ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μεταπτυχιακή Φοιτήτρια

Καραγιαννάκη Ιουλία

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Ι. Τσαμαρδίνος

Παρασκευή, 28 Αυγούστου 2020, ώρα 10:00 π.μ

Τηλεδιάσκεψη (μέσω του συστήματος e:Presence), Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Κανάλι YouTube του Τμήματος
https://www.youtube.com/channel/UC7uE3QiMTQjkrpByB_Gnt6Q/live

"Μαθαίνοντας Βιολογικά Ερμηνεύσιμες Κρυφές Αναπαραστάσεις από Βιολογικά Δεδομένα"

Περίληψη

Τα δεδομένα γονιδιακής έκφρασης είναι κατά κύριο λόγο πολυδιάστατα με πολύ μικρό αριθμό δειγμάτων. Αυτό οδηγεί σε στατιστικές και μεθοδολογικές προκλήσεις, οι οποίες πρέπει να αντιμετωπιστούν για την περεταίρω ανάλυση και κατανόηση των υποκείμενων βιολογικών μηχανισμών που υπάρχουν σε δεδομένα αυτού του τύπου. Γι' αυτό το σκοπό, έχουν προταθεί μέθοδοι μείωσης διαστάσεων, οι οποίες μαθαίνουν ένα χώρο χαμηλότερης διάστασης (κρυφός διανυσματικός χώρος) που αποτελείται από νέες μεταβλητές και αναπαριστούν τα αρχικά δεδομένα ως άθροισμα αυτών (κρυφές αναπαραστάσεις). Η προβολή των αρχικών δεδομένων στον κρυφό διανυσματικό χώρο συμπιέζει τα δεδομένα, ενώ διατηρεί τη σημαντική τους πληροφορία και μειώνει το θόρυβο.

Οι κλασσικές τεχνικές μείωσης διαστάσεων, όπως η ανάλυση κυρίων συνιστωσών (PCA), βρίσκουν κρυφές αναπαραστάσεις οι οποίες δεν είναι βιολογικά ερμηνεύσιμες. Για την καλύτερη βιολογική ερμηνεία, έχουν προταθεί νέες μέθοδοι οι οποίες βρίσκουν αραιές κρυφές αναπαραστάσεις. Συγκεκριμένα, οι νέες μεταβλητες μπορούν να αναπαρασταθούν ως γραμμικός συνδυασμός μικρού πλήθους των αρχικών μοριακών ποσοτήτων. Όμως και πάλι η ερμηνεία των αραιών κρυφών

αναπαραστάσεων δεν είναι πλήρως κατανοητή, διότι οι ήδη υπάρχουσες τεχνικές μαθαίνουν αραιές κρυφές αναπαραστάσεις, οι οποίες δεν αντιστοιχούν άμεσα στα ήδη γνωστά βιολογικά μονοπάτια ή σε άλλα γνωστά σύνολα γονιδίων.

Σε αυτή την εργασία, θα παρουσιάσουμε μία νέα τεχνική δημιουργία νέων μεταβλητών και μείωση διαστάσεων που ονομάζεται Pathway Activity Score Learning (PASL). Η βασική καινοτομία της μεθόδου PASL είναι ότι ο κρυφός διανυσματικός χώρος που επιστρέφει, είναι βιολογικά ερμηνεύσιμος καθώς εφαρμόζονται περιορισμοί έτσι ώστε να αντιστοιχεί σε γνωστά βιολογικά μονοπάτια. Ο έλεγχος της ορθότητας της μεθόδου γίνεται τόσο σε συνθετικά, όσο και σε πραγματικά δεδομένα. Δείχνουμε ότι η μέθοδος PASL διατηρεί την προβλεπτική ικανότητα των αρχικών δεδομένων. Επίσης η εύρεση διαφορικά εκφραζόμενων βιολογικών μονοπατιών δίνει επιπρόσθετη πληροφορία στην ανάλυση εμπλουτισμού γονιδίων.

University of Crete

Computer Science Department

M.Sc. Thesis presentation / examination

Karagiannaki Ioulia

Master's Thesis Supervisor: Professor Ioannis Tsamardinos

Friday, 28 August 2020, 10:00 a.m

Teleconference (will use the e: Presence system), Computer Science Department, University of Crete

YouTube channel : https://www.youtube.com/channel/UC7uE3QiMTQjkrpByB_Gnt6Q/live

“Learning Biologically Interpretable Latent Representations from Gene Expression Data”

Abstract

Gene expression data are typically high dimensional with low sample size. This leads to several statistical and analytical challenges that one needs to overcome in order to analyze and infer the underlying biological mechanisms of such data. To this end, several dimensionality reduction techniques have been proposed. Dimensionality reduction techniques learn a lower dimensional space (latent space), of newly constructed features and represent the data as a sum of those (latent representations). The projection of the data to the latent feature space compresses the data, retains the significant information and reduces noise.

Typical dimensionality reduction techniques, such as Principal Component Analysis, derive latent representations that are uninterpretable biologically. In order to regain a degree of interpretability, other methods return sparse latent representations. Particularly, the new features are constructed as linear combinations of only a few of the molecular quantities. However, sparse latent representations are still hard to interpret biologically as they do not directly correspond to the known biological pathways or other known genesets.

In this thesis, we present a novel algorithm for feature construction and dimensionality reduction called Pathway Activity Score Learning (PASL). The major novelty of PASL is that the constructed features are constrained to directly correspond to known molecular pathways and can be interpreted as pathway activity scores. PASL is evaluated both on simulated and real data. We show that PASL retains the predictive information for disease classification on new, unseen datasets. We also show that differential activation analysis provides complementary information to standard geneset enrichment analysis.