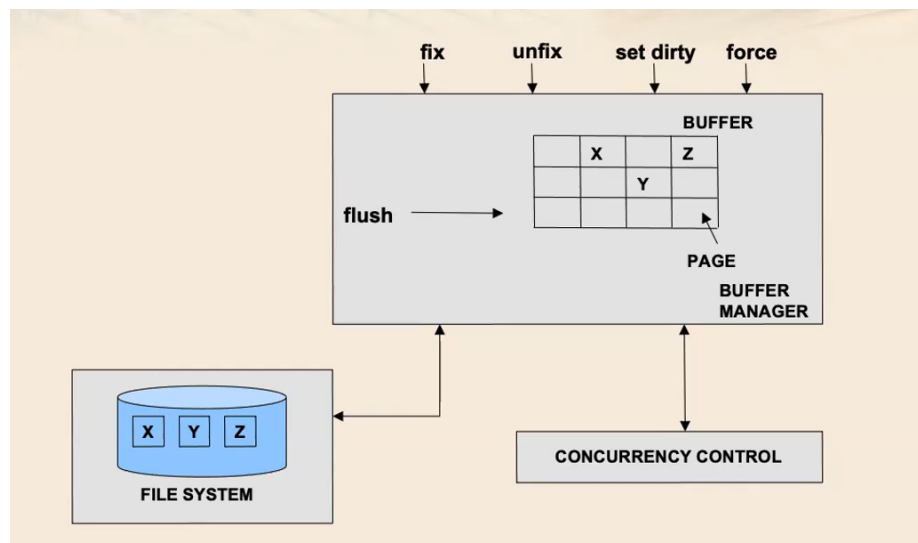


Figure 6: Buffer Manager

Il buffer manager ha a disposizione una parte della memoria principale in cui salva delle pagine di dati del DB quando viene fatta query. Esistono delle primitive per poter accedere a questa memoria:

- **fix**: richiesta dalla transazione, se la pagina è disponibile allora viene passata, altrimenti si fanno delle operazioni di IO per recuperare la pagina, se il buffer è pieno un'altra pagina deve essere rimpiazzata, solitamente sarà una pagina con il **count**=0, questa pagina viene detta **vittima** e verrà rimpiazzata con

qualla richiesta dalla transazione, se durante la transazione una pagina ha il **dirty biy** = 1 allora deve essere fatta un'operazione di IO per sincronizzare la pagina. Prima che la lettura inizi il count viene incrementato di 1;

- **unfix**: viene richiesta dalla transazione quando la pagina non deve essere utilizzata;
- **set dirty**: setta il valore dirty ad una pagina quando viene fatta un'operazione di scrittura;
- **force**: richiede un trasferimento sincrono della pagina al disco, quindi si scrive la pagina, non è detto che la pagina venga rimossa;
- **flush**: richiede un trasferimento asincrono della pagina al disco, questo traferimento;
- **steal**: quando si utilizza steal il buffer manager può selezionare una pagina locked con un count=0, detta pagina **vittima**, e la rimuove dal buffer, se non-steal è settato non si può selezionare una pagina su cui sta avvenendo una transazione;
- **force**: tutta la pagine di una transazione sono scritte in modo sincrono sul disco;
- **no-force**: utilizza flush per scrivere le pagine in modo asincrono sul disco;

Inoltre il buffer manager mette a disposizione:

- sequential read;
- write and sequential write;
- directory;

15.2 Accesso alla Memoria

L'access methods manager riceve dall'ottimizzatore il piano di esecuzione e decide quale metodo di accesso utilizzare per leggere o scirvere i dati, vengono poi selezionati i blocchi dei file da prendere dalla memoria principale, questi dati sono richiesti al buffer manager.

Le strutture fisiche che si possono utilizzare sono dipendenti dal tipo di operazione che si effettuano su di loro (select, update, delete, ...), queste strutture sono dette di tipo **accessorio**, in particolare gli **indici**, che sono sturtture definite all'interno del db per velocizzare le operazioni di IO. I dati fisici posso essere salvati in strutture sequenziali o strutture di heap, mentre gli indici utilizzano strutture ad albero, bitmap, unclustered hash.

