Università degli studi di Firenze

Dipartimento di Ingegneria
Corso di Laurea in Ingegneria Informatica

Elaborato Intelligenza Artificiale
Implementazione di una strategia di apprendimento e pruning di alberi di decisione

Leonardo Gori

February 6, 2021
Anno Accademico 2019/2020

1 Introduzione

L'elaborato consiste nella realizzazione di un applicativo per l'apprendimento di un albero di decisione e il confronto del'accuratezza prima e dopo l'esecuzione di una strategia di pruning.

1.1 Alberi di decisione

Un albero di decisione è un grafo non diretto, connesso e aciclico i quali nodi sono collegati da un numero di archi finito. Ogni albero di decisione è costituito da un nodo di partenza (detto 'radice') dal quale si diramano un numero finito di sottoalberi, fino a concludersi con nodi senza successori (detti 'foglie'). Ogni nodo interno compresa la radice racchiude un test su un parametro i quali risultati lo collegano ai successivi. Le foglie racchiudono il valore della funzione racchiusa nell'albero, ovvero una decisione oppure, nel caso di questa relazione, una classificazione. Il fine di questa struttura è quella di definire una serie di test per il raggiungimento di una decisione, con i vantaggi relativi all'utilizzo di una struttura a albero.

1.2 Apprendimento di alberi di decisione

Per apprendimento di alberi di decisione si intende la generazione di un albero da parte di un algoritmo, tramite l'utilizzo di un insieme di 'esempi' (denominato training set), dai quali sia possibile dedurre un modello di decisione per poter predire la classificazione di un altro insieme di esempi distinto dal primo ma della stessa categoria (denominato test set). In questo elaborato per l'apprendimento di alberi di decisione è stato implementato l'algoritmo denominato Decision-Tree-Learning (descritto in R&N 2009, §18.3).

1.3 Potatura di alberi di decisione

Poichè la generazione di un albero di decisione si basa sull'utilizzo di un set di esempi non prestabilito, l'accuratezza delle relative predizioni potrebbe non essere quella ottimale. Ciò è dovuto al verificarsi di una condizione denominata **Overfitting**, causata da un'errata generalizzazione dovuta a uno sbilanciamento tra il numero di attributi e il numero di esempi utilizzati per la fase di training. Il risultato di questo fenomeno si traduce in un'accuratezza delle predizioni non ottimale, e quindi in una generalizzazione non corretta. Per risolvere questo problema è necessaria una modifica dell'albero di decisione che si rispecchia nella fase di pruning. Il concetto base dell'algoritmo di potatura è quello di rimuovere gli elementi della struttura che non ne migliorano l'accuratezza delle predizioni, e si divide in più tipologie. IN questo elaborato è stata implementato una strategia di pruning sulle regole corrispondenti all'albero basata sull'errore sul validation set denominata Rule post-pruning (descritta in Mitchell 1997, §3.7.1.2).

2 Informazioni sull'elaborato

L'elaborato allegato insieme alla relazione utilizza il dataset Nursery reperito online, il quale contiene esempi basati su parametri e classificazioni per l'accettazione di richieste di adesione a un asilo sloveno. Il dataset è costituito da tutte le possibili combinazioni dei parametri che motivano la classificazione (chiamati attributi) e le corrette decisioni. Gli attributi tenuti in considerazione per la classificazione sono:

- Parents: valuta in termini qualitativi l'occupazione dei genitori che presentano la richiesta. La qualità della situazione lavorativa familiare viene riassunta in 3 valori: usual, pretentious, great_pret.
- Has_Nurs: valuta in termini qualitativi l'asilo alternativo in cui la famiglia potrebbe permettersi di iscriversi se la richiesta fosse rifiutata. La qualità dell'asilo è riassunta in 4 valori: proper, less_proper, improper, critical, very_crit.
- Form: descrive la struttura familiare in termini di completezza e di discendenza biologica. La descrizione viene riassunta in 4 valori: complete, completed, incomplete, foster
- Children: valuta in termini quantitativi il numero di figli che compongono il numero familiare. Le quantità considerate rilevanti vengono riassunte in 4 valori: 1, 2, 3, more
- Housing: valuta in termini qualitativi le condizioni strutturali dell'alloggio del nucleo familiare. La qualità strutturale viene riassunta in 3 valori: convenient, less_conv, critical
- Finance: valuta in termini qualitativi la situazione finanziaria della famiglia, essa viene riassunta in 2 valori: convenient, inconv
- Social: descrive in termini qualitativi l'immagine sociale della famiglia, la quale qualità viene riassunta in 3 valori: non-prob, slightly_prob, problematic
- Health: valuta in termini qualitativi le condizioni sanitarie della famiglia, esso viene riassunto in 3 valori: recommended, priority, not_recom

La decisione finale per una generica richiesta di adesione è rappresentata da una classificazione di 5 valori: not_recom, recommend, very_recom, priority, spec_prior. Nel corso dei test il dataset Nursery viene suddiviso in modo casuale in tre sottoinsiemi disgiunti:

- Un insieme denominato **Training Set** utilizzato per la fase di apprendimento di un albero di decisione contenente il 60% del dataset originale

- Un insieme denominato **Validation Set** utilizzato per la fase di potatura di un albero di decisione contenente il 20% del dataset originale
- Un insieme denominato **Test Set** utilizzato per calcolare l'accuratezza di un albero di decisione prima e dopo la fase di potatura, contenente il rimanente 20% del dataset originale

3 Test

I test effettuati per confrontare l'accuratezza prima e dopo il pruning sono eseguiti su 10 alberi generati sulla base di training set diversi estratti in modo casuale dal dataset. Per ogni albero generato in fase di apprendimento viene poi eseguito l'algoritmo di potatura. Di seguito vengono mostrati un esempio di albero di decisione e i risultati dell'accuratezza del set di regole prima e dopo la potatura.

