

Quality Threshold Clustering

Balice Davide, del Vescovo Samuele, Lopopolo Antonio

E-mail: {d.balice1, s.delvescovo2, a.lopopolo4}@studenti.uniba.it

1. Introduzione

Il data mining definisce l'insieme di tecniche e metodologie che hanno come scopo l'estrazione (semi) automatica di conoscenza nascosta in voluminose basi di dati al fine di renderla disponibile e direttamente utilizzabile. Una delle possibili aree di applicazione inerente al data mining è la segmentazione (o clustering), che consiste nell'individuazione di gruppi con elementi omogenei all'interno del gruppo e diversi da gruppo a gruppo (es. individuazione di gruppi di consumatori con comportamenti simili). In particolare, il software realizzato si avvale dell'algoritmo del Quality Threshold (QT), più recente rispetto al K-means. Rispetto a quest'ultimo, l'algoritmo del QT è computazionalmente più pesante, non necessita di un input del numero di cluster e i risultati ottenuti rimangono coerenti al dataset fornito in input.

2. Quality Threshold Clustering

L'algoritmo viene suddiviso nelle seguenti operazioni:

1. Viene definito un raggio massimo per la definizione dei cluster (dato in input dall'utente)
2. Viene costruito un cluster candidato per ogni transazione nel dataset, raggruppando le transazioni all'interno del raggio
3. Viene scelto il cluster candidato più popoloso da aggiungere al clusterSet e conseguentemente vengono eliminate le transazioni in esso incluse
4. Torna al passo 2 fino a quando ci sono ancora transazioni nel dataset

Il caso di studio è strutturato come un'applicazione client/server scritta in java.

Il client non ha accesso al tipo di dati utilizzato dal server e non è a conoscenza delle operazioni effettuate durante il clustering.

Il server deve essere eseguito specificando un indirizzo e una porta (preimpostati rispettivamente a 127.0.0.1 e 12346). La porta 12346 è stata modificata per evitare eventuali conflitti con porte più comunemente utilizzate (es. 8080). Sono accettati solo indirizzi e porte valide, questo avviene grazie a un controllo basato su un'espressione regolare per l'IP.

Quando un client si connette, il server può eseguire le sue richieste tramite le seguenti operazioni:

1. Caricamento da database
 - 1.a. Il server carica una tabella da database richiesta dal client, tramite il suo nome
 - 1.b. Il server riceve il raggio che verrà utilizzato nell'operazione di clustering
 - 1.c. Una volta finita l'operazione di clustering, vengono visualizzati i risultati
 - 1.d. Può salvare i risultati ottenuti dall'operazione di clustering su file scelto dall'utente

2. Caricamento da file

2.a Il server carica un file scelto dall'utente salvato in precedenza

2.b Il server mostra i risultati relativi al file scelto

Il client può invece effettuare le seguenti operazioni:

1. Input dati per la connessione

1.a. Input dell'indirizzo

1.b. Input della porta

2. Input dati per il database (nuova operazione di clustering)

2.a. Input del nome della tabella

2.b. Input del raggio

3. Input del nome del file (caricamento di risultati precedenti)

2.1 Estensione

L'estensione riguardante questo caso di studio è un'interfaccia grafica per il client.

Essa è stata realizzata utilizzando le JavaFX, tramite l'utilizzo di file .FXML creati con Scene Builder.

È stato utilizzato come design pattern il Model View Controller (MVC).

Esso prevede la suddivisione fra i componenti software in tre ruoli principali:

- il **model** gestisce i dati necessari per il client e può metterli a disposizione alle altre componenti
- il **view**, nel nostro caso, gestisce il caricamento del file .FXML attraverso l'FXMLLoader e la creazione di un nuovo stage
- il **controller** gestisce gli eventi sollevati dai componenti grafici presenti all'interno delle relative view, interfacciandosi col model quando necessario

Il client ha quindi la possibilità di visualizzare i risultati in maniera tabulare e tramite un pie-chart.

La visualizzazione è facilitata tramite l'uso di colori diversi per ogni cluster, mentre il pie-chart aiuta l'utente a comprendere la distribuzione del dataset, tramite la grandezza dei cluster nel grafico.

