Implementa il metodo di classificazione di base a due livelli per i dati a più istanze, senza selezione di attributi.

<https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/kais07.pdf>  
  
Weidmann et al. (2003) hanno proposto Tlc per affrontare i problemi multiistanza generalizzati, che costruisce una metaistanza per ogni borsa e quindi passa la metaistanza e l'etichetta di classe della borsa corrispondente insieme a un classificatore comune.

Tlc utilizza un albero decisionale standard per imporre una struttura allo spazio dell'istanza, che viene addestrato sull'insieme di tutte le istanze contenute in tutte le borse dove le istanze sono etichettate con l'etichetta della classe del loro bagaglio, in modo tale che venga generata una metaistanza per ogni bag.

Tuttavia, vale la pena notare che Tlc genera solo una metaistanza per ogni borsa.

https://link.springer.com/content/pdf/10.1007%2F978-3-540-39857-8\_42.pdf

Nell'apprendimento tradizionale multiistanza (MI), un unico positivo l'istanza in una borsa produce un'etichetta di classe positiva. Quindi, lo studente sa come l'etichetta della classe della borsa dipende dalle etichette delle istanze nella borsa e può utilizzare esplicitamente queste informazioni per risolvere il compito di apprendimento. In questo articolo esaminiamo una visione generalizzata del problema dell'IM in cui questo il semplice presupposto non è più valido. Nella prima fase, questo metodo costruisce una singola istanza da un sacchetto. Questa cosiddetta metaistanza rappresenta le regioni nello spazio delle istanze e ha un attributo per ciascuna di queste regioni. Ogni attributo indica il numero di istanze nel bag che possono essere trovate nella regione corrispondente. Insieme all'etichetta della classe della borsa, la metaistanza può essere trasmessa a uno studente propositivo standard al fine di apprendere l'influenza delle regioni sulla classificazione di una borsa. Dando uno sguardo più da vicino al modo in cui vengono creati i dati MI, ci sono due funzioni che determinano l'etichetta della classe di una borsa. In primo luogo, c'è una funzione che assegna un'istanza in una borsa a un concetto, una funzione di concetto mono-istanza. In secondo luogo, c'è la funzione del concetto MI che calcola un'etichetta di classe dalle istanze in un sacchetto, data la loro appartenenza al concetto dalla prima funzione. Quindi, a due livelli

l'approccio all'apprendimento sembra appropriato. Al primo livello, cerchiamo di apprendere la struttura dello spazio di istanza X, e al secondo livello cerchiamo di scoprire l'interazione che porta all'etichetta di classe di una borsa. Parliamo prima del secondo livello. Nel primo livello della nostra procedura di classificazione, costruiamo una singola istanza da una borsa MI. Gli attributi in questa istanza rappresentano le regioni dello spazio dell'istanza e il valore di un attributo è semplicemente il numero di istanze nella borsa che appartengono alla regione corrispondente. Di conseguenza, utilizziamo un albero decisionale standard per imporre una struttura allo spazio dell'istanza perché è in grado di rilevare questi cambiamenti. L'albero decisionale si basa sull'insieme di tutte le istanze contenute in tutte le borse, etichettate con l'etichetta della classe della borsa

<https://link.springer.com/chapter/10.1007/978-3-319-03680-9_37>

The multi-instance learning method presented in this paper is a simple extension of the Two-Level Classification (TLC) method as presented in [18]. TLC uses a heuristic approach to partition the instance space into regions. Once this has been done, a bag of instances (i.e. an example for learning in a multi-instance dataset) is propositionalised by counting how many instances of the bag fall into each region. These counts are attribute values in the propositionalised problem (i.e. there is one attribute for each region in the partition). Once each bag of data has been propositionalised in this form, and each bag’s classification

La domanda è come definire le regioni nello spazio delle istanze. TLC utilizza un albero decisionale standard a istanza singola per ottenere una partizione. Per conoscere questo albero, tutte le istanze di tutte le borse vengono unite in un unico set di dati, scartando le informazioni sull'appartenenza alla borsa ed etichettate dall'etichetta della classe della borsa. Per assicurarsi che le borse grandi ricevano lo stesso peso delle borse piccole, ogni istanza in questo set di dati è ponderata di 1\|X| \* N\b where X is is the bag the instance comes from, N is the number of instances in the joined data, and b is the number of bags in the original dataset. In this manner, the sum of weights for the instances in the new dataset is N.