Nelle strutture sequenziali le tuple sono storicizzate nella pagina seguendo l'ordine di inserimento, il vantaggio è che lo spazio occupato è massimo all'interno della pagina e che l'inserimento è molto veloce, quando si fanno delle operazioni di delete si vengono a creare dei buchi, mentre con l'update si rischia di scrivere più dati di quelli che avevo prima.

Gli **heap file** utilizzano delle strutture secondarie per avvalersi di ricerca veloce, tutti i tipi di operazione sono identiche a quelle delle strutture sequenziali.

Le **ordered sequential structures** sono delle strutture sequenziali ordinate in base ad una chiave di ordinamento. I dati ordinati servono per velocizzare le operazioni per quelle interrogazioni che usano la stessa chiave di ordinamento. Come contro le operazioni di delete e inserimento sono molto onerose. Viene lasciato dello spazio in più per eventuali operazioni di update, oppure viene implementato un file di **overflow**, dove vengono salvati la parte di dati che dopo un'update o insert non entra nella pagina.

Nelle **strutture ad albero** nei nodi si trovano solo le chiavi per effettuare la ricerca, negli indici unclustered le foglie si trovano parte dei valori della tuple e il puntatore alla tupla originale, mentre in una struttura clustered le foglie corrispondono ai puntatori della tupla, inoltre da una foglia si può passare all'altra essendo tutte ordinate in modo sequenziale (come nei **B+-Tree**).

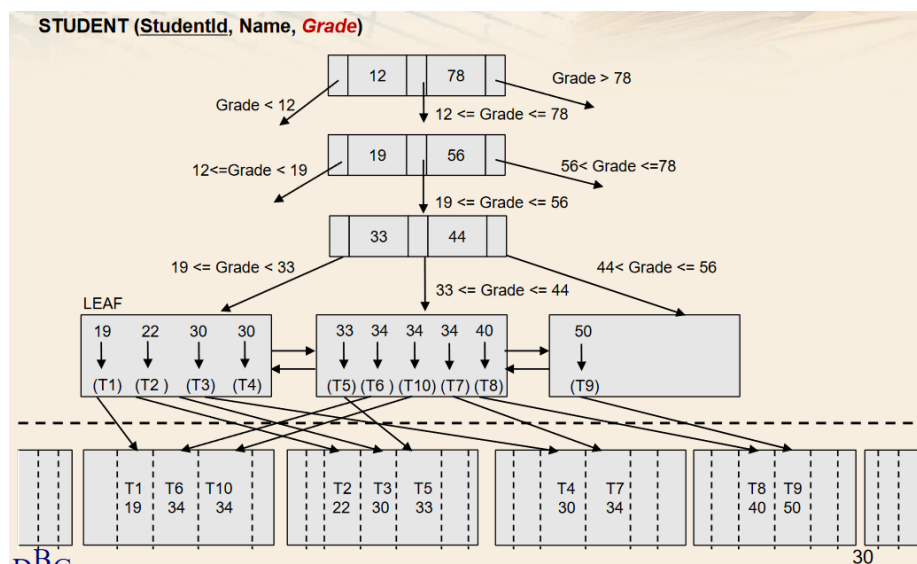


Figure 7: Unclustered B+-Tree Index

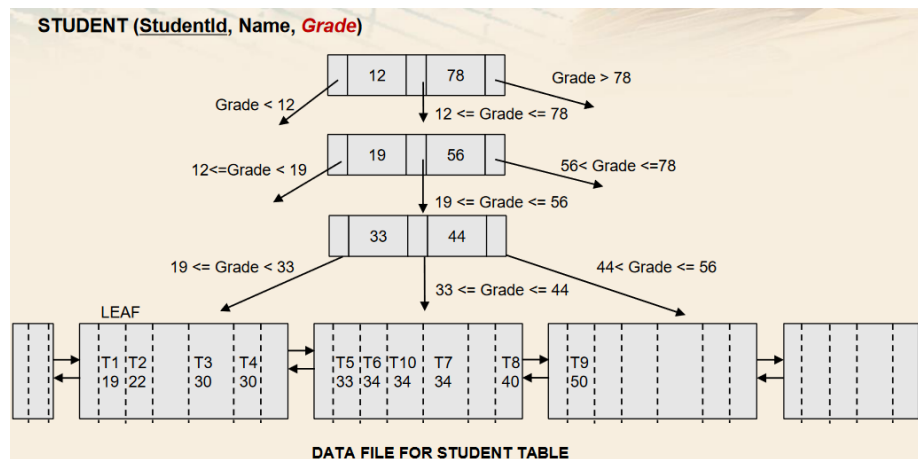


Figure 8: Clustered B+-Tree Index

Su una tabella può essere creato un unico indice di tipo clustered, mentre possono essere create più strutture unclustered.