All'interno della visualizzazione tabulare, dopo l'esecuzione dell'operazione di clustering, è possibile visualizzare la distanza di ogni transazione dal suo centroide, quest'ultimo invece, avrà distanza da sé stesso pari a 0.

Il server ha subito dei leggeri cambiamenti per adattarsi alla gestione delle richieste tramite interfaccia grafica, con una logica orientata agli eventi prodotti dal client tramite la sua interfaccia.

La rappresentazione dei dati ricevuti dal client avviene tramite l'ausilio di liste, così facendo infatti, il client non ha bisogno di implementare o importare le classi relative alle rappresentazioni dei dati, che vengono invece utilizzate dal server insieme alla logica per il mining.

Per ridurre il numero possibile di errori nell'input da parte dell'utente, si è cercato di limitare i possibili errori disabilitando gli appositi pulsanti e le relative finestre in caso di input sbagliati, incompleti o che non producessero risultati rilevanti. Alcuni esempi includono l'utilizzo di espressioni regolari, come per l'input di indirizzo IP e numero di porta, oppure della impossibilità di effettuare l'operazione di clustering senza aver inserito sia nome della tabella che raggio, e più in generale quando una textfield è vuota. Gli errori comunque derivabili da input teoricamente corretti, vengono gestiti tramite le relative eccezioni.

3. Guida di installazione

Il software è stato scritto in java e ha quindi bisogno della Java Virtual Machine per essere eseguito. Sono presenti degli script in formato .sql che creano un utente chiamato 'MapUser' e un database 'MapDB', al cui interno saranno create delle tabelle preimpostate, che possono essere utilizzate per testare il software.

Sono inoltre presenti degli script in formato .bat che serviranno per eseguire il server e il client.

3.1. Prerequisiti

- Java e tramite esso la Java Virtual Machine (scarica [qui](#))
- MySql per l'esecuzione degli script (scarica [qui](#))
- Il processo (servizio) di MySql in esecuzione sulla porta 3306

3.2. Come eseguire il software

- Eseguire lo script .sql "script.sql" (nella root directory)
- Avviare il server tramite "StartServer.bat" (in /setup/)
- Avviare il client tramite "StartClient.bat" (in /setup/)

4. Guida utente


L'utente, all'avvio del software, troverà la seguente interfaccia:

The screenshot shows the main window of the QT-Miner JavaFX application. The title bar reads "QT-Miner JavaFX". The interface is divided into two main sections: "Load From Database" and "Load From File". The "Load From Database" section has a "Table Name" field with the value "playtennis" and a "Radius" field with the value "2.0". The "Load From File" section has a "File Name" field with the value "filename". At the bottom, there are "Load" and "Reset" buttons. A "Connection Status" indicator at the bottom right shows a red dot. A gear icon in the top right corner indicates settings. Red numbers in parentheses are used as annotations: (1) points to the gear icon, (2) points to the "Connection Status" indicator, (3) points to the "Load From Database" section, (4) points to the "Table Name" field, (5) points to the "File Name" field, and (6) points to the "Load" button.

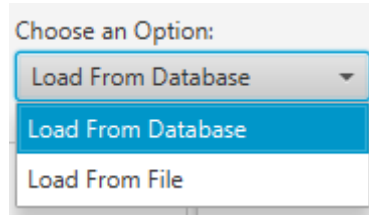
1. L'utente può accedere alle opzioni di connessione, con le quali potrà immettere l'indirizzo IP e la porta da lui scelta.

The image shows two side-by-side screenshots of the "SETTINGS" window, specifically the "CONNECTION OPTIONS" section. The left screenshot shows the "IP address" field empty and the "Port" field with the value "12346". The right screenshot shows the "IP address" field with the value "127.0.0.1" and the "Port" field with the value "12346". Both screenshots have "Connect" and "Reset" buttons at the bottom.