<https://www.researchgate.net/figure/A-multi-instance-dataset-with-a-presence-based-MI-concept-Instances-in-a-two-dimensional_fig1_2870939>

Nei problemi di MIL generalizzato, lo studente ha una conoscenza molto minore del modo in cui l'etichetta della classe è determinata dalle istanze in una borsa, rendendo questo tipo di problema più difficile. Introduciamo l'idea della classificazione a due livelli (TLC) per affrontare i problemi di IM generalizzata. Nella prima fase, questo metodo costruisce un istanza singola istanza da una borsa. Questa cosiddetta metaistanza rappresenta le regioni nello spazio delle istanze e ha un attributo per ciascuna di queste regioni. Ogni attributo indica il numero di istanze nella bag che possono essere trovate nella regione corrispondente. Insieme all'etichetta della classe della borsa, la metaistanza può essere trasmessa a uno studente propositivo standard al fine di apprendere l'influenza delle regioni sulla classificazione di una borsa.

Questo documento è strutturato come segue. La sezione 2 offre una panoramica sul problema multiistanza standard e introduce le convenzioni notazionali. Nella sezione 3, diamo le definizioni per le nostre tre generalizzazioni del problema MI. Il metodo di classificazione a due livelli è delineato nella Sezione 4 e gli esperimenti con l'algoritmo su dati artificiali e i problemi di Musk sono descritti nella Sezione 5.

Riassumiamo i nostri risultati nella Sezione 6.

Pertanto, un approccio all'apprendimento a due livelli sembra appropriato. Al primo livello, cerchiamo di apprendere la struttura dello spazio delle istanze X, e al secondo livello cerchiamo di scoprire l'interazione che porta all'etichetta di classe di una borsa.

4.2 Primo livello: strutturare lo spazio delle istanze

Nel primo livello della nostra procedura di classificazione, costruiamo una singola istanza da una borsa MI. Gli attributi in questa istanza rappresentano le regioni dello spazio dell'istanza e il valore di un attributo è semplicemente il numero di istanze nella borsa che appartengono alla regione corrispondente. Un possibile approccio per identificare le regioni sarebbe il raggruppamento. Tuttavia, questo elimina le informazioni fornite dalla borsa etichetta. Se etichettiamo ogni istanza con l'etichetta della sua borsa, le regioni con un'elevata percentuale osservata di istanze positive saranno componenti candidati per i concetti (vedi Figura 1).Di conseguenza, utilizziamo un albero decisionale standard per imporre una struttura allo spazio dell'istanza perché è in grado di rilevare questi cambiamenti. L'albero decisionale si basa sull'insieme di tutte le istanze contenute in tutte le borse, etichettate con l'etichetta della classe della loro borsa. Il peso di ogni istanza in un sacchetto X è impostato su 1 / | X | · N / b, dove N indica la somma di tutte le dimensioni dei bagagli e b il numero di bagagli nel set di dati. Questo attribuisce a borse di dimensioni diverse lo stesso peso e rende il peso totale pari al numero di istanze. Il guadagno di informazioni viene utilizzato come misura di selezione del test.

Un nodo nell'albero non viene ulteriormente suddiviso quando il peso delle sue istanze è inferiore a 2 e non viene utilizzata nessun'altra forma di potatura. Usando questo albero, convertiamo un sacchetto in una singola istanza con un attributo numerico per ogni nodo dell'albero. Ogni attributo conta quante istanze nel bag sono assegnate al nodo corrispondente nell'albero.

J48 https://stackoverflow.com/questions/42116255/how-to-interpret-weka-classification-result-j48