Documentazione Progetto
Corso di Ingegneria della Conoscenza
ICON1 A.A. 2023-2024

Gruppo di lavoro

- Nicola Guida, 745142, n.guida2@studenti.uniba.it

URL repository associato:
https://github.com/drivecode94/progetto_Icon-23-24.git

Indice

INTRODUZIONE

Il progetto ha come obiettivo primario quello di esaminare l'efficacia dei diversi metodi di apprendimento automatico per analizzare il dataset "FILM".

Attraverso il **Preprocessing** dei dati, ho preparato le informazioni per l'elaborazione e l'analisi successiva. In seguito, ho utilizzato tecniche di **Clustering** per raggruppare i film in base a determinate caratteristiche. Inoltre, ho creato una **Knowledge Base(Base di conoscenza)** per rappresentare le informazioni sul dataset in modo coerente e accessibile. Infine, ho applicato la **Bayesian Network(Rete Bayesiana)** per effettuare previsioni sulle caratteristiche dei film e per comprendere le relazioni causali tra le diverse variabili.

In questo modo sono stato in grado di ottenere una comprensione più approfondita del dataset e di utilizzare l'apprendimento automatico per sviluppare nuove conoscenze e previsioni sul mondo dei film.

Elenco argomenti di interesse

- **PREPROCESSING DEI DATI:** PCA, MinMaxScaler
- **CLUSTERING:** algoritmo K-Means per la creazione di cluster.
- **KNOWLEDGE BASE**: Prolog per il ragionamento su una base di conoscenza partendo dai dati contenuti nel dataset, permettendo di inferire nuove informazioni.
- **BAYESIAN NETWORK:** tecnica Variable Elimination, Massima Verosimiglianza.

Strumenti adottati

Il linguaggio utilizzato per sviluppare il progetto è **Python**, data la grande potenzialità nel manipolare dati con le numerose librerie come **Pandas** per la gestione del dataset, **PySwip** per poter usare i comandi dell'applicativo **SWI-Prolog**, **Sklearn** per la parte relativa al clustering e **Pgmpy** per l'inferenza probabilistica effettuata con la Bayesian Network. **Matplotlib** per visualizzare i dati dei film, **Numpy** per lavorare con array. **Sys** per la gestione degli argomenti da riga di comando.

PREPROCESSING DEI DATI

Il **preprocessing** del dataset dei film è una fase fondamentale nell'elaborazione delle informazioni. Serve per pulire, trasformare e preparare i dati in modo che siano adatti all'analisi e all'utilizzo Successivo, inoltre, mediante il preprocessing vi è la possibilità di tenere bassa la complessità del modello, andando, magari, ad eliminare tutte quelle informazioni che per l'eleborazione potrebbero risultare inutili e quindi vanno ad aumentare i tempi di elaborazione. Il preprocessing è stato effettuato prima di applicare il Clustering, Knowledge Base e la Bayesian Network perché c'era la necessità di avere dati precisi e fornire dati affidabili. Grazie al preprocessing ho rimosso eventuali valori mancanti o duplicati, normalizzato e discretizzato i valori. Inoltre ho cancellato gli spazi tra le stringhe.

Decisioni di Progetto

1. Inizialmente ho dato una **pulizia generale** al dataset film.csv: ho deciso di eliminare le colonne inutili ai nostri scopi finali come 'voto_critica','voto_pubblico','voti_totali','descrizione','note','titolo_italiano' e mi sono accertato che

nel dataset non ci fossero film duplicati o film con

informazioni essenziali mancanti come ad esempio l'anno di uscita o la durata. In seguito, ho fatto in modo di avere nelle colonne "attori", "registi", "paese" solo un valore, questo perché complicava le fasi successive del progetto.

