Celiachion: riduzione dimensionale, relative visualizzazioni e classificazione

Salvatore Calderaro Simone Contini Dario Curreri

Abstract

Abbiamo condotto uno studio empirico per determinare le relazioni presenti tra quattro diverse tecniche di visualizzazione (scatterplot 2D per dati bidimensionali, scatterplot 2D per dati tridimensionali, scatterplot 3D interattivo e plot di Draftman) e quattro diverse tecniche per la riduzione della dimensionalità (PCA, kernel PCA, MDS e t-SNE). L'analisi è stata effettuata su un dataset virtuale contenente dati biomedici utilizzati per la diagnosi della celiachia. Dalle rappresentazioni grafiche si evince come PCA e kernel PCA siano più efficienti per la visualizzazione delle due classi. Inoltre si nota che nella maggior parte dei casi gli scatterplot 2D sono già sufficienti per rilevare una buona distinzione tra le classi. Sono stati creati, dunque, degli alberi decisionali visualizzati oltre che con alberi binari, anche attraverso treemap. Infine i risultati ottenuti sono stati confrontati - in termini di accuratezza - con quelli del classificatore addestrato sul dataset non ridotto.

1 Introduzione

Ridurre la dimensionalità dei dati è uno dei task fondamentali per l'analisi dei dati e per le eventuali elaborazioni successive (regressione, classificazione, etc.). Il dataset utilizzato per gli esperimenti è il "celiachion", un insieme di dati relativi a pazienti virtuali generato per addestrare un classificatore fuzzy per la diagnosi della celiachia nei bambini siciliani e maltesi.
Nella sezione due sono descritte le tecniche di riduzione dimensionale applicate al nostro dataset: PCA, kernel PCA, MDS e t-SNE.
Nella sezione tre vengono dapprima definite le quattro tecniche di visualizzazione impiegate: scatterplot 2D per dati bidimensionali, scatterplot 2D per dati tridimensionali, scatterplot 3D interattivo e plot di Draftman; e successivamente vengono applicate per ciascuna tecnica di riduzione.
Nella quarta sezione, invece, viene prima addestrato l'albero decisionale con il dataset non ridotto e in seguito viene addestrato con i dataset ottenuti attraverso le quattro tecniche di riduzione usate.
Nella sezione cinque, infine, vengono tratte le conclusioni riguardo le migliori visualizzazioni prodotte e confrontate le metriche dei classificatori addestrati.

2 Tecniche di riduzione della dimensionalità

3 Tecniche di visualizzazione

4 Alberi decisionali

5 Conclusioni