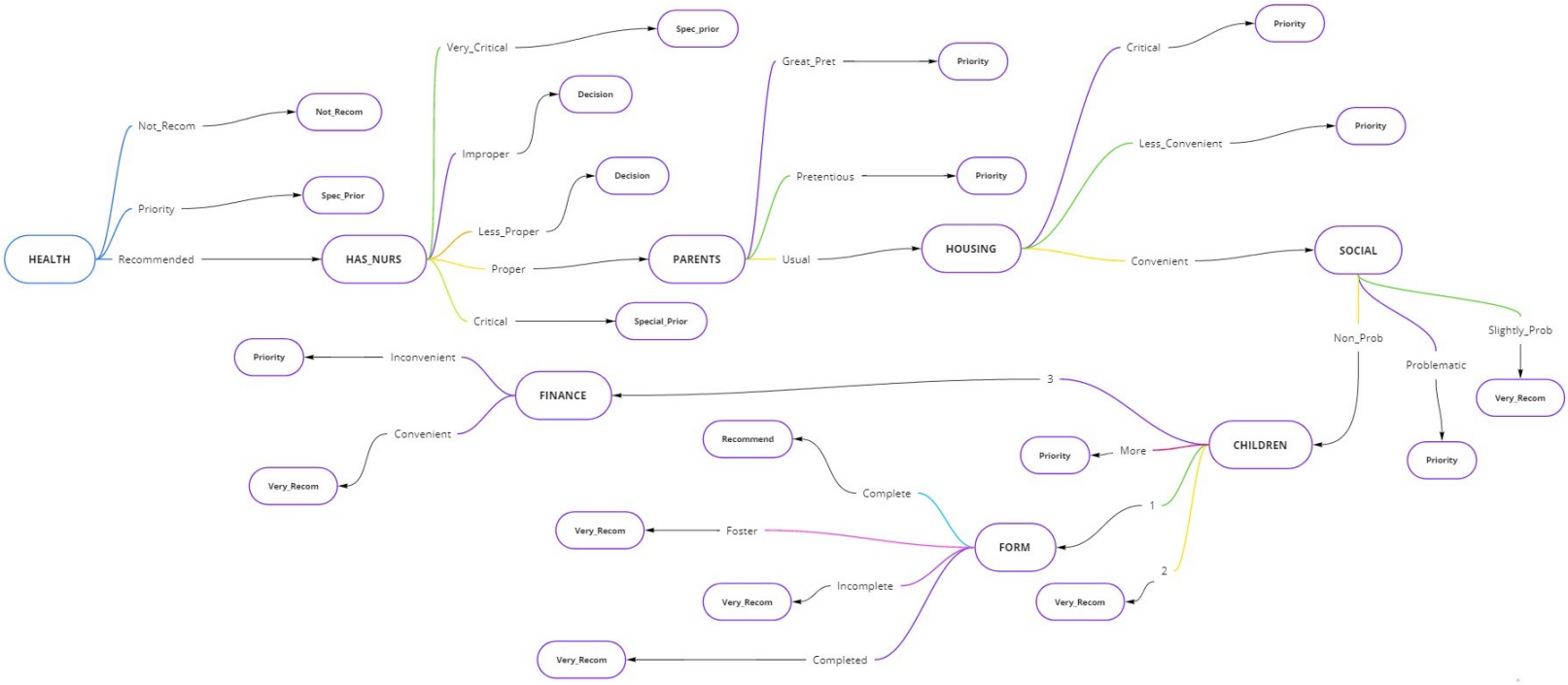

Figure 1: Esempio grafico dell'albero di decisione generato dall'algoritmo di apprendimento, prima dell'esecuzione della strategia di pruning

Test n.	Accuratezza pre pruning	Dimensione Rule-Set pre pruning	Accuratezza post pruning	Dimensione Rule-Set post pruning
1	75.617 %	20	75.694 %	20
2	75.424 %	20	75.540 %	19
3	76.157 %	20	76.157 %	19
4	75.308 %	20	75.385 %	19
5	74.922 %	20	74.961 %	20
6	75.887 %	20	75.925 %	20
7	75.231 %	20	75.347 %	19
8	75.771 %	20	75.810 %	20
9	74.614 %	20	74.652 %	18
10	74.575 %	20	74.421 %	20

Table 1: Confronto di accuratezza sul Test Set prima e dopo l'esecuzione della strategia di pruning e relative dimensioni del Rule Set.

4 Conclusioni

L'algoritmo di potatura Rule-Post Pruning si basa sull'idea di potare i letterali delle regole fino a quando la loro ***qualità*** non migliora, utilizzando un approccio greedy (potando i letterali che restituiscono un maggiore guadagno di qualità). Ci sono molti metodi per valutare la qualità di una regola; in questa relazione è stata utilizzato il concetto di **accuratezza**: la frazione del numero di esempi predetti correttamente nell'insieme di esempi che soddisfano le precondizioni della regola all'interno del Validation Set. Dopo l'applicazione della strategia di pruning il Rule Set (non più mutualmente esclusivo) è stato ordinato in modo decrescente del valore di accuratezza delle regole. Come si vede dalla tabella dopo l'applicazione della strategia di pruning l'accuratezza del set di regole aumenta nella quasi totalità dei casi, con l'ulteriore guadagno di un insieme di regole generalmente ridotto e composto da un minor numero di letterali.