2. Per quanto riguarda il **preprocessing per il cluster**, mi sono limitato a considerare le principali feature per considerare la similarità dei film come genere, anno, durata, paese, humor, ritmo, impegno, tensione, erotismo. Per fare ciò è stato necessario applicare la Principal Component Analysis (**PCA**), per ridurre il numero di features, e uno dei metodi più utilizzati per scalare i valori dei features, ovvero **MinMaxScaler**: trasforma i dati nell'intervallo tra 0 e 1.
3. Dopo aver effettuato il clustering, si va a definire il dataset che verrà utilizzato per **la Knowledge Base**, ottenuto semplicemente aggiungendo all'interno del dataset ottenuto dall'operazione di cleaning, una colonna con il relativo cluster per ciascun film.
4. Per quanto riguarda il **preprocessing per la Bayesian Network**, è stato necessario eliminare le colonne non funzionali all'inferenza probabilistica. La procedura importante è stata quella di discretizzazione di varie colonne, trasformate in valori testuali per rendere più comprensibile e

naturale il linguaggio con cui l'utente andrà ad interagire con la Bayesian Network. I valori sono stati eliminati col l'utilizzo del metodo di inferenza "**Variable Elimination**" il quale richiamato sui valori non ancora discretizzati, va ad indicare gli intervalli per suddividere il dataset in maniera omogenea.

CLUSTERING

La tecnica del clustering è stata scelta per poter andare ad individuare e raggruppare film che presentano similarità tra loro, sfruttando ciò anche per la knowledge base. Questa tecnica, infatti, è atta all'individuazione dei cosiddetti cluster, ovvero esempi più simili a dei centroidi calcolati automaticamente.

Il clustering può essere di due tipi:

· Hard clustering, che prevede un'assegnazione statica di ogni esempio ad una classe di appartenenza

· Soft clustering, che adotta delle distribuzioni di probabilità sulle classi associate ad ogni esempio

Nel nostro caso, è stato adottata la prima tipologia: nello specifico, la tecnica da noi utilizzata è quella del **k-means**. Tale algoritmo va ad inizializzare in maniera casuale un dato numero di centroidi (pari al numero di

cluster che l’utente ha inserito), per poi fare una prima assegnazione degli esempi del dataset. A questo punto, per ciascun centroide verrà calcolata la media dei valori delle features, i quali verranno utilizzati per il successivo ricalcolo delle assegnazioni: il k-means, infatti, ha come condizione per la convergenza la minimizzazione della sommatoria degli errori quadratici.

Decisioni del progetto

Numero di clusters ed iterazioni sono parametri richiesti all’utente, ma, nonostante ciò, il numero migliore di clusters è determinato attraverso il “**metodo del gomito**”, il quale permette di individuare il numero di cluster minimo per poter diminuire significativamente l’errore associato al modello attraverso il plot di un grafico che vada a rappresentare l’errore stesso in relazione ad un crescente numero di clusters.

KNOWLEDGE BASE

Una base di conoscenza (Knowledge Base) in logica di primo ordine è costituita da un insieme di proposizioni, dette assiomi, che sono assunte essere vere senza dimostrazione. La base di conoscenza è utile per rappresentare, all'interno di una macchina, la conoscenza riguardo un particolare mondo.

Decisioni di Progetto

In particolare, la creazione di un kb prevede i seguenti passaggi:

- Decidere il dominio da rappresentare: esso può includere aspetti del mondo reale, un mondo immaginario o un mondo astratto (numeri e insiemi)

- Il progettista deve poi scegliere le proposizioni atomiche per rappresentare il mondo

- In seguito, il progettista deve definire le proposizioni che saranno vere nell'interpretazione intesa (Assiomatizzazione del dominio)

- Infine, si possono porre al sistema delle query, ovvero, determinare se specifiche proposizioni sono conseguenze logiche della KB (proposizione vera in tutti i modelli della KB)

Il sistema, a differenza del progettista, non comprende il significato dei simboli, bensì, è in grado di decidere se una particolare proposizione sia conseguenza logica oppure no, in base agli assiomi presenti nella base di conoscenza. Successivamente il progettista, in base all'interpretazione intesa, comprende se il risultato ottenuto dal sistema è valido oppure no.

Quindi, come ulteriore metodo per la rappresentazione di conoscenza è stato scelto di definire una base di conoscenza in Prolog, ragionando, perciò, in logica di primo ordine. Per poter definire una KB, il dataset è stato prima soggetto ad una fase di preprocessing (descritto nel file cleaning_dataset.py) e successivamente si sono scelte le features da assiomatizzare come fatti nella KB. Inoltre, sono state aggiunte alcune regole che permettono all'utente di sottomettere al sistema query più complesse.