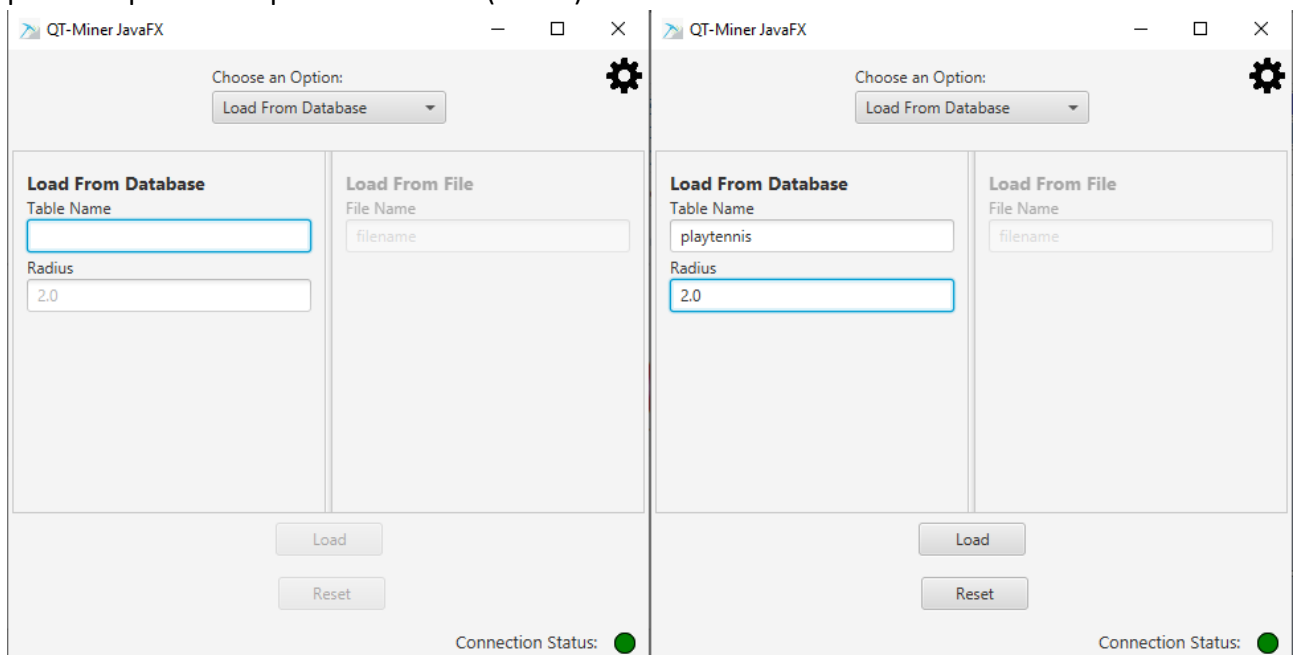
2. Segnala all'utente lo stato della connessione, diventa verde quando ci si connette

Connection Status: 

3. Permette di scegliere se effettuare un'operazione di clustering da database oppure di caricare dei risultati ottenuti precedentemente da file.
Saranno abilitate solo le textfield relative all'input selezionato



4. Permette di scegliere il nome della tabella e il raggio da usare durante l'operazione di clustering. Entrambi i campi sono obbligatori, in quanto con un input incompleto non sarà possibile premere il pulsante "Load" (vedi 6).



Premendo il pulsante “Load”, con dati validi, verrà mostrata la seguente schermata:

DATA VISUALIZATION							
Data Table:							
Cluster ID	outlook	temperature	umidity	wind	play	Distance	
0	sunny	30.3	high	weak	no	0.0	
0	sunny	30.3	high	strong	no	1.0	
0	sunny	13.0	high	weak	no	0.570957095709571	
1	overcast	30.0	high	weak	yes	1.5775577557755776	
1	overcast	0.1	normal	strong	yes	1.4092409240924093	
1	sunny	12.5	normal	strong	yes	2.0	
1	overcast	12.5	high	strong	yes	0.0	
1	rain	12.5	high	strong	no	2.0	
2	rain	13.0	high	weak	yes	1.4290429042904291	
2	rain	0.0	normal	weak	yes	0.0	
2	rain	0.0	normal	strong	no	2.0	