Un'altra struttura utilizzabile è la **struttura di hash**, utilizza una funzione di hash per assegnare ad ogni tupla secondo una chiave, un blocco tra quelli predefiniti della struttura, anche le strutture di hash possono essere clustered o unclustered.

I **bitmap index** sono delle strutture unclustered, che si portano bene a rappresentare gli attributi categorici.

15.3 Progettazione Fisica

Per fare una buona progettazione fisica sono:

- il progetto logico;
- il carico di lavoro in termine di query e frequenza di esecuzione;
- engine del database;

Nella progettazione fisica dato il carico delle query vengono definiti uno o più indici e le strutture dati.

Euristiche:

- mai indicizzare una tabella piccola;
- mai indicizzare un attributo con bassa cardinalità;
- analizzare i predicati nelle clausole **where**;
- quando si creano degli indici composti si deve considerare il costo del mantenimento;

- per migliorare le operazioni di **join** si usano il **nested loop** (quando si joina un tabella piccola con una grande), ed il **merge scan** (vengono ordinate le tabelle prima di fare il join);
- per le **group by** si usa un ordinamento una struttura di hash, l'ottimizzatore spesso anticipa le operazioni di **group by** (**group by push down**) diminuendo di molto la cardinalità dei dati da analizzare;

Nelle condizioni di join vengono definite due tabelle:

- **outer**: tabella letta in modo sequenziale;
- **inner**: tabella letta n volte, tante quante le righe della tabella outer;

Le alternative per il join sono:

- hash join;
- nested loop: se la tabella è piccola comunque non si indicizza;

Se si decide di creare un indice composto è meglio che l'indice sia **coprente**, ovvero che leggendo l'indice si riesce a rispondere alla query senza leggere la tabella, altrimenti diventa troppo oneroso mantenerlo, diventa più efficace un indice su un solo attributo.

15.4 Ottimizzatore delle query

L'ottimizzatore delle query garantisce efficienza ed indipendenza dai dati. L'ottimizzatore genera un piano di esecuzione, basandosi su delle statistiche, come sulla distribuzione dei valori e di come quei valori sono distribuiti nei vari blocchi fisici, le soluzioni sono dinamiche, infatti al cambiamento dei dati può cambiare anche il piano di esecuzione.

L'ottimizzatore fa un controllo su:

- errori lessicali: misspelled keywords;
- errori sintattici: errori nella grammatica dell'sql;
- errori semantici: viene utilizzato il data dictionary per controllare se un oggetto esiste;

Effettuati i controlli genera una rappresentazione interna in algebra relazionale.

Il passaggio successivo è l'ottimizzazione algebrica, creano un albero di esecuzione.

Una volta creato l'albero viene fatta un'ottimizzazione basata sui costi di accesso, quindi ad ogni parte dell'albero vengono assegnati dei punti di accesso.

L'ottimizzatore può essere utilizzato in due modalità:

- **compile and go**: una query non viene salvata, ma viene ricompilata ogni volta, utile quando i dati variano abbastanza frequentemente nel tempo;
- **compile and store**: la query viene compilata e salvata per riutilizzi successivi;

15.4.1 Ottimizzazione algebrica

Si basa su regole di ottimizzazione dell'algebra relazionale e sulle statistiche dei dati.

15.4.2 Ottimizzazione basata sui costi

Quando riceve un albero relazionale in base al fatto che

Nel **data profile** sono presenti:

- cardinalità delle tuple;
- numero di byte delle tuple;
- numero di byte degli attributi di una tupla;
- numero di valori distinti;
- valori minimi e massimi di un attributo;