L'utente tramite queste query può sapere se due film sono simili, o se magari lo stesso regista ha lavorato per 2 film diversi, inoltre può sapere se un film appartiene a una particolare categoria ad esempio "horror" oppure no.

BAYESIAN NETWORK

La rete Bayesiana, invece, è stata utilizzata per analizzare le relazioni tra variabili del dataset per capire meglio come queste variabili influiscano sulle performance diun film. In generale mi ha aiutato a scoprire relazioni

nascoste e a prendere decisioni basate su dati quantitativi. La qualità dei risultati dipende dalla qualità e dalla quantità del dataset.

Ho considerato diverse feature quali anno, voto, paese, durata, genere ecc…

Ad esempio, dalla relazione voto e durata si scopre che i film più brevi tendono ad avere voti più alti rispetto ai film più lunghi.

Un altro esempio è la relazione anno e voto, infatti, mediante questa relazione, si scopre che, infatti, i film più recenti tendono ad avere voti più alti rispetto ai film più vecchi.

Decisioni di Progetto

L'apprendimento e la realizzazione della Belief Network sono stati effettuati con lo scopo di evidenziare, attraverso la struttura ottenuta dai dati, il modo in cui le features sono tra loro correlate, e permettere all'utente di eseguire interrogazioni basate sull'inferenza probabilistica.

Ho analizzato il dominio che la rete deve rappresentare:

- ho elencato tutti i fattori del problema per trasformarli in variabili. Ogni variabile identifica un nodo della rete Bayesiana. Si individuano i nodi

rilevanti di un problema e si eliminano quelli non rilevanti.

- Ho individuato le variabili che influenzano altre variabili del problema. I nodi genitori hanno maggiore importanza rispetto ai nodi figli, poiché condizionano gli eventi del dominio.
- Ho individuato tra i nodi genitori quelli che non sono influenzati da altre variabili del problema. Questi nodi sono detti nodi-radice o nodi-causa.
- Ho sviluppato la rete a partire dai nodi- radice del problema seguendo le relazioni padre-figlio, fino agli ultimi nodi del problema (nodi- foglie). A ciascun nodo della rete è associata la probabilità condizionata.

  Ad esempio P(voto|durata) dove voto è il nodo figlio di durata e quindi abbiamo scoperto che i film più brevi tendono ad avere voti più alti rispetto ai film più lunghi.

Ho addestrato la rete utilizzando il metodo fit.Il metodo fit utilizza un oggetto estimator per stimare le probabilità condizionali dei nodi nella rete Bayesiana. Nel caso in questione viene utilizzato “MaximumLikelihoodEstimator” per stimare le

probabilità condizionali utilizzando la massima verosimiglianza.

Ho utilizzato l'algoritmo **Variable Elimination**per effettuare inferenze probabilità delle variabili (target) data l'evidenza di altre variabili nella rete. L'algoritmo rimuove variabili non essenziali e le sostituisce con tabelle di probabilità condizionale; quindi, effettua somme e prodotti per eliminare le variabili una alla volta finche non rimane solo la variabile target.

Infine, l'utente può finalmente sottoporgli delle **query** basate sulle sue preferenze.

Conclusioni del progetto

Lo scopo principale del progetto è stato quello di analizzare la conoscenza sotto varie forme: la definizione di una base di conoscenza (logica di primo ordine), tecniche di apprendimento in forma probabilistica (Belief network) e infine tecniche di apprendimento non supervisionato (clustering).

L’utilizzo di tali tecniche permette di scoprire pattern interessanti riguardo diversi.

Inoltre, la tecnica del clustering (che utilizza l’algoritmo k-means) è stata sfruttata dalla base di conoscenza per poter ottenere informazioni su film simili.

VALUTAZIONE

Per quanto riguarda la valutazione si scopre che in base alla relazione voto e durata, i film più brevi tendono ad avere voti più alti rispetto ai film più lunghi.

Un altro esempio è la relazione tra anno e voto, infatti, mediante questa relazione, si scopre che i film più recenti tendono ad avere voti più alti rispetto ai film più vecchi, magari dovuti al fatto che si è più vicini alla concezione del mondo reale non prendendo il considerazione il bias, dove per bias s’intende una distorsione che viene utilizzata per generare delle opinioni su qualcosa sulla quale non si è avuta un’esperienza passata personale.