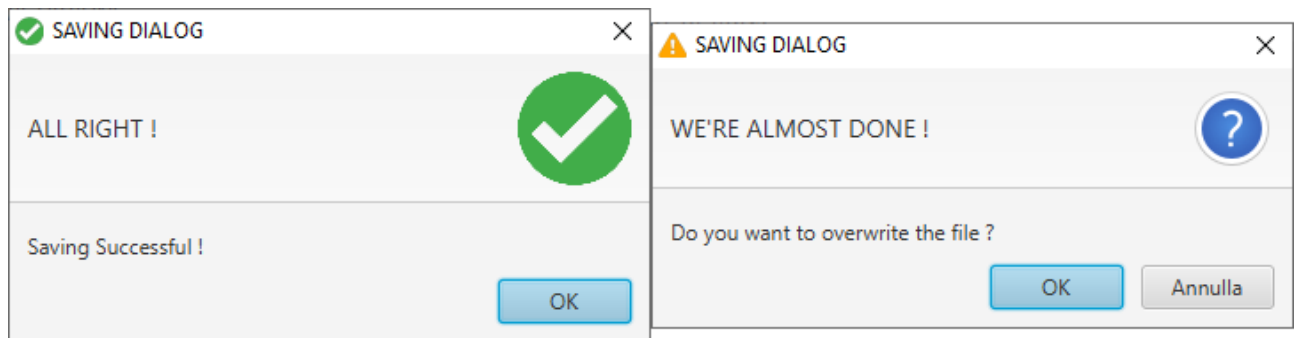
Save

Plot data

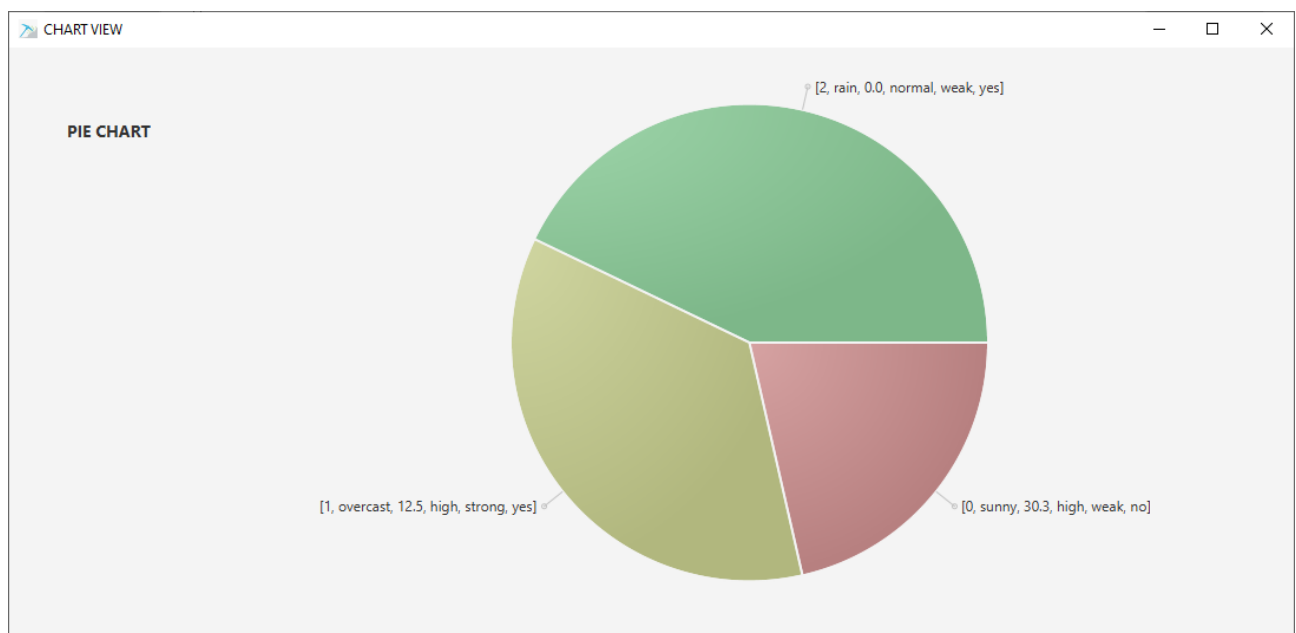
Tramite il pulsante “Save”, ci verrà richiesto di immettere il nome del file sul quale verranno salvati i risultati dell’operazione di clustering

