Guadagno Informativo pesato: Documento dei requisiti

Descrizione del problema

I normali algoritmi per la creazione di alberi decisionali basati sul guadagno informativo (di cui c4.5 è forse l'esponente più illustre) nel momento in cui devono decidere in base a quale attributo dividere gli elementi presenti in un nodo procedono calcolando il guadagno informativo (dato, si ricorda, dalla differenza tra l'entropia del nodo prima della divisione e dalla somma delle etropie dei nodi risultanti dalla divisione, pesate in base al numero di oggetti che finirebbero nei vari nodi) e scegliendo di partizionare il nodo in base all'attributo che offre il maggior guadagno informativo. In questo modo utilizzano per primi gli attributi che più sono utili a determinare la classe risultato. Ciò consente di leggere un albero decisionale anche come una serie di caratteristiche che sono correlate alla classe di importanza, in ordine di importanza partendo dalla radice. Quindi si può usare un albero decisionale anche per stabilire un ordine di test da effettuare per determinare se un'entità appartiene o meno ad una classe risultato.

I sistemi attuali, tuttavia, fanno un'implicita assunzione: che tutti i test sui vari attributi abbiano lo stesso costo, cosa non sempre vera in tutti i casi. Un esempio può aiutare a chiarire la questione.

Immaginiamo che sia stata scoperta una nuova malattia, per la quale non esistono procedure di diagnosi valide. I pazienti vengono sottoposti a vari esami, solo alcuni pazienti presentano la malattia, mentre altri sono sani. Abbiamo quindi un database contenente i risultati degli esami svolti che possiamo usare come training set da dare in pasto ad un classificatore, in modo da scoprire quale sequenza di esami effettuare per arrivare ad una diagnosi corretta.

idPaziente	Pressione Arteriosa	Presenza proteina XK59 nel sangue	Gastroscopia	Elettrocardiogramma alterato	Malato?
1	150	Pos	Pos	Neg	SI
2	130	Neg	Neg	Pos	No
3	142	Pos	Neg	Pos	No
4	115	Neg	Neg	Neg	No
5	128	Pos	Pos	Neg	Si
…	…	…	…	…	…

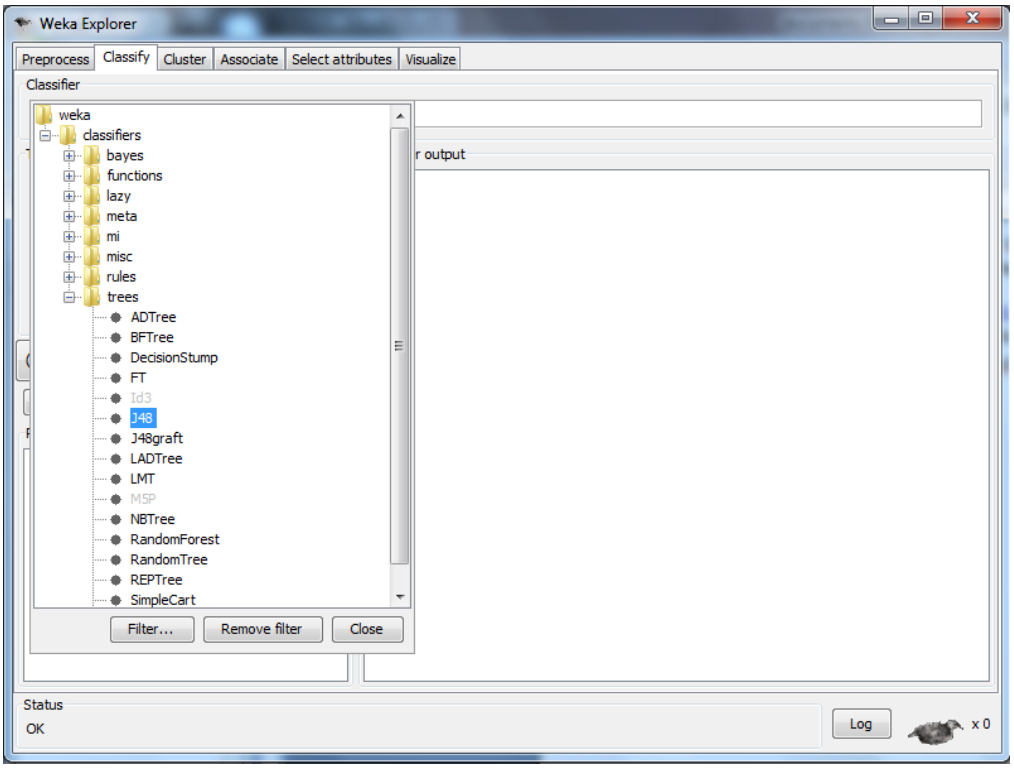
Supponiamo che al primo passo dell'algoritmo i due attributi che offrono il maggiore guadagno informativo siano Gastroscopia e Presenza proteina XK59 nel sangue e che il guadagno informativo di questi due attributi sia quasi identico, ma leggermente superiore per la gastroscopia. Un albero decisionale classico suggerirebbe quindi che il primo passo da fare verso una corretta diagnosi della malattia sia eseguire una gastroscopia a tutta la popolazione, mentre il buon senso impone di preferire di fare come prima cosa le analisi del sangue, molto meno costose ed invasive. La causa di questo errore dell'albero decisionale è legata al fatto che gli algoritmi esistenti non sono in grado di "pesare" i guadagni informativi sulla base di altre indicazioni dovute al costo degli attributi in esame.

Lo scopo del progetto è quello di creare un algoritmo che crei un albero decisionale tenendo conto dei pesi dati dall'utente agli attributi.

Requisiti

Il classificatore dovrà accettare come input una serie di pesi, uno per ogni attributo, il cui valore deve essere compreso tra zero e uno. L'algoritmo deve essere analogo a C4.5, con la differenza che, al momento della scelta dell'attributo in base al quale fare lo split, si dovrà prediligere l'attributo che offre il miglior prodotto tra il peso fornito dall'utente ed il guadagno informativo.

Tale algoritmo deve essere inserito in un'estensione di Weka (http://www.cs.waikato.ac.nz/ml/weka/), un software per il data ming sviluppato dall'università di Waikato, NZ.



Il classificatore deve supportare dati categorici ed è desiderabile sia in grado di trattare anche dati con attributi non completamente specificati .

Una volta prodotto l'output si deve poter visualizzare l'albero prodotto sia in forma testuale sia in forma grafica. Si deve anche poter utilizzare le funzionalità di weka relative all'utilizzo di test set per verificare l'accuratezza del risultato prodotto