<div>SAVE OPTIONS</div> <div>Choose the Name of the File</div> <div>Name : <input type="text"/></div> <div><div>Save</div><div>Reset</div></div>	<div>SAVE OPTIONS</div> <div>Choose the Name of the File</div> <div>Name : <input type="text" value="risultati_playtennis"/></div> <div><div>Save</div><div>Reset</div></div>
--	---

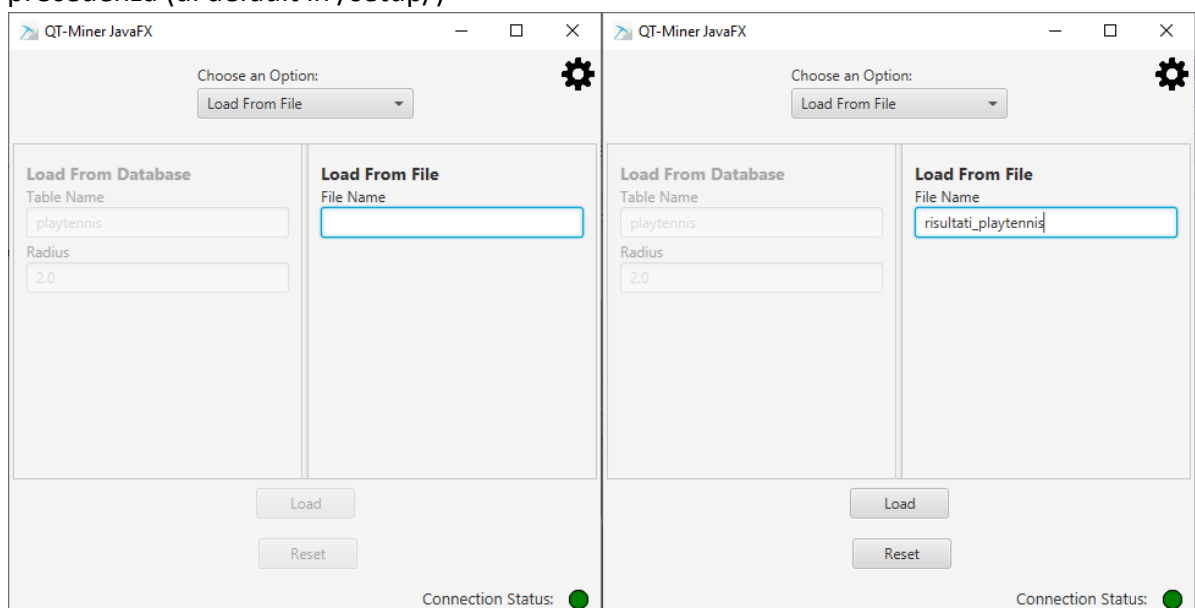
Il file verrà quindi salvato, oppure, in caso di file già presente, verrà chiesto se sovrascriverlo o meno



Premendo il pulsante “Plot data”, verranno visualizzati i risultati tramite pie-chart.



5. Permette di scegliere il nome del file da cui verranno caricati dei risultati salvati in precedenza (di default in /setup/)



Scegliendo un file presente (di default in /setup/), verranno visualizzati salvati in precedenza tramite un'operazione di clustering

[illegible]

6. Il pulsante “Load” serve a caricare la tabella ed il raggio scelto per effettuare un’operazione di clustering, oppure a caricare da file dei risultati salvati in precedenza, in base all’opzione scelta dal menu a tendina (vedi 3). Il pulsante “Reset” serve invece a resettare le textfield nelle quali si è scritto. Non è possibile premere tali pulsanti fino a quando non è stata instaurata una connessione e se le textfield nelle quali si vuole scrivere sono già